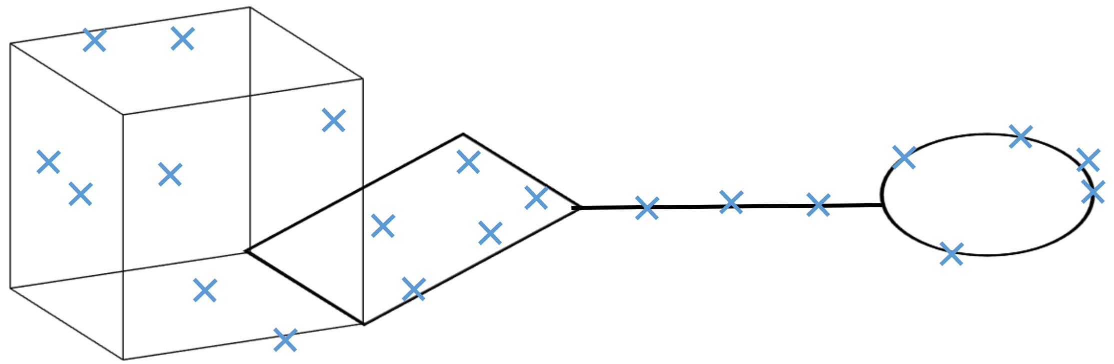
Data visualization

Dimension Reduction - Principal Components Analysis (PCA)

David Hoksza
http://siret..ms.mff.cuni.cz/hoksza

Motivation

- **Nominal** (**observed**) dimensionality = number of **measurements** for each observation
- **Intrinsic** (**true**) dimensionality = dimension of the **space actually covered** by the observations (number of dimensions needed to describe an observation)



- **Nominal** dimensionality of a set is **higher** or equal to the **intrinsic** dimensionality → finding a **projection** from the nominal space to the intrinsic space

Nominal vs intrinsic dimensionality in real data

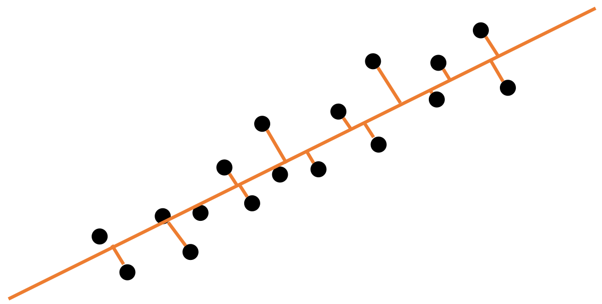
- Patients observations
  - Number of operations
  - Insurance company costs
  - Blood preassure
  - Wake-up time
  - Number of days spent in hospital

Age

Principal components analysis

- PCA is a nonparametric **tool** for **extracting relevant information** from (usually highly dimensional data) data

- **Goal** of PCA is to find the **linear subspace** in which the data reside
  - The subspace should fit the data as best as possible
  - E.g., cloud of points along a diagonal is a linear subspace of a 2D space



Application domains

- **Machine learning**
  - Dimension reduction pre-step
- **Visualization**
  - Objects represented by many descriptors
  - PCA helps to find structure among objects which could not be visualized otherwise
- **Compression**
  - Representation of objects only by their coordinates in the respective subspace
  - E.g. in the eigenfaces (see later), each face can be reasonable approximated by only 10 coordinates

Linear algebra review

Matrices, norm, trace, eigendecomposition, spectral decomposition, SVD

Variance, covariance (1)

- **Variance** measures the spread of data in a dataset from the mean

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

- **Covariance** measures how each of the dimensions varies from the mean with respect to each other

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Variance, covariance (2)

- **Positive** covariance of two dimensions indicates that they change together (number of hours spent studying – grade)
- **Negative** covariance indicates that change in one dimension causes inverse change in the other (number of hours spent in a pub – balance of your bank account)

- **Covariance matrix** is a matrix of all pairwise covariences, e.g., for 3 dimensions X, Y, Z:

$$\begin{pmatrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & cov(Z,Z) \end{pmatrix}$$

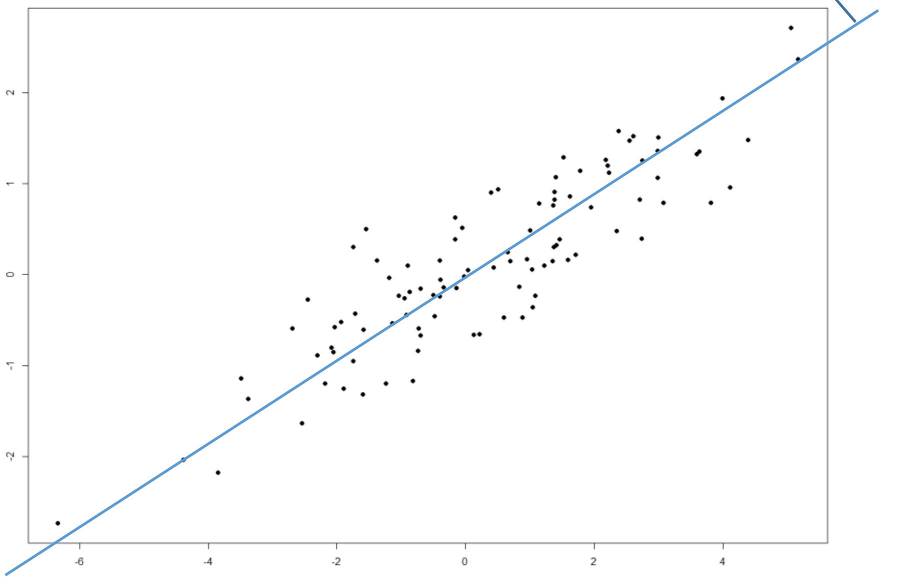
PCA formulation (1)

If we project the data onto this line, we lose as little information as possible = we keep as much variance as possible.

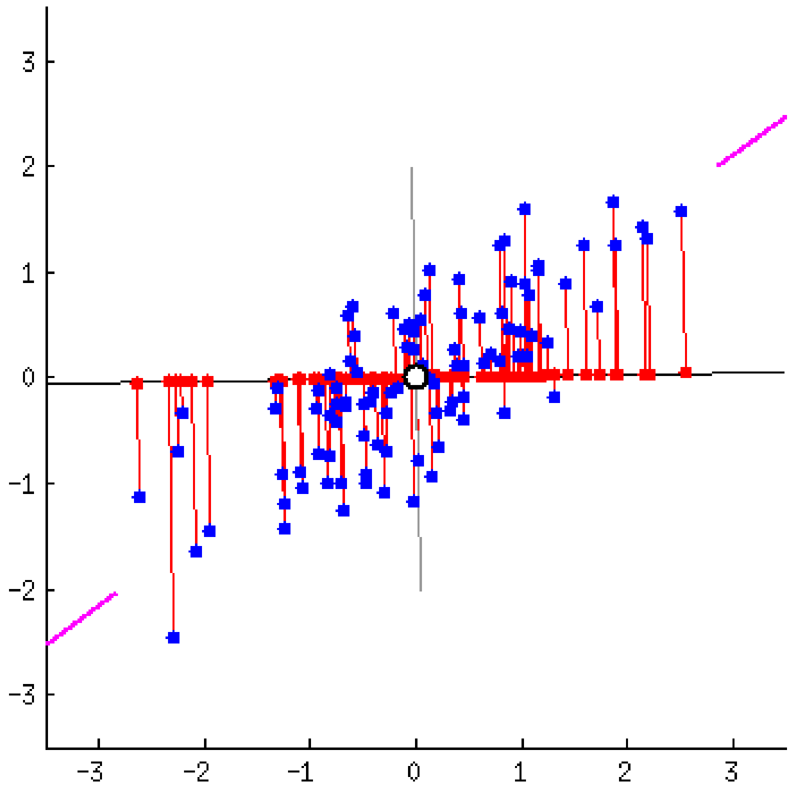
- Let us have a random variable (observations) $x^T = (x_1, \dots, x_p)$ with mean μ and covariance matrix Σ

- First PC is the linear combination

$$y_1 = a_1^T x = \sum_{i=1}^{p} a_{1i} x_i$$



where a_1 is chosen such that $\boldsymbol{var(y_1)}$ **is maximum**
subject to $\boldsymbol{a_1^T a_1 = 1}$ (normalization constraint)

