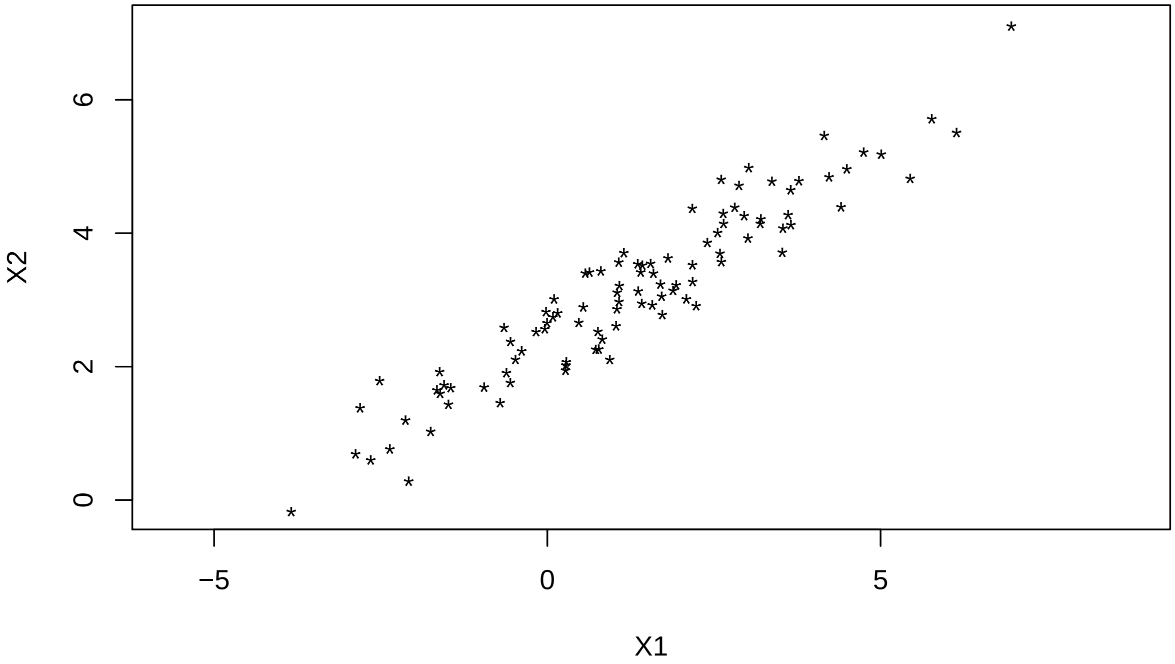
MAST 90138: Multivariate Statistical Techniques

See Härdle and Simar, chapter 11.

5 Principal component analysis

5.1 Introduction

Visualizing 1, 2 or 3 dimensional data is easy: use scatterplot.



When the data are in higher dimension, it is difficult to visualize them.

☛ Can we find a way to summarise the data?

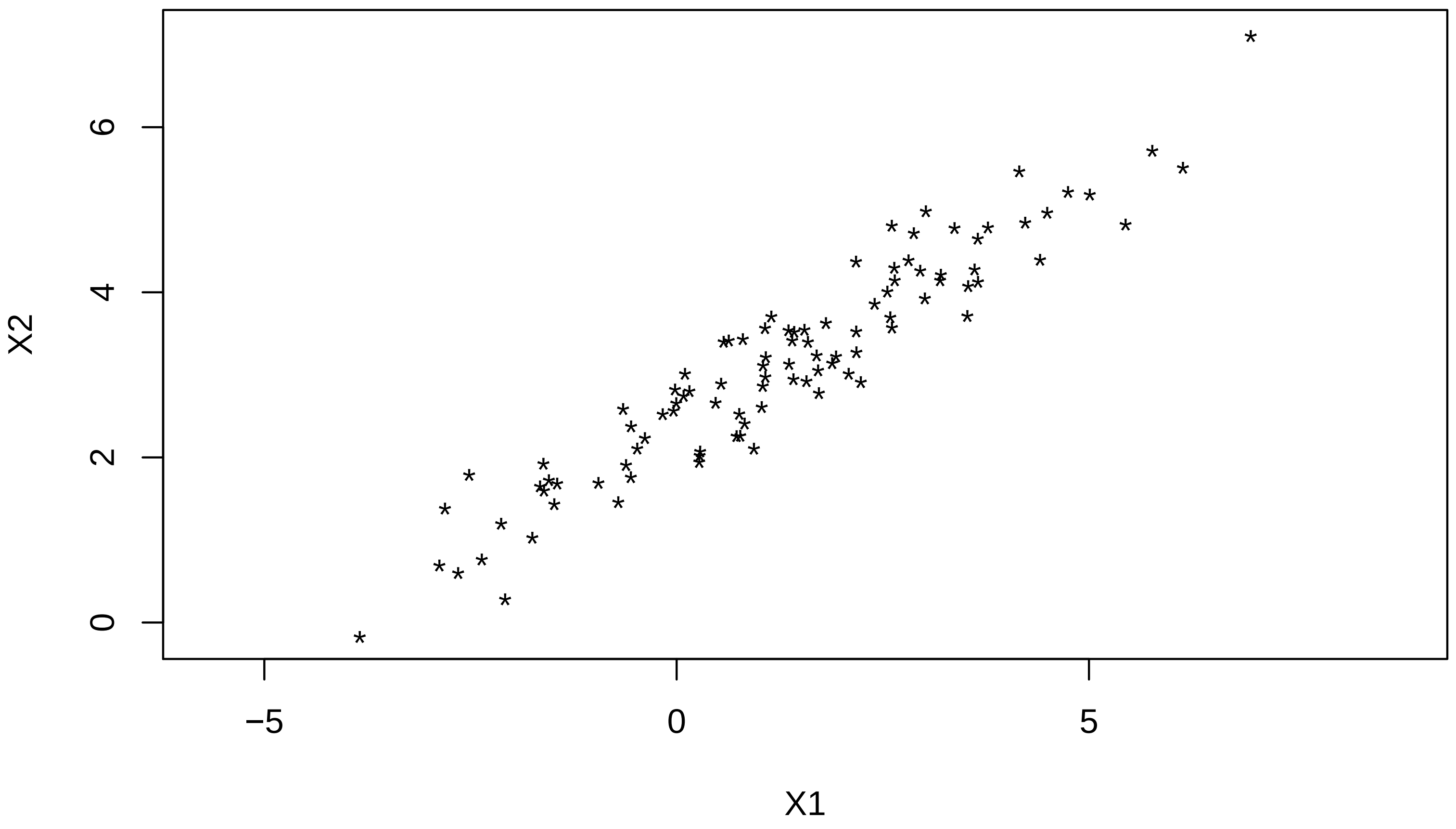
☛ Summaries should be easy to represent graphically.

☛ Summaries should still contain as much information as possible about the original data.

☛ Often achieved through dimension reduction.

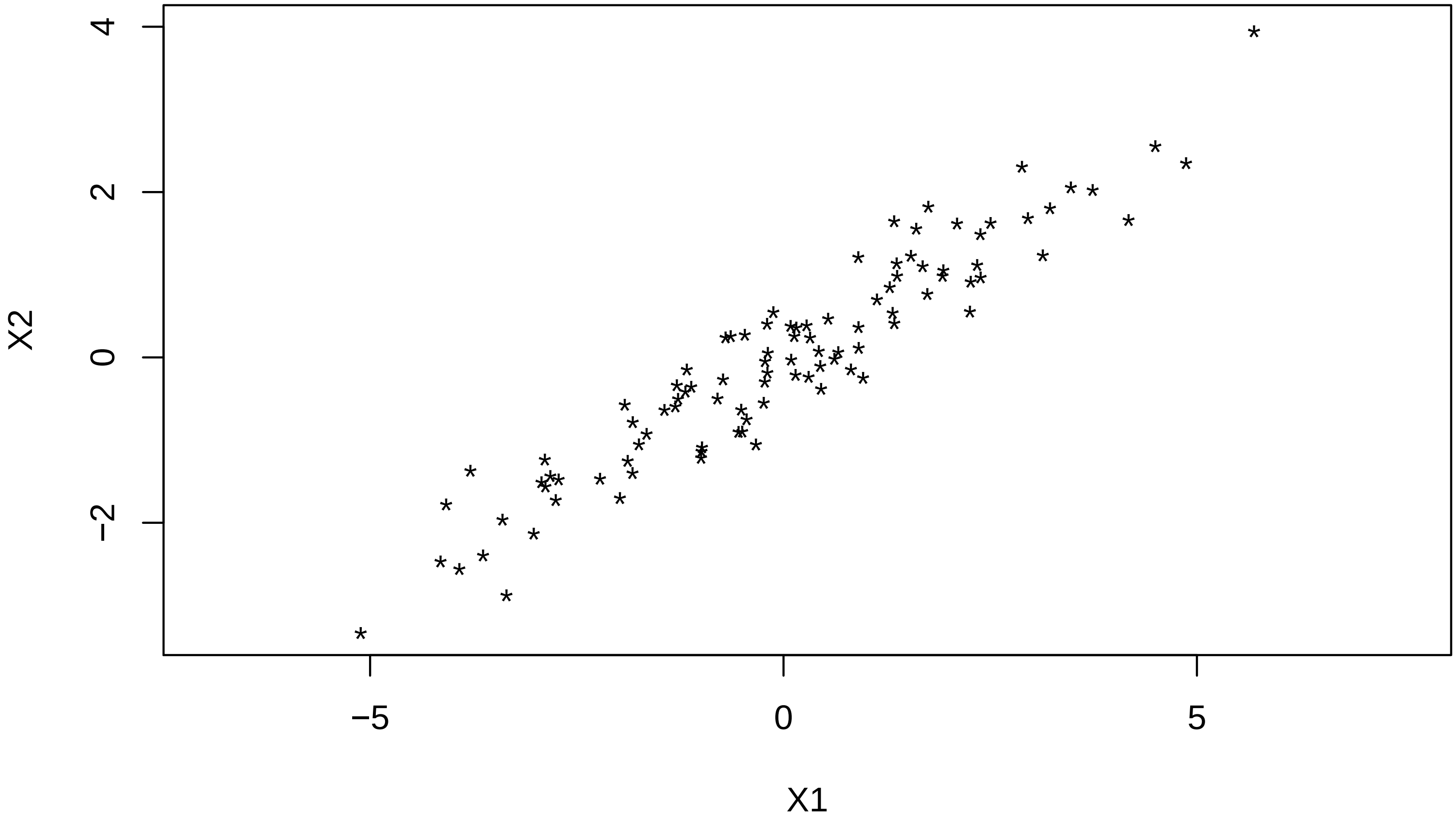
Toy example: Reduce the following 2-dimensional data to 1 dimension .

Data: A collection of i.i.d. pairs $(X_{i1}, X_{i2})^T \sim (\mu, \Sigma)$, for $i = 1, \ldots, n$, shown in the scatter plot.



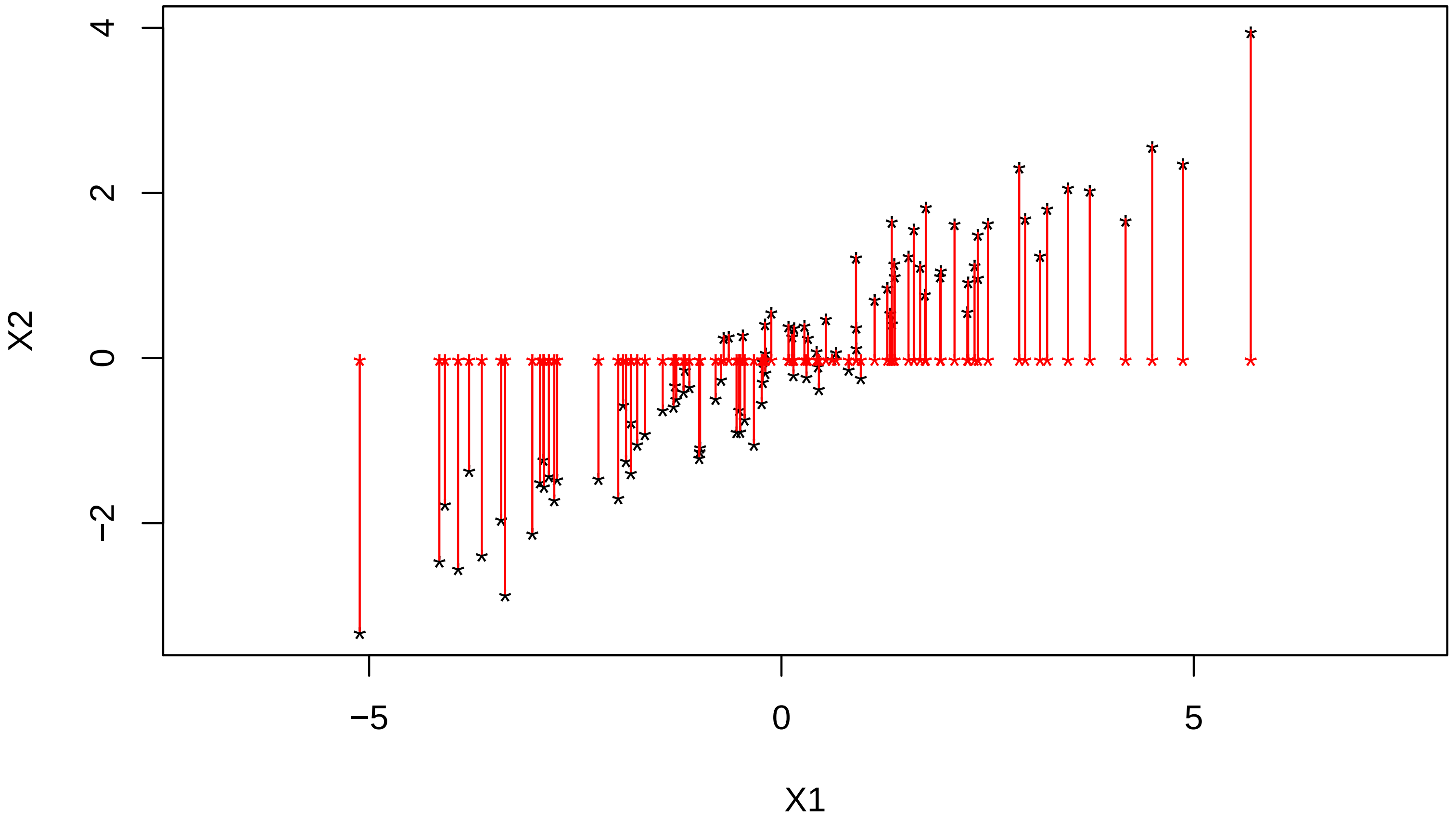
The first thing usually done in these problems is to center the data (easier to understand the geometry for centered data):

