

## Examples of Principal Component Plots

This handout presents examples of two types of principal component plots.

The first illustrates the use of principal components based on an estimated variance matrix of *residuals* from a linear model, that is on purely random variation. Such analysis is useful in studying the structure of covariance matrices and for assessing difference between groups of covariance matrices. It is not useful in finding linear combinations that differ between groups.

The second and third illustrate the use of principal components based on a variance matrix computed from heterogeneous data, ignoring any groups. Among other things, this illustrates the usefulness of principal components in displaying differences in means between groups.

The first two examples are based on the Fisher Iris data in matrix T11\_05 in file JWData5.txt. This consists of 4-dimensional vectors of measurements of blossoms from three iris varieties. There are data from 50 blossoms of each variety in the data set.

### Principal component plots based on residuals from a linear model

The coefficients of the first principal component are the coefficients of the linear combination of the four variables that has the greatest *residual* variance after removing the variety means, and similarly for the remaining PC's. The coefficients used to compute PC's are either elements the eigenvectors of the pooled variance matrix **S** or of the right singular vectors of the data matrix after subtracting the sample mean.

The variables (PC's) plotted are the corresponding linear combinations of the responses  $x_1, \dots, x_4$ . This uses information about the within group covariance matrix to compute principal components, but preserves the between group information. Where there are differences among variety means of the original responses, there will generally also be difference in variety means among the principal components.

```
Cmd> irisdata <- read("", "t11_05", quiet:T) #read jwdata5.txt
Read from file "TP1:Stat5401:Data:JWData5.txt"

Cmd> varieties <- irisdata[,1]; y <- irisdata[,-1]

Cmd> setosa <- y[varieties == 1,]; versicolor <- y[varieties == 2,]
Cmd> virginica <- y[varieties == 3,]

Cmd> list(setosa, versicolor, virginica)
setosa          REAL    50    4    (labels)
versicolor      REAL    50    4    (labels)
virginica        REAL    50    4    (labels)

Cmd> s1 <- tabs(setosa, covar:T)
Cmd> s2 <- tabs(versicolor, covar:T)
Cmd> s3 <- tabs(virginica, covar:T)
```

## Principal Components Example

```
Cmd> spooled <- (49*s1+49*s2+49*s3)/147;spooled # pooled var matrix
(1,1)      0.26501      0.092721      0.16751      0.038401
(2,1)      0.092721      0.11539      0.055244      0.03271
(3,1)      0.16751      0.055244      0.18519      0.042665
(4,1)      0.038401      0.03271      0.042665      0.041882
```

Or you could compute spooled from the one-way MANOVA error matrix

```
Cmd> varieties <- factor(varieties) # make sure varieties is a factor
Cmd> manova("y = varieties",silent:T) # silent:T suppresses output
Cmd> e <- matrix(SS[3,,]); fe <- DF[3]
```

```
Cmd> spooled <- e/fe; spooled
      SepLen      SepWid      PetLen      PetWid
SepLen  0.26501      0.092721      0.16751      0.038401
SepWid  0.092721      0.11539      0.055244      0.03271
PetLen  0.16751      0.055244      0.18519      0.042665
PetWid  0.038401      0.03271      0.042665      0.041882
```

```
Cmd> r <- cor(RESIDUALS) # correlations of residuals needed below
```

```
Cmd> eigs <- eigen(spooled); eigs
component: values
(1)      0.44357      0.086183      0.055352      0.022364
component: vectors
      (1)      (2)      (3)      (4)
SepLen  0.73775      0.056086      0.63238      0.22951
SepWid  0.32057     -0.87323     -0.18057     -0.31953
PetLen  0.57285      0.45883     -0.58182     -0.35042
PetWid  0.15748     -0.15425     -0.47851      0.84996
```

```
Cmd> pcomp <- y %*% eigs$vectors # columns are principal components
```

```
Cmd> list(pcomp)
pcomp      REAL      150      4      (labels)
```

If you do MANOVA using the principal components as response, the error matrix is diagonal. This is the case because the PCs were computed from the pooled within group variance matrix.

```
Cmd> manova("pcomp = varieties", silent:T)
Cmd> SS[3,,]/DF[3] # Diagonal with eigenvalues on the diagonal
      (1)      (2)      (3)      (4)
ERROR1 (1)      0.44357 -1.0441e-16 -2.3602e-17 -1.195e-16
      (2) -1.0441e-16      0.086183  4.9091e-18 -9.2991e-18
      (3) -2.3602e-17  4.9091e-18      0.055352  7.1655e-17
      (4) -1.195e-16 -9.2991e-18  7.1655e-17      0.022364
```

