Computer Practical Exercise on MDS in PLINK (and R)

Overview

Purpose

In this exercise you will explore Multidimensional Scaling (MDS) as implemented in the program PLINK, for detecting population outliers and adjusting for population stratification. The program R will be used for plotting and visualisation of the results.

Program documentation

PLINK documentation:

PLINK has an extensive set of docmentation including a pdf manual, a web-based tutorial and web-based documentation:

http://pngu.mgh.harvard.edu/~purcell/plink/

Data overview

We will analyse data derived from a set of families that will be used tomorrow in the family-based association exercise. The data comes from a number of families with two or more affected children plus parents (of varying disease status, mostly unknown). All individuals have been typed at a genome-wide set of around 262,000 SNP markers. In the interests of time, we will analyse the data from just three chromosomes, but in principal one could use the same approach to analyse all 22 autosomal chromosomes as well as chromosome X.

Instructions

Data files

You should save the following file to an appropriate directory (folder) on your machine:

familydata3chrs.zip

You then need to unpack/unzip the file. This should create a new directory 'familydata3chrs' containing 8 files: a file named 'files2merge.txt', a file named 'threesnps.txt' and a pedigree file and a map file for each of three chromosomes (10, 13, 14). The files C**.ped refer to the pedigree files and the files C**.recode.map are the corresponding map files (where ** denotes chromosome).

For this exercise you will also need some additional files, which you should download into the same folder as the pedigree and map files:

allpopsIDs.out
outlier.txt
updatedmap.txt
prunedhapmap.bed
prunedhapmap.bim

prunedhapmap.fam

Data format

The data is in standard pedigree file format, with columns corresponding to family id, subject id (within family), father's id, mother's id, sex (1=m, 2=f), affection status (1=unaffected, 2=affected) and one column for each allele for each locus genotype. Missing data is coded with a zero.

If you are unfamiliar with the standard pedigree file format (which is a commonly-used format for many linkage analysis programs) and need more explanation, please ask an instructor.

The map files are in PLINK format. These files describe the markers (in order) in the pedigree file. The PLINK-format map file contains exactly 4 columns:

chromosome (1-22, X, Y or 0 if unplaced)
rs number or snp identifier
Genetic distance (morgans)
Base-pair position (bp units)

Most analyses in PLINK do not require a genetic map distance to be specified, therefore the genetic distance column can be set at 0.

Take a look at the data files, and check that you understand how the data is coded.

Step-by-step instructions

Open up an MSDOS window and move into the directory (folder) where you saved the data files e.g. by typing

```
cd xxxxx
```

(where `xxxxx` is replaced by the name of the appropriate folder).

To start with, we will use PLINK to merge together the data from all 3 chromosomes. We will output this merged data as a special binary format genotype file, which will take up less disc space and be quicker to read into PLINK when performing various subsequent analyses.

To do this, type:

```
plink --noweb --ped C10.ped --map C10.recode.map --merge-list files2merge.txt --make-bed --out merged
```

Here the `--noweb` command tells PLINK to run without bothering to check via the web to see whether there are updated versions of PLINK. The `--ped xxxx` command tells PLINK that the name of the first pedigree file is `xxxx` and the `--map yyyy` command tells PLINK that the name of the first map file is `yyyy`. The `--merge-list files2merge.txt` command tells PLINK to merge in the remaining pedigree and map files listed in `files2merge.txt`. The `--make-bed` command tells PLINK to output the resulting merged file in binary format and the `--out merged` command tells PLINK to use the stem 'merged' for the name of all its output files.

PLINK outputs some useful messages (you should always read these to make sure they match up with what you expect!) and also outputs a copy of these messages to the file `merged.log` .

You should now find 4 new files in your directory: `merged.bed`, `merged.bim`, `merged.fam` and `merged.log`. The file `merged.bed` is the binary genotype file which will not be human readable. The file `merged.bim` is a map file for the merged data. You can take a look at this (e.g. by typing `more merged.bim`). The file `merged.fam` gives the pedigree structure in a format that is compatible with the binary genotype file. You can take a look at this (e.g. by typing `more merged.fam`). Note this file is the same as the first six columns of the original pedigree file, except that unknown disease status has been coded as '-9'.