For $i = 1, \ldots, n$, we replace $(X_{i1}, X_{i2})^T$ by $(X_{i1} - \bar{X}_1, X_{i2} - \bar{X}_2)^T$:

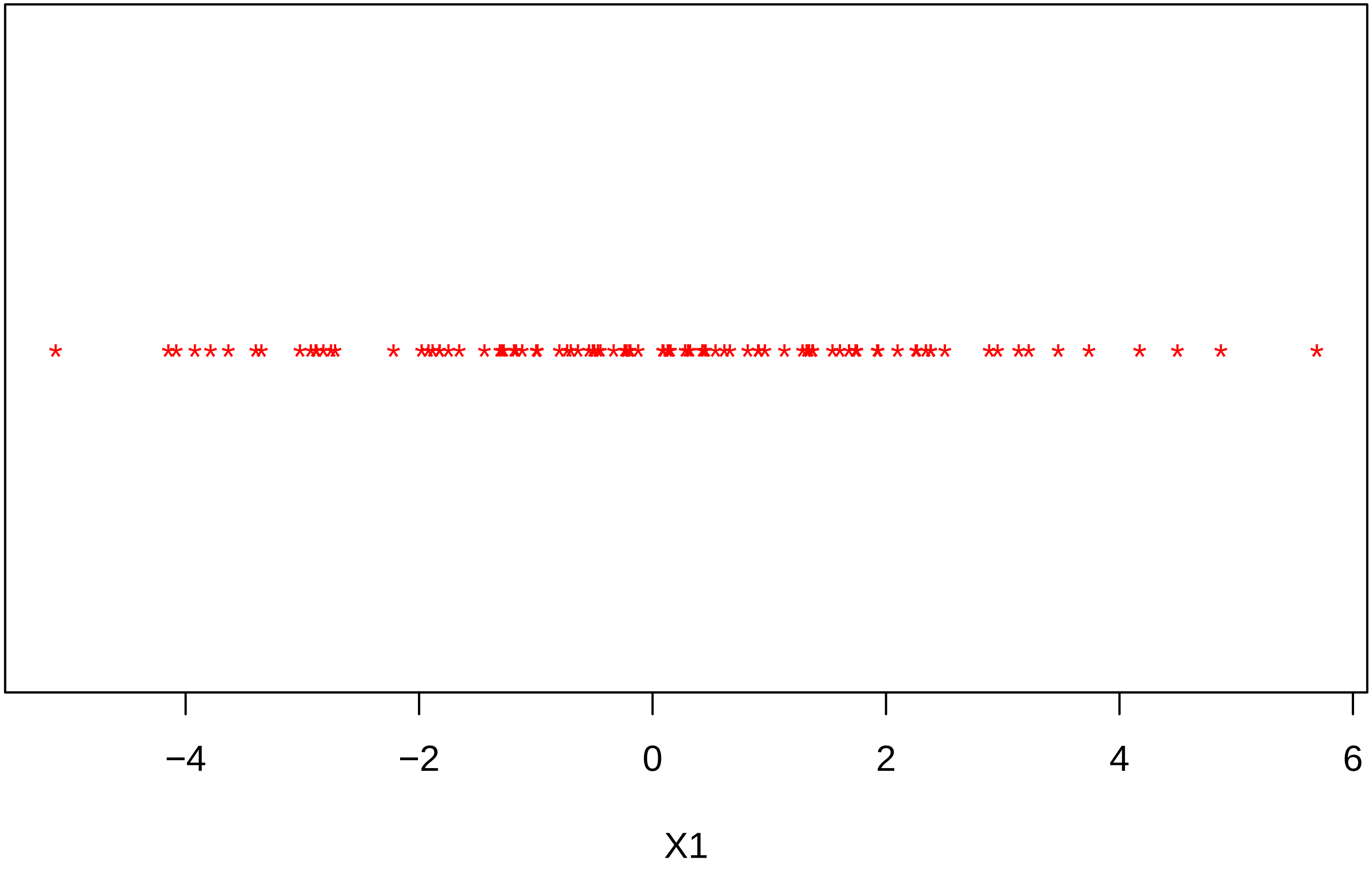


(Unless otherwise specified, for the rest of this chapter, to simplify notation when we use X_{ij} to implicitly mean $X_{ij} - \bar{X}_j$.)

To reduce these data to a single dimension we could, for example, keep only the first component X_{i1} of each data point.



Only the first component:



☛ Not very interesting: lose all the information about the second component X_2.

☛ Suppose the data contain the age (X_1) and the height (X_2) of $n = 100$ individuals. This amounts to keeping only the age and drop completely the data about height.

Why not instead create a new variable that contains information about both age and height?

Simple approach: take a linear combination of the age and the height.

☛ For $i = 1, \ldots, n$ we could create a new variable

$$Y_i = \text{age}_i/2 + \text{height}_i/2,$$

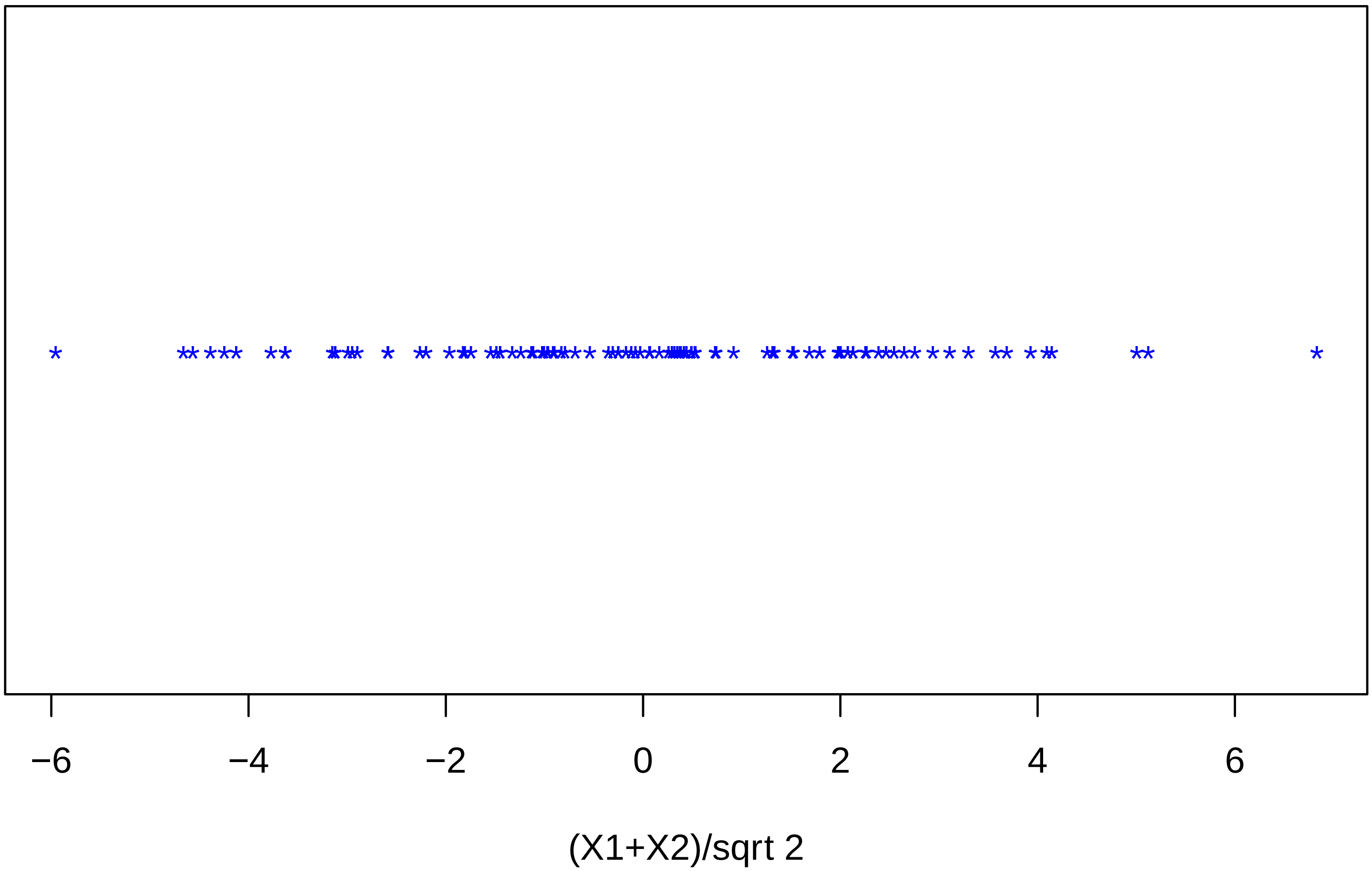
i.e. the average of the age and the height. $1/2$ and $1/2$ are the *weights* of age and height, respectively.

☛ Often prefer to rescale linear combinations so that *the sum of the square of the weights equals 1*, for example,

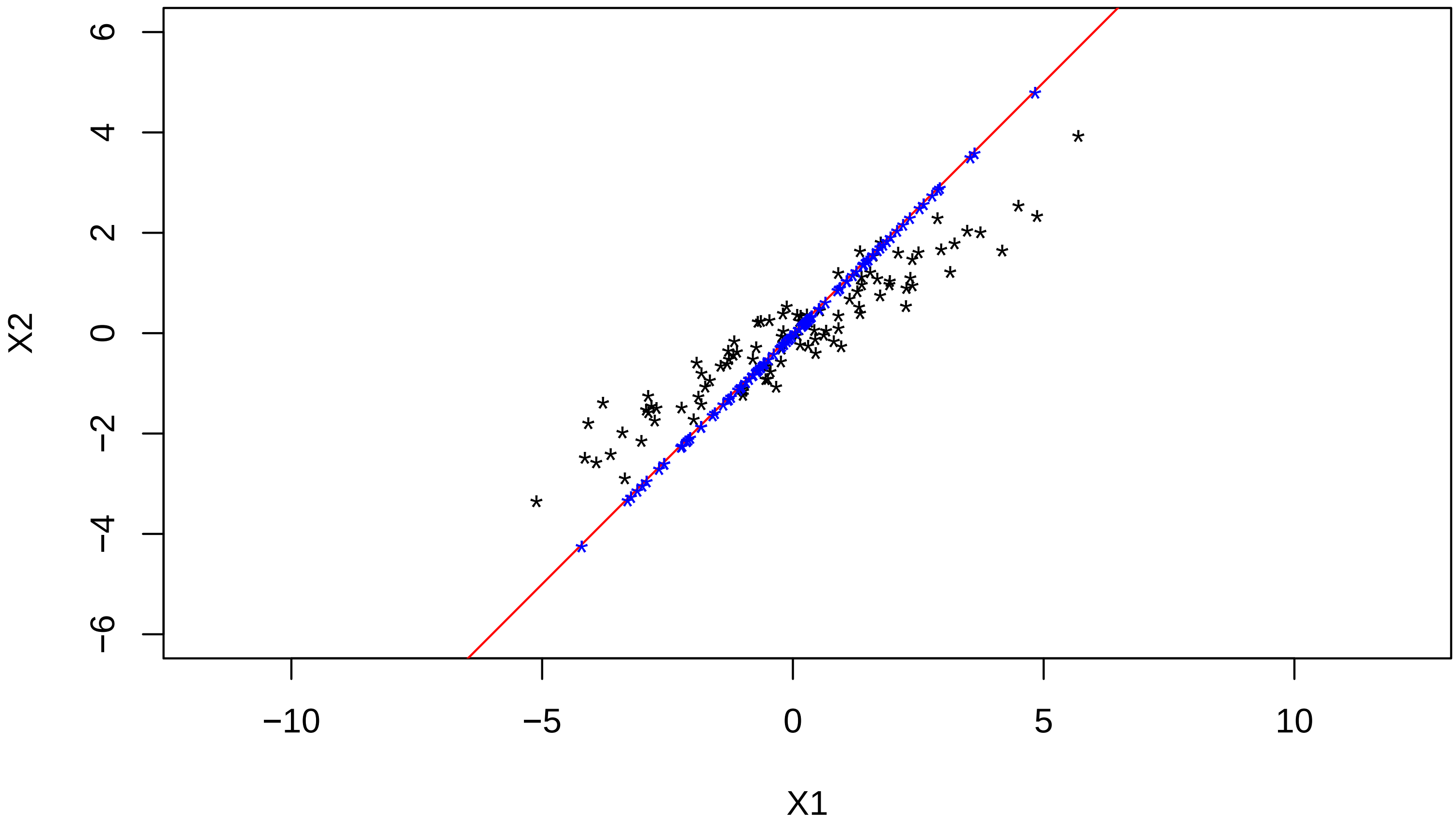
$$Y_i = \text{age}_i/\sqrt{2} + \text{height}_i/\sqrt{2}.$$

The values

$$Y_i = X_{i1}/\sqrt{2} + X_{i2}/\sqrt{2}:$$



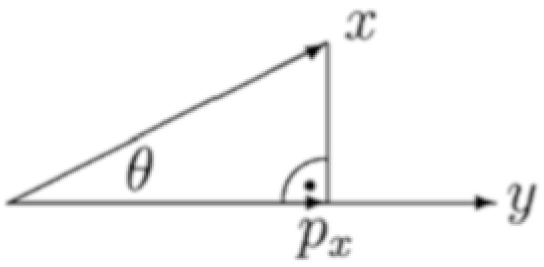
Taking a scaled average of the two components = projecting the data onto the 45 degree line (red) and keeping only the projected values.