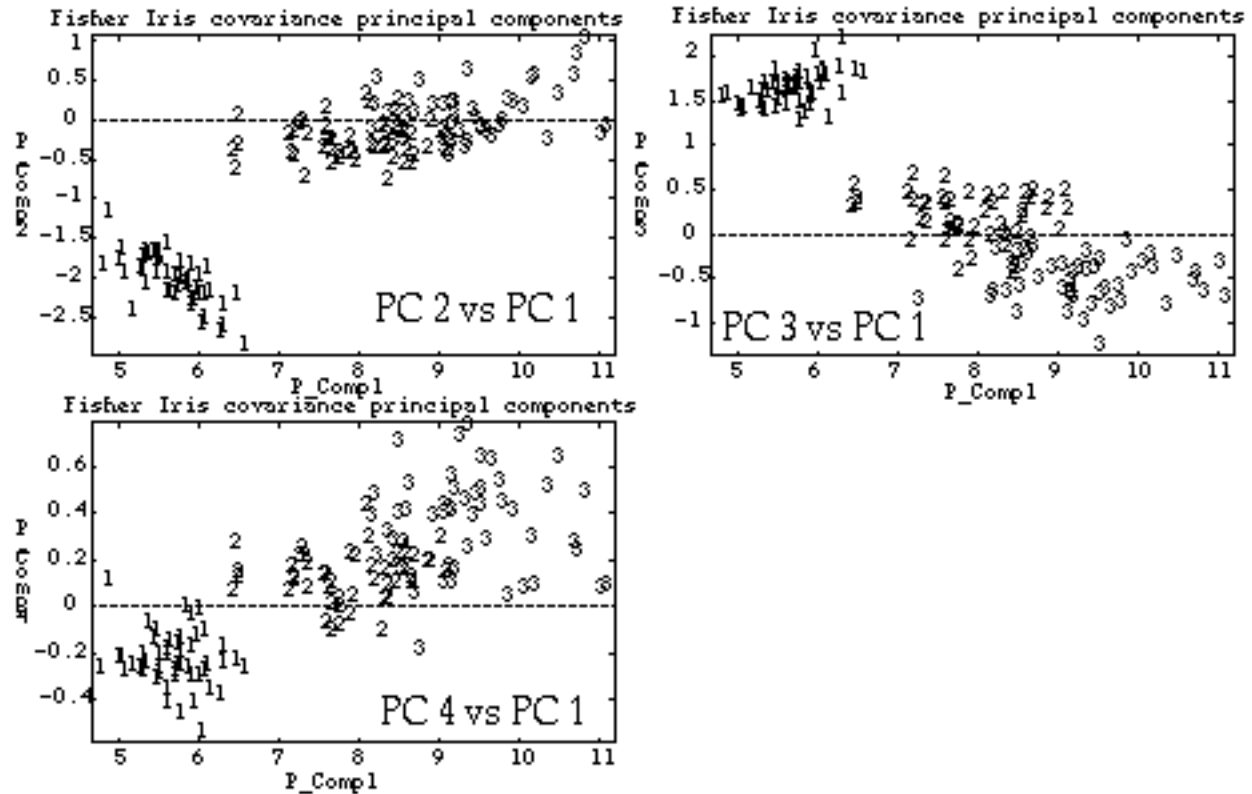
## Principal Components Example

Now make plots of the first 3 PC's vs each other.

```
Cmd> plot(P_Compl:pcomp[,1],P_Comp2:pcomp[,2],symbols:varieties,\
         title:"Fisher Iris covariance principal components")
```

```
Cmd> plot(P_Compl:pcomp[,1],P_Comp3:pcomp[,3],symbols:varieties,\
         title:"Fisher Iris covariance principal components")
```

```
Cmd> plot(P_Compl:pcomp[,1],P_Comp4:pcomp[,4],symbols:varieties,\
         title:"Fisher Iris covariance principal components")
```



Note the different orientations of the clouds of points. This indicates that the covariance matrices are almost certainly not identical.

You can also compute principal components from the eigenvectors of the correlation matrix  $r$ . They are different from the covariance principal components. We computed  $r$  previously.

```
Cmd> r
      SepLen  SepWid  PetLen  PetWid
SepLen      1    0.53024  0.75616  0.36451
SepWid    0.53024      1    0.37792  0.47053
PetLen    0.75616    0.37792      1    0.48446
PetWid    0.36451    0.47053    0.48446      1
```

You could also compute  $r$  from  $\text{spooled}$  by pre-multiplying and post-multiplying  $S$  by a diagonal matrix with  $1/\sqrt{s_{ii}}$  on the diagonals.

