**4.6 - Additional Examples of PCA**

Example 4.3: Weekly Stock Market Returns

The weekly rates of return for five stocks (JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, and ExxonMobil) listed on the New York Stock Exchange were determined over 103 successive weeks. The weekly rate of return was computed as



and is adjusted for stock splits and dividends.

The data in R…

JPMorgan\_Bank Citibank\_Bank WellsFargo\_Bank RoyalDutchShell\_Oil ExxonMobil\_Oil

1 0.0130338 -0.0078431 -0.0031889 -0.0447693 0.0052151

2 0.0084862 0.0166886 -0.0062100 0.0119560 0.0134890

3 -0.0179153 -0.0086393 0.0100360 0.0000000 -0.0061428

4 0.0215589 -0.0034858 0.0174353 -0.0285917 -0.0069534

5 0.0108225 0.0037167 -0.0101345 0.0291900 0.0409751

6 0.0101713 -0.0121978 -0.0083768 0.0137083 0.0029895

7 0.0111288 0.0280044 0.0080721 0.0305433 0.0032290

8 0.0484801 -0.0051480 0.0182495 0.0063348 0.0076752

9 -0.0344914 -0.0137991 -0.0080468 -0.0299011 -0.0108108

10 -0.0046596 0.0209882 -0.0060841 -0.0203940 -0.0126677

:

:

:

94 0.0373241 0.0359281 0.0252751 0.0581879 0.0169708

95 0.0238029 0.0031125 -0.0068757 0.0122545 0.0281715

96 0.0256826 0.0525266 0.0406957 -0.0316623 -0.0188482

97 -0.0060622 0.0086334 0.0058413 0.0445584 0.0305941

98 0.0217449 0.0229645 0.0291983 0.0084395 0.0319296

99 0.0033740 -0.0153061 -0.0238245 -0.0016738 -0.0172270

100 0.0033626 0.0029016 -0.0030507 -0.0012193 -0.0097005

101 0.0170147 0.0095061 0.0181994 -0.0161758 -0.0075614

102 0.0103929 -0.0026612 0.0044290 -0.0024818 -0.0164502

103 -0.0127948 -0.0143678 -0.0187402 -0.0049759 -0.0163732

The principal component analysis could be done on either the variance/covariance matrix or the correlation matrix.

PC Analysis based on variance/covariance matrix

> stock.pca=princomp(Stock[,2:6])

> summary(stock.pca)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5

Standard deviation 0.03680217 0.02635056 0.01585365 0.01188352 0.01085046

Proportion of Variance 0.52926066 0.27133298 0.09821584 0.05518400 0.04600652

Cumulative Proportion 0.52926066 0.80059364 0.89880948 0.95399348 1.00000000

PC Analysis based on correlation matrix

> stock.pca.cor=princomp(Stock[,2:6],cor=TRUE)

> summary(stock.pca.cor)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5

Standard deviation 1.5611768 1.1861756 0.7074693 0.63248050 0.50514343

Proportion of Variance 0.4874546 0.2814025 0.1001025 0.08000632 0.05103398

Cumulative Proportion 0.4874546 0.7688572 0.8689597 0.94896602 1.00000000

The outcomes are similar in this case because the variables (i.e. weekly rate of return) are similar for these stocks. We will continue with the analysis based on the correlation matrix.

> stock.pca.cor$loadings

Loadings:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5

JPMorgan\_Bank 0.469 0.368 0.604 0.363 0.384

Citibank\_Bank 0.532 0.236 0.136 -0.629 -0.496

WellsFargo\_Bank 0.465 0.315 -0.772 0.289

RoyalDutchShell\_Oil 0.387 -0.585 -0.381 0.595

ExxonMobil\_Oil 0.361 -0.606 0.109 0.493 -0.498

> loadplot(stock.pca.cor$loadings, variables = 1:3)