How was this figure constructed?

☛ Recap: the projection p_x of a vector x, onto a vector y, is the vector

$$p_x = \frac{x^T y}{\|y\|^2} y, .$$



☛ Another more transparent way of viewing this:

$$p_x = \underbrace{\frac{x^T y}{\|y\|}}_{\substack{\text{length of} \\ \text{the projection} \\ \text{(“projected value”)}}} \underbrace{\frac{y}{\|y\|}}_{\substack{\text{unit vector} \\ \text{in the direction of} \\ \text{the projection}}}$$

☛ The linear combination $Y_i = X_{i1}/\sqrt{2} + X_{i2}/\sqrt{2}$ is the same as

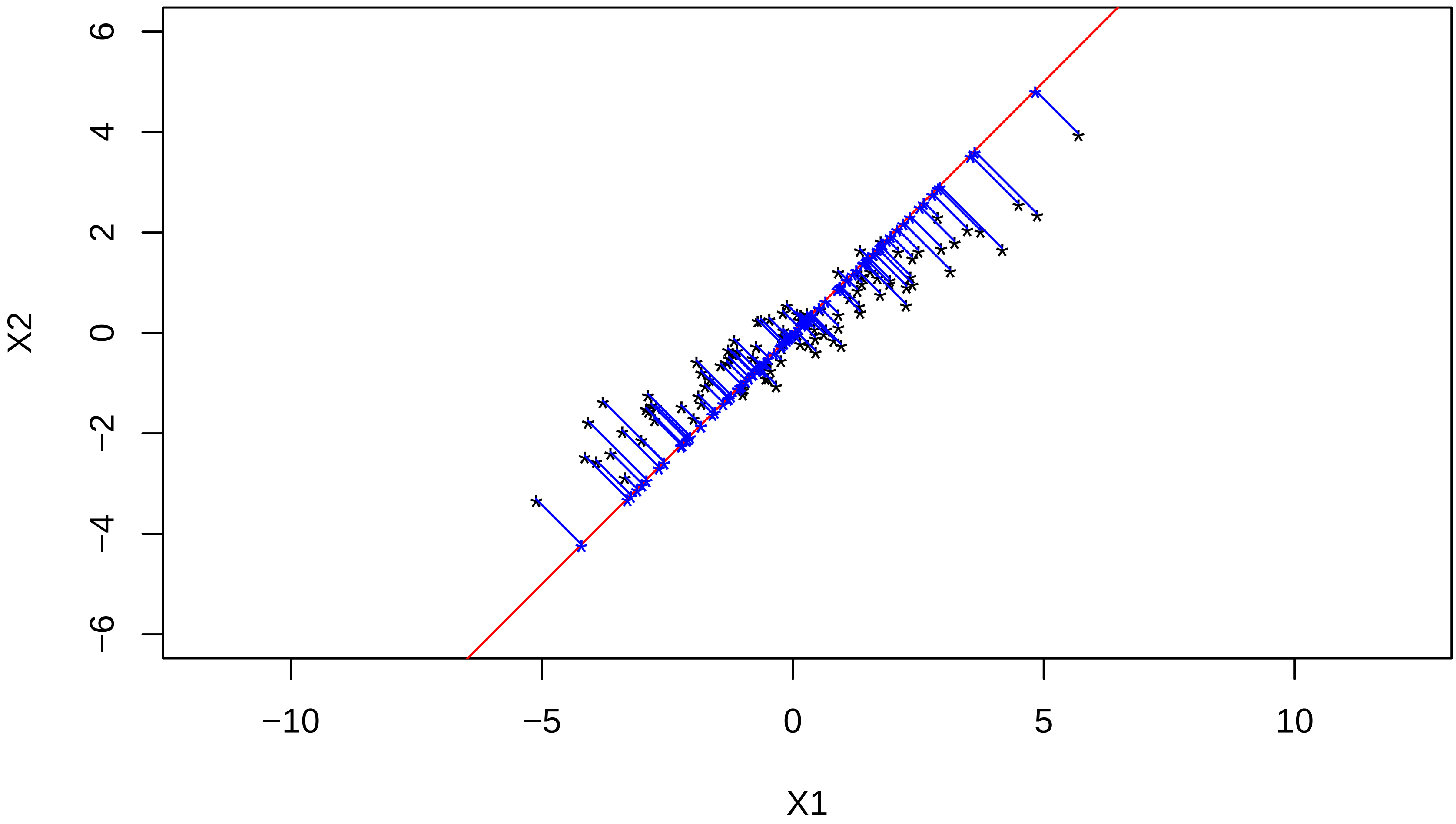
$$Y_i = X_i^T a$$

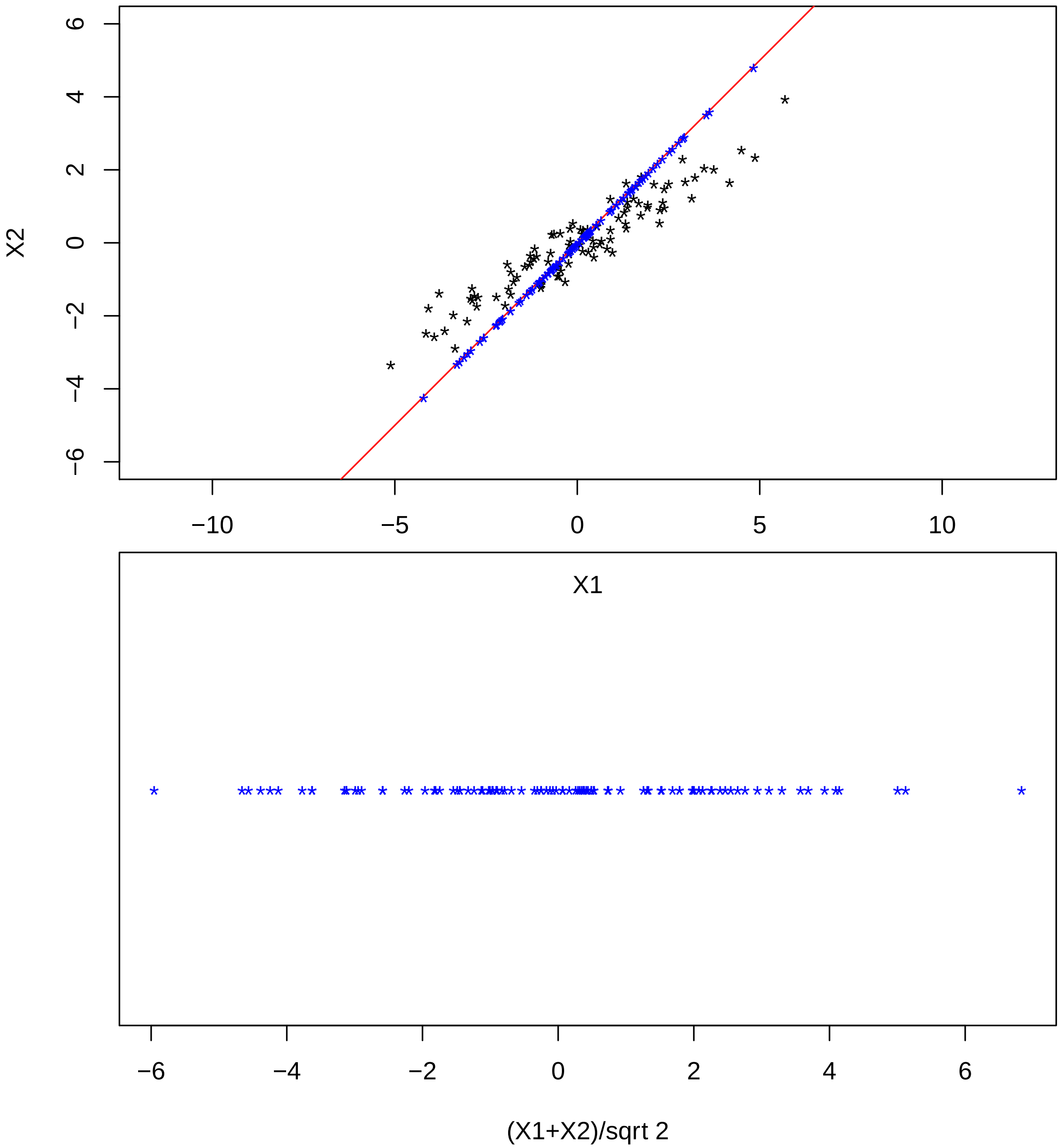
where

$$X_i = (X_{i1}, X_{i2})^T, \;\; a = (1/\sqrt{2}, 1/\sqrt{2})^T.$$

☛ So Y_i is the projection value (length of the projection vector) of X_i onto the 45 degree line passing through the origin.

The the line passing through the origin and $a = (1/\sqrt{2}, 1/\sqrt{2})^T$ is shown in red. The projection of each X_i on that line is shown in blue.




X2
6
4
2
0
−2
−4
−6
−10
−5
0
5
10
X1
−6
−4
−2
0
2
4
6
(X1+X2)/sqrt 2

☛ Instead of giving equal weight to each component of X_i, when reducing dimension we would like to lose as little information about the original data as possible.

☛ How we define “lose information” ?

☛ In principal component analysis (PCA), we reduce dimension by projecting the data onto lines.

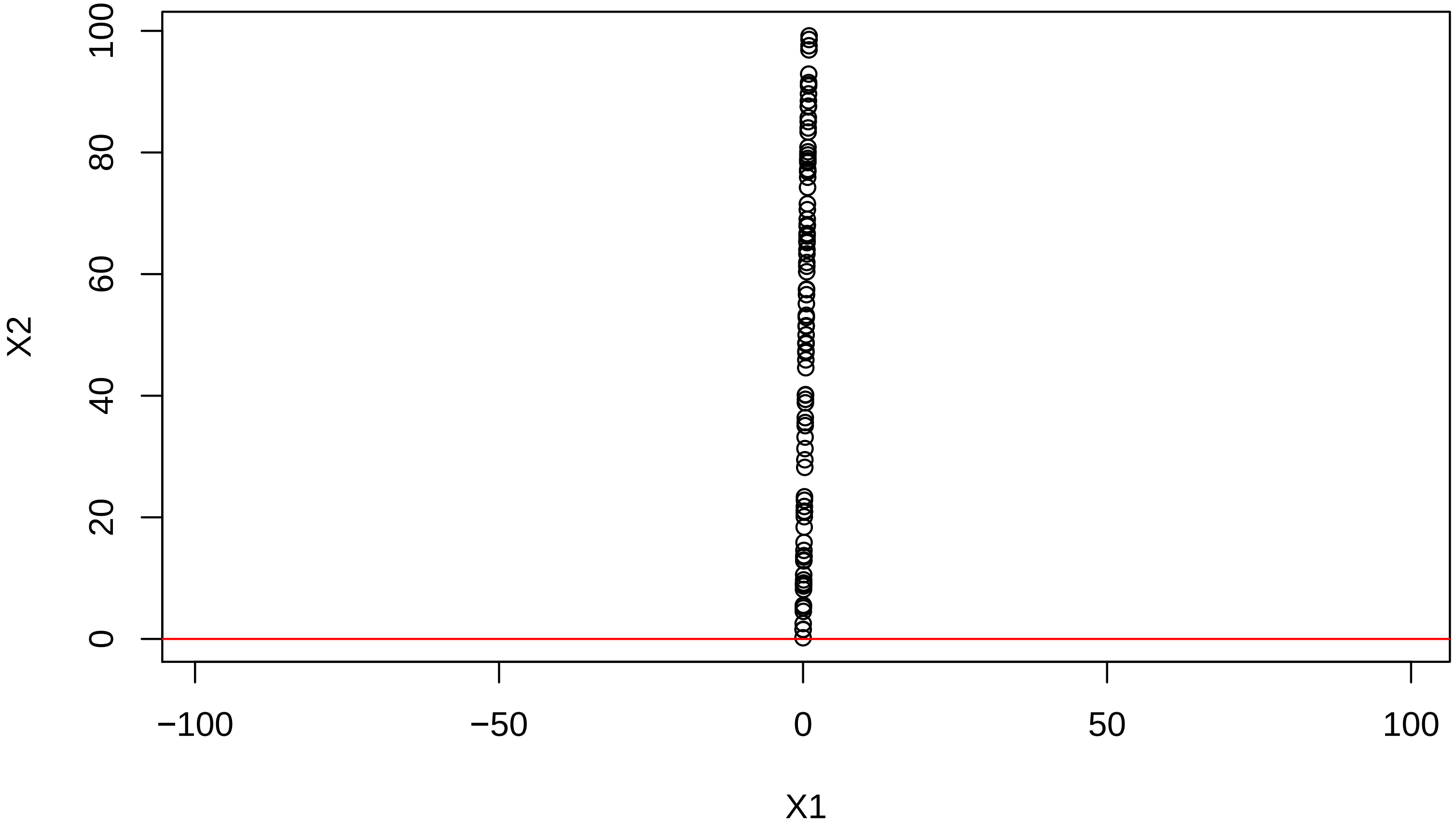
☛ Moreover, in PCA, “lose as little information as possible” is defined as “keep as much of the variability of the original data as possible”.