## Principal Components Example

```

Cmd> eigsr <- eigen(r);eigsr # find eigenvalues and vectors of r
component: values
(1)      2.5038      0.72514      0.5824      0.1887
component: vectors
              (1)      (2)      (3)      (4)
SepLen      0.5424     -0.45697     0.21498     0.67139
SepWid      0.46638     0.46647     0.69656    -0.28232
PetLen      0.53483    -0.45341    -0.31393    -0.64017
PetWid      0.44971     0.60663    -0.60831     0.24436

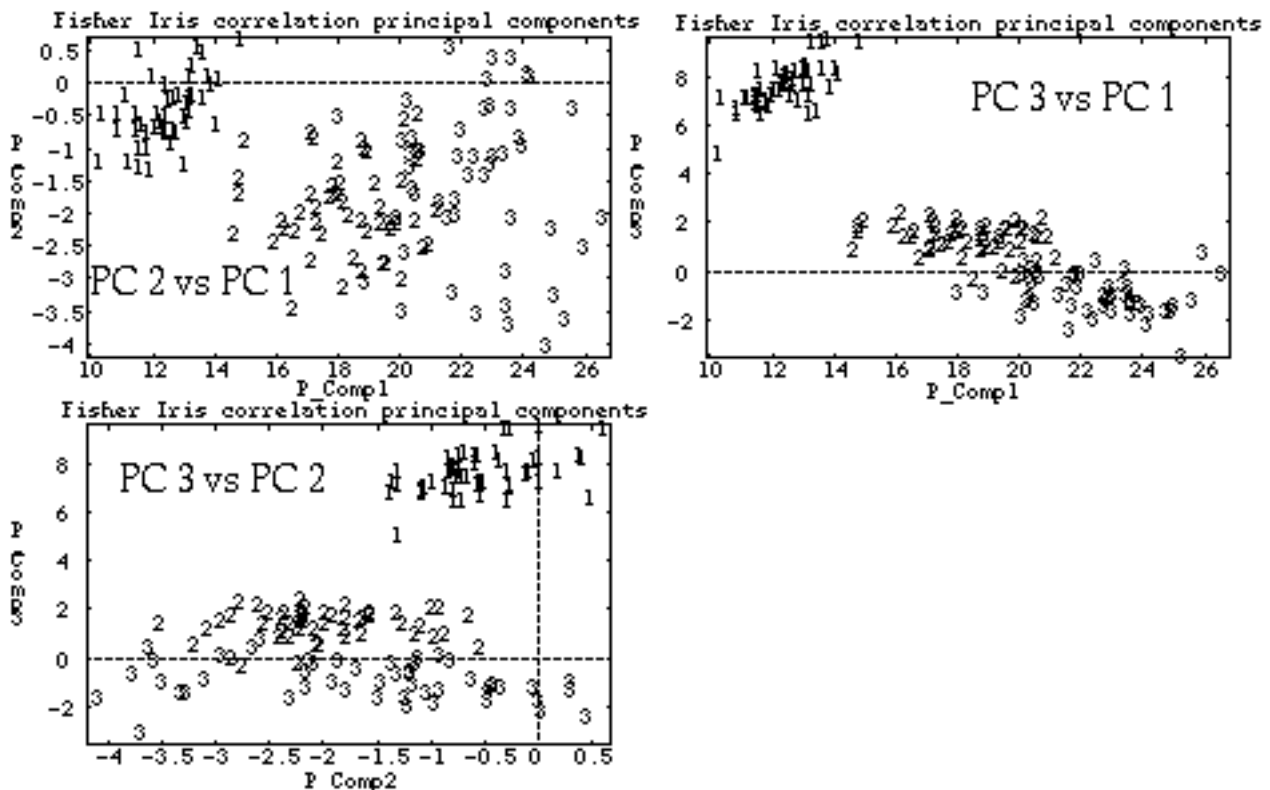
```

Now compute the *correlation* principal components. You need to divide the rows of the eigenvectors by the standard deviations to compensate for the fact that the eigenvectors come from the correlation matrix, that is the variance matrix of standardized variables. This is needed to get coefficients that can be directly applied to the data matrix. Alternatively you could divide the columns of  $y$  by the standard deviations and use the eigenvectors as they are computed. That's what we do here.

```

Cmd> sd <- sqrt(diag(spoiled)) # find standard deviations
Cmd> pcompsr <- (y/sd') %%% eigsr$vectors
Cmd> plot(P_Comp1:pcompsr[,1],P_Comp2:pcompsr[,2],symbols:varieties,\
         title:"Fisher Iris correlation principal components")
Cmd> plot(P_Comp1:pcompsr[,1],P_Comp3:pcompsr[,3],symbols:varieties,\
         title:"Fisher Iris correlation principal components")
Cmd> plot(P_Comp2:pcompsr[,2],P_Comp3:pcompsr[,3],symbols:varieties,\
         title:"Fisher Iris correlation principal components")

```



## Principal Components Example

### Principal component plots based on heterogeneous data

Principal components analysis is often viewed as a method to analyze an estimated covariance or correlation matrix, derived either from a homogeneous sample, or from residuals from a linear model as in the first example. However, the basic computations are often informative when applied to data that may be heterogeneous data, perhaps sampled from a mixture of several populations.

Consider a situation in which there may be several distinct but *non-identified* sub-populations of the population being sampled, perhaps subspecies or varieties. Then a sample will contain observations from some or all of these sub-populations.

To be specific, suppose you have observations  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N$ ,  $n_1$  of which come from a population with mean  $\boldsymbol{\mu}_1$  and variance matrix  $\boldsymbol{\Sigma}_1$ ,  $n_2$  from a population with mean  $\boldsymbol{\mu}_2$  and variance matrix  $\boldsymbol{\Sigma}_2$ , and so on. What is different from an ordinary one-way MANOVA situation is that (a) you do not know the number  $g$  of subpopulations, and (b) you cannot *a priori* associate a data point with a specific subpopulation. This further implies that you do not know  $n_1, n_2, \dots$ . What you would like to do is display the data in a way that is likely to expose the differences between groups, perhaps with the aim of identifying the subpopulations.

If  $\mathbf{S} \equiv \frac{1}{N-1} \sum_{i=1}^N (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$  is the sample variance matrix computed from the undifferentiated data, then its expectation is  $E[\mathbf{S}] = \boldsymbol{\Sigma}^* + \boldsymbol{\Delta}$ , where

$$\boldsymbol{\Sigma}^* = \frac{1}{N-1} \sum_{j=1}^g (n_j - 1) \boldsymbol{\Sigma}_j, \boldsymbol{\Delta} = \frac{1}{N-1} \sum_{j=1}^g n_j (\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})', \bar{\boldsymbol{\mu}} = \frac{1}{N} \sum_{j=1}^g n_j \boldsymbol{\mu}_j.$$

Note  $\boldsymbol{\Sigma}^*$  involves only the covariance matrices  $\boldsymbol{\Sigma}_j$ , while  $\boldsymbol{\Delta}$  involves only the  $\boldsymbol{\mu}$ 's. Moreover,  $\boldsymbol{\Delta}$  is  $\mathbf{0}$  only if all the subpopulation means are the same, and the more the means are separated, the larger is  $\boldsymbol{\Delta}$ .

