**P8139**

**Homework #3**

Ruixi Li

2024-03-11

# Answers were denoted yellow.

# Data Preparation and Conduct PCA analysis

id can not provide information/variance for student’s performance, and was used as thr row names here.

mydata = read.table(file="PCA\_data.csv", header=TRUE, row.names=1, sep=",")  
  
mydata.pca = prcomp(mydata, retx=TRUE, center=TRUE, scale=TRUE)  
# variable means set to zero, and variances set to one "scale=TRUE"  
# PCA scores for each sample store in mydata.pca$x  
# loadings stored in mydata.pca$rotation  
# square roots of eigenvalues score in mydata.pca$sdev (note that eigenvalues are variances of principal compoments)  
# variable means stored in mydata.pca$center  
# variable standard deviations stored in mydata.pca$scale  
sd = mydata.pca$sdev  
loadings = mydata.pca$rotation  
rownames(loadings) = colnames(mydata)

# Q1a.

# PCA scores from the R package prcomp.

#output PCA scores from prcmp package  
scores = mydata.pca$x  
scores

## PC1 PC2 PC3  
## 1 -1.6579369 0.571682476 0.27659132  
## 2 -2.5759191 0.171154076 -0.39465359  
## 3 -1.8500938 0.561523155 -0.32895929  
## 4 -2.0029634 -0.279232617 0.20909576  
## 5 -1.7019579 -0.553776982 0.53149553  
## 6 -1.0237890 0.746472175 0.42698709  
## 7 -0.9343334 0.599025432 0.24836634  
## 8 -2.0297283 -0.031871212 -0.13372583  
## 9 -2.2666815 0.246059593 -0.25427541  
## 10 -2.6579342 -0.005870397 -0.16437182  
## 11 1.3079701 0.359969524 0.41720262  
## 12 0.9980941 1.269880082 -0.28446005  
## 13 1.3110519 1.182549991 -0.11825137  
## 14 1.6046169 1.232476842 0.15248740  
## 15 1.2290368 1.005525517 0.11203039  
## 16 0.8027869 0.651507049 0.04113399  
## 17 0.8884538 0.880662669 0.23237339  
## 18 1.5032773 0.978342557 0.09160858  
## 19 1.1241139 0.484893331 -0.76606024  
## 20 0.7141229 1.472874891 -0.01392787  
## 21 1.1276809 -1.336079451 0.05473737  
## 22 0.3206321 -1.229551905 -0.01435777  
## 23 0.9474764 -1.159024498 -0.70700433  
## 24 0.5478545 -1.331670861 0.35630217  
## 25 1.3013734 -0.943475598 -0.26024587  
## 26 0.6022446 -1.366839327 0.41257208  
## 27 1.0307847 -0.864161491 0.18169711  
## 28 0.4413646 -1.112674562 -0.05800904  
## 29 0.4666995 -1.049140991 -0.04278934  
## 30 0.4317024 -1.151229468 -0.20358933

#PCA scores from loadings and the original data  
mydata\_scaled = scale(mydata)  
scores\_manual <- as.matrix(mydata\_scaled) %\*% loadings  
scores\_manual

