- The Wishart distribution is a generalisation to multiple dimensions of the chi square distribution.

  It depends on 3 parameters: p, a $p \times p$ scale matrix Σ and the number of degrees of freedom n:

$$W_p(\Sigma, n)\,.$$

- Recall that if $Z_1, \ldots, Z_n$ are independent $N(0,1)$ then

$$X = \sum_{k=1}^{n} Z_k^2 \sim \chi_n^2$$

  is a chi square with n degrees of freedom.

- If M is an $p \times n$ matrix whose columns are independent and all have a $N_p(0, \Sigma)$ distribution, then the matrix

$$MM^T \sim W_p(\Sigma, n)\,,$$

  i.e. MM^T has a Wishart distribution with parameters p, Σ and n.

- When σ is a scalar, a $W_1(\sigma^2, n)$ is the same as σ^2 times a χ^2_n.

- If a $p \times p$ random matrix $\mathcal{Y} \sim W_p(\Sigma, n)$ and B is a $q \times p$ matrix then

$$B\mathcal{Y}B^T \sim W_q(B\Sigma B^T, n)\,.$$

- If a $p \times p$ random matrix $\mathcal{Y} \sim W_p(\Sigma, n)$ and a is a $p \times 1$ vector such that $a^T\Sigma a \neq 0$, then

$$a^T\mathcal{Y}a/a^T\Sigma a \sim \chi^2_n\,.$$

- Recall the empirical covariance matrix constructed from a sample $X_1, \ldots, X_n$ of p-vectors:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T\,.$$

  It can be proved that

$$(n-1)S \sim W_p(\Sigma, n-1).$$

- The Hotelling $T^2_{p,n}$ distribution is a generalisation to multiple dimensions of the student t_n distribution with n degrees of freedom.

- If $X \sim N_p(0, I_p)$ is independent of $M \sim W_p(I_p, n)$, then

$$nX^T M^{-1} X \sim T^2_{p,n}\,.$$

- It can be proved that if $X_1, \ldots, X_n$ are i.i.d.$\sim N_p(\mu, \Sigma)$, then the sample mean vector $\bar{X}$ and the sample covariance matrix S are such that

$$n(\bar{X} - \mu)^T S^{-1} (\bar{X} - \mu) \sim T^2_{p,n-1}.$$

- In the univariate case, a variable $T \sim t_n$ if

$$T = X\sqrt{n/Y}$$

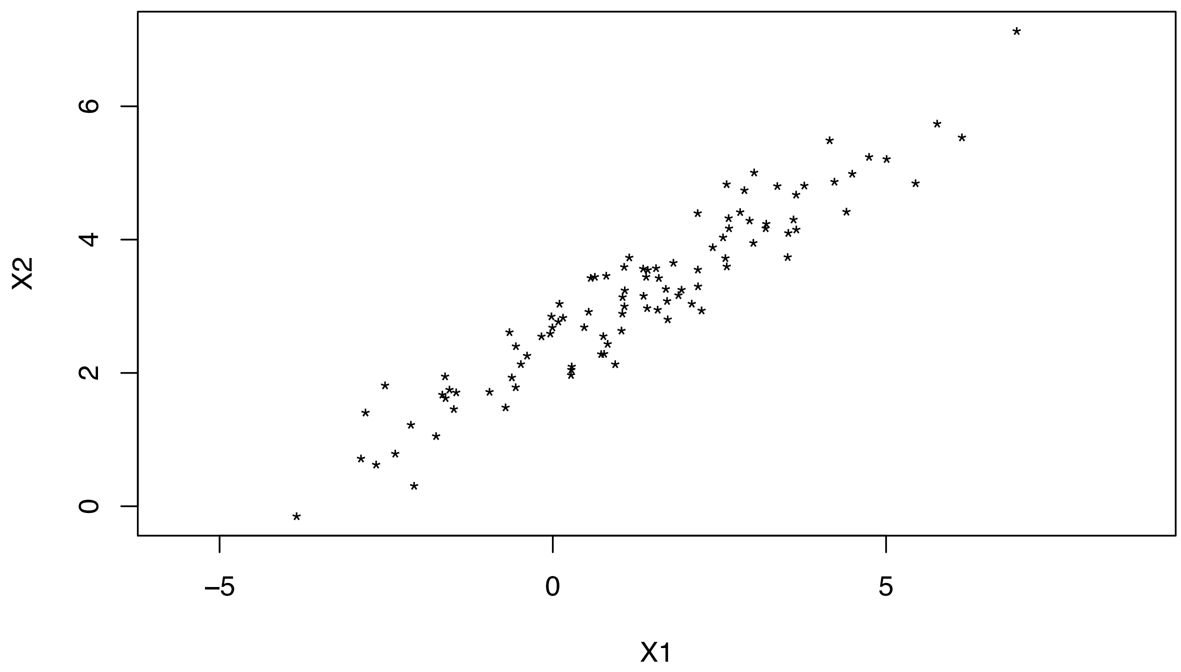
  where $X \sim N(0,1)$ is independent of $Y \sim \chi^2_n$. Thus T^2 is a $T^2_{1,n}$.