For subsequent analysis of these data, the files `merged.bed`, `merged.bim` and `merged.fam` will always need to

be read into PLINK together, as together they provide all the required information. This can be done using the command `--bfile merged`.

If you take a look at `merged.log`, you should see that your merged data set consists of 1336 individuals and 24766 SNPs. For genome-wide studies, it is important to do some quality control checks and filter out individuals and SNPs that are likely to be of poor-quality. By default, PLINK does not impose any filters on things like minor allele frequency or genotyping rate. If you want to impose such filters, you can either do it at the same time as doing the analysis, or you can first do some QC filtering and output a new 'filtered' version of the data set for subsequent analysis.

(For example, to exclude individuals with more than 10% missing genotypes (over all SNPs), we can use the command `--mind 0.1`. To then exclude SNPs with more than 10% missing genotypes (over all individuals), we can use the command `--geno 0.1`. To exclude SNPs with minor allele frequency (in founders) less than 0.05, we can use the command `--maf 0.05`. To exclude markers that fail a Hardy-Weinberg test (in founders) at significance threshold 0.000001, we can use the command `--hwe 0.000001`. To finally exclude individuals and/or markers on the basis of Mendelian inheritance error rates, we can use the option `--me 0.05 0.1`, where the first parameter determines that families with more than 5% Mendelian errors (considering all SNPs) will be discarded, and the second parameter indicates that SNPs with more than 10% Mendelian error rate (based on the number of trios) will be excluded).

We will use PLINK to generate a new data set where some people and some SNPs have been removed according to the QC checks we think most appropriate. We do this by typing:

```
plink --noweb --bfile merged --mind 0.05 --geno 0.05 --maf 0.05 --hwe 0.000005 --me 0.05 0.05 --make-bed --out QCed
```

This command should produce a new set of "QCed" (quality-controlled) data (19680 SNPs) contained in the 3 files `QCed.bed` `QCed.bim` `QCed.fam`

To perform eigenvector and MDS analysis, the individuals should be unrelated. We will therefore just perform the analysis in the founders (parents). We can get PLINK to produce a new set of files `founders.bed` `founders.bim` `founders.fam` that just contain data for founders by typing:

```
plink --noweb --bfile QCed --filter-founders --make-bed --out founders
```

To perform principal components or MDS analysis, it is also important that we use SNPs that are not too correlated. We can use plink to prune the SNPs down to get an approximately independent set of SNPs:

```
plink --noweb --bfile founders --indep 50 5 2 --out prunedsnps
```

This command should produce a file `prunedsnps.prune.in` listing the SNPs that should be retained in the analysis. (Check the PLINK website to get more details about how this option works). To produce a new set of files for the founders, containing only these SNPs, type:

```
plink --noweb --bfile founders --extract prunedsnps.prune.in --make-bed --out prunedfounders
```

1. Analysis to detect population outliers

We will start by trying to detect population outliers by adding in the HapMap data. Unfortunately the SNP map we have been using for our own data is based on older annotation information than the HapMap files we have. So we will first update the map information that is being used in our own data, to correspond to the map files being used for the HapMap data:

```
plink --noweb --bfile prunedfounders --update-map updatedmap.txt --make-bed --out updatedfounders
```

This command made use of a file `updatedmap.txt` that gives the updated map required. Take a look at this file and check you understand how it is coded.

Next we will merge in the HapMap data which has been pre-prepared for you in PLINK-format files `prunedhapmap.bed` `prunedhapmap.bim` `prunedhapmap.fam` , using the pruned set of SNPs and the same allele-

coding as used in our own data.

```
plink --noweb --bfile prunedhapmap --bmerge updatedfounders.bed updatedfounders.bim updatedfounders.fam --make-bed
--out allpops
```

Now we have our final set of files `allpops.bed` `allpops.bim` `allpops.fam` which contain data at the required set of SNPs, in all 5 populations (our own population, and the 4 HapMap populations from Europe (CEU), China (CHB), Japan (JPT) and Yoruba (YRI)).

To perform MDS analysis in PLINK, we first calculate a file `allpopsIBD.genome` of genome-wide estimates of IBD sharing:
(WARNING - THIS ANALYSIS CAN BE VERY SLOW!)

```
plink --noweb --bfile allpops --genome --out allpopsIBD
```

Then we use PLINK to calculate the MDS scores. To calculate the first two scores for everyone in the data set, type:

```
plink --noweb --bfile allpops --read-genome allpopsIBD.genome --cluster --mds-plot 2 --out allpopsmds
```

This produces a file `allpopsmds.mds`. Take a look at the file.