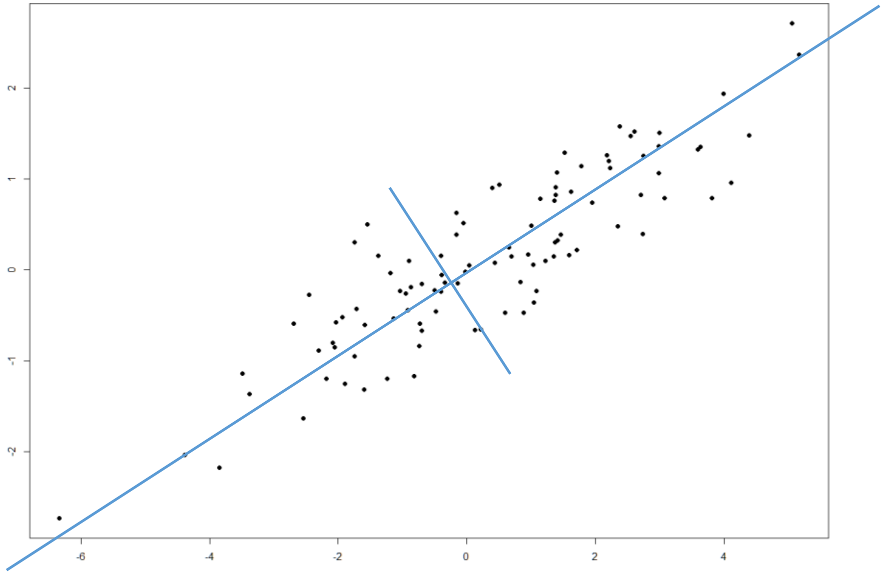


source: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

PCA formulation (2)

- Second PC is the linear combination

$$y_2 = a_2^T x = \sum_{i=1}^{p} a_{2i} x_i$$



where a_k is chosen to **maximize** $\boldsymbol{var(y_2)}$

subject to $\boldsymbol{a_2^T a_2 = 1}$ and $\boldsymbol{cov(y_1, y_2) = 0}$

PCA formulation (3)

- Generally, k-th PC is the linear combination

$$y_k = a_k^T x = \sum_{i=1}^{p} a_k x_i$$

where a_k is chosen such that $\boldsymbol{var(y_k)}$ **is maximum**

subject to $\boldsymbol{a_k^T a_k = 1}$ and $\mathbf{\forall l, l < k}$**:** $\boldsymbol{cov(y_k, y_l) = 0}$

Searching for the first PC (1)

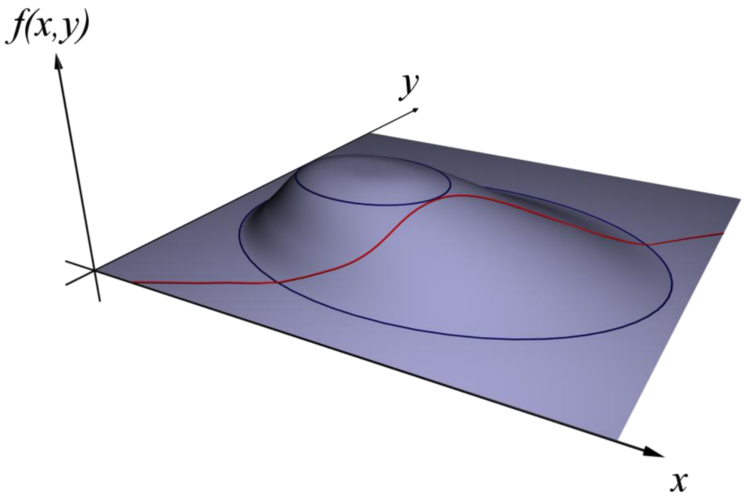
- Assumption that the data are normalized, i.e., the **mean is subtracted**

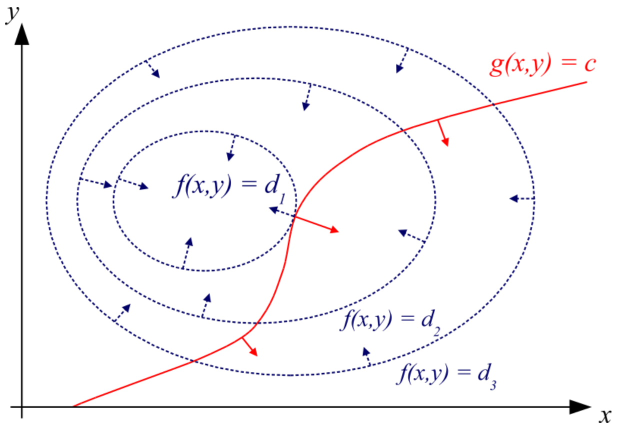
- Find a **1D subspace** so that the observations have **maximum spread** in it ⟶ maximizing variance

$$\boldsymbol{var(y_1)} = \boldsymbol{var\left(a_1^T X\right)} = E[(a_1^T X - E[a_1^T X])(a_1^T X - E[a_1^T X])^T]$$
$$= E[(a_1^T X)(a_1^T X)^T] = E[a_1^T X X^T a_1] = E[a_1^T \Sigma a_1] = \boldsymbol{a_1^T \Sigma a_1}$$

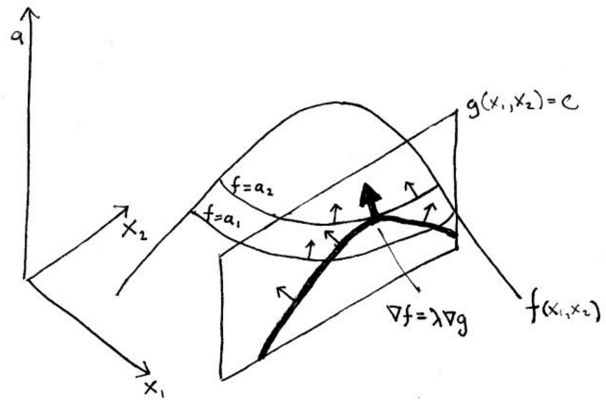
- The goal is to **maximize** variance given $a_1^T a_1 = 1$ ⟶ **Lagrange multipliers**

Lagrange multipliers





source: Wikipedia



source: Andrew Chamberlain (The Idea Shop)

- Maximize $f(x, y)$ subject to $g(x, y) = c$ → introduction of a new variable - Lagrange multiplier λ ($\nabla f = \lambda \nabla g \rightarrow \nabla f - \lambda \nabla g = 0$)

Lagrangian

$$\Lambda(x, y, \lambda) = f(x, y) + \lambda(g(x, y) - c) \rightarrow \frac{\Delta\Lambda(x, y, \lambda)}{\Delta x, y, \lambda} = 0$$

Searching for the first PC (2)

- Transcription into the Lagrangian form

$$\Lambda(a_1, \lambda) = a_1^T \Sigma a_1 - \lambda(a_1^T a_1 - 1)$$

- Now we need to differentiate the Lagrangian

$$\frac{\partial \Lambda(a_1, \lambda)}{\partial a_1} = \frac{\partial \Lambda(a_1, \lambda)}{\partial \begin{bmatrix} a_{11} \\ \cdots \\ a_{1k} \end{bmatrix}} = 2\Sigma a_1 - 2\lambda a_1 = 0$$

Searching for the first PC (3)

$$2\Sigma a_1 - 2\lambda a_1 = 0$$

- This leads to the eigenproblem $\boldsymbol{\Sigma a_1 = \lambda a_1} \rightarrow a_1$ is an eigenvector of Σ with eigenvalue λ

$$\boldsymbol{var(y_1)} = \boldsymbol{var\left(a_1^T X\right)} = a_1^T \Sigma a_1 = \lambda a_1^T a_1 = \boldsymbol{\lambda}$$

- Suppose that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \rightarrow$ to **maximize** $\boldsymbol{var(y_1)}$ we must choose $\boldsymbol{\lambda = \lambda_1}$

Searching for the next PCs

- The principle is similar, but due to the uncorrelation requirement we must extend the constraint with

$$0 = cov(y_1, y_2) = cov(a_1^T x, a_2^T x) = a_1^T \Sigma a_2 = a_2^T \Sigma a_1 = a_2^T \lambda a_1 = \lambda a_2^T a_1$$

- Leading to a modified Lagrangian

$$\Lambda(a_2, \lambda, \kappa) = a_2^T \Sigma a_2 - \lambda(a_2^T a_2 - 1) - \kappa(a_2^T a_1)$$
$$\left(a_2^T \Sigma a_2 - \lambda(a_2^T a_2 - 1) - \kappa(a_2^T a_1)\right) \frac{d}{da_2} = 0$$
$$\Sigma a_2 - \lambda a_2 - \kappa a_1 = 0$$
$$a_1^T \Sigma a_2 - \lambda a_1^T a_2 - \kappa a_1^T a_1 = 0$$
$$0 \; - 0 \; - \kappa = 0$$

$$\Sigma a_2 - \lambda a_2 = 0$$
$$\boldsymbol{\Sigma a_2 = \lambda a_2 \Rightarrow \lambda = \lambda_2}$$