☛ In our two dimensional case example, when choosing the projection $Y_i = X_i^T a$ on a line, this means we want to find a such that

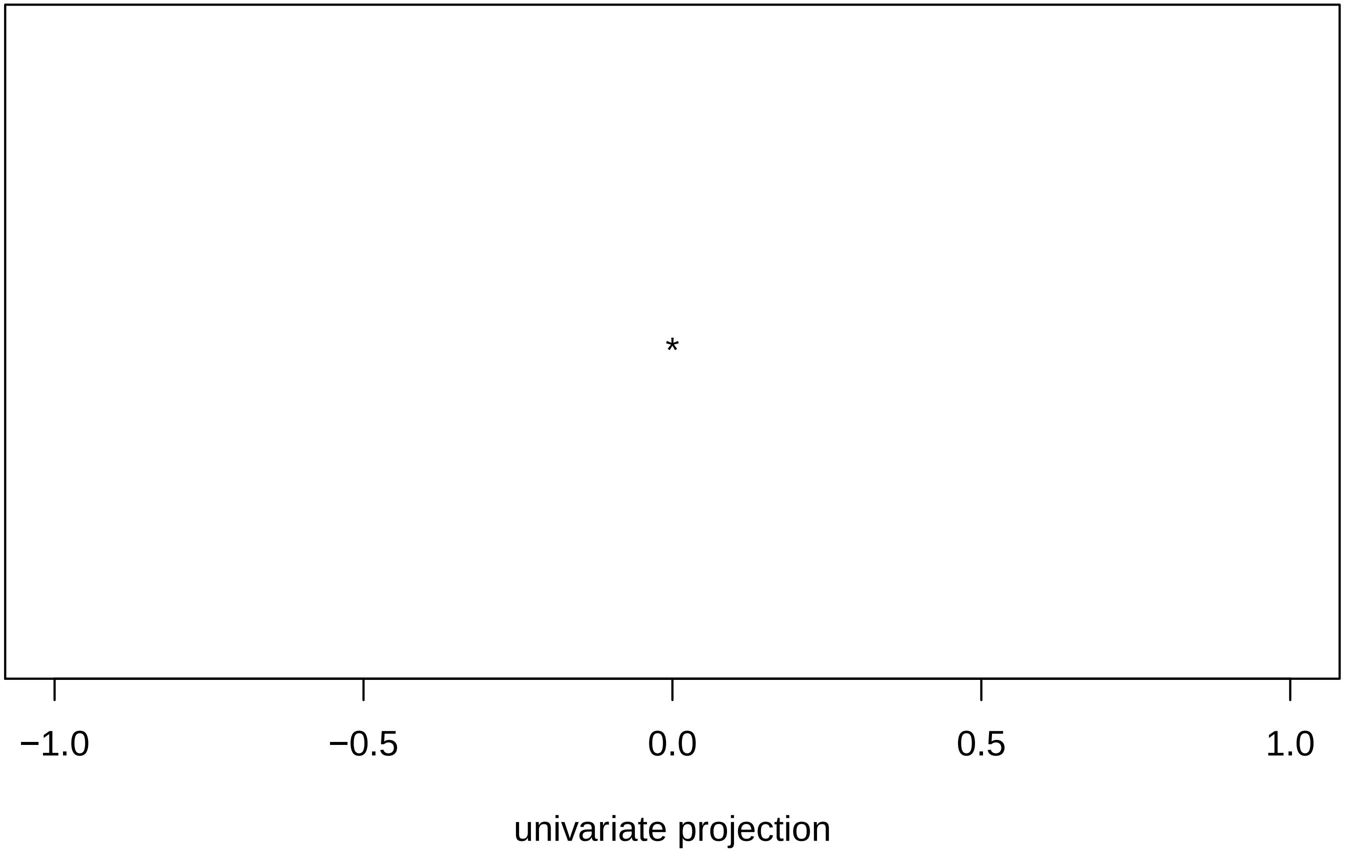
$$var(Y_i)$$

is as large as possible.

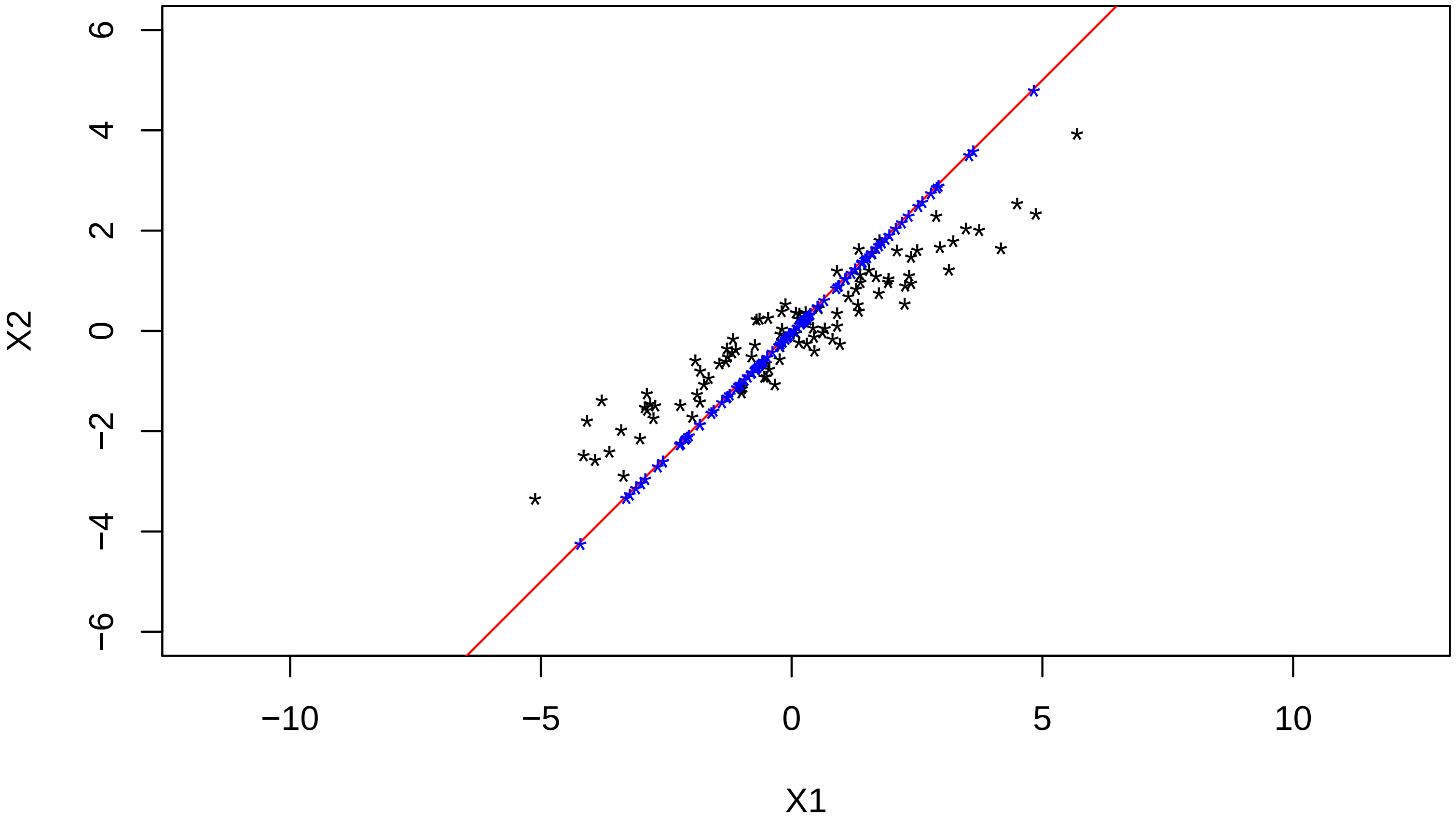
Why do we want to maximise variance? Here is an example where the projected data have *no variability*: project the data on the red line



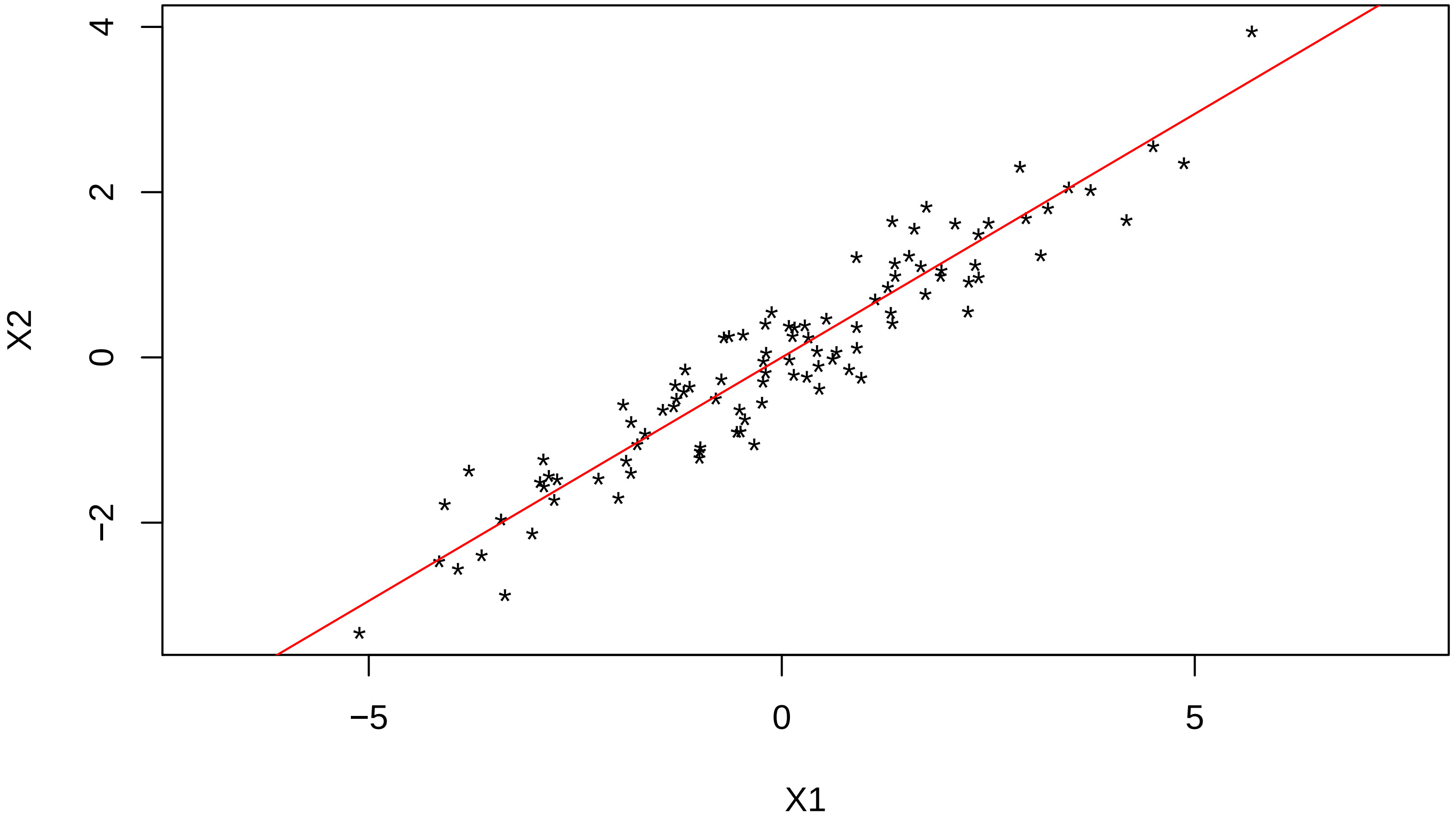
Projected data land on the same point and have zero variance; don't learn anything about the data.