What matters, of course, is how separated the means are *relative* to within group variance as measured by  $\boldsymbol{\Sigma}^*$ . When  $\boldsymbol{\Delta}$  is large in comparison with  $\boldsymbol{\Sigma}^*$ , the eigenvalues and eigenvectors of  $\mathbf{S}$  or  $\mathbf{R}$  will primarily reflect the structure of  $\boldsymbol{\Delta}$  and scatter plots of the first few principal components against each other will tend to emphasize differences among the unknown group means. On the other hand, when  $\boldsymbol{\Delta}$  is not large in comparison with  $\boldsymbol{\Sigma}^*$ , the eigenvalues and eigenvectors of  $\mathbf{S}$  will primarily reflect the eigen structure of  $\boldsymbol{\Sigma}^*$  which depends only on the within groups variance matrices. In this case, because  $\boldsymbol{\Sigma}^*$  is a weighted average of the separate variance matrices, the eigen structure will tell you little.

When the variables differ substantially in scale, it is almost always better to plot the principal components based on the correlation matrix  $\mathbf{R} = \mathbf{D} \mathbf{S} \mathbf{D}$ , where  $\mathbf{D} = \text{diag}[1/\sqrt{s_{11}}, 1/\sqrt{s_{22}}, \dots, 1/\sqrt{s_{pp}}]$ . If  $\mathbf{u}_j = [u_{1j}, \dots, u_{pj}]'$  is the  $j$ -th eigenvector of  $\mathbf{R}$ , then the coefficients of the  $j$ -th *correlation* principal component are the elements of the vector  $\tilde{\mathbf{u}}_j = \mathbf{D} \mathbf{u}_j = [u_{1j}/\sqrt{s_{11}}, u_{2j}/\sqrt{s_{22}}, \dots, u_{pj}/\sqrt{s_{pp}}]'$ , that is the rows of  $\mathbf{u}_j$  are divided by the sample standard deviations  $\sqrt{s_{kk}}$  of the  $Y$ 's.