PCA transformation

- Thus, the **coefficients of the linear combination which transform the observations onto the PCs are formed by eigenvalues of the covariance matrix**

- Let A contain the eigenvectors a_i as its columns and let x be a p-dimensional vector representing an observation, then

$$\boldsymbol{y} = \boldsymbol{A}^{\boldsymbol{T}}(\boldsymbol{x} - \boldsymbol{\mu})$$

Variance

- **PCs** are **components of variance explaining** the total variation in the data
  - The sum of variances of the original variables $var(X)$ and of the PCs $var(Y) = var(AX)$ are the same

$$\Sigma = A\Lambda A^T$$

$$tr(\Sigma) = tr(A\Lambda A^T) = tr(\Lambda A^T A) = tr(\Lambda)$$

  - Therefore

$$\frac{\boldsymbol{\lambda_i}}{\boldsymbol{\lambda_1 + \cdots + \lambda_p}}$$

can be interpreted as the total variation in the original data explained by the i-th principal component

Scores and loadings

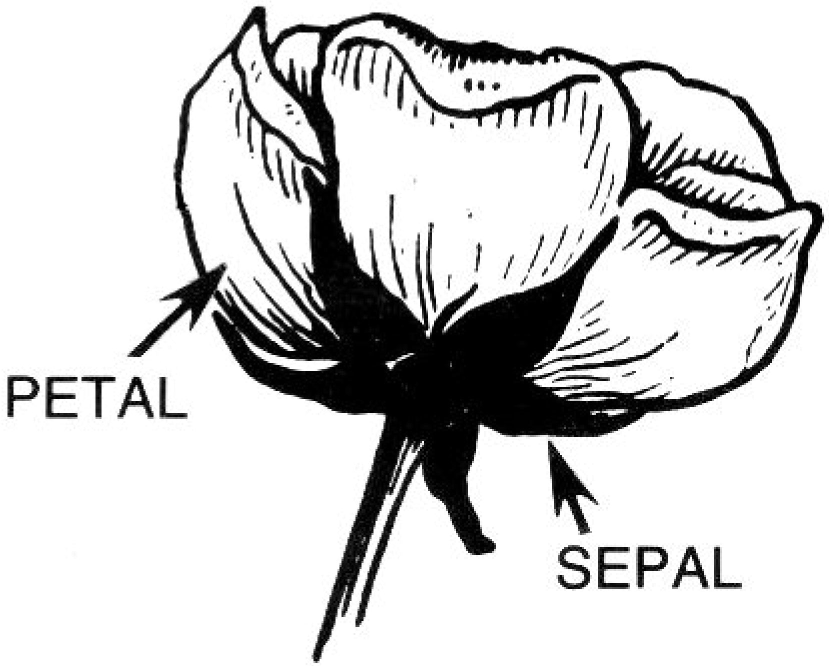
- **Scores**
  - **Transformed variable values** corresponding to a particular observation
    - Original data multiplied by the loadings
  - Geometrically, scores are the **coordinates** of each observation with respect to the **new axis**
- **Loadings**
  - Weight by which each standardized original variable should be multiplied to get the component score → separate loadings for each component
  - Expresses which variables have high **loading** in which PCs
    - Loadings close to zero indicate which variables do not contribute much to given component
  - Extent to which given **variable** is **correlated with given component**

Scale invariance

- PCA is **NOT scale invariant** → variance in consistently large variable will dominate the spectrum of eigenvalues → variables should be of comparable scale
  - E.g., if height of a person was expressed in nanometers, the first PC would probably be identical with the height dimensions (highest variance)
- Often the variables are divided by the square root of its variance → **correlation matrix** instead of covariance matrix

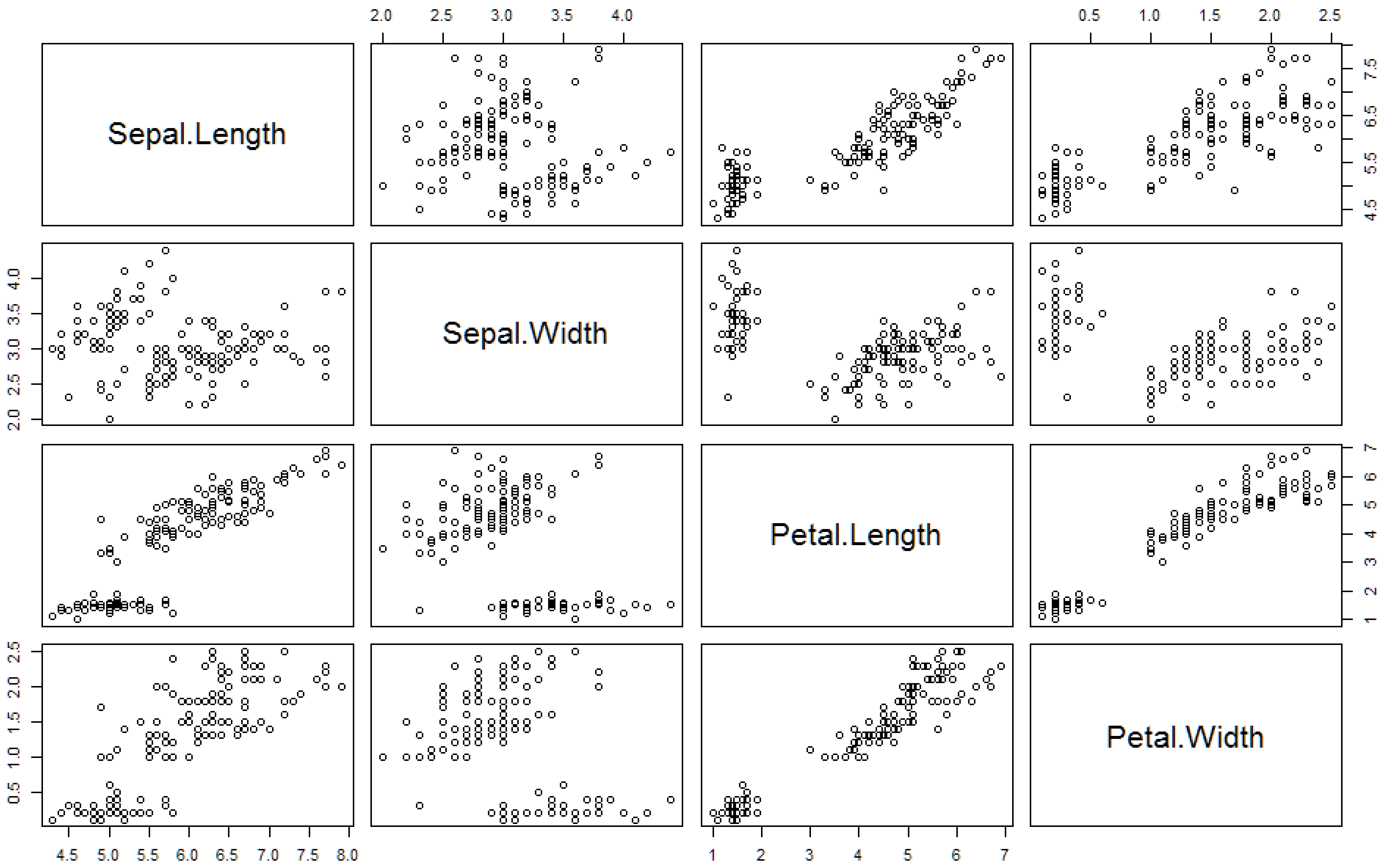
$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{cov(X,X)cov(Y,Y)}}$$

Iris dataset



- One of the R datasets

- The measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris → $n = 150,\ p = 4$


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
2.0 2.5 3.0 3.5 4.0
1 2 3 4 5 6 7
0.5 1.0 1.5 2.0 2.5
4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
1 2 3 4 5 6 7

PCA in R

- The most common ways to conduct PCA in R is `prcomp` (stats), `princomp` (stats) or `PCA` (FactoMineR)

```
data(iris)
ir.descriptors <- iris[, 1:4]
ir.species <- iris[, 5]
ir.pca <- prcomp(ir. descriptors, center = TRUE, scale. = TRUE)
```

```
print(ir.pca)
```

```
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