Graphical user interface, timeline

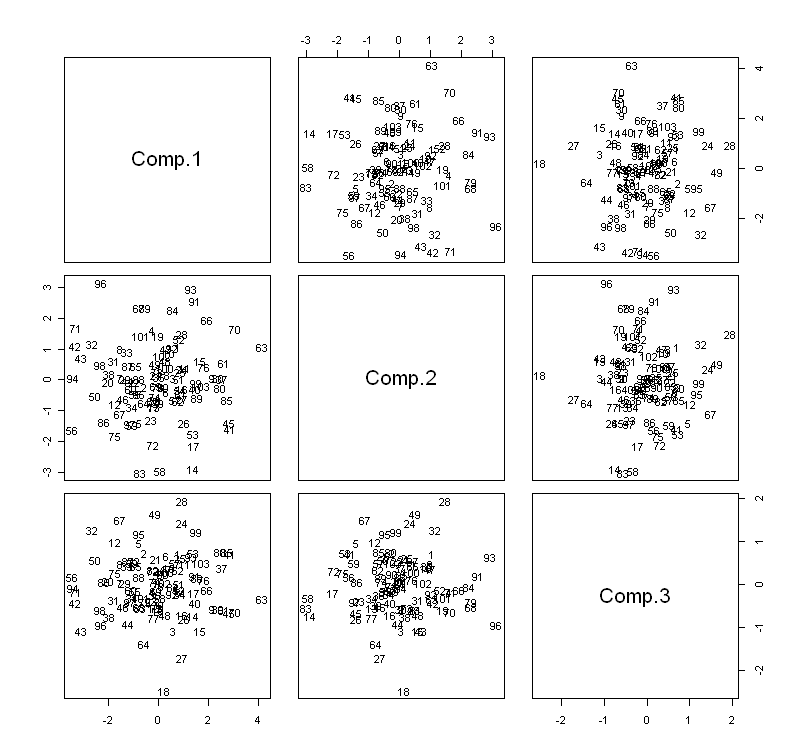
Description automatically generated

**Questions:**

1. What is the make-up of the 1st principal component?
2. What is the make-up of the 2nd principal component?
3. In this analysis it appears we gain some additional information by considering the third principal component. What is the make-up of the 3rd principal component?

Plotting the first three principal components in a pairwise fashion.

> pairs(stock.pca.cor$scores[,1:3],panel=function(x,y){text(x,y,1:103)})



Getting the biplot for this example…

> biplot(stock.pca.cor)

Chart, scatter chart

Description automatically generated

Understanding the extreme points in this plot…

> outlying=c(96,93,91,63,45,41,14,58,83,56,86,94,42,71)

> Stock[outlying,]

JPMorgan\_Bank Citibank\_Bank WellsFargo\_Bank RoyalDutchShell\_Oil ExxonMobil\_Oil

96 0.0256826 0.0525266 0.0406957 -0.0316623 -0.0188482

93 0.0307783 -0.0160888 0.0031045 -0.0539478 -0.0556609

91 0.0090063 -0.0022422 0.0000000 -0.0429774 -0.0620229

63 -0.0300058 -0.0497446 -0.0167189 -0.0507510 -0.0583157

45 -0.0446763 -0.0408118 -0.0163225 -0.0035049 -0.0008137

41 -0.0329997 -0.0313480 -0.0362141 0.0055866 -0.0065208

14 -0.0458668 -0.0278243 -0.0142696 0.0374776 0.0332022

58 -0.0260532 -0.0168492 -0.0080604 0.0432676 0.0587486

83 -0.0265896 -0.0002339 -0.0033698 0.0615551 0.0561091

56 0.0188149 0.0379692 0.0154985 0.0510400 0.0784157

86 0.0136240 0.0182335 0.0086520 0.0568640 0.0387476

94 0.0373241 0.0359281 0.0252751 0.0581879 0.0169708

42 0.0302098 0.0522778 0.0317662 0.0267857 0.0105865

71 0.0465672 0.0410867 0.0349723 0.0152170 0.0056721

Example 4.4: Nutritional Information for Three Fast Food Chains

For this example, consider Burger King, Wendy’s and McDonald’s from the Nutritional dataset.