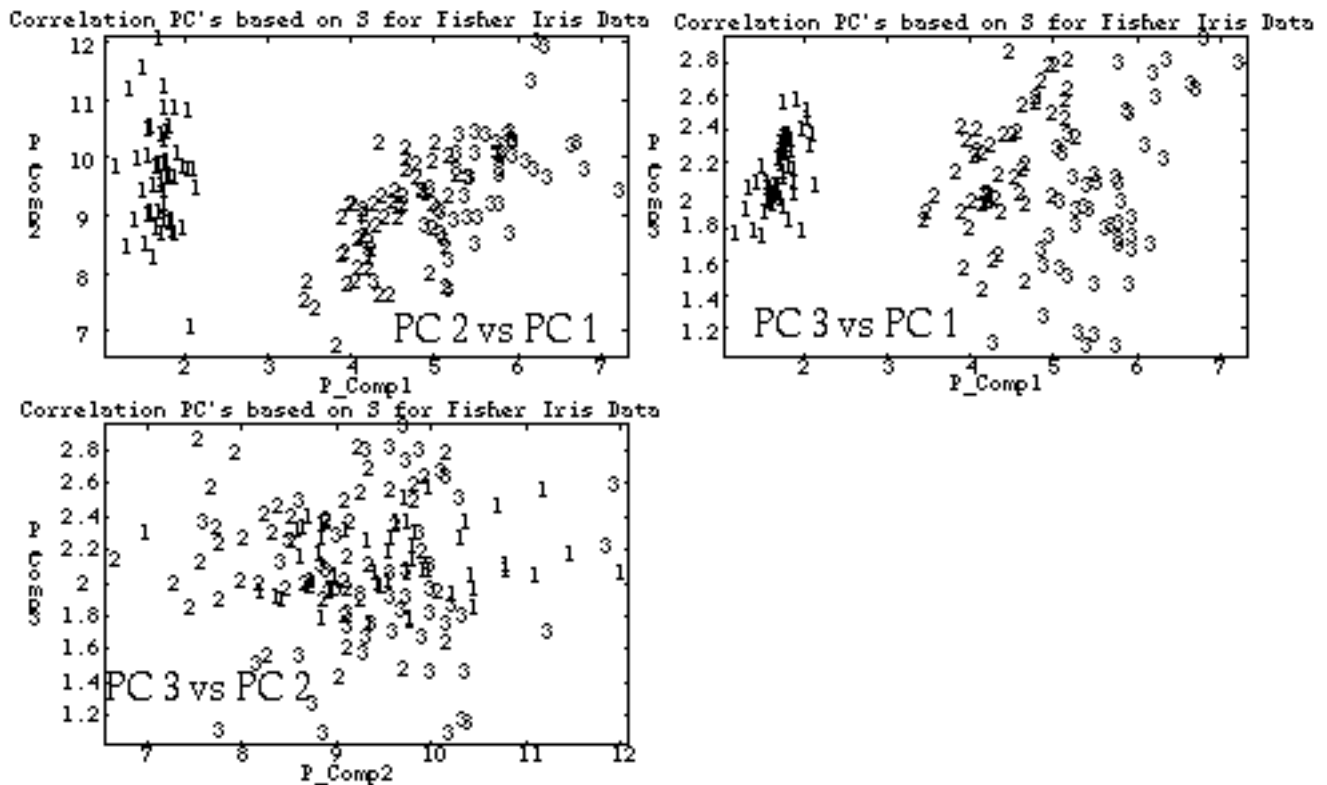
## Principal Components Example

Here is a continuation of the analysis of the Fisher Iris data.

```
Cmd> s <- tabs(y,covar:T) # var. of all data, ignoring varieties
Cmd> r <- cor(y); r # correlation matrix, ignoring varieties
      SepLen      SepWid      PetLen      PetWid
SepLen      1      -0.11757      0.87175      0.81794
SepWid     -0.11757      1      -0.42844     -0.36613
PetLen      0.87175     -0.42844      1      0.96287
PetWid      0.81794     -0.36613      0.96287      1
Cmd> eigs <- eigen(r);eigs # eigen structure of correlation matrix
component: values
(1)      2.9185      0.91403      0.14676      0.020715
component: vectors
      (1)      (2)      (3)      (4)
SepLen      0.52107      0.37742      0.71957      0.26129
SepWid     -0.26935      0.9233      -0.24438     -0.12351
PetLen      0.58041      0.024492     -0.14213     -0.80145
PetWid      0.56486      0.066942     -0.63427      0.5236
Cmd> sd <- sqrt(diag(s)) # standard deviations
Cmd> princomps <- (y/sd') %*% eigs$vectors
```

## Principal Components Example

```
Cmd> plot(P_Comp1:princomps[,1],P_Comp2:princomps[,2],\
        symbols:varieties,title:\
        "Correlation principal components based on S for Fisher Iris Data")
Cmd> plot(P_Comp1:princomps[,1],P_Comp3:princomps[,3],\
        symbols:varieties,title:\
        "Correlation principal components based on S for Fisher Iris Data")
Cmd> plot(P_Comp2:princomps[,2],P_Comp3:princomps[,3],\
        symbols:varieties,title:\
        "Correlation principal components based on S for Fisher Iris Data")
```

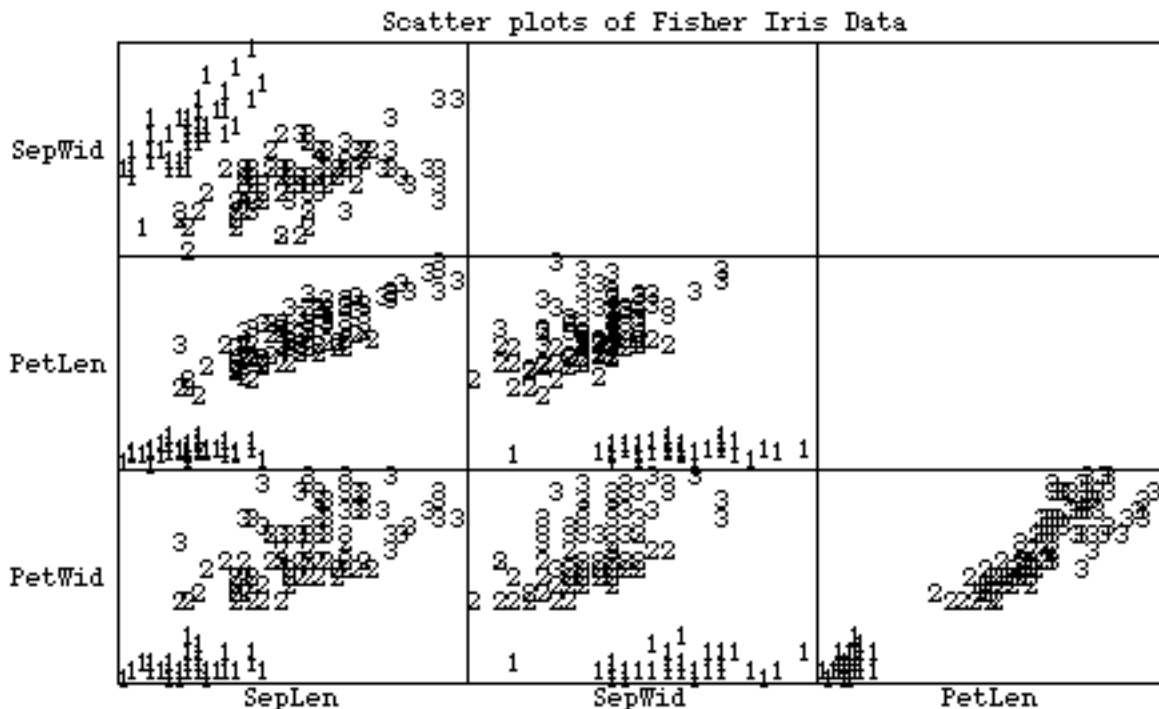


The upper left plot, of the first and second principal components, does a pretty good job of separating the three varieties.

By way of comparison, here are all 6 scatter plots of pairs of the actual variables in the data set created using `macro plotmatrix()`.