```
summary(ir.pca)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

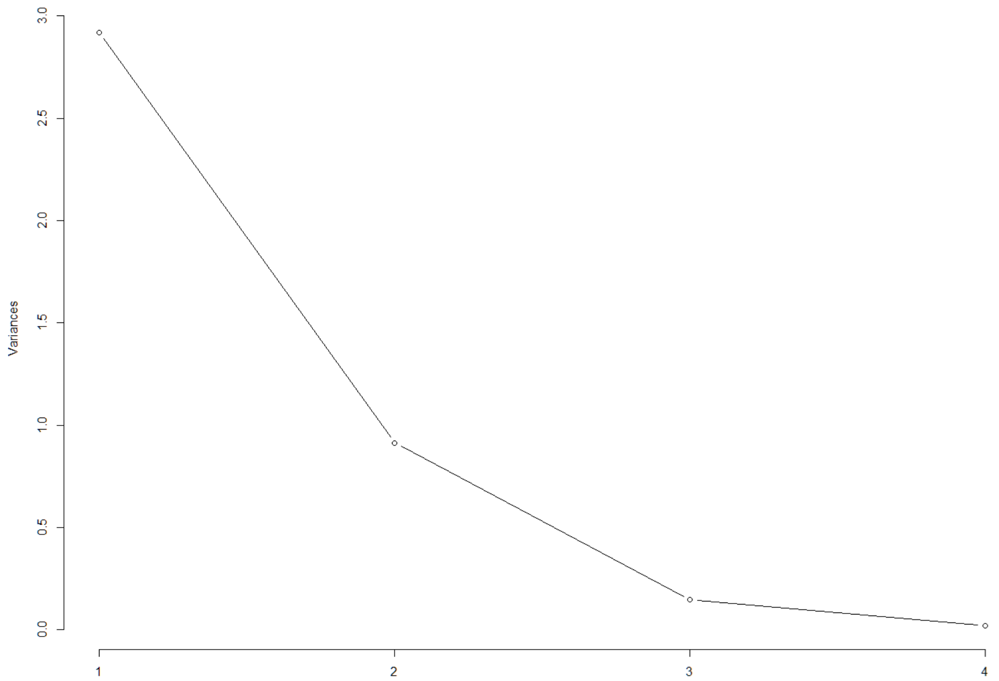

http://www.r-fiddle.org/#/fiddle?id=L46pBuU5

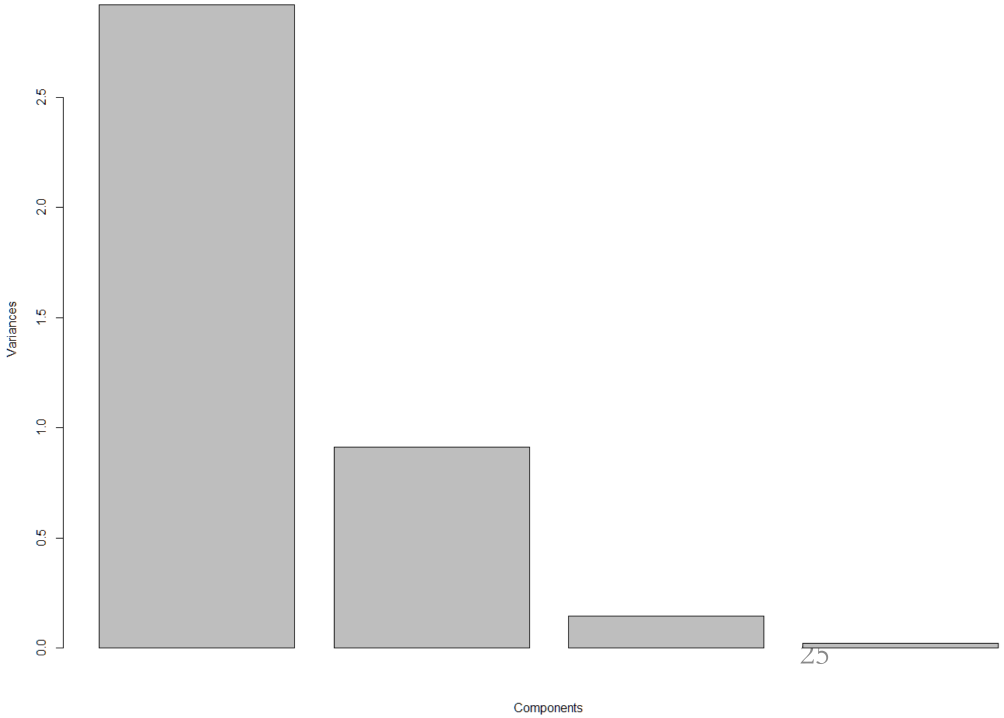
Scree plot

- Display of variance of each of the component

- Plot of magnitudes of eigenvalues

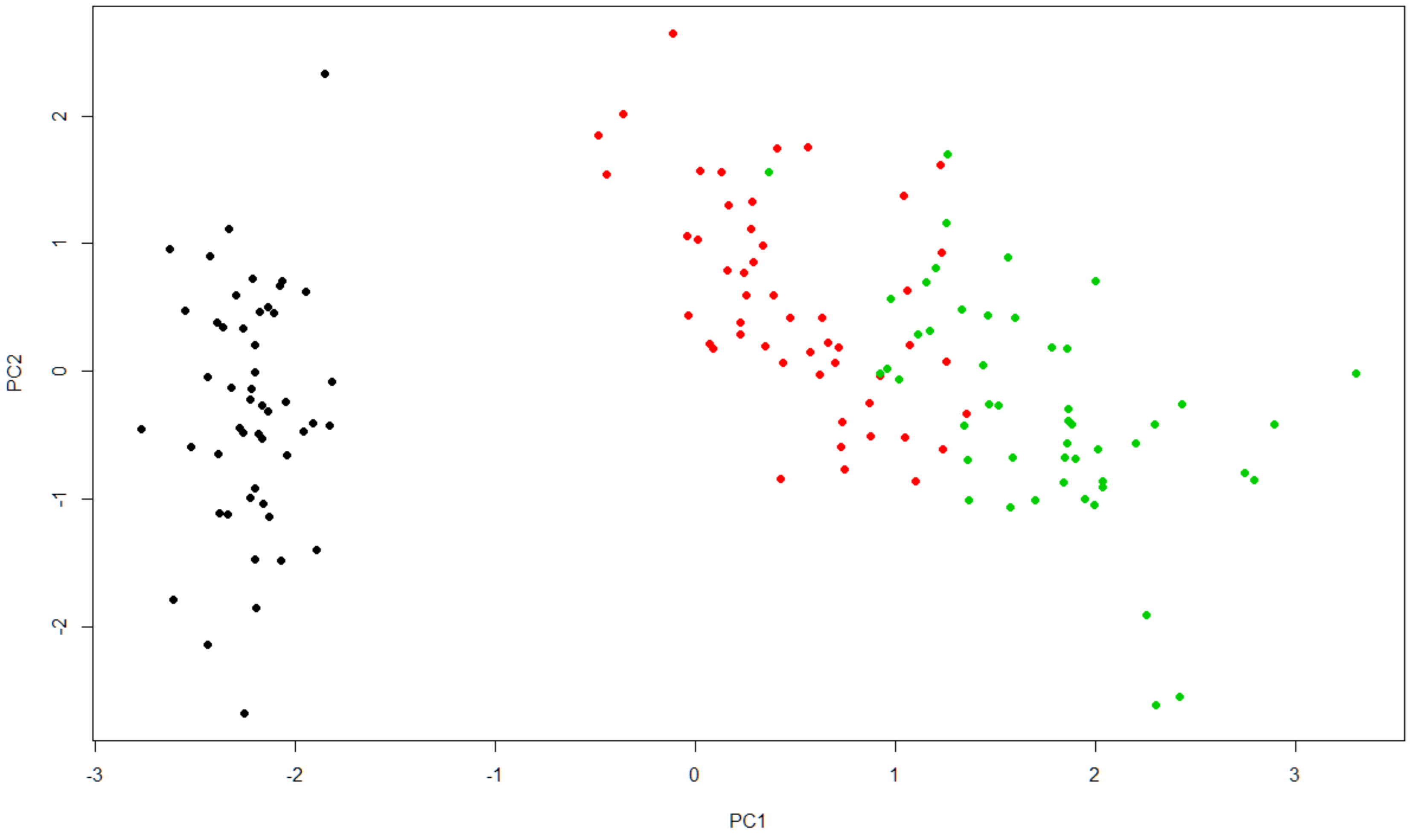
- Gives impression of the intrinsic dimensionality