Table

Description automatically generated

Doing an **incorrect** PC analysis (i.e. using the covariance matrix instead of the correlation matrix).

> Nutritional.pca=princomp(Nutritional[,6:14])

> Nutritional.pca

Call:

princomp(x = Nutritional[, 6:14])

Standard deviations:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9

486.249802 129.003847 42.707897 12.253591 7.421000 5.163179 2.254073 1.528761 1.069378

9 variables and 80 observations.

> summary(Nutritional.pca)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9

Standard deviation 486.2498018 129.00384739 42.707896941 1.225359e+01 7.4210000586 5.1631787406 2.254073e+00 1.528761e+00 1.069378e+00

Proportion of Variance 0.9266833 0.06522556 0.007148729 5.884902e-04 0.0002158427 0.0001044833 1.991355e-05 9.159922e-06 4.482033e-06

Cumulative Proportion 0.9266833 0.99190890 0.999057628 9.996461e-01 0.9998619612 0.9999664445 9.999864e-01 9.999955e-01 1.000000e+00

> Nutritional.pca$loadings

Loadings:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9

Calories 0.361 0.918 0.126

TotalFat 0.258 0.410 0.171 -0.172 -0.838

SaturatedFat -0.128 0.102 -0.962 0.129 -0.142

Cholesterol 0.106 -0.975 -0.153

Sodium 0.928 -0.369

TotalCarbs 0.144 -0.854 0.211 -0.216 -0.387

Fiber -0.125 -0.158 -0.974

Sugar -0.296 -0.932 0.177 0.112

Protein 0.288 -0.255 -0.852 -0.330

**Questions:**

1. What is the make-up of the 1st principal component?
2. What is the make-up of the 2nd principal component?
3. Why is Fiber absent from the first five principal components?   
   **Hint**: Compare the scale for Fiber to Sodium or Calories.

Doing a **correct** PC analysis (i.e. using the correlation matrix which again is equivalent to standardizing your data and running your PC analysis).

> Nutritional.pca.cor=princomp(Nutritional[,6:14],cor=TRUE)

> summary(Nutritional.pca.cor)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6

Standard deviation 2.2601049 1.1631822 0.93581548 0.91377871 0.61020487 0.50978426

Proportion of Variance 0.5675638 0.1503325 0.09730562 0.09277684 0.04137222 0.02887555

Cumulative Proportion 0.5675638 0.7178963 0.81520195 0.90797879 0.94935101 0.97822656

Comp.7 Comp.8 Comp.9

Standard deviation 0.34841363 0.269726592 0.0426195789

Proportion of Variance 0.01348801 0.008083604 0.0002018254

Cumulative Proportion 0.99171457 0.999798175 1.0000000000

> plot(Nutritional.pca.cor,type="line",main="Scree Plot")

Chart, line chart

Description automatically generated

**Question:**

1. How many principal components are necessary for this analysis?

Plotting the first four principal components…

First, creating a label for the plot

> mylabel=as.factor(c(rep("W",28),rep("M",28),rep("B",24)))

Using the pairs function to create the pairwise plots…

> pairs(Nutritional.pca.cor$scores[,1:4],

panel=function(x,y){text(x,y,mylabel,col=as.numeric(mylabel))})

Calendar

Description automatically generated

Using the pairs function to create the pairwise plots (using item as a label)

> pairs(Nutritional.pca.cor$scores[,1:4],

panel=function(x,y){text(x,y,Nutritional[,2],cex=0.50)})

Diagram, engineering drawing

Description automatically generated

> pairs(Nutritional.pca.cor$scores[,1:4],

panel=function(x,y){text(x,y,Nutritional[,2],col=as.numeric(as.factor(Nutritional[,3])),cex=0.50)})

Calendar

Description automatically generated

Obtaining the biplot for this analysis…

> biplot(Nutritional.pca.cor)

