Assignment: Predict Missing Age Values

Huang Ao 17338053

1. Use PCA to Reduct Dimensions

Run PCA R scripts on the server using PBS job management system.

```r
library(psych)
library(limma)

setwd("~/assignments/age_interpolate/")
### input data
sample <- read.table("~/assignments/age_interpolate/sample_inform.txt",header = T)
rnaseq <- read.table("~/assignments/age_interpolate/RNAseq_hg19_refGene_counts.txt",
        + header = F, row.names = 1)
rnaseq_inv = t(rnaseq)

### remove batch effect
batch = t(matrix(sample[,3])) # row vector
rnaseq_rmbatch = removeBatchEffect(rnaseq,batch = batch) # row:genes
rnaseq_rmbatch_inv = t(rnaseq_rmbatch) #col:genes(variables)

### PCA (with batch effect)
pca_rna = principal(rnaseq_inv,nfactors = 3)
pca_rna

write.csv(pca_rna$scores,file = "pca_rna_scores.csv")

### PCA (without batch effect)
pca_rna_rmbatch = principal(rnaseq_rmbatch_inv,nfactors = 3)
pca_rna_rmbatch

write.csv(pca_rna_rmbatch$scores,file = "pca_rna_scores_rmbatch.csv")
```

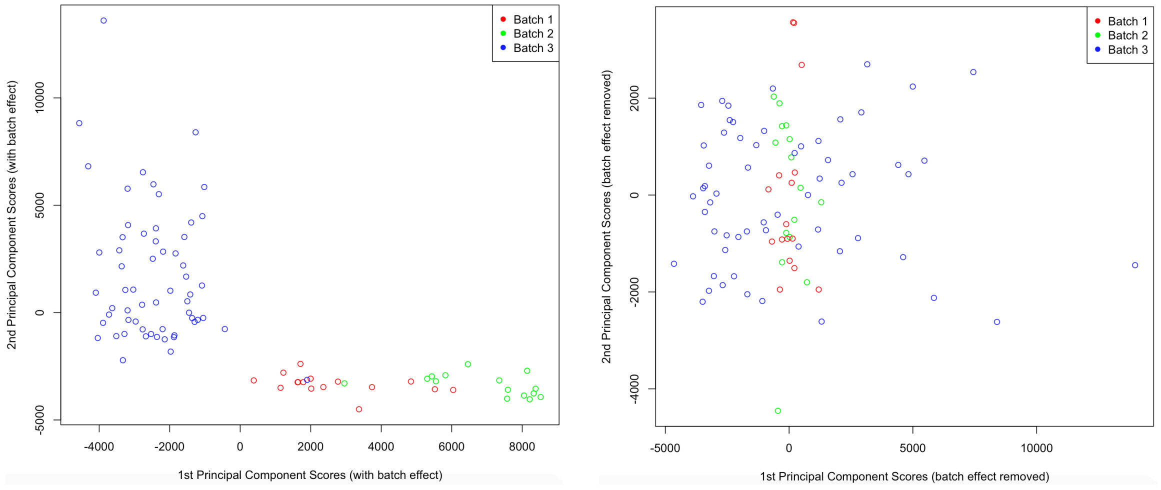
The PCA results are exported in .csv files and analyzed on personal computer. Note that the original .csv files were integrated with sample information (ID, Age and Batch) using EXCEL before the analysis.

PCA results are visualized as in Figure 1. It can be conspicuously seen that the batch effect have great influence on sample clustering (Figure 1-a), and that the removal of batch effect using the *removeBatchEffect* of the *Limma* package have positive performance on reducing the clusterings caused by different batches, preserving only effects from experiment designs, sample characteristics (e.g. age, gender) or whatsoever.

```
### input data
pca_rna_scores = read.csv("~/Desktop/pca_rna_scores.csv",header = T, row.names = 1)
pca_rna_scores_rmbatch = read.csv("~/Desktop/pca_rna_scores_rmbatch.csv",header = T,
        + row.names = 1)
### color matrix
mycol = rep(rgb(1,0,0),length(pca_rna_scores$RC1))
mycol[pca_rna_scores$Batch == 2] = rgb(0,1,0)
mycol[pca_rna_scores$Batch == 3] = rgb(0,0,1)

### result visualization
plot(pca_rna_scores$RC1,pca_rna_scores$RC2, col = mycol,
        xlab = "1st Principal Component Scores (with batch effect)",
        ylab = "2nd Principal Component Scores (with batch effect)")
legend(x="topright",legend = c("Batch 1","Batch 2","Batch 3"),
        col = c("red","green","blue"), pch = 16)

plot(pca_rna_scores_rmbatch$RC1,pca_rna_scores_rmbatch$RC2, col = mycol,
        xlab = "1st Principal Component Scores (batch effect removed)",
        ylab = "2nd Principal Component Scores (batch effect removed)")
legend(x="topright",legend = c("Batch 1","Batch 2","Batch 3"),
        col = c("red","green","blue"), pch = 16)
```



(a) With Batch Effect

(b) Batch Effect Removed

图 1: Comparisons of PCA Results with or without Batch Effect

2. Use 2-D Interpolation to Predict Missing Values

The first two components of PCA (batch effect removed) are selected to represent the expression profile of the 17320-dimensional RNA-seq data, measured by PCA scores. The score data are inputed and analyzed with the interpolation function *griddata* in Matlab.

```
%%% input the PCA scores of the known age samples (77 samples)
x1 = RC1';
x2 = RC2';
y = Age';
%%% input the PCA scores of the unknown age samples (14 samples)
xi1 = RC4';
xi2 = RC5';
%%% generate the grid sample points
[xi1,xi2] = meshgrid(xi1,xi2);
%%% data interpolation using v4 method
yi = griddata(x1,x2,y,xi1,xi2,'v4')
```

The results of y_i is a 14×14 matrix, and our targeted age values are in the main diagonal (which corresponds to the PCA scores of x_{i1} and x_{i2}). The predicted age values are organized as in Table 1.

表 1: Predicted Age values

ID	Age	ID	Age	ID	Age	ID	Age
GEP_013	52	GEP_104	38	GEP_045	45	GEP_077	39
GEP_046	29	GEP_111	70	GEP_064	24	GEP_089	43
GEP_079	86	GEP_012	67	GEP_072	47		
GEP_114	60	GEP_035	69	GEP_073	44		