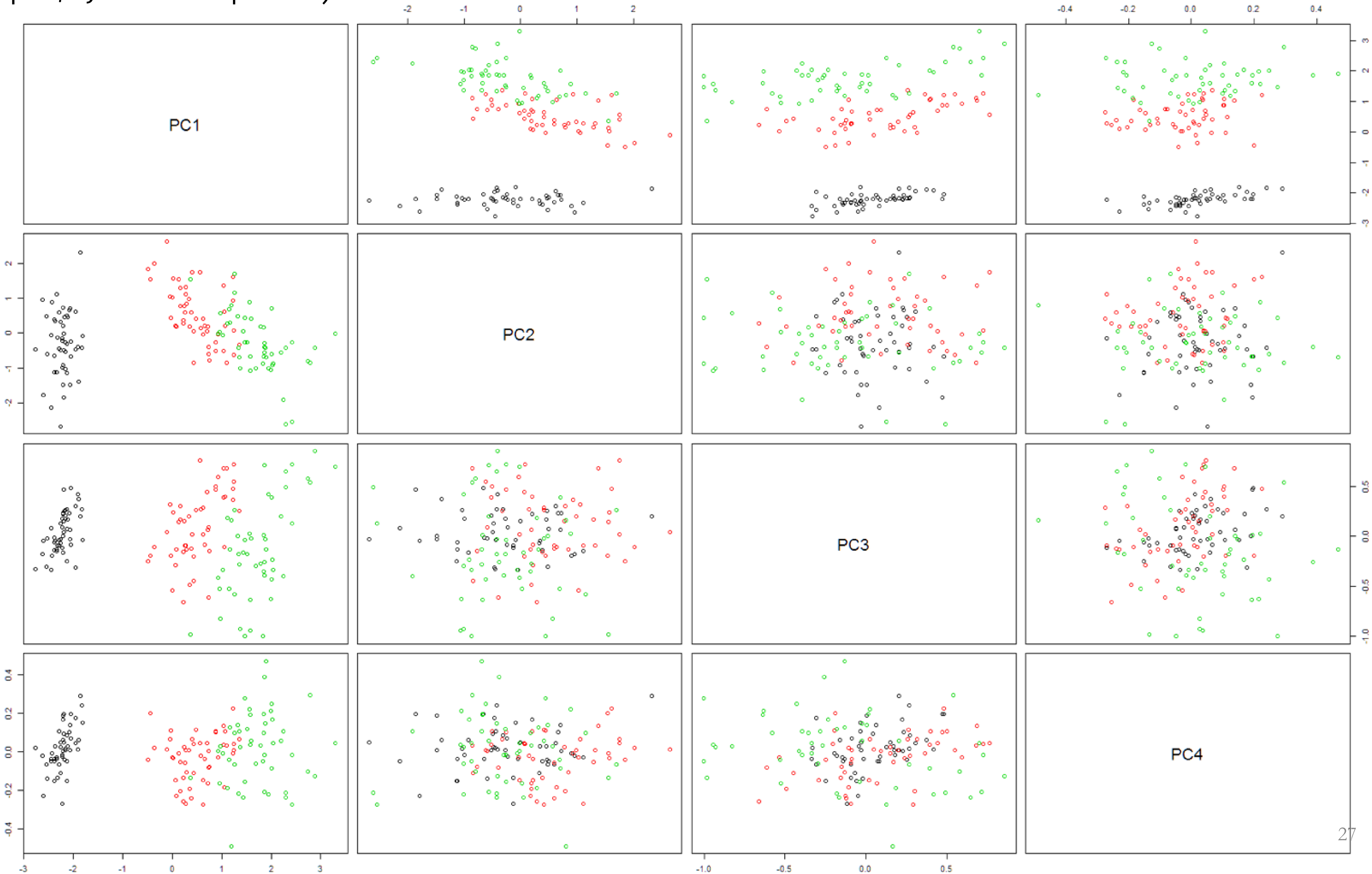
Score plot

- Closeness in the score plot indicates similar “behavior” between samples



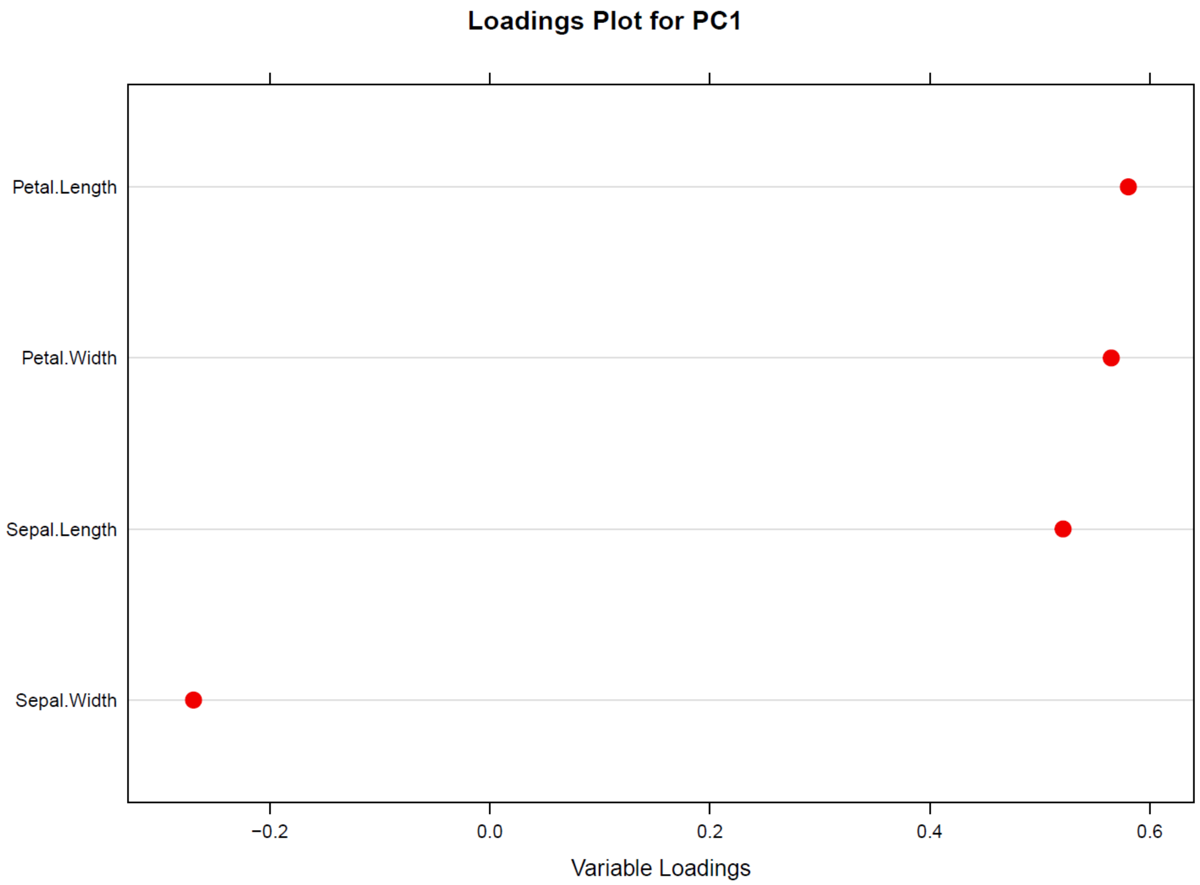

```
ir.pca$x

plot(ir.pca$x, col=ir.species,
    pch = c(16))
```