Chart, scatter chart

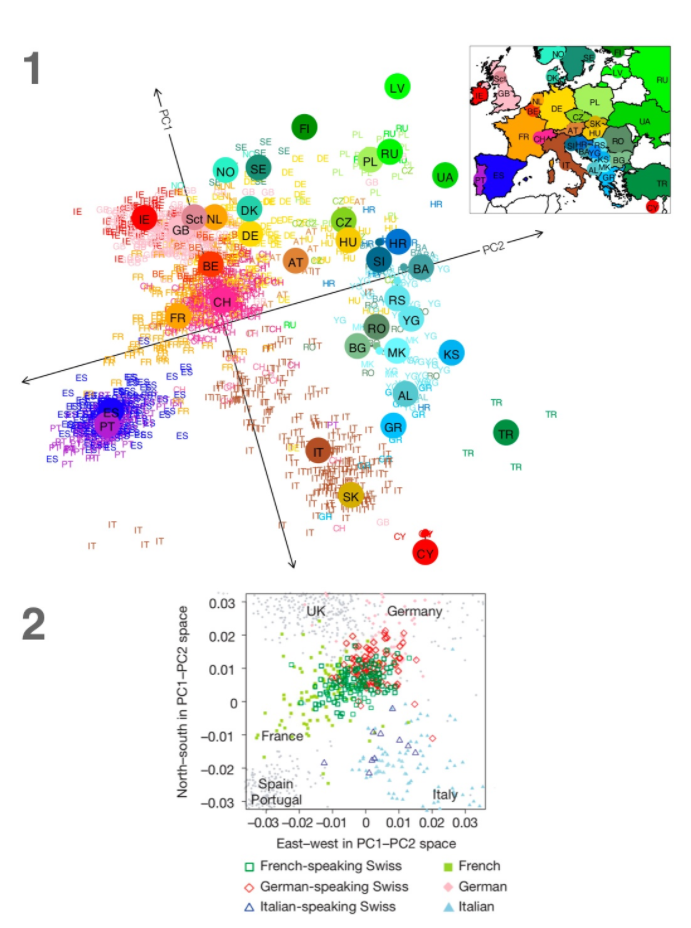
Description automatically generated

> outlying <- c(21,47,56,53,45,12,30,54,58,6,3,69)

Table

Description automatically generated

Example 4.5: Genetic Analysis of Europeans



Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/>

Example 4.6: Acetaminophen Liver Toxicity Study (**Liver Toxicity Study.csv**)

The data come from a liver toxicity study (Bushel et al., 2007) in which male rats of the inbred strain Fisher 344 were exposed to non-toxic (50 or 150 mg/kg), moderately toxic (1500 mg/kg), or severely toxic (2000 mg/kg) doses of acetaminophen (paracetamol) in a controlled experiment. Necropsies were performed at 6, 18, 24 and 48 hours after exposure and the mRNA from the liver was extracted. Ten clinical chemistry measurements of variables containing markers for liver injury are available for each subject and the serum enzymes levels are measured numerically. The data were further normalized and pre-processed by Bushel et al. (2007).

Variables

* *Animal ID* – ID number for rat
* *Treatment.Group* – conveys both dose level and time animal was sacrificed
* *Dose.Group* – 50, 150, 1500, or 2000 mg/kg
* *Time.Group* – time sacrificed (6,18, 24, or 48 hrs.)

10 clinical markers for liver injury and serum enzyme levels

* *BUN.mg.dL, Creat.mg.dL, TP.g.dL, ALB.g.dL, ALT.IU.L, SDH.IU.L, AST.IU.I, ALP.IU.L, TBA.umol.L, Cholesterol.mg.dL*

3,116 genetic expression levels

* *A\_43\_P14555, A\_45\_P2220, … , A\_42\_P720241*

Clearly these data are high dimensional and therefore a good candidate for employing dimension reduction methods.