## Principal Components Example

```
Cmd> plotmatrix(y,lower:T,symbols:varieties,\
  title:"Scatter plots of Fisher Iris Data",xlab:"",ylab:"",\
  labels:vector("SepLen","SepWid","PetLen","PetWid"))
```



For every pair of variables, variety 1 is well separated from the other two varieties, so we should not be surprised that the principal components reveal this separation. And for every pair there is substantial overlap between varieties 2 and 3. It is not clear there is any gain from the use of principal components.

As a second example, we analyze the multivariate data in file `cbspots.txt`. Each of the  $p = 19$  variables is the density of a spot on an autoradiograph of a sample of blood from a rat. The density is presumed proportional to the amount of a particular protein in the blood. The rats had been subjected to 10 treatment, one of which, treatment 2 or B, was a control (no treatment). In this analysis, the known treatment structure is ignored in computing principal components. To display how well principal components reveal the underlying structure, the points in the principal component plots are labelled A, B, ..., J to show group membership. The analysis is in terms of  $\log_{10}(y+1)$  because  $y$  is very skew.

```
Cmd> spots <- read("", "spots", quiet:T) # read from cbspots.txt
Read from file "TP1:Stat5401:Data:cbspots.txt"

Cmd> treatment <- spots[,1]; y <- log10(spots[,-1] + 1)

Cmd> s <- tabs(y, covar:T) # covariance matrix ignoring varieties

Cmd> eigs <- eigen(s); eigs$values
(1)      1.896      1.0798      0.29923      0.19009      0.12589
(6)      0.076357    0.069458    0.051192    0.034265    0.025372
(11)     0.023382    0.017485    0.012756    0.0087761   0.0085185
(16)     0.006892    0.0052675   0.0049177   0.0034336

Cmd> zc <- y %*% eigs$vectors # Compute principal components
```



## Principal Components Example

```
Cmd> print(format:"7.4f",tabs(zc,covar:T)[run(8),run(8)])
```

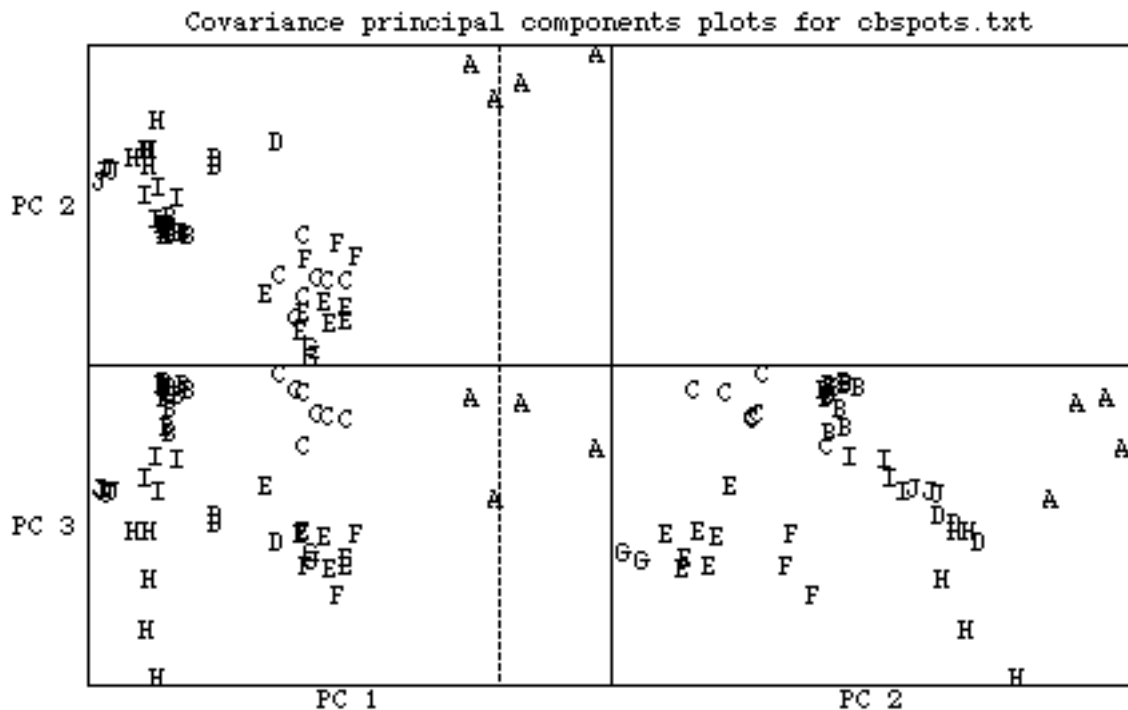
MATRIX:

```
(1,1)  1.8960  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
(2,1)  0.0000  1.0798  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
(3,1)  0.0000  0.0000  0.2992  0.0000  0.0000  0.0000  0.0000  0.0000
(4,1)  0.0000  0.0000  0.0000  0.1901  0.0000  0.0000  0.0000  0.0000
(5,1)  0.0000  0.0000  0.0000  0.0000  0.1259  0.0000  0.0000  0.0000
(6,1)  0.0000  0.0000  0.0000  0.0000  0.0000  0.0764  0.0000  0.0000
(7,1)  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0695  0.0000
(8,1)  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0512
```

```
Cmd> # note PC's are uncorrelated and have eigenvalues as variance
```

```
Cmd> labs <-\
      vector("A","B","C","D","E","F","G","H","I","J")[treatment]
```

```
Cmd> plotmatrix(zc[,run(3)],lower:T,symbols:labs,xlab:"",ylab:"",\
labels:vector("PC 1","PC 2","PC 3"),\
title:"Covariance principal components plots for cbspots.txt")
```



```
Cmd> r <- cor(y) # compute correlation matrix
```

```
Cmd> print(format:"7.4f",r[run(8),run(8)]) # part of corr matrix
```