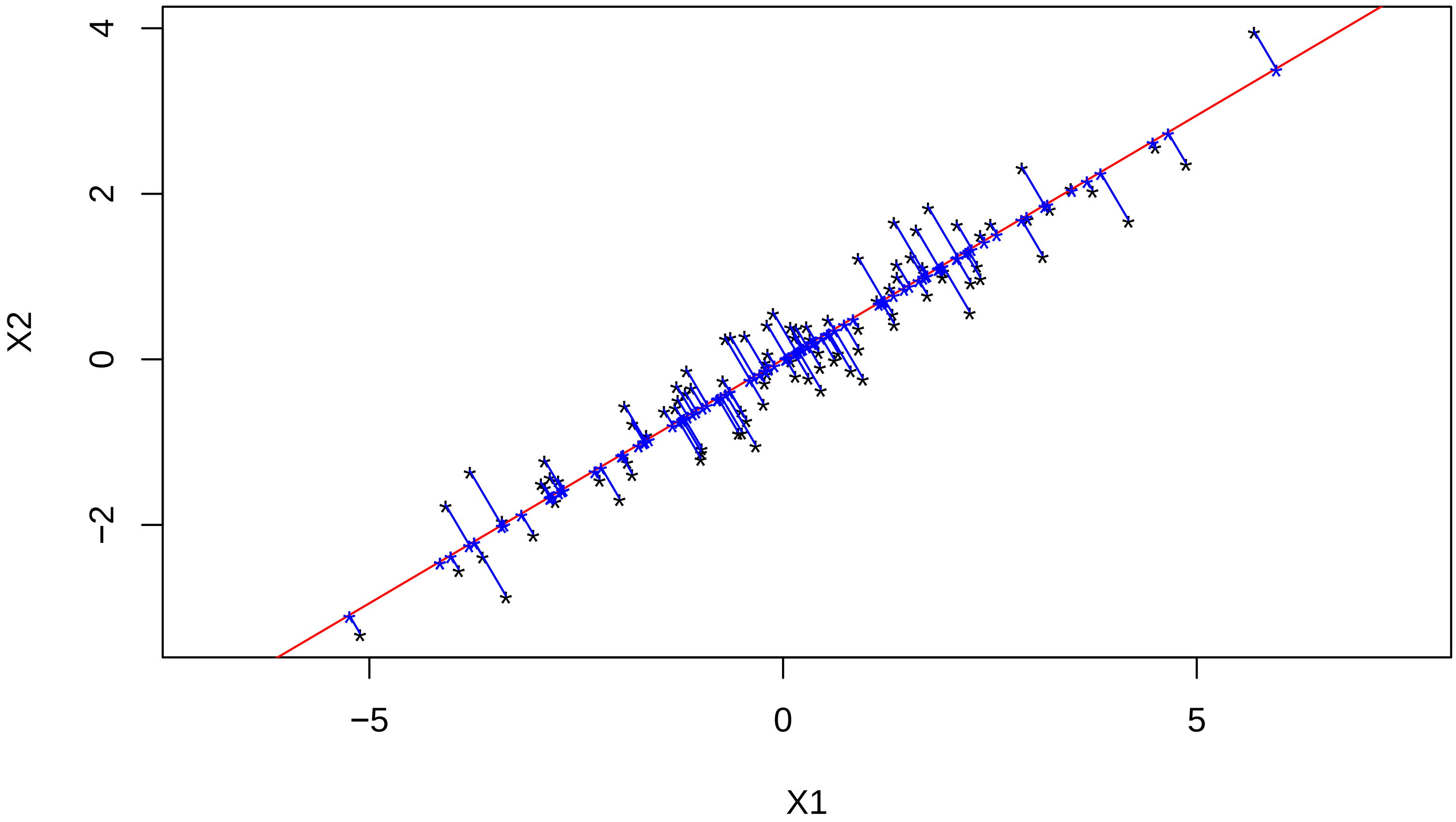


Recall in the example, we've projected onto the 45-degree line through the origin:



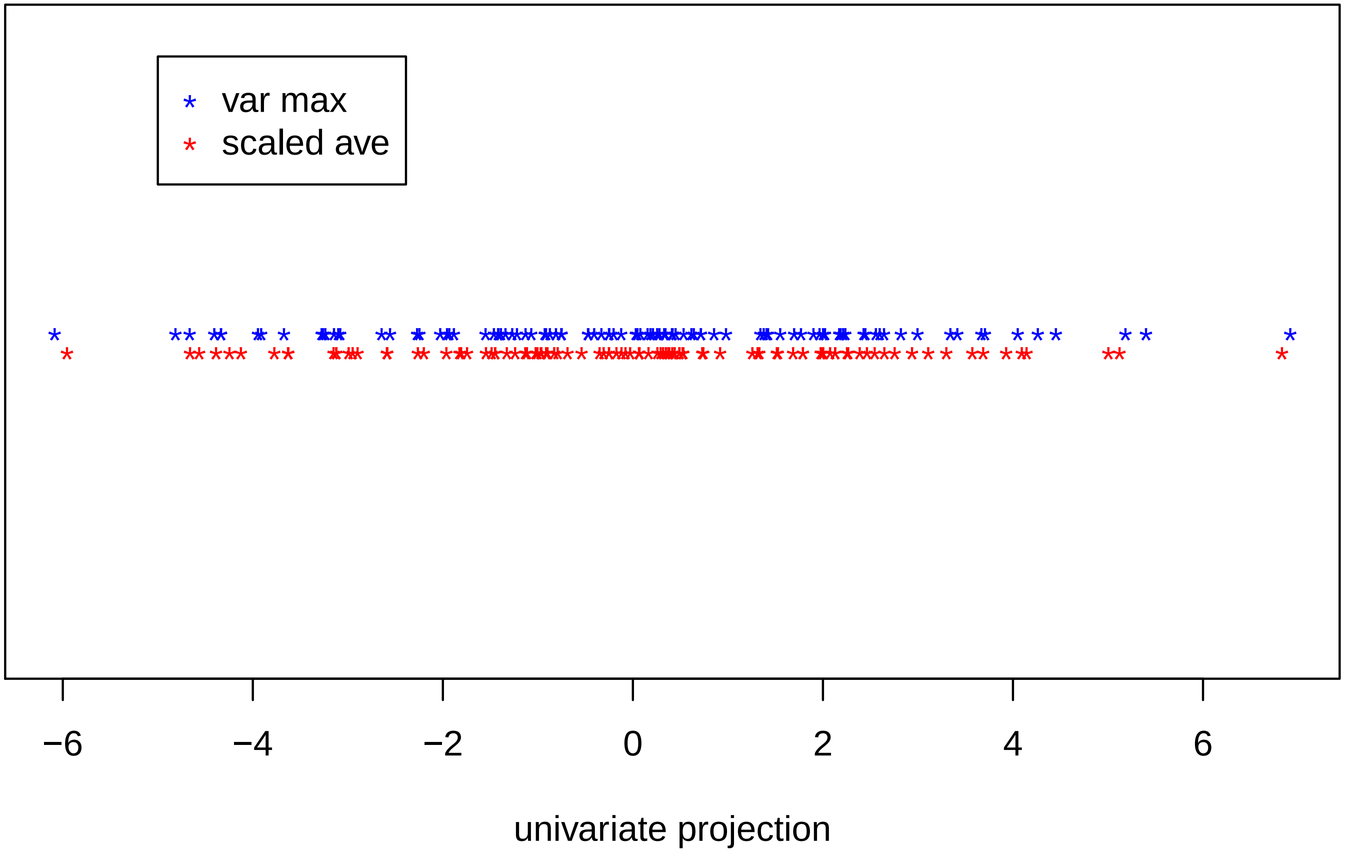
However we would have kept more information if we had instead projected the data on the following line:





Indeed, on this line, the projected data are more variable than on the previous line.

The scaled average (in red) is less variable than the last suggested projected values (in blue):



The blue projections in fact maximize the variance of the projected values of the data.

Formally, in PCA, when reducing the p-variate X_i's to univariate Y_{i1}'s, for $i = 1, \ldots, n$, where the X_i's are i.i.d.$\sim (0, \Sigma)$, the goal is to find the linear combination

$$Y_{i1} = a_1 X_{i1} + \ldots + a_p X_{ip} = X_i^T a \, ,$$

where $a = (a_1, \ldots, a_p)^T$ is such that

$$\|a\|^2 = \sum_{j=1}^{p} a_j^2 = 1$$

and

$$var(Y_{i1})$$

is as large as possible.

☛ We use Y_{i1} instead of Y_i because there will be more than one projection.

☛ The unit-length constraint on a makes the problem well-defined; otherwise, $var(Y_{i1})$ can be made as large as one wants by multiplying a with an arbitrary large scalar.

☛ Let $\gamma_1, \ldots, \gamma_p$ denote the p unit-length eigenvectors (i.e., $\|\gamma_j\| = 1$) of the covariance matrix Σ, respectively associated with the eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p.$$

(Note: The orthonormal γ_j's are only defined up to a change of sign, so each γ_j can be replaced by $-\gamma_j$.)

☛ The a that maximises the variance is equal to

$$\gamma_1,$$

the eigenvector with largest eigenvalue ("first eigenvector") .

☛For $a = \gamma_1$, the variable

$$Y_{i1} = a_1 X_{i1} + \ldots + a_p X_{ip} = a^T X_i = \gamma_1^T X_i$$

is called the first principal component of X_i, or "PC1" for short. This is precisely the projected value of X_i in the direction of γ_1.

☛ More generally, if the data are i.i.d.$\sim (\mu, \Sigma)$ and not already centered,

$$Y_{i1} = \gamma_1^T \{X_i - E(X_i)\} = \gamma_1^T (X_i - \mu)$$

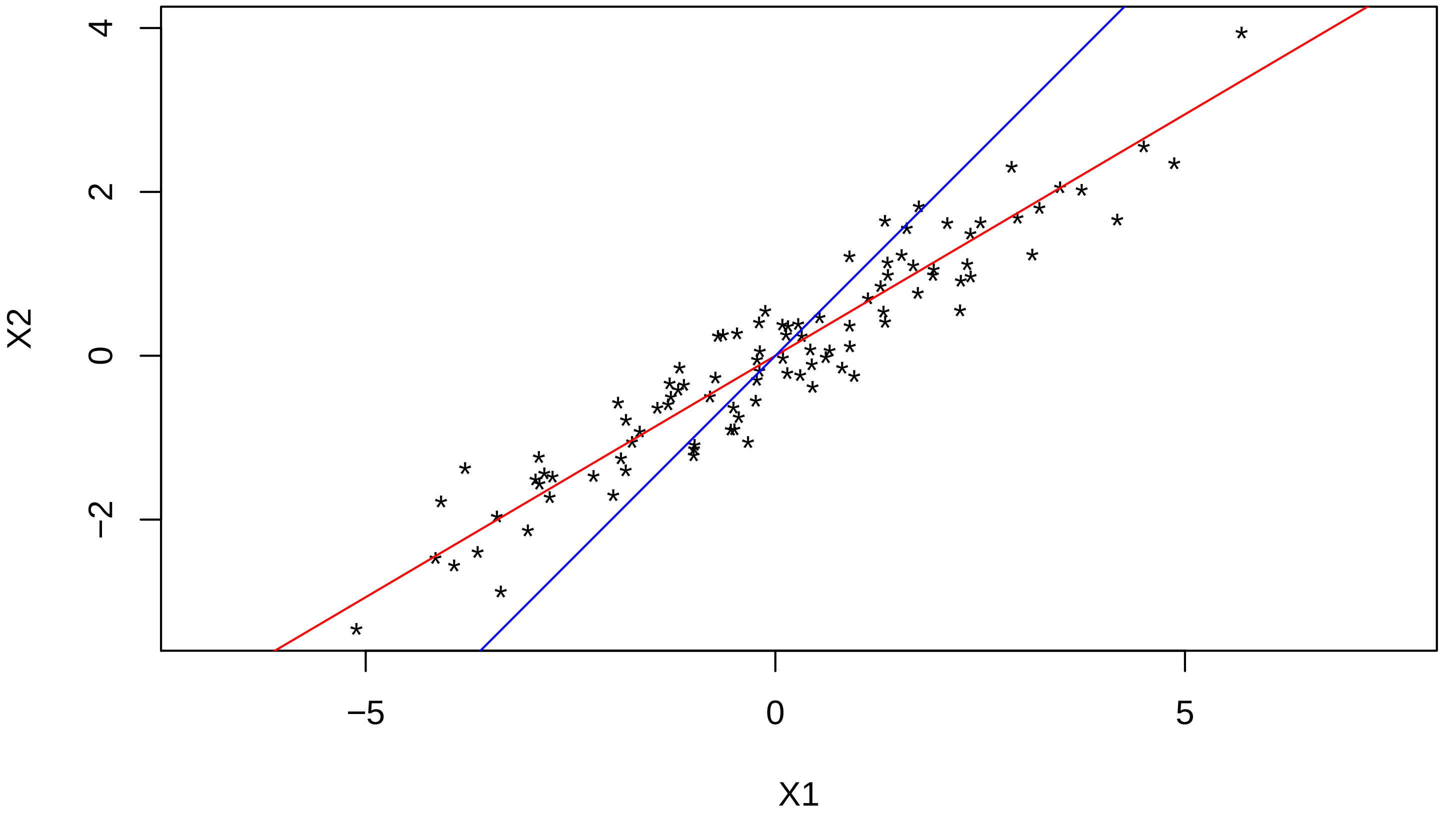
is called the first principal component of X_i.

☛ It is the linear projection of the data that has maximum variance.