## PC1 PC2 PC3  
## 1 -1.6579369 0.571682476 0.27659132  
## 2 -2.5759191 0.171154076 -0.39465359  
## 3 -1.8500938 0.561523155 -0.32895929  
## 4 -2.0029634 -0.279232617 0.20909576  
## 5 -1.7019579 -0.553776982 0.53149553  
## 6 -1.0237890 0.746472175 0.42698709  
## 7 -0.9343334 0.599025432 0.24836634  
## 8 -2.0297283 -0.031871212 -0.13372583  
## 9 -2.2666815 0.246059593 -0.25427541  
## 10 -2.6579342 -0.005870397 -0.16437182  
## 11 1.3079701 0.359969524 0.41720262  
## 12 0.9980941 1.269880082 -0.28446005  
## 13 1.3110519 1.182549991 -0.11825137  
## 14 1.6046169 1.232476842 0.15248740  
## 15 1.2290368 1.005525517 0.11203039  
## 16 0.8027869 0.651507049 0.04113399  
## 17 0.8884538 0.880662669 0.23237339  
## 18 1.5032773 0.978342557 0.09160858  
## 19 1.1241139 0.484893331 -0.76606024  
## 20 0.7141229 1.472874891 -0.01392787  
## 21 1.1276809 -1.336079451 0.05473737  
## 22 0.3206321 -1.229551905 -0.01435777  
## 23 0.9474764 -1.159024498 -0.70700433  
## 24 0.5478545 -1.331670861 0.35630217  
## 25 1.3013734 -0.943475598 -0.26024587  
## 26 0.6022446 -1.366839327 0.41257208  
## 27 1.0307847 -0.864161491 0.18169711  
## 28 0.4413646 -1.112674562 -0.05800904  
## 29 0.4666995 -1.049140991 -0.04278934  
## 30 0.4317024 -1.151229468 -0.20358933

# This calculation uses matrix multiplication (%\*%) to project the standardized data onto the PCA loadings.

PCA scores using loadings and PCA scores original math/chem/bio scores are the same.

# Q1b.

Since eigenvalues are variances of principal compoments, we can calculate the percent variance explained of each component by diving the eigenvalue of each component by the sum of all eigenvalues. Given that mydata.ppca$sdev is square roots of eigenvalues store…

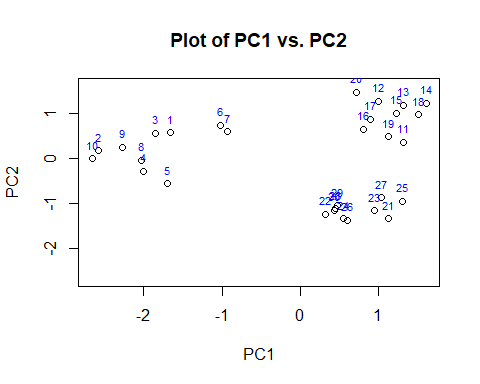
eigenvalues <- mydata.pca$sdev^2  
total\_variance <- sum(eigenvalues)  
percent\_variance\_explained <- (eigenvalues / total\_variance) \* 100  
percent\_variance\_explained

## [1] 66.928022 29.711364 3.360614

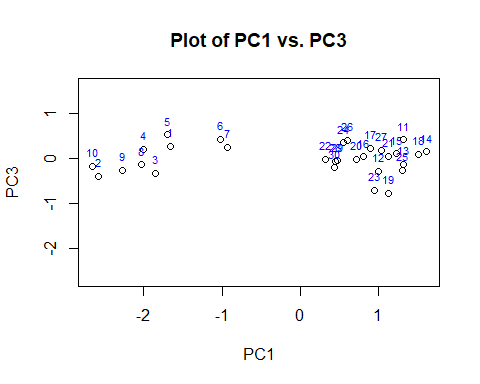
* The percent variance explained of PC1 is 66.9280223%，the percent variance explained of PC2 is 29.7113637%, the percent variance explained of PC3 is 3.360614%.

# Q1c.

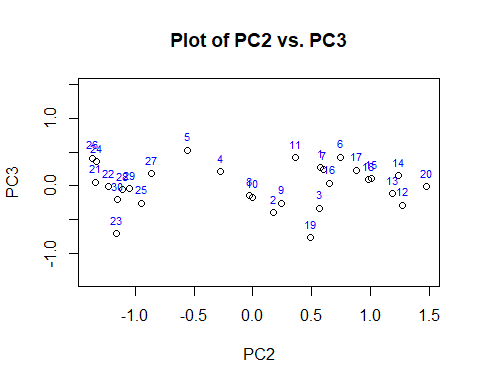
plot(scores[,1:2],  
xlim=c(min(scores[,1:2]),max(scores[,1:2])),ylim=c(min(scores[,1:2]),max(scores[,1:2])))  
text(scores[,1], scores[,2], rownames(scores), col="blue", cex=0.7, pos=3)  
title("Plot of PC1 vs. PC2")