See Härdle and Simar, chapter 11.

5 PRINCIPAL COMPONENT ANALYSIS

5.1 INTRODUCTION

Visualizing 1, 2 or 3 dimensional data is relatively easy: we can represent them on a scatterplot, from which we can learn a lot about the structure/properties of the data.



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

☛ Can we find a way to summarise the data?

☛ Summaries should be easier to represent graphically.

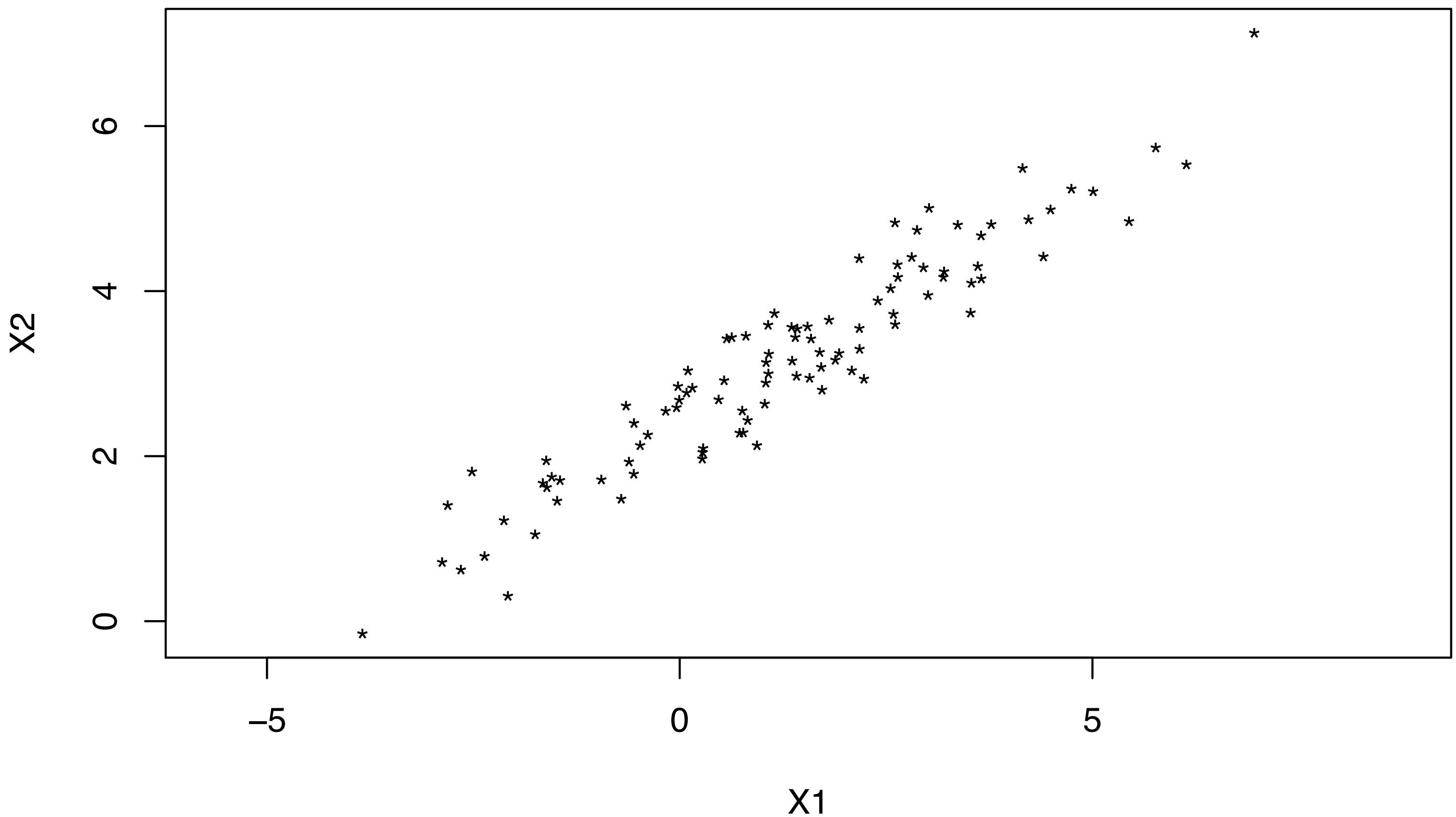
☛ Summaries should still contain as much information as possible about the original data.