To plot the scores you can read this file into R. Start up R and, within R, move to the directory where you saved the files. Then type:

```
mds <- read.table("allpopsmds.mds", header=T)
popnids <- read.table("allpopsIDs.out", header=T)
```

These commands read the MDS file you created into an R data frame, together with an additional file `allpopsIDs.out` that we have prepared for you, giving an indicator of which population each person in the data set belongs to. Take a look at the top of the data frames you have created in R by typing:

```
head(mds)
head(popnids)
```

We can create a new variable in R that codes for the different populations. We will call this variable "colour" since we will use it to colour the different populations in our plot:

```
colour <- popnids$CEUind+2*popnids$CHBind+3*popnids$JPTind+4*popnids$YRIind+5*popnids$OURSind
```

To look at the variable you have just made, type

```
colour
```

Now display the MDS plot in colour:

```
plot(mds$C1, mds$C2, col=colour)
```

This plots the 1st and 2nd component for each person, with each point representing a different person. The colouring corresponds to the different populations. You should be able to see the three different HapMap populations in different positions on the plot, together with our data. The CEU samples are shown in black, the YRI in dark blue, the CHB in red and the JPT in green. Our samples (the largest data set) are shown in turquoise blue. The HapMap CEU data points have been largely covered up by our data, so colour them in again in black by typing:

```
points(mds$C1, mds$C2, col=popnids$CEUind)
```

You can see that our data appears to be predominently Caucasian (since it overlaps with the CEU HapMap data) but there is one outlier. You can check who this is by typing:

```
mds[mds$C1>0,]
```

This outputs a list of people who scored more than 0 on the first component. These are mostly the HapMap samples, but also include the outlier from our data, person 1 in family region1c_002.

2. Analysis to detect population stratification

To see whether there might be more subtle population stratification in our own samples, we can repeat the MDS analysis but just using our own data (i.e. not merging in the HapMap data).

First we need to delete the population outlier from our own data. We have prepared a file `outlier.txt` listing the family number and ID of the person to remove. Take a look at the file and check you understand it.

To prepare new PLINK files with the outlier removed, type:

```
plink --noweb --bfile prunedfounders --remove outlier.txt --make-bed --out nooutliers
```

Then perform the genome-wide estimatation of IBD sharing:

```
plink --noweb --bfile nooutliers --genome --out nooutliersIBD
```

Then we use PLINK to calculate the MDS scores. To calculate the first two scores for everyone in the data set, type:

```
plink --noweb --bfile nooutliers --read-genome nooutliersIBD.genome --cluster --mds-plot 2 --out nooutliersmds
```

This produces a file `nooutliersmds.mds`. Take a look at the file.

To plot the scores you can read this file into R. Start up R and, within R, move to the directory where you saved the files. Then type:

```
mds2 <- read.table("nooutliersmds.mds", header=T)
```

Now display the MDS plot:

```
plot(mds2$C1, mds2$C2)
```

This plots the first two principal components from the MDS analysis, with each point representing a different person. It is unclear from the plot whether there is any underlying population structure/population stratification. One could always use the first few principal components from the MDS analysis as covariates in any case/control association analysis, in order to adjust for possible unmeasured population stratification.

In fact, our data comes from two different geographical locations. Take a look at the data by typing:

```
mds2
```

It turns out that the first 284 people come from one location (region 1), while the remainder (persons 285-556) come from another location (region 2). We can colour the MDS plot according to geographical location by using the following commands:

```
line<-(1:556)
place<-(line>284)+1
plot(mds2$C1, mds2$C2, col=place)
```

You can see that the points do indeed cluster by geographical location. This population stratification could cause problems in case/control analysis if we did not adjust for it. If we did not know which geographical location each person came from, we could use the first few principal components from the MDS analysis as covariates in any case/control association analysis, in order to adjust for possible population stratification. Alternatively, we can use family-based methods of association analysis, as we shall do for these data tomorrow.

References

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904-9.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661-78.

Exercises prepared by: Heather Cordell
Checked by:
Programs used: PLINK, smartpca, GenABEL
Last updated: 06/29/2010 10:19:27