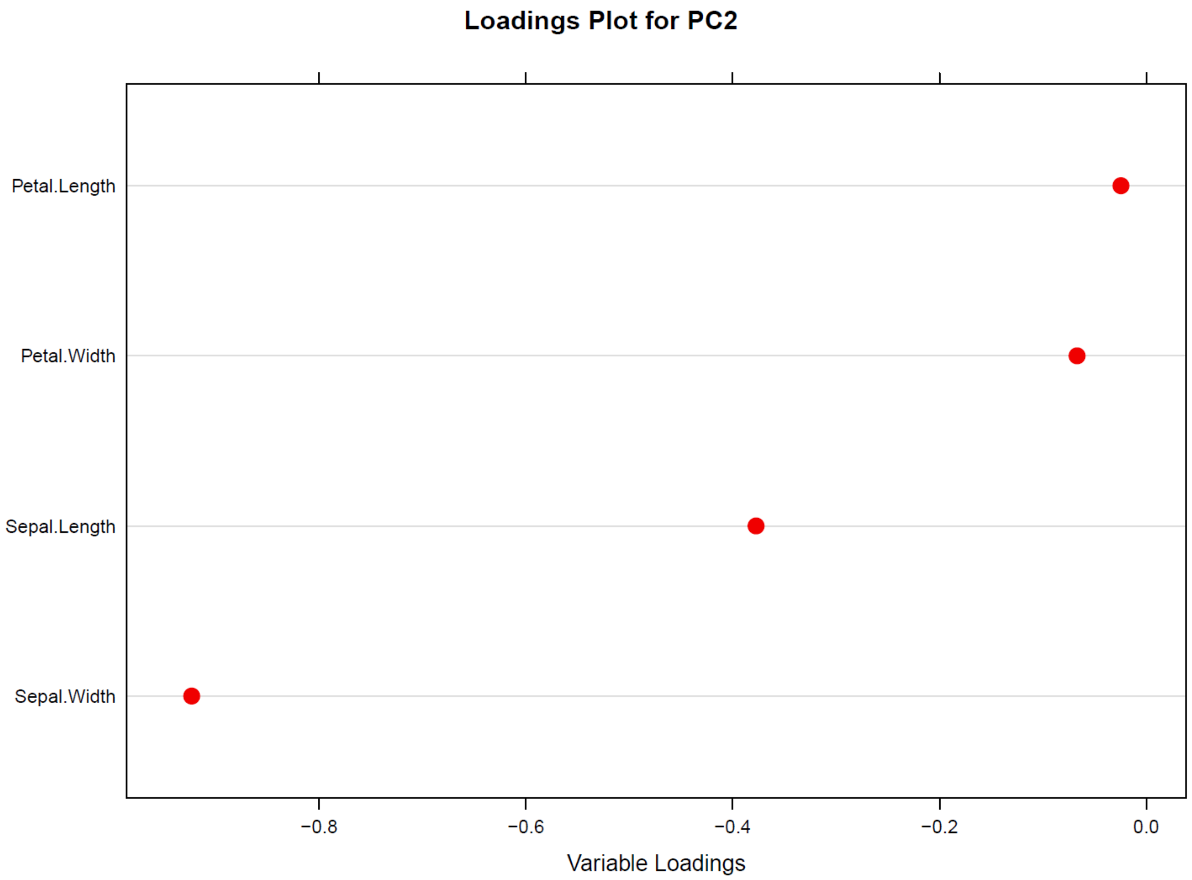
```
pairs(ir.pca$x, col=ir.species)
```



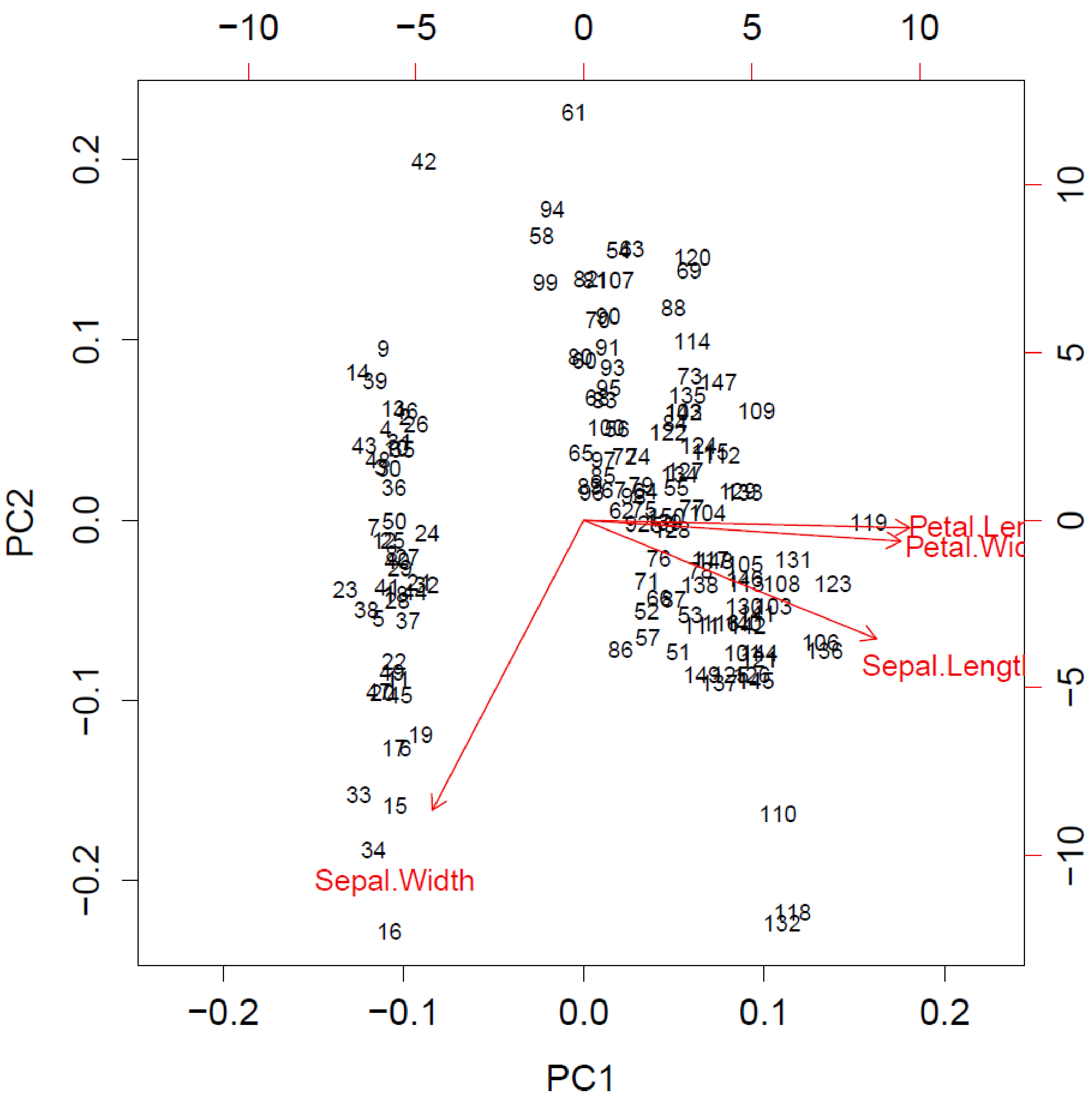
Loadings plot

- Closeness in the score plot indicates similar “behavior” between variables