plot(scores[,-2],  
xlim=c(min(scores[,-2]),max(scores[,-2])),ylim=c(min(scores[,-2]),max(scores[,-2])))  
text(scores[,1], scores[,3], rownames(scores), col="blue", cex=0.7, pos=3)  
title("Plot of PC1 vs. PC3")



plot(scores[,-1],  
xlim=c(min(scores[,-1]),max(scores[,-1])),ylim=c(min(scores[,-1]),max(scores[,-1])))  
text(scores[,2], scores[,3], rownames(scores), col="blue", cex=0.7, pos=3)  
title("Plot of PC2 vs. PC3")



2. Read the paper “Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and

Insulin-dependent Diabetes Mellitus (IDDM)” by Spielman et al. published on American Journal

of Human Genetics (AJHG) in 1993 and write one page summarizing what you learned.

Other than your understanding of the test, in your summary, you can cite some sentences from the

paper that clarifies some of your understandings in concepts we have been discussing in the class,

such as linkage and linkage disequilibrium, samples used to examine these, etc.

As means for identifying genes for complex diseases, both the association (spurious association due to population substructures) and the affected-sib-pairs approaches (lack power due to modest number of ASP and linkage heterogeneity) have limitations. To solve the above problems, the author proposed an alternative method to test for linkage with a genetic marker when population association has been found.

表格

描述已自动生成

Since only the data from heterozygous M1M2 parents should be used in the test (discordant pairs) and covariances should be considered, "McNemar's test" rather than chi-squared test is valid.

If a marker allele is not associated with the disease, then for a heterozygous parent(M1M2), the probability of transmitting ‘M1’ is the same as the probability of transmitting ‘M2’, we have Thus, under the null hypothesis of no association, conditional on the number of heterozygous parents, we have: , so the test statistics is:

In the context of genetics, for a recessive disease, the probabilities corresponding to the four cells of table 2 are as shown in our table 3.

表格

描述已自动生成

This means TDT tests for both linkage and association. The probabilities of the various possible combinations of transmitted and non-transmitted marker locus alleles are determined by the association (disequilibrium) parameter ­ and the recombination fraction between the loci.

*It is instructive to consider the case of families with two children, both of whom are affected*. This will allow not only a straightforward extension of expression (1) but also a comparison to be made of this TDT test, and a conventional test (model-free linkage analysis in the simple ASPs design) for linkage between D and M loci, the ­ for haplotype sharing by affected sibs, which we denote as

​Consider then the alleles transmitted to the affected children from each of these h heterozygous parents. The information obtained can be summarized by defining the following three categories:

i = number of parents who transmit Ml to both children.

h-i-j = number of parents who transmit M1 to one child and M2 to the other.

j = number of parents who transmit M2 to both children.

is a test statistic used for linkage disequilibrium, derived as , where h is the total number of heterozygous parents. [application of expression (1)]

compares the number of times a parent transmits the identical marker allele(i+j) to both members of a pair of affected children with the number of times a parent transmits different alleles(h-i-j), using the formula [application of expression (1)]

is a combined test statistic that includes and ​, and is calculated with the formula [application of ],

This paper also mentioned the t2 and Y statistics described by Blackwelder and Elston (1985) for testing for linkage between the D and M loci. Blackwelder and Elston focus attention on statistics calculated from the number of families where the two affected sibs share k (k = 0,1,2) parental marker genes.

reflects the excess or deficit of observed IBD sharing () relative to the expected sharing(n) if there were no linkage. The square root term in the denominator,, is the standard error of the observed count under the null hypothesis. The square of is , indicating that use of is equivalent to the use of

*These methods can be generalized to sibships with three, four or more affected.*

For families with varying numbers of affected children, we recommend simply combining all affected children in the data, irrespective of number of affected in the family, in one overall transmission/disequilibrium statistic of the form , where B is the total number of transmissions of M1 to affected children and C is the total number of transmissions of M2. In the case where segregation distortion at the M locus is a possibility, an aggregate 2 X 2-table is appropriate, corresponding to that discussed above for the case of one affected child per family.