More on population substructure

Cavan Reilly

September 30, 2019

Recalling previous code

```
url="http://www.biostat.umn.edu/~cavanr/FMS_data.txt"
fms <- read.delim("url, header=T, sep="\t")

# Necessary code from Example 3.9:
attach(fms)
NamesAkt1Snps <- names(fms)[substr(names(fms),1,4)==
  "akt1"]
FMSgeno <- fms[,is.element(names(fms),NamesAkt1Snps)]
FMSgenoNum <- data.matrix(FMSgeno)
FMSgenoNum[is.na(FMSgenoNum)] <- 4
DistFmsGeno <- as.matrix(dist(FMSgenoNum))
```

Multidimensional scaling

Next, one just uses the figure to set up indicator variables

```
mds=cmdscale(DistFmsGeno)
ind1=c(mds[,1]>0 & mds[,2]<4)
ind2=c(mds[,1]>0 & mds[,2]>4)
ind3=c(mds[,1]<0 & mds[,2]>4)

m1=glm(Met_syn~nr3c1_rs4582314+ind1+ind2+ind3,
  family=binomial)
```

Principal components analysis

Same set up as first slide

```
PCFMS <- prcomp(FMSgenoNum)

sind1=c(PCFMS$x[,1]<0 & PCFMS$x[,2]<4)
sind2=c(PCFMS$x[,1]<0 & PCFMS$x[,2]>4)
sind3=c(PCFMS$x[,1]>0 & PCFMS$x[,2]>4)
```

Principal components analysis

We can see that these are the same indicator variables

```
> table(ind1,sind1)
       sind1
ind1    FALSE TRUE
  FALSE   870    0
  TRUE      0  527
```

Cluster analysis

We can also use a clustering algorithm on this data set.

```
library(mclust)

m1=Mclust(FMSgenoNum)
summary(m1)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 2

components:

 log.likelihood    n  df       BIC       ICL
       1676.917 1397 626 -1179.709 -1179.709
```

Cluster analysis

```
> table(m1$class)

  1   2
774 623
> table(m1$class,ind1)
   ind1
    FALSE TRUE
  1   774    0
  2    96  527
```

Cluster analysis

```
> table(m1$class,ind2)
   ind2
    FALSE TRUE
  1   774    0
  2   527   96
> table(m1$class,ind3)
   ind3
    FALSE TRUE
  1   737   37
  2   623    0
```