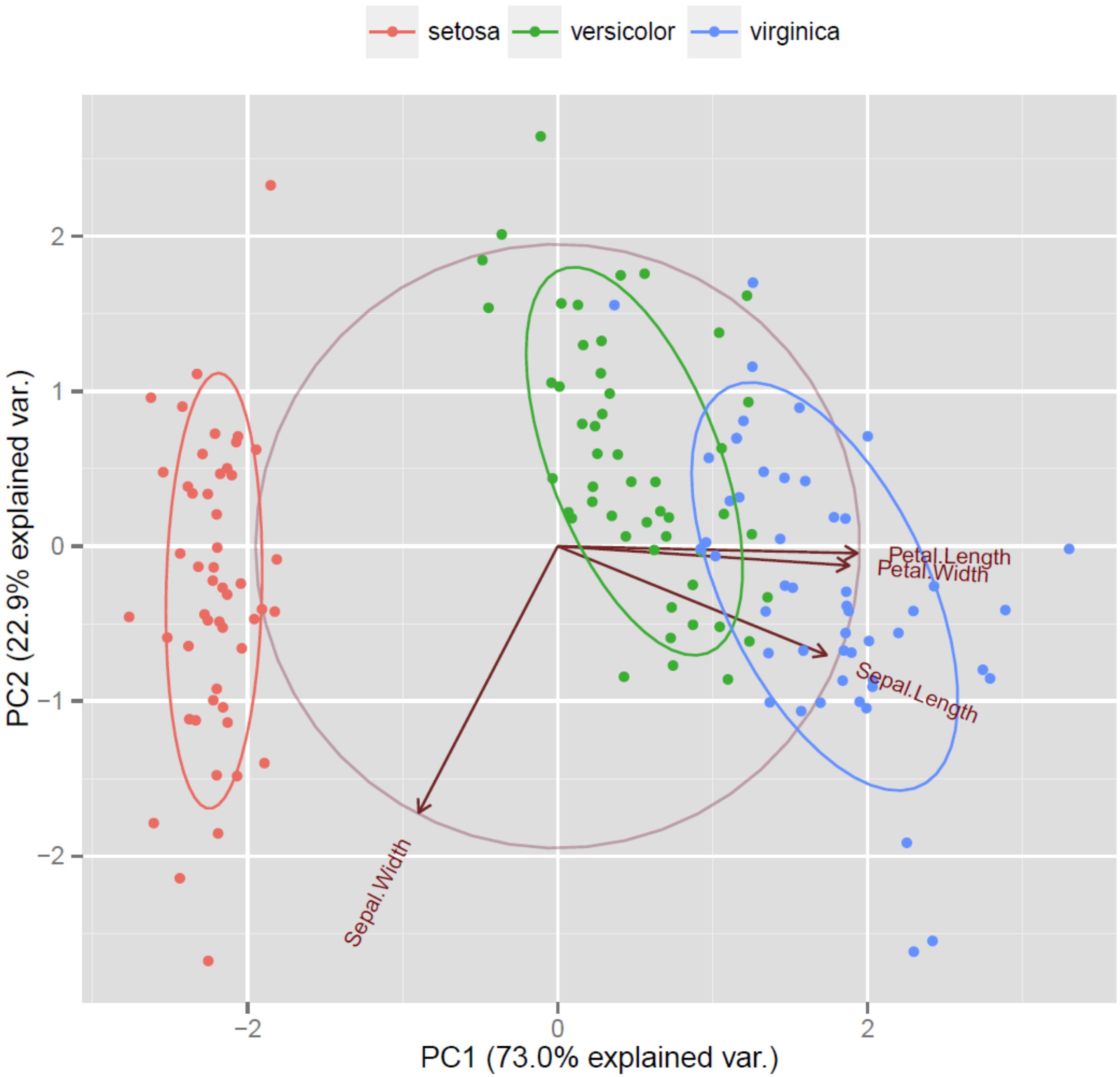


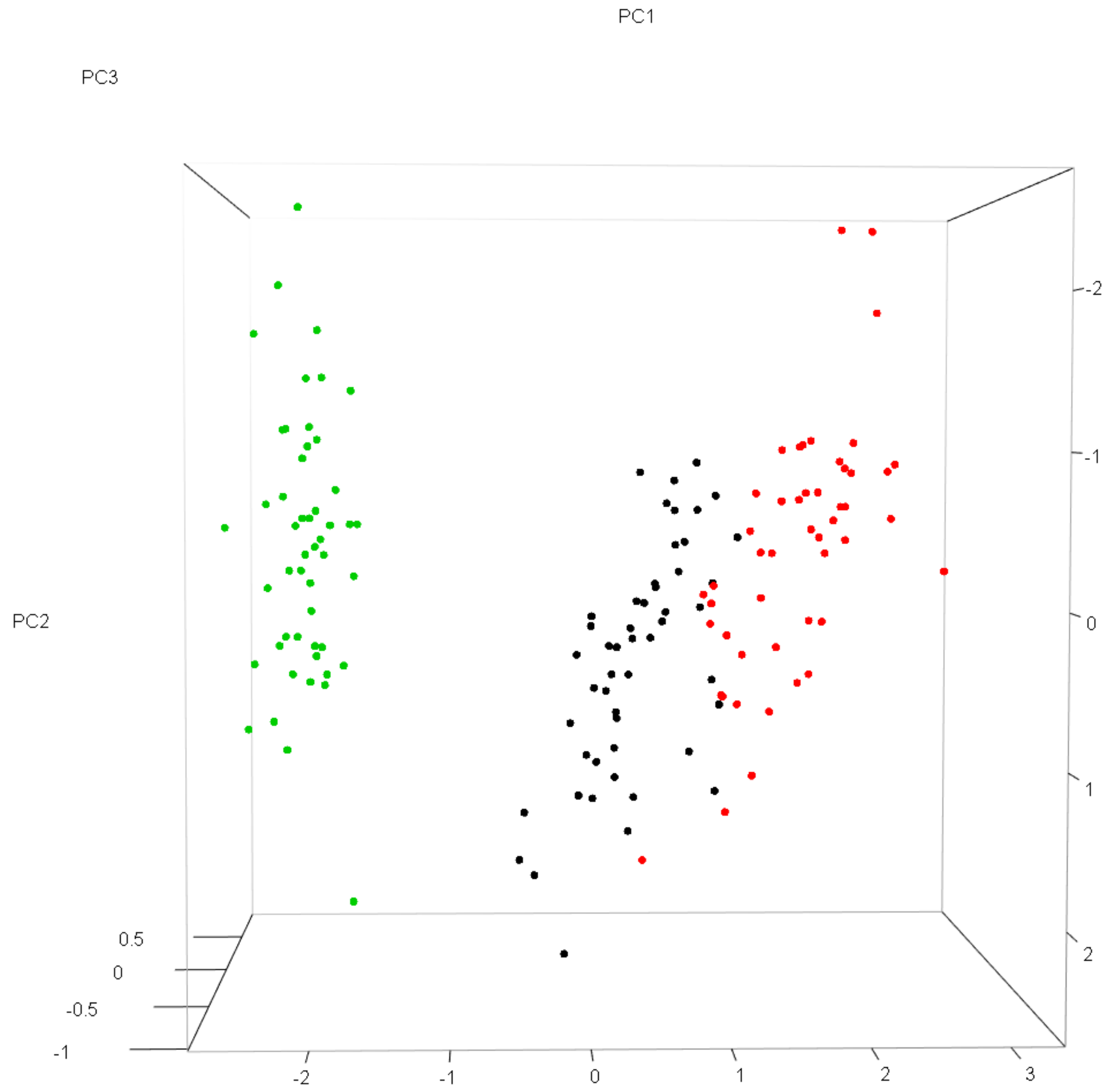


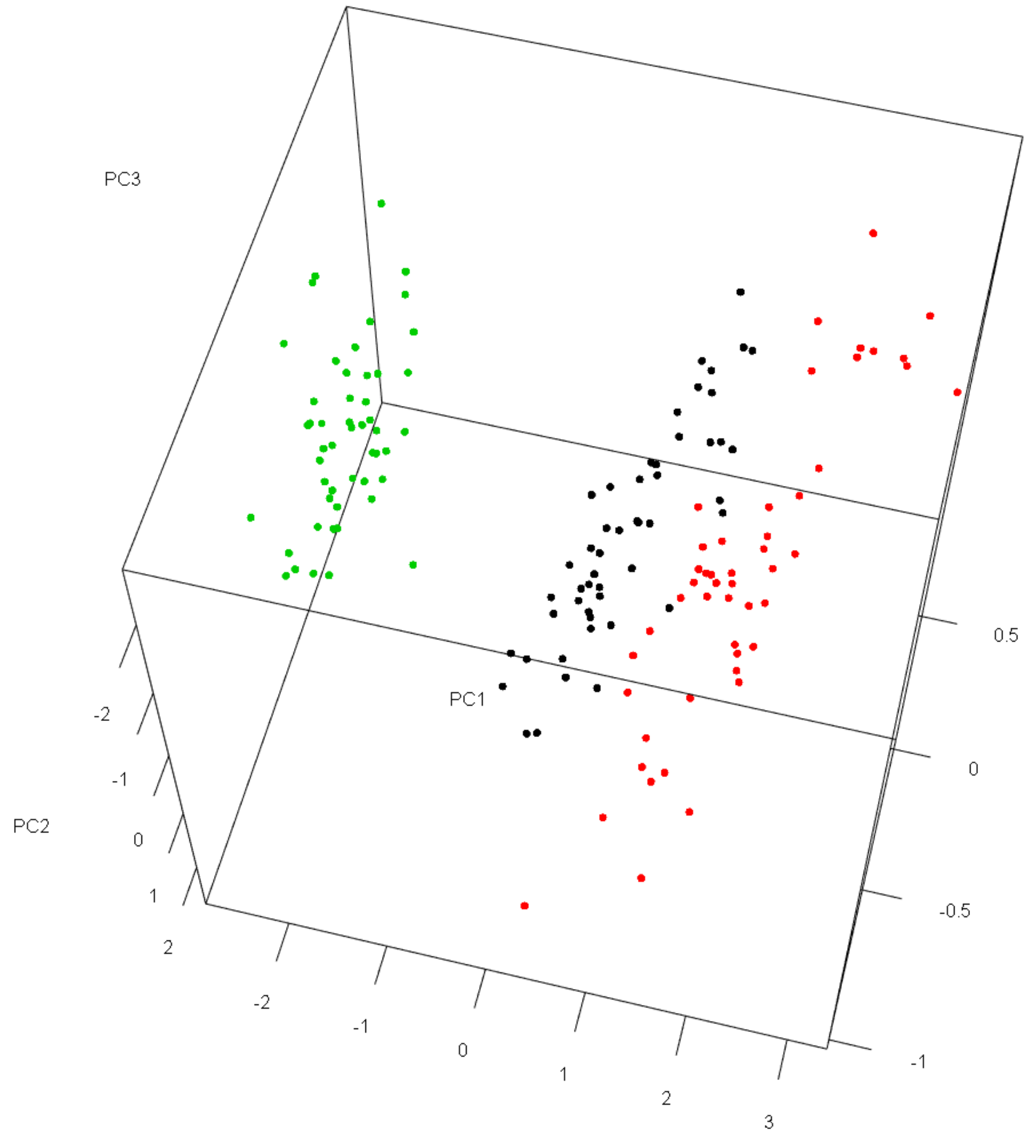
Biplot




```
biplot(ir.pca,
    cex=c(0.6, 0.8))
```