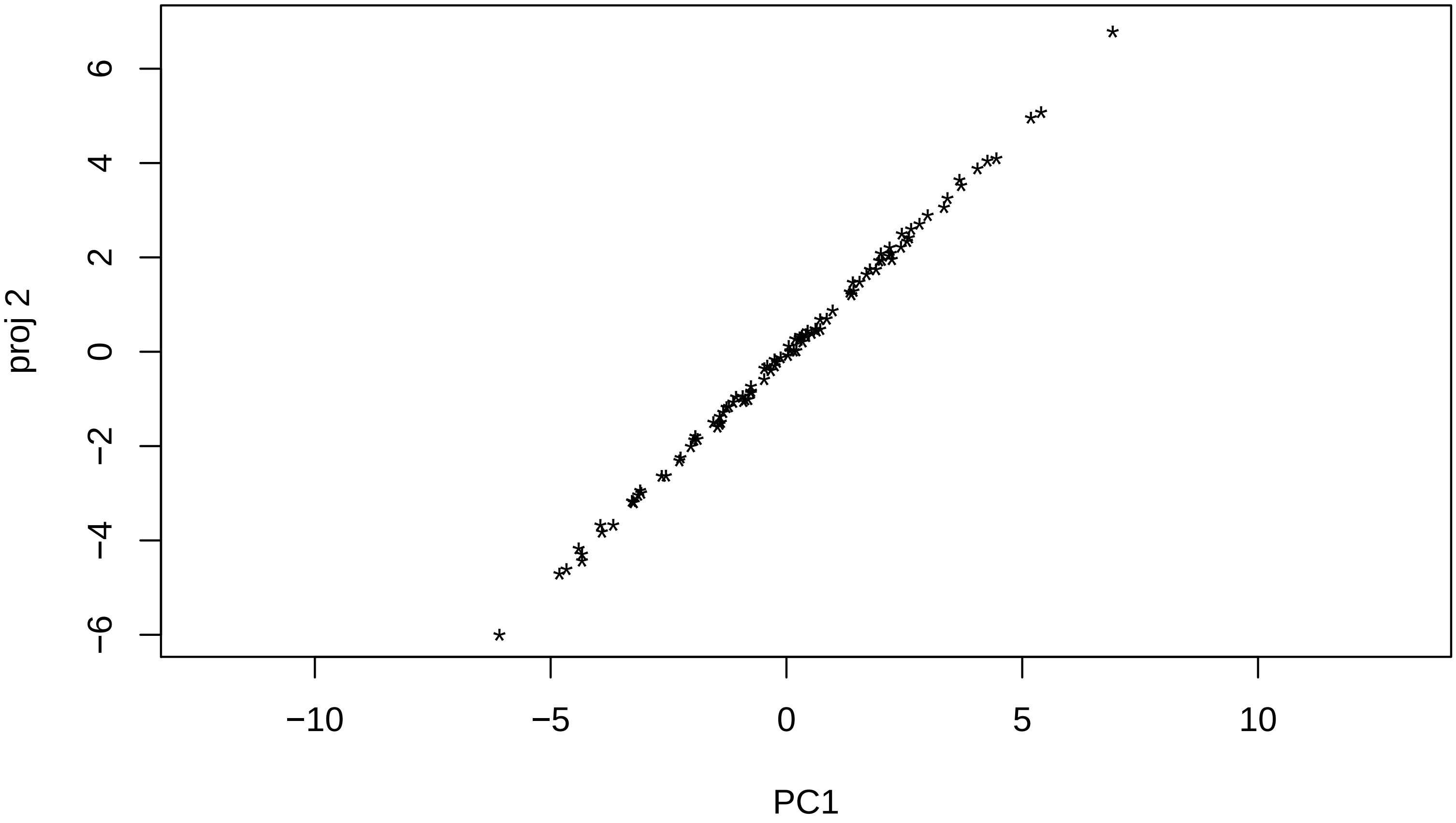
☛ We always center the data before projecting.

☛ In PCA, once we have found a univariate projection, how do we add a second projection?

☛ One possibility: on the good old 45-degree line (blue below)

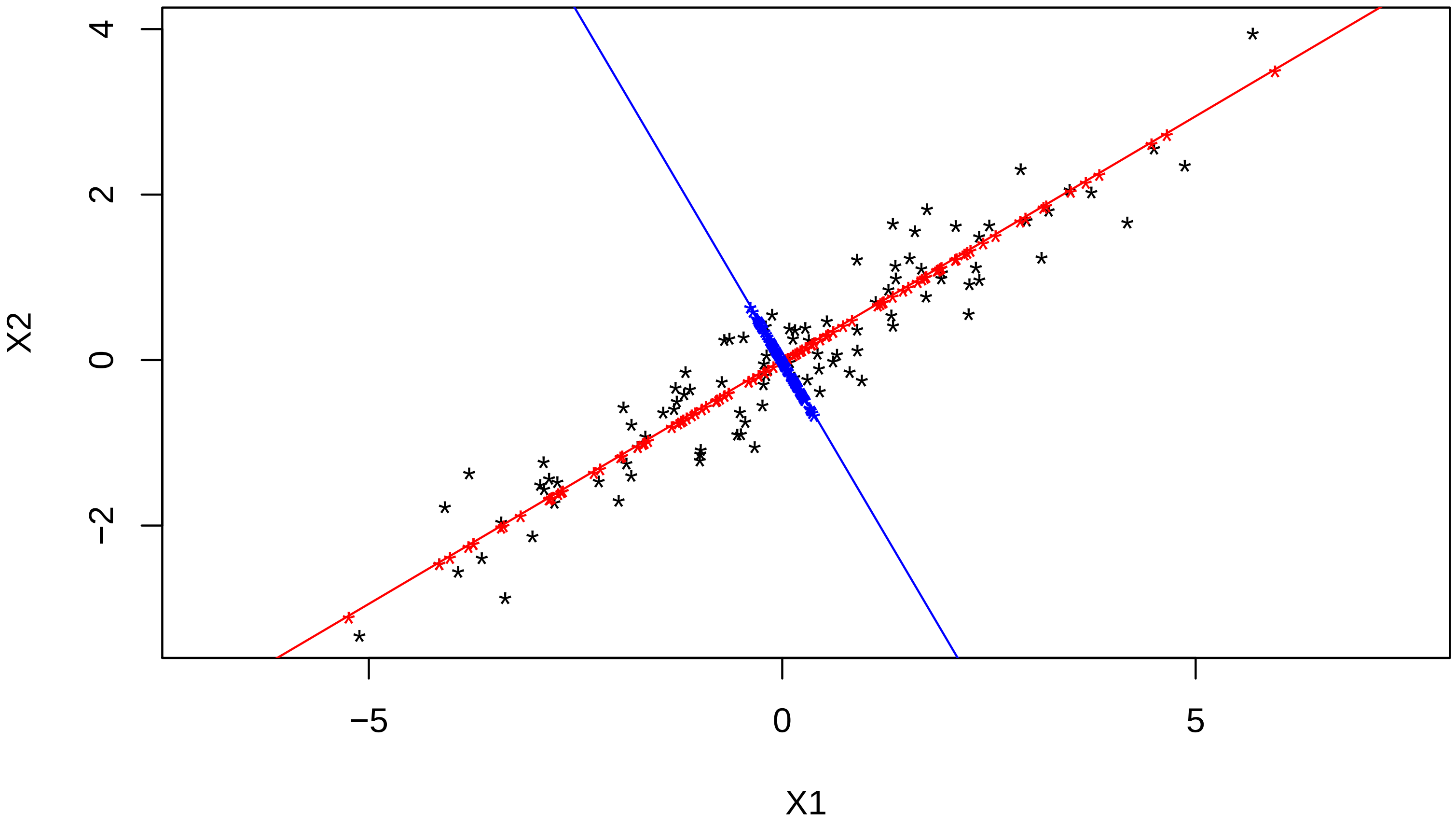


Those two projections are essentially redundant, we don’t learn much more:

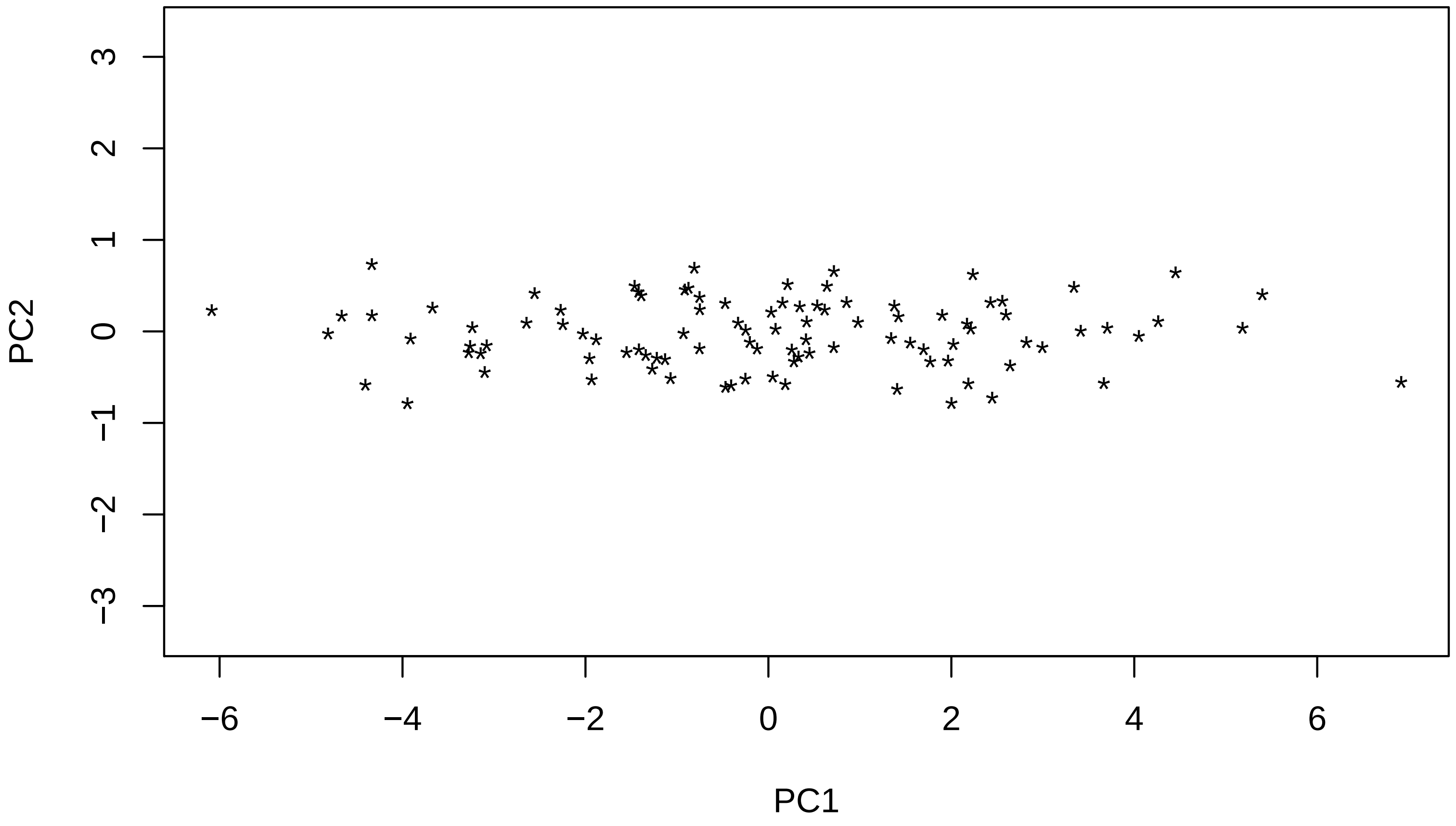


☛ We should project onto a line as different as possible, to learn complementary information. How?

☛ Project onto a perpendicular direction to that of the PC1. The variable obtained is called the *second principal component* ("PC2" for short).



The data projected on the two lines are just the same as the original data, but the axes have been rotated to match the blue and the red lines.



More generally, suppose we have p-dimensional data $X_i \sim (\mu, \Sigma)$, where $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues of Σ and $\gamma_1, \ldots, \gamma_p$ are the corresponding eigenvectors. In PCA, we reduce these X_i's into q-dimensional data for $q \leq p$ via these steps:

- Take the first principal component of X_i

$$Y_{i1} = \gamma_1^T \{X_i - E(X_i)\} = \gamma_1^T (X_i - \mu);$$

γ_1 the eigenvector of Σ corresponding to the largest eigenvalue λ_1.

- Then for $k = 2, \ldots, q$, we take the kth principal component of X_i

$$Y_{ik} = \gamma_k^T \{X_i - E(X_i)\} = \gamma_k^T (X_i - \mu) \tag{1}$$

where γ_k is the eigenvector of Σ corresponding to the kth largest eigenvalue, λ_k.

- γ_j's are orthonormal $\Rightarrow$ the projection directions are orthogonal to each other.

☛ In matrix notation, letting $Y_i = (Y_{i1}, \ldots, Y_{ip})^T$ and $\Gamma = [\gamma_1 | \ldots | \gamma_p]$, we have

$$Y_i = \Gamma^T (X_i - \mu).$$

☛ Suppose we construct $Y_{i1}, \ldots, Y_{ip}$ as described above. Then we have

$$E(Y_{ij}) = 0, \quad \text{for } j = 1, \ldots, p$$

$$var(Y_{ij}) = \lambda_j, \quad \text{for } j = 1, \ldots, p$$

$$cov(Y_{ik}, Y_{ij}) = 0, \quad k \neq j$$

$$var(Y_{i1}) \geq var(Y_{i2}) \geq \ldots \geq var(Y_{ip})$$

$$\sum_{j=1}^{p} var(Y_{ij}) = tr(\Sigma)$$

$$\prod_{j=1}^{p} var(Y_{ij}) = |\Sigma|.$$

☛ Property 1: it is not possible to construct a linear combination

$$V_i = a^T X_i \text{ where } \|a\| = 1$$

which has larger variance than $\lambda_1 = var(Y_{i1})$.

☛ Property 2: if we take a variable

$$V_i = a^T X_i \text{ where } \|a\| = 1$$

which is not correlated with the first k PCs of X_i, then the variance of V_i is maximised by taking $V_i = Y_{i,k+1}$, the $(k+1)$-th PC of X_i.

☛ Note: q is generally chosen to be much less than p if p is large.

☛ With these properties, we hope to gather as much information as possible about the original data by projecting them onto a few PCs.

Proof of Property 1: (Assuming $E[X_i] = 0$ without loss of generality)

☛ For any a such that $\|a\| = 1$, by spectral decomposition, we know that

$$var(a^T X_i) = a^T \Gamma \Lambda \Gamma^T a,$$

where $\Sigma = \Gamma \Lambda \Gamma^T$ is the spectral decomposition of Σ, $\Lambda = diag(\lambda_1, \dots, \lambda_p)$ and $\Gamma = (\gamma_1 | \cdots | \gamma_p)$.

☛ Let $b = (b_1, \dots, b_p)^T = \Gamma^T a$. Since multiplication by an orthogonal matrix is norm preserving, we know that $\|b\| = 1$.