> LiverTox = read.csv(<http://course1.winona.edu/bdeppa/DSCI%20415/Data/Liver%20Toxicity%20Study.csv>)

> dim(LiverTox)

[1] 64 3130

> names(LiverTox)[1:30]

[1] "Animal.ID" "Treatment.Group" "Dose.Group"

[4] "Time.Group" "BUN.mg.dL" "Creat.mg.dL"

[7] "TP.g.dL" "ALB.g.dL" "ALT.IU.L"

[10] "SDH.IU.L" "AST.IU.L" "ALP.IU.L"

[13] "TBA.umol.L" "Cholesterol.mg.dL" "A\_43\_P14555"

[16] "A\_43\_P22290" "A\_43\_P20792" "A\_43\_P21286"

[19] "A\_43\_P12995" "A\_43\_P15834" "A\_43\_P12356"

[22] "A\_42\_P564516" "A\_43\_P22018" "A\_43\_P21075"

[25] "A\_43\_P23125" "A\_43\_P17653" "A\_43\_P18129"

[28] "A\_43\_P16337" "A\_43\_P22160" "A\_43\_P23348"

… LOTS MORE mRNA VARIABLES !!!

> gene.mat = LiverTox[,-c(1:14)]

> dim(gene.mat)

[1] 64 3116

> gene.mat = scale(gene.mat)

> gene.cor = cor(gene.mat)

> gene.cor2 = apply(gene.cor,2,rev)

> par(pty="s")

> image(1:3116,1:3116,gene.cor2)

The image plot of the correlations is shown on the following page. Does there appear to be some correlation structure that could be useful in reducing the dimensionality of the mRNA levels?

A picture containing chart

Description automatically generated

We will now try to use principal components to reduce the dimensionality of these data.

> gene.PCA = prcomp(gene.mat)

> summary(gene.PCA)

Importance of components%s:

PC1 PC2 PC3 PC4 PC5 PC6 PC7

Standard deviation 29.577 21.520 15.7528 12.9452 11.9146 11.0600 10.4290

Proportion of Variance 0.281 0.149 0.0796 0.0538 0.0456 0.0393 0.0349

Cumulative Proportion 0.281 0.429 0.5090 0.5628 0.6083 0.6476 0.6825

PC8 PC9 PC10 PC11 PC12 PC13 PC14 PC15

Standard deviation 8.8174 8.6130 7.5137 7.3379 6.7104 6.4615 5.7705 5.49484

Proportion of Variance 0.0249 0.0238 0.0181 0.0173 0.0144 0.0134 0.0107 0.00969

The first principal components together explain 90.14% of the total variation in the dimensional data! Nice!!!

Using only dimensions gets us to over 50% which if you think about it is not bad.