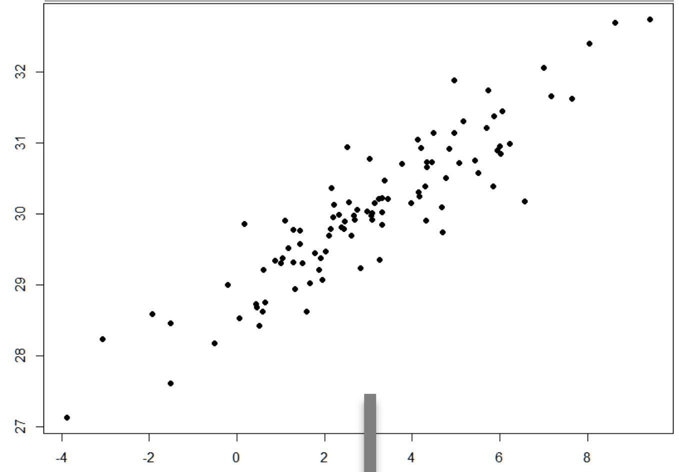

setosa
versicolor
virginica
PC2 (22.9% explained var.)
PC1 (73.0% explained var.)
2
1
0
−1
−2
−2
0
2
Petal.Length
Petal.Width
Sepal.Length
Sepal.Width

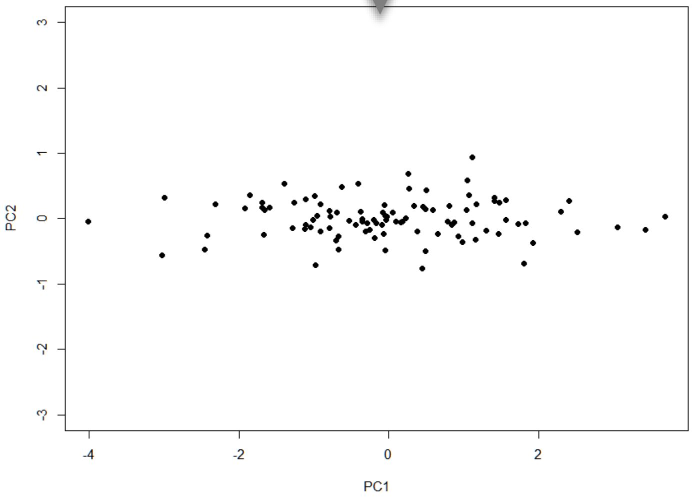





```
plot3d(ir.pca$x[,1:3], col=as.numeric(factor(iris$Species,levels =
c("versicolor","virginica","setosa"))), size=7)
```

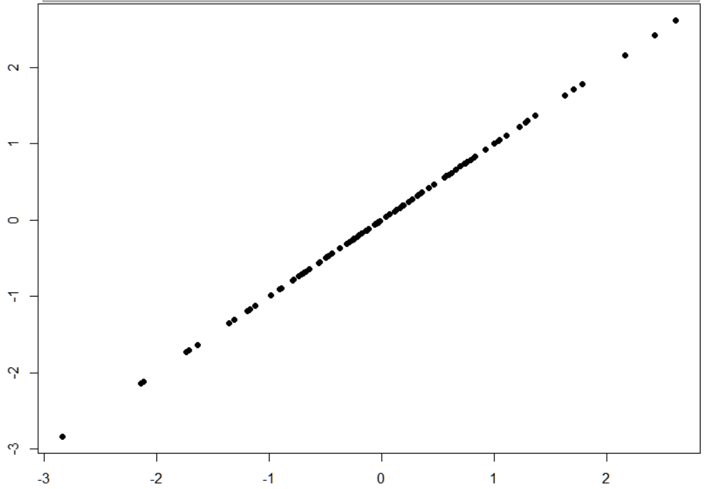
```r
data.orig = rmvnorm(100, mean =
c(3, 30), sigma = matrix(c(5, 2,
2, 1), nrow = 2))
plot(data.orig, pch=c(16))
```

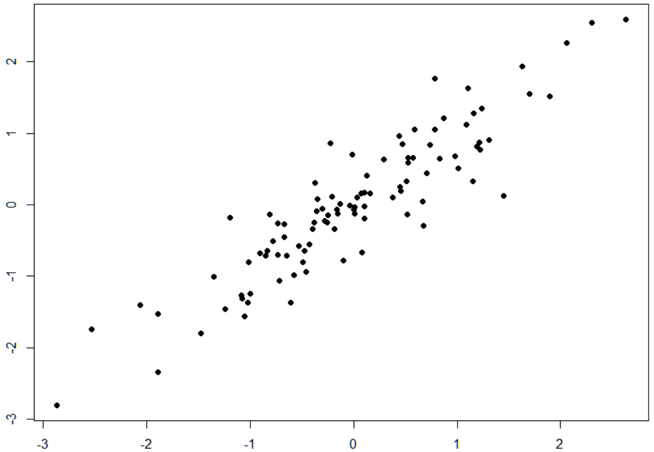





```r
data.pca <- prcomp(data.orig,
center = TRUE, scale = TRUE)
plot(data.pca$x, pch=c(16),
ylim=c(-3,3))
```

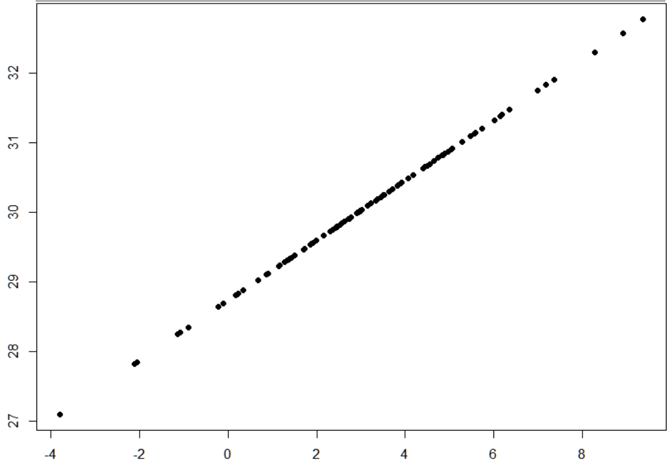
```r
restr = data.pca$x[,1] %*%
t(data.pca$rotation[,1])
plot(restr, pch=16)
```

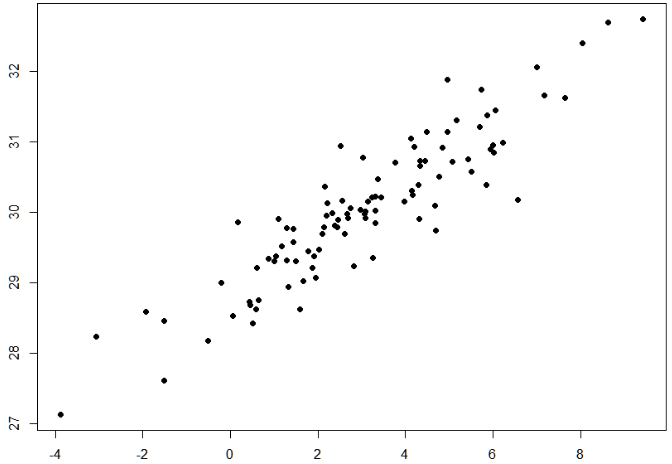




```r
restr = data.pca$x %*%
t(data.pca$rotation)
plot(restr, pch=16)
```

```r
restr <- scale(restr, center =
FALSE , scale=1/data.pca$scale)
restr <- scale(restr, center = -1
* data.pca$center, scale=FALSE)
plot(restr, pch=16)
```



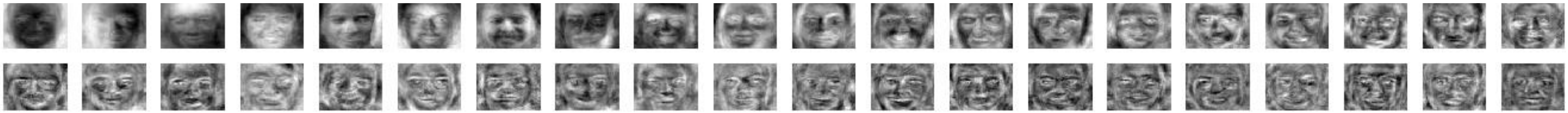


```r
restr <- scale(restr, center =
FALSE , scale=1/data.pca$scale)
restr <- scale(restr, center = -1
* data.pca$center, scale=FALSE)
plot(restr, pch=16)
```

PCA on grayscale images

- Dataset of 96x96 grayscale images
- PCA allows to compress images by representing the original pixels by few linear combinations (scores)

1. Convert each image into a 9216-long (96x96) vector of numbers (0-255) → each image is a point in a 9216-dimensional space
2. Run PCA on the 9216-dimensional objects
3. Take first k PCs (first k columns of the matrix $A \rightarrow A_k$) so that enough variability is captured
4. Convert each object x into the new k-dimensional space using $A_k x$



Principal components

Approximations

Literature

- Jolliffe, I.T. (2002) Principal Component Analysis, Second Edition. Springer-Verlag New York