☛ Write

$$var(a^T X_i) = b_1^2 \lambda_1 + \cdots + b_p^2 \lambda_p.$$

Since $b_1^2 + \cdots + b_p^2 = 1$, these $b_1^2, \dots, b_p^2$ can be thought of as non-negative weights that sum to 1. Since $\lambda_1 \geq \cdots \geq \lambda_p$, $var(a^T X_i)$ is maximized if we can "invest" all the weights on λ_1 to give $var(a^T X_i) = \lambda_1$, i.e. when $b_1^2 = 1$, and $b_2^2 = \cdots = b_p^2 = 0$.

☛ In particular, $b = (1, 0, \dots, 0)$ precisely happens when we pick $a = \gamma_1$ to begin with.

Proof of Property 2: (Assuming $E[X_i] = 0$ without loss of generality)

☛ Suppose we have already formed the first k PC with $\gamma_1, \ldots, \gamma_k$, and want to find a linear combination $a^T X_i$ with maximized variance, subject to $\|a\| = 1$ and

$$Cov(a^T X_i, \gamma_j^T X_i) = a^T \Sigma \gamma_j = 0 \text{ for all } j = 1, \ldots, k.$$

☛ By defining $c = \Gamma^T a$ and the spectral decomposition $\Sigma = \Gamma \Lambda \Gamma^T$, we know that

$$a^T \Sigma \gamma_j = a^T \Gamma \Lambda e_j = \lambda_j a^T \Gamma e_j = \lambda_j c^T e_j,$$

where $e_j = (0 \ldots, 0, \underbrace{1}_{j \text{ th}}, 0 \ldots, 0)^T$ is a column vector with a "1" at its j-th position and zeros elsewhere.

☛ Hence, for $a^T \Sigma \gamma_j$ to be zero for all $j = 1, \ldots, k$, the vector c needs to have zeros in its 1-th, $\ldots$, k-th positions. Recalling the definition of c, this is precisely true when a is orthogonal to all of $\gamma_1, \ldots, \gamma_k$.

☛ In other words, our problem is equivalent to:

$$\max var(a^T X_i)$$

subject to $a^T \gamma_1 = \ldots a^T \gamma_k = 0$ and $\|a\| = 1$.

Proof of Property 2 (Con’t):

- Now we let $b = \Gamma^T a$, where a respects all the constraints above.

- Subject to $a^T\gamma_1 = \ldots, a^T\gamma_k = 0$, the first k positions of b must be zero, i.e. $b = (0, \ldots, 0, b_{k+1}, \ldots, b_p)^T$. Moreover, since transforming by Γ^T, an orthogonal matrix, is norm-preserving, subject to $\|a\| = 1$ we must have $b_{k+1}^2 + \cdots + b_p^2 = 1$

- By spectral decomposition, write

$$\begin{aligned} var(a^T X_i) &= a^T \Sigma a \\ &= b^T \Lambda b = \lambda_{k+1} b_{k+1}^2 + \cdots + \lambda_p b_p^2. \end{aligned}$$

  By the“weight investment” argument similar to the proof of Property 1, we know that this quantity can be maximized when $b_{k+1}^2 = 1$ and $b_{k+2} = \cdots = b_p = 0$.

- The latter amounts to $b = (0, \ldots, 0, \underbrace{1}_{k+1 \text{ th}}, 0, \ldots, 0)^T$, which is precisely true when $a = \gamma_{k+1}$.

5.3 In practice

In practice: We do not know Σ nor $\mu = E(X_i)$. Instead we use their empirical counterparts S and $\bar{X}$, i.e.:

☛ We start by taking the first principal component of X_i

$$Y_{i1} = g_1^T(X_i - \bar{X})$$

where g_1 is the eigenvector of S corresponding to the largest eigenvalue ℓ_1 of S.

☛ Then for $k = 2, \dots, q$, we take the kth principal component of X_i

$$Y_{ik} = g_k^T(X_i - \bar{X})$$

where g_k is the eigenvector of S corresponding to the kth largest eigenvalue ℓ_k of S.

☛ In matrix notation, letting

$$Y_i = (Y_{i1}, \ldots, Y_{ip})^T \text{ [a p-vector]}$$

and

$$\mathcal{Y} = (Y_1, \ldots, Y_n)^T, \text{ [an n-by-p matrix]}$$

we have

$$\mathcal{Y} = (\mathcal{X} - 1_n \bar{X}^T)G$$

for $G = [g_1| \ldots |g_p]$.

☛ Once we have computed the PC's we can:

$*$ plot them to see if we can detect clusters
$*$ see influential observations (outliers)
$*$ see if we can get any insight about the data.

☛ When we detect something in the PC plots, we can:

$*$ go back to the original data and try to make the connection,
$*$ check if our interpretation seems correct.

☛ Example: Swiss bank notes data.

☛ Data: variables measured on 200 Swiss 1000-franc banknotes, of which 100 were genuine and 100 were counterfeit. Found in the R package `mclust` by typing `data(banknote)`.

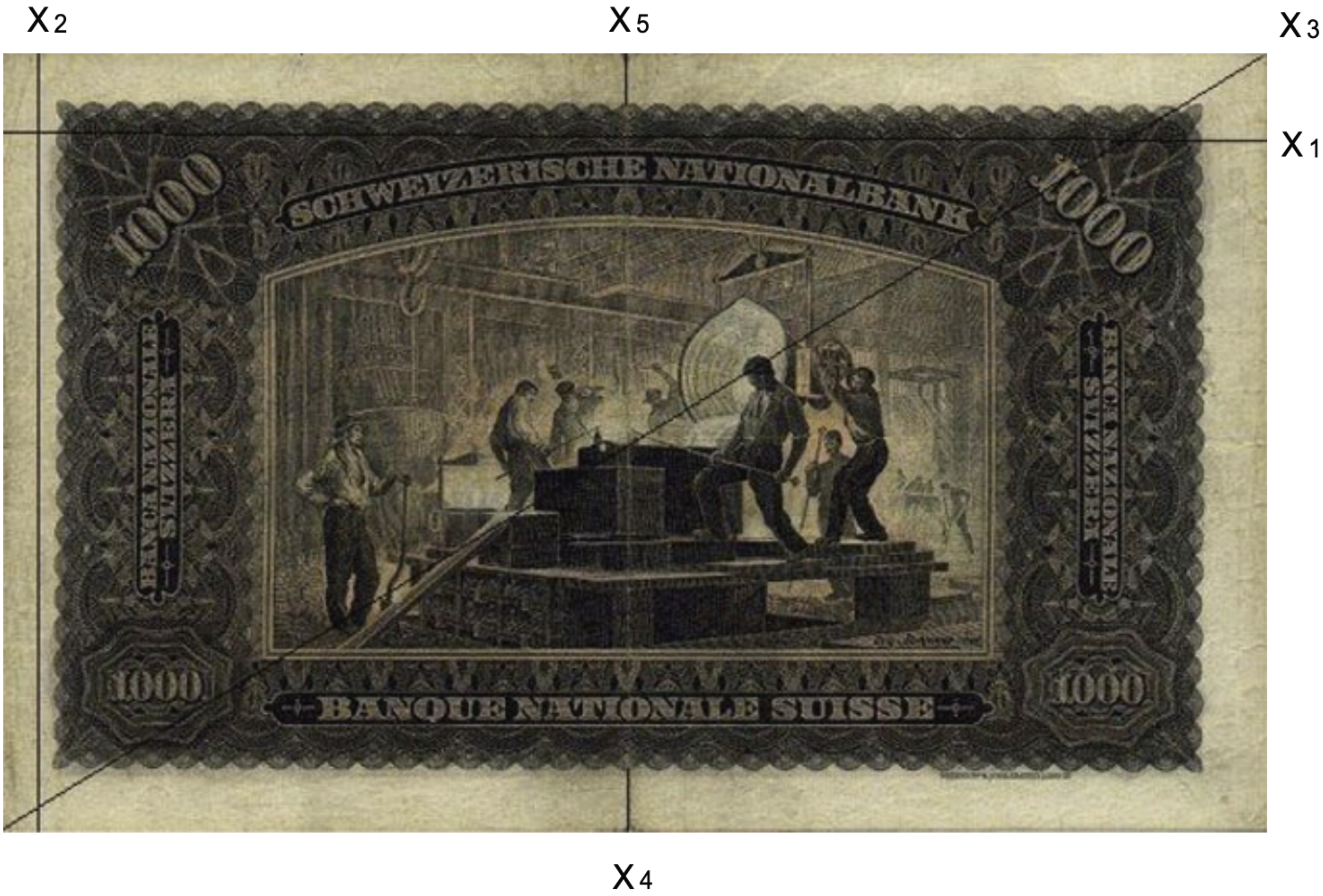
(Source: Flury, B. and Riedwyl, H. (1988). Multivariate Statistics: A practical approach. London: Chapman & Hall, Tables 1.1 and 1.2, pp. 5–8.)

☛ The variables measured are:

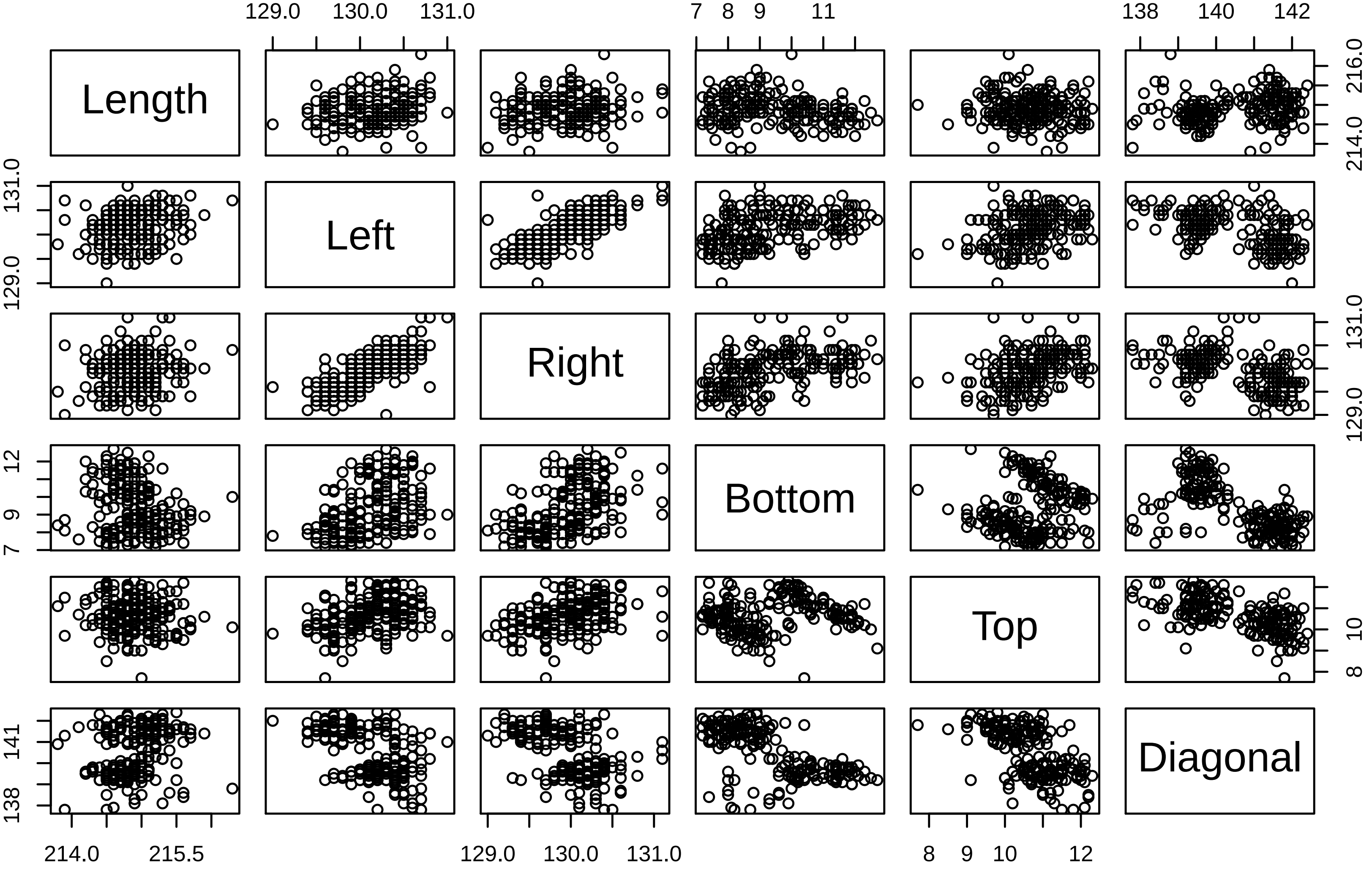
X_1: Length of bill (mm)
X_2: Width of left edge (mm)
X_3: Width of right edge (mm)
X_4: Bottom margin width (mm)
X_5: Top margin width (mm)
X_6: Length of diagonal (mm)

☛ The first 100 banknotes are genuine and the next 100 are counterfeit.