Cumulative Proportion 0.7075 0.7312 0.7494 0.7667 0.7811 0.7945 0.8052 0.81488

PC16 PC17 PC18 PC19 PC20 PC21 PC22

Standard deviation 5.32941 5.09056 4.94436 4.89087 4.73424 4.61732 4.54974

Proportion of Variance 0.00912 0.00832 0.00785 0.00768 0.00719 0.00684 0.00664

Cumulative Proportion 0.82399 0.83231 0.84015 0.84783 0.85502 0.86187 0.86851

PC23 PC24 PC25 PC26 PC27 PC28 PC29

Standard deviation 4.2869 4.23196 4.20055 4.06784 4.00914 3.99912 3.83156

Proportion of Variance 0.0059 0.00575 0.00566 0.00531 0.00516 0.00513 0.00471

Cumulative Proportion 0.8744 0.88015 0.88582 0.89113 0.89629 0.90142 0.90613

PC30 PC31 PC32 PC33 PC34 PC35 PC36

Standard deviation 3.70055 3.68072 3.60669 3.58842 3.56746 3.46103 3.42442

Proportion of Variance 0.00439 0.00435 0.00417 0.00413 0.00408 0.00384 0.00376

Cumulative Proportion 0.91052 0.91487 0.91905 0.92318 0.92726 0.93111 0.93487

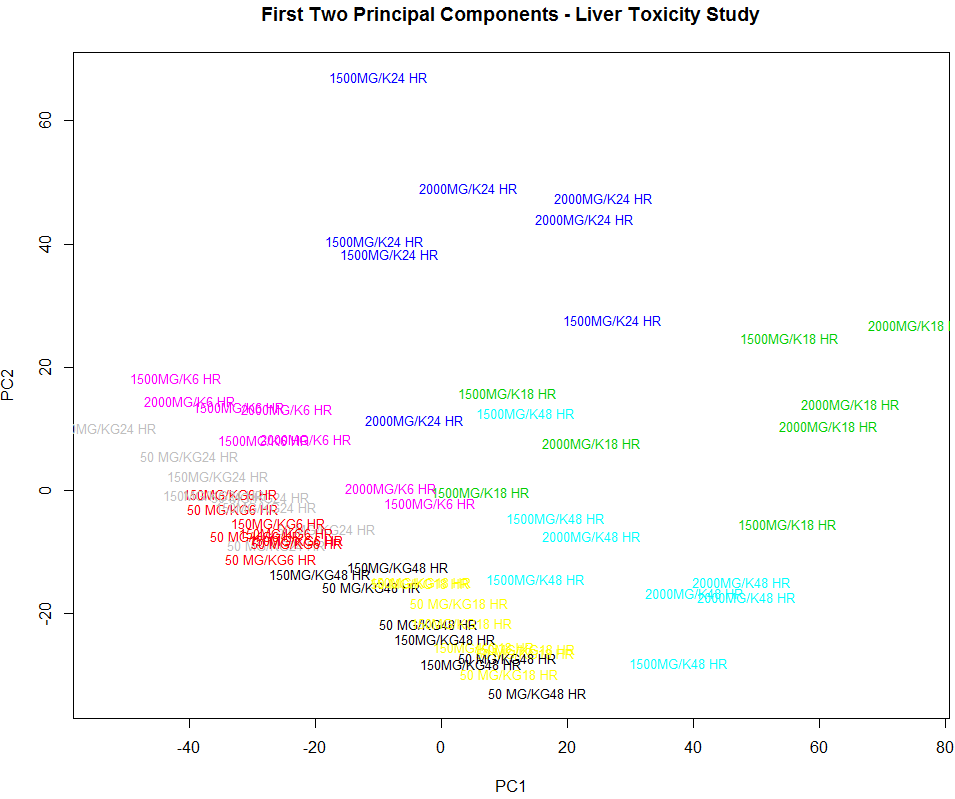
> PC1 = gene.PCA$x[,1]

> PC2 = gene.PCA$x[,2]

> PC3 = gene.PCA$x[,3]

> plot(PC1,PC2,type="n")

> text(PC1,PC2,as.character(LiverTox$Treatment.Group),col=as.numeric(LiverTox$Dose.Group)+3,cex=.8)