MATRIX:

	Sp01	Sp02	Sp03	Sp04	Sp05	Sp06	Sp07	Sp08
Sp01	1.0000	0.7777	0.7129	0.8913	-0.0812	0.6958	0.7880	-0.3736
Sp02	0.7777	1.0000	0.9134	0.7711	-0.2579	0.8507	0.7318	-0.5619
Sp03	0.7129	0.9134	1.0000	0.6288	-0.3898	0.8645	0.6154	-0.5166
Sp04	0.8913	0.7711	0.6288	1.0000	-0.0128	0.6298	0.8320	-0.3870
Sp05	-0.0812	-0.2579	-0.3898	-0.0128	1.0000	-0.5055	-0.1085	0.3062
Sp06	0.6958	0.8507	0.8645	0.6298	-0.5055	1.0000	0.6362	-0.5377
Sp07	0.7880	0.7318	0.6154	0.8320	-0.1085	0.6362	1.0000	-0.3585
Sp08	-0.3736	-0.5619	-0.5166	-0.3870	0.3062	-0.5377	-0.3585	1.0000

## Principal Components Example

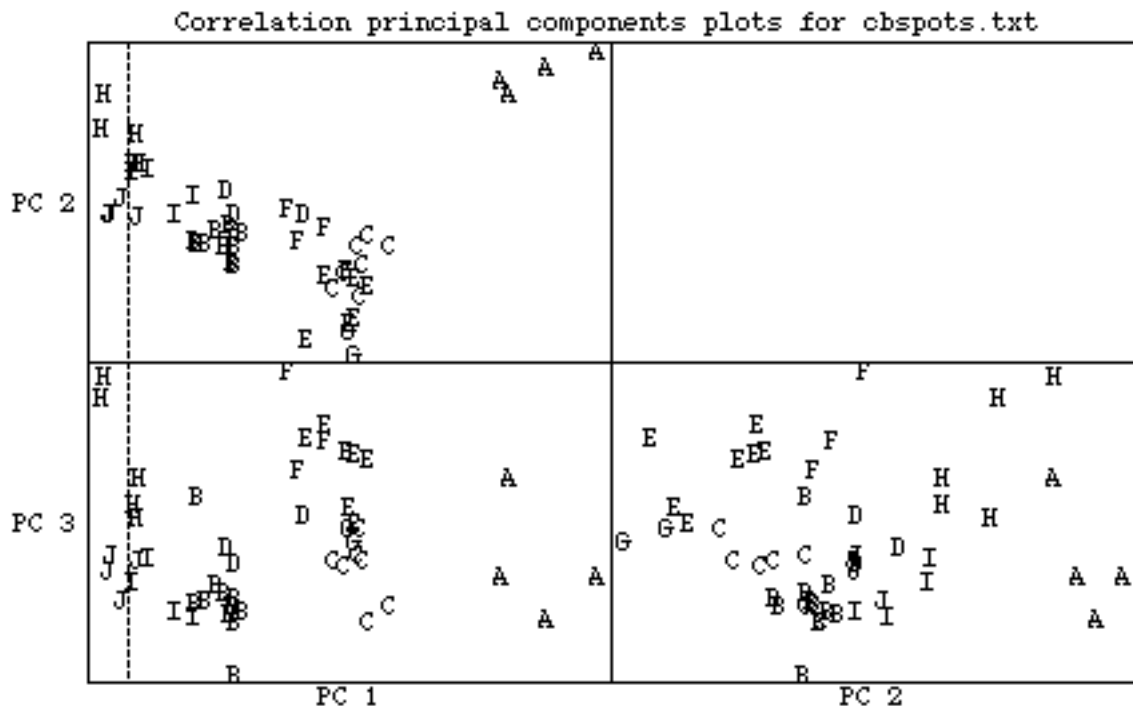
```

Cmd> # Note: labels came from labels in cbspots.txt
Cmd> eigs <- eigen(r);eigs$values # correlation eigenvalues
(1)      8.1623      3.7179      1.7374      1.1274      0.89645
(6)      0.76408     0.56048     0.53438     0.36238     0.2947
(11)     0.20819     0.16422     0.12944     0.10691     0.080667
(16)     0.054977    0.050571    0.02898     0.018455

Cmd> # There is less dominance of first 2 principal components
Cmd> sd <- sqrt(diag(s)) # standard deviations
Cmd> coeffs <- eigs$vectors / sd # divide rows by standard deviations
Cmd> zr <- y %%% coeffs # coefficients directly applicable to y's
Cmd> print(format:"7.4f",tabs(zr,covar:T)[run(8),run(8)])
MATRIX:      Part of covariance matrix of principal components
(1,1)  8.1623  0.0000  0.0000  0.0000 -0.0000 -0.0000  0.0000 -0.0000
(2,1)  0.0000  3.7179 -0.0000 -0.0000 -0.0000  0.0000  0.0000 -0.0000
(3,1)  0.0000 -0.0000  1.7374 -0.0000  0.0000 -0.0000 -0.0000 -0.0000
(4,1)  0.0000 -0.0000 -0.0000  1.1274  0.0000  0.0000  0.0000 -0.0000
(5,1) -0.0000 -0.0000  0.0000  0.0000  0.8964 -0.0000  0.0000 -0.0000
(6,1) -0.0000  0.0000 -0.0000  0.0000 -0.0000  0.7641 -0.0000 -0.0000