An old Swiss 1000-franc banknote:



Scatterplots:



In R, read the data and produce the scatterplots:

```
library(mclust)
data(banknote)
StatusX=banknote[,1]
plot(banknote[,2:7])
```

`Status` contains the info about whether a note is genuine or counterfeit. Center the data and perform the PC analysis:

```
XCbank=scale(banknote[, 2:7], scale = FALSE)
PCX=prcomp(XCbank,retx=T)
PCX
```

Let's take a closer look at the two PCs for the banknote data.
The eigenvalues and eigenvectors are given in R in the following form:

```
Standard deviations (1, .., p=6):
[1] 1.7321388 0.9672748 0.4933697 0.4412015 0.2919107 0.1884534

Rotation (n x k) = (6 x 6):
           PC1    PC2   PC3    PC4    PC5    PC6
Length    0.044 -0.011 0.326 -0.562 -0.753  0.098
Left     -0.112 -0.071 0.259 -0.455  0.347 -0.767
Right    -0.139 -0.066 0.345 -0.415  0.535  0.632
Bottom   -0.768  0.563 0.218  0.186 -0.100 -0.022
Top      -0.202 -0.659 0.557  0.451 -0.102 -0.035
Diagonal  0.579  0.489 0.592  0.258  0.084 -0.046
```

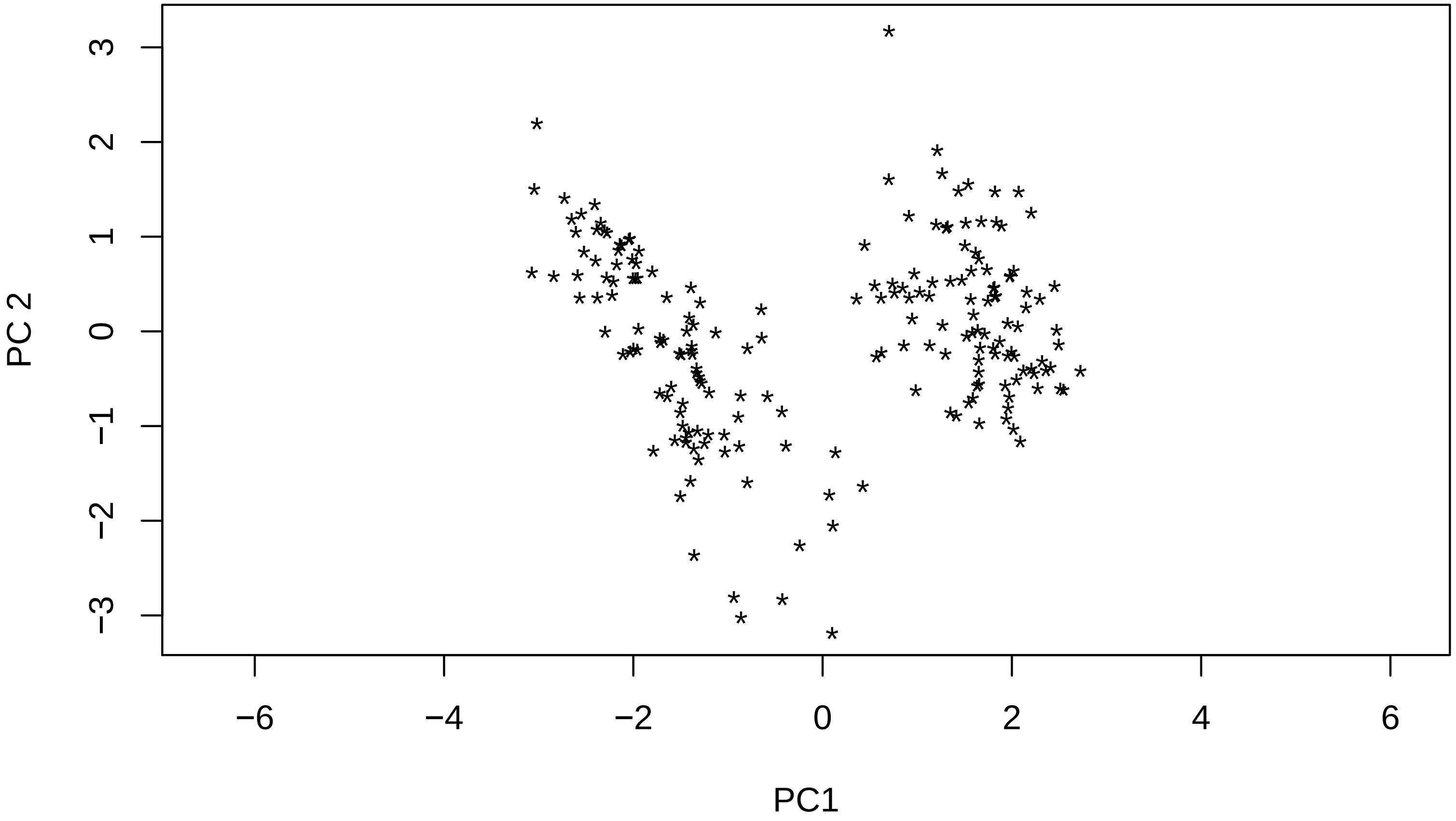
The eigenvectors are the columns of the so called rotation matrix and the eigenvalues are the square of the so-called standard deviations.

Keep eigenvectors in `gamma` and eigenvalues in `lambda`.

```
gamma=PCX$rotation
lambda=PCX$sdev^2
```

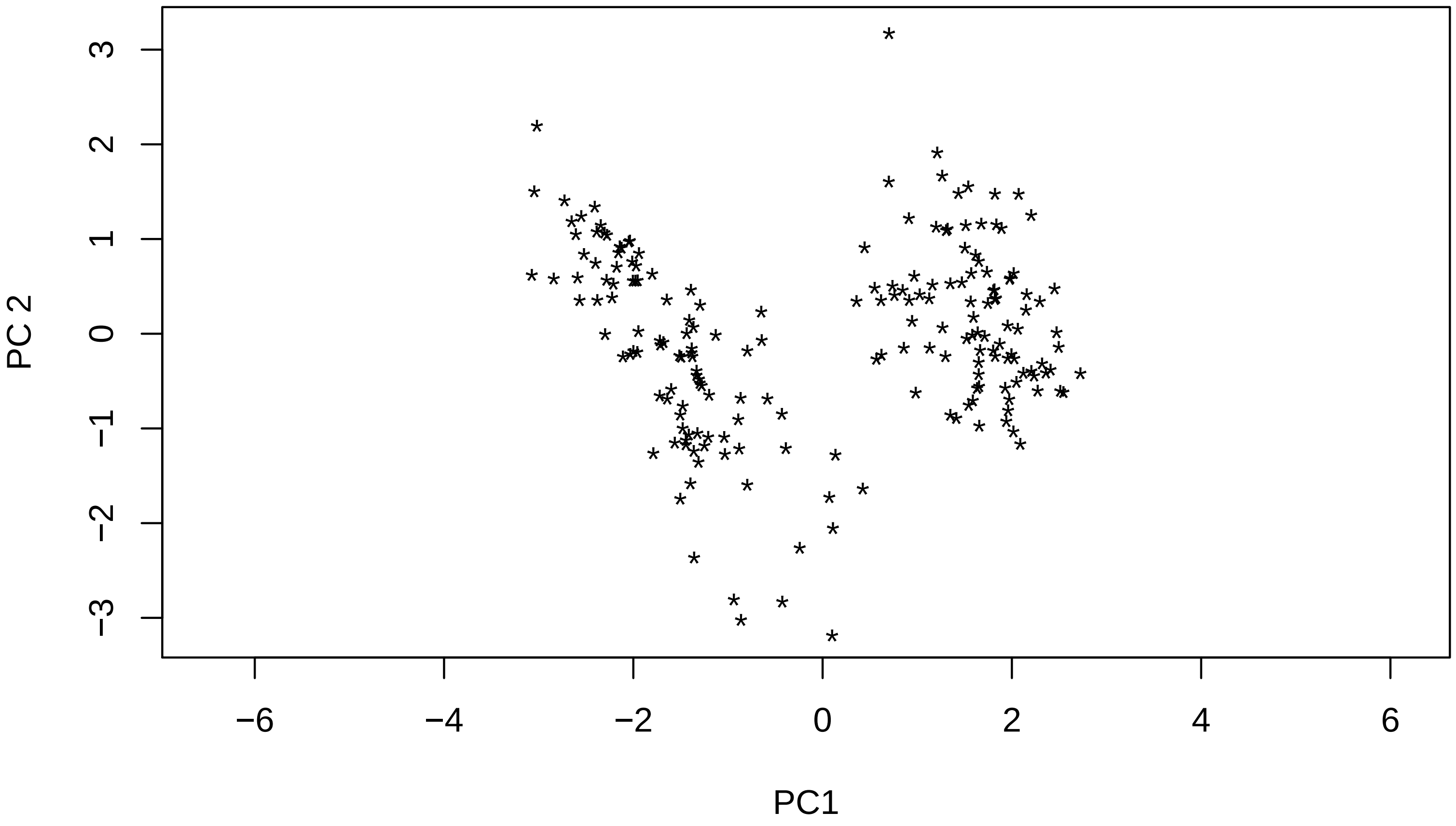
Let’s looks at the first 2 PCs: the data clearly separate into two groups

```
pX=XCbank%*%gamma
plot(pX[,1],pX[,2],pch="*",xlab="PC1",ylab="PC 2",asp=1)
```

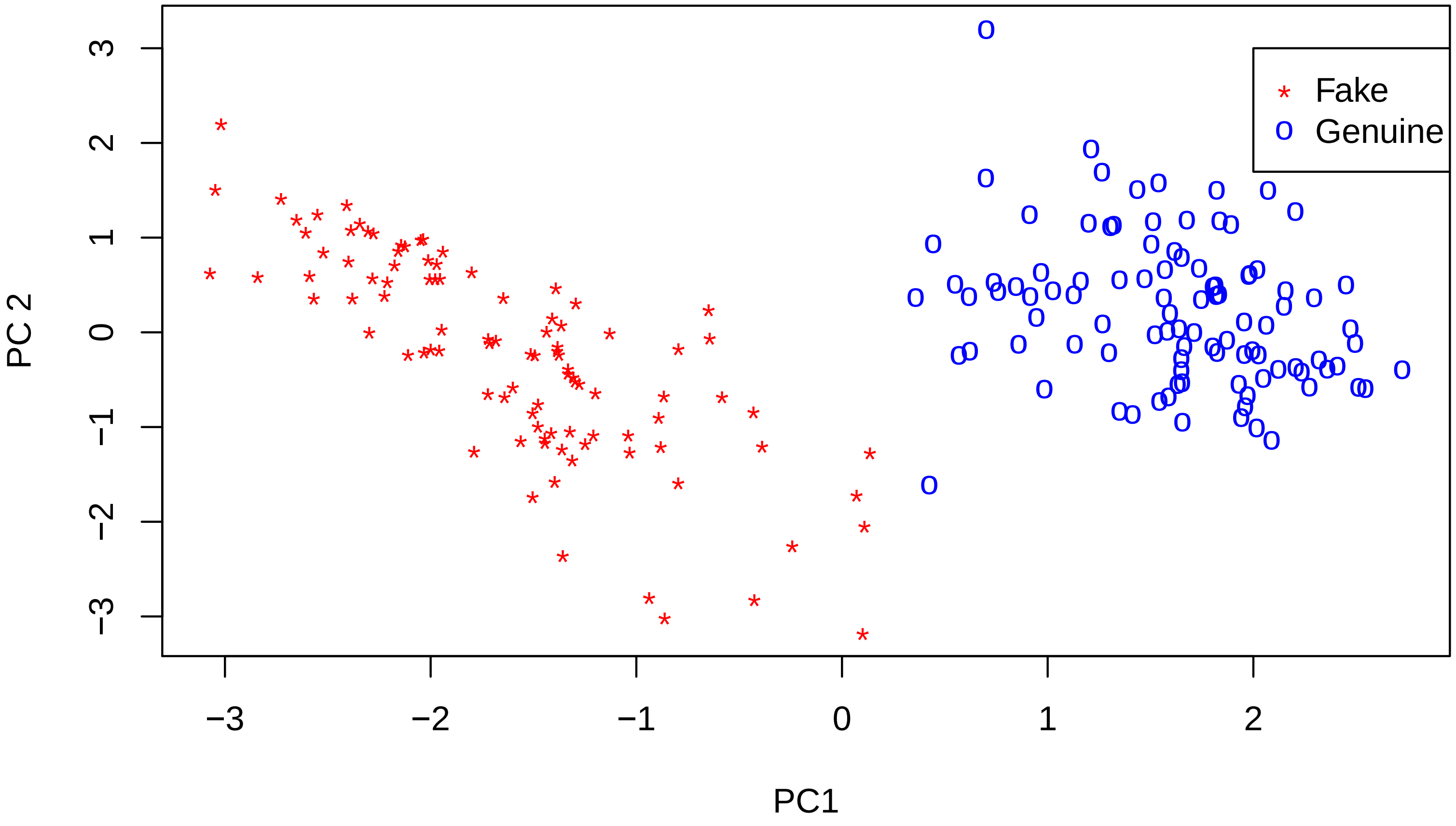


The "x" in the `prcomp` object `PCX` already has the PC values:

```
Y=PCX$x
plot(Y[,1],Y[,2],pch="*",xlab="PC1",ylab="PC 2",asp=1)
```



The two groups actually correspond to the genuine and the fake banknotes. The first two PCs have captured that information! We don't need to keep all 6 dimensions to see this. The genuine notes (blues) tend to have large PC1 and PC2 values.



We have

$$\begin{aligned}
Y_{i1} &= 0.044X_{i1} - 0.112X_{i2} - 0.139X_{i3} - 0.768X_{i4} - 0.202X_{i5} + 0.579X_{i6} \\
Y_{i2} &= -0.011X_{i1} - 0.071X_{i2} - 0.066X_{i3} + 0.563X_{i4} - 0.659X_{i5} + 0.489X_{i6}.
\end{aligned}$$

Thus

- the first PC is roughly the difference between the 6th (length of diagonal) and the 4th component (bottom margin);

- the second PC is roughly the difference between the 5th (top margin) and the sum of the 6th (length of diagonal) and the 4th component (bottom margin).