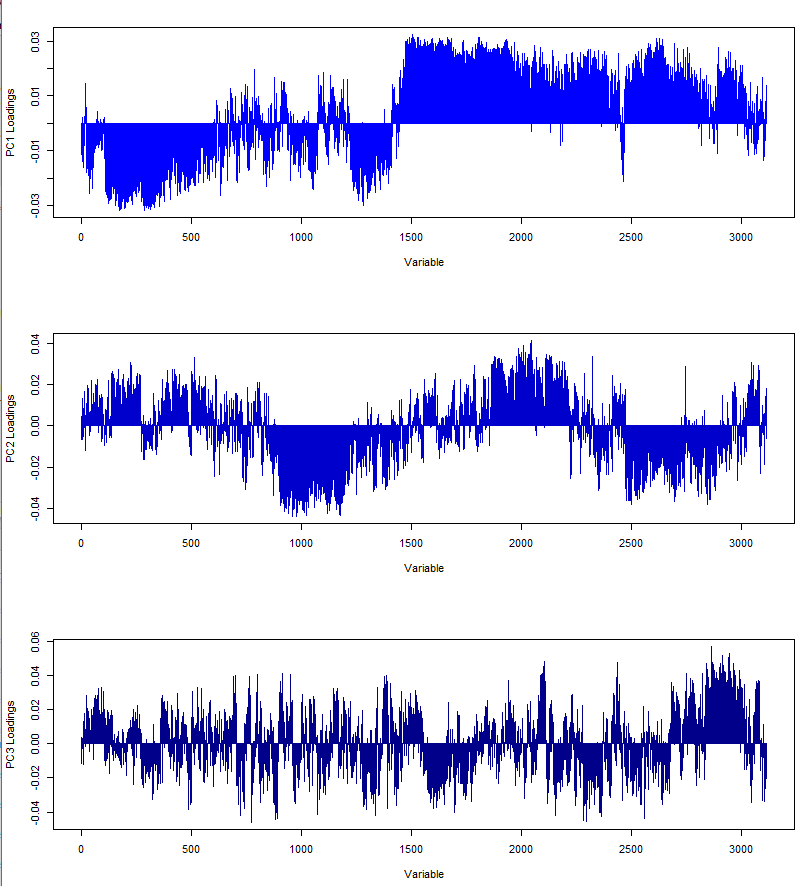
  
  
The treatment and/or dose groups are fairly well separated in the plot of the first two principal component scores. The low dose rats all lie along a diagonal in the lower left portion of this plot with higher doses and sacrifice times located away from this diagonal.

Interpreting loadings with 3,116 variables is very difficult. Below are loading plots for the first three principal components drawn as a needle chart.  
> par(mfrow=c(3,1))

> plot(gene.PCA$rotation[,1],type="h",col="blue",xlab="Variable",ylab="PC1 Loadings")

> plot(gene.PCA$rotation[,2],type="h",col="mediumblue",xlab="Variable",ylab="PC2 Loadings")

> plot(gene.PCA$rotation[,3],type="h",col="darkblue",xlab="Variable",ylab="PC3 Loadings")



We could also sort the absolute loadings and look at the top few loadings.  
> pc1.load = gene.PCA$rotation[,1]

> pc1.load = abs(pc1.load)

> pc1.load = sort(pc1.load,decreasing=T)  
> pc1.load[1:50]

A\_43\_P19261 A\_43\_P10804 A\_42\_P681326 A\_42\_P765304 A\_42\_P562025

0.0322200 0.0317527 0.0316614 0.0316062 0.0314743

A\_42\_P825183 A\_43\_P13921 A\_42\_P749154 A\_43\_P13006 A\_42\_P555125

0.0314720 0.0314507 0.0314278 0.0314070 0.0313965

A\_42\_P578238 A\_43\_P22882 A\_42\_P504523 A\_42\_P562984 A\_42\_P771848

0.0313338 0.0312428 0.0311493 0.0311346 0.0311342

A\_42\_P563582 A\_42\_P513461 A\_43\_P12802 A\_43\_P14985 A\_43\_P14959

0.0311277 0.0311150 0.0311108 0.0310841 0.0309893

A\_43\_P20773 A\_42\_P616351 A\_43\_P12673 A\_43\_P22062 A\_42\_P478951

0.0309586 0.0309263 0.0308861 0.0308531 0.0308105

A\_43\_P11789 A\_42\_P664274 A\_42\_P732544 A\_42\_P793765 A\_43\_P21298

0.0307789 0.0307784 0.0307217 0.0307033 0.0306839

A\_43\_P22469 A\_42\_P460185 A\_43\_P12214 A\_43\_P20311 A\_42\_P795467

0.0306154 0.0306086 0.0305359 0.0305311 0.0305301

A\_42\_P756235 A\_43\_P22032 A\_42\_P529681 A\_42\_P790848 A\_43\_P13008

0.0305173 0.0304938 0.0304722 0.0304449 0.0304402

A\_43\_P15527 A\_42\_P654031 A\_43\_P17696 A\_42\_P722904 A\_42\_P828727

0.0304205 0.0304180 0.0303869 0.0303857 0.0303731

A\_42\_P815698 A\_43\_P12723 A\_42\_P601568 A\_42\_P520211 A\_43\_P10744

0.0303482 0.0303413 0.0303365 0.0303205 0.0302957

> plot3d(PC1,PC2,PC3,type="n")  
> text3d(PC1,PC2,PC3,texts=as.character(Treatment.Group),col=as.numeric(Dose.Group),cex=.6)

Chart

Description automatically generated with medium confidence