(7,1)  0.0000  0.0000 -0.0000  0.0000  0.0000 -0.0000  0.5605  0.0000
(8,1) -0.0000 -0.0000 -0.0000 -0.0000 -0.0000 -0.0000  0.0000  0.5344

Cmd> plotmatrix(zr[,run(3)],lower:T,symbols:labs,xlab:"",ylab:"",\
labels:vector("PC 1","PC 2","PC 3"),\
title:"Correlation principal components plots for cbspots.txt")

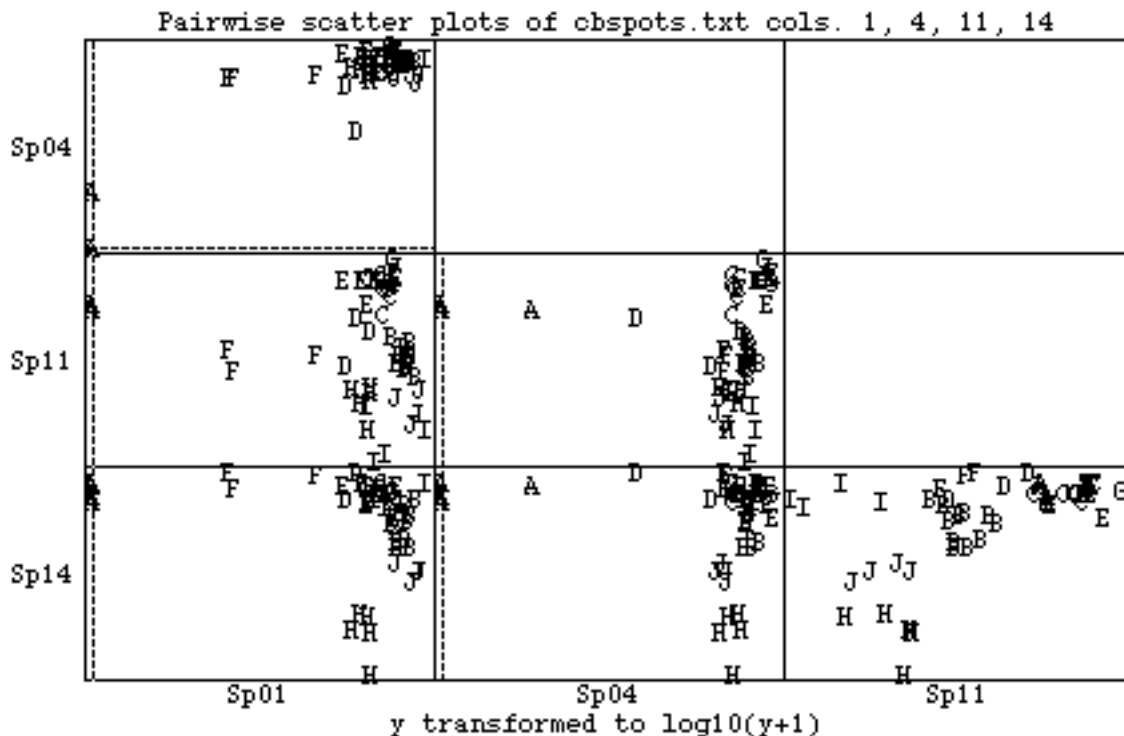
```



## Principal Components Example

Let's see how this analysis compares with making scatter plots of the original variables. It make sense to concentrate on the variables with the largest univariate F-statistics because you might expect that these would do a good job of displaying the differences among the treatment groups.

```
Cmd> treatment <- factor(treatment) # turn vector into factor
Cmd> manova("y=treatment",silent:T) # suppress output
Cmd> print(paste(TERMNAMES)) # names of the three MANOVA terms
CONSTANT treatment ERROR1
Cmd> fh <- DF[2]; fe <- DF[3] # hypothesis and error d.f.
Cmd> fstats <- (diag(SS[2,,])/fh)/(diag(SS[3,,])/fe); fstats
(1)      146.07      49.278      24.806      102.51      26.041
(6)      48.442      46.134      5.9153      52.793      20.797
(11)     68.431      20.052      20.809      57.768      7.6594
(16)     36.844       7.2989      5.3226      2.6324
Cmd> J <- grade(fstats,down:T)[run(4)];J# indices of 4 largest F's
(1)          1          4          11          14
Cmd> fstats[J] # the 4 largest F-statistics
(1)      146.07      102.51      68.431      57.768
Cmd> plotmatrix(y[,J],lower:T,symbols:labs,\
  xlab:"y transformed to log10(y+1)",ylab:"",\
  title:"Pairwise scatter plots of cbspots.txt cols. 1, 4, 11, 14")
```



Group A is very different from all the treated groups on both variables  $Y_1$  and  $Y_4$  and the plots show relatively little beside this. The plots show less intergroup separation than do the principal components plots.