☛ Often we can achieve this through dimension reduction.

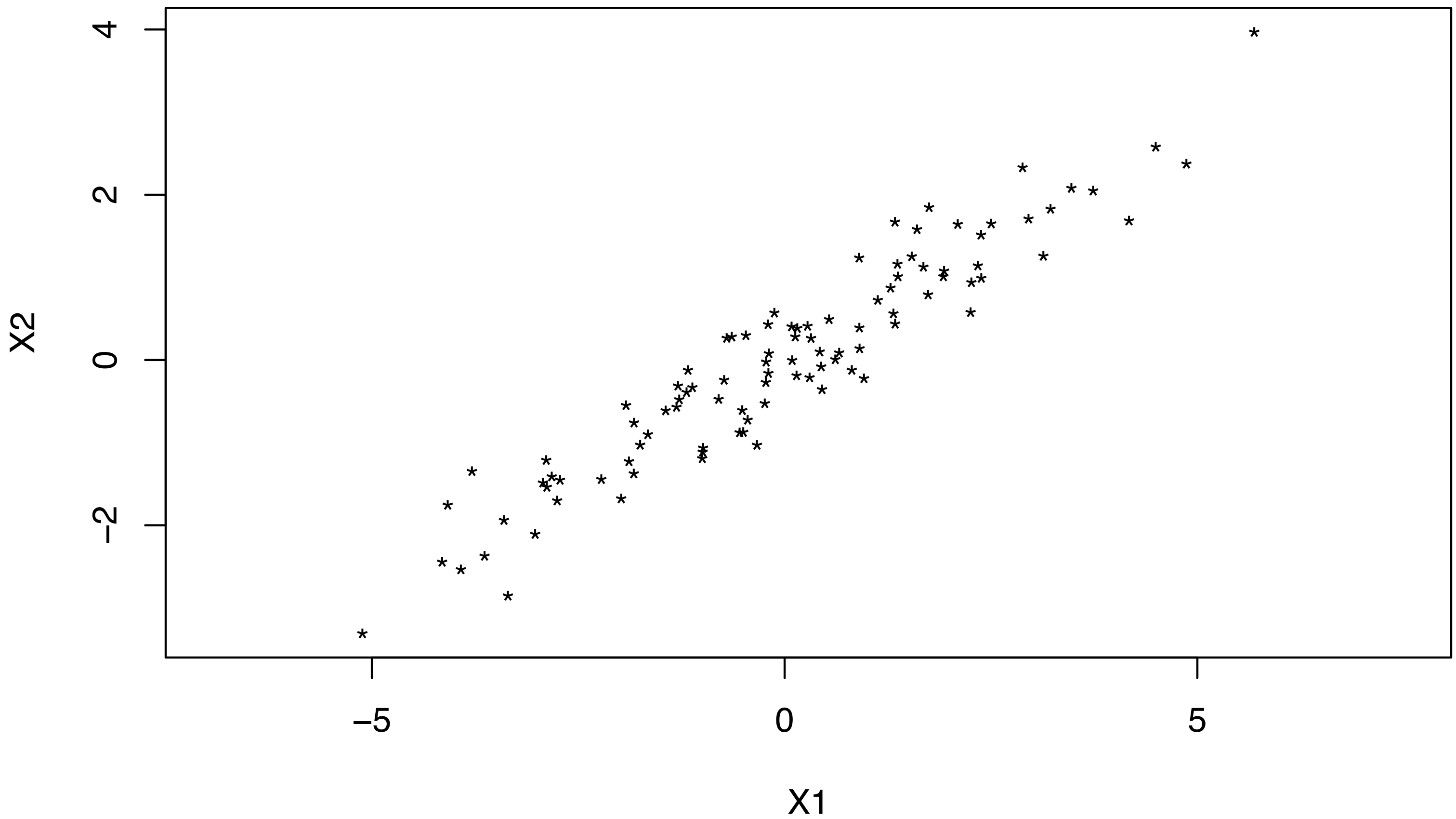
☛ Next we explain the ideas of dimension reduction by reducing data of dimension 2 to a single dimension.

As a toy example, we will first see how to reduce to 1 dimension the following 2-dimensional data.

The data are a collection of i.i.d. pairs $(X_{i1}, X_{i2})^T \sim (\mu, \Sigma)$, for $i = 1, \ldots, n$ which are shown in the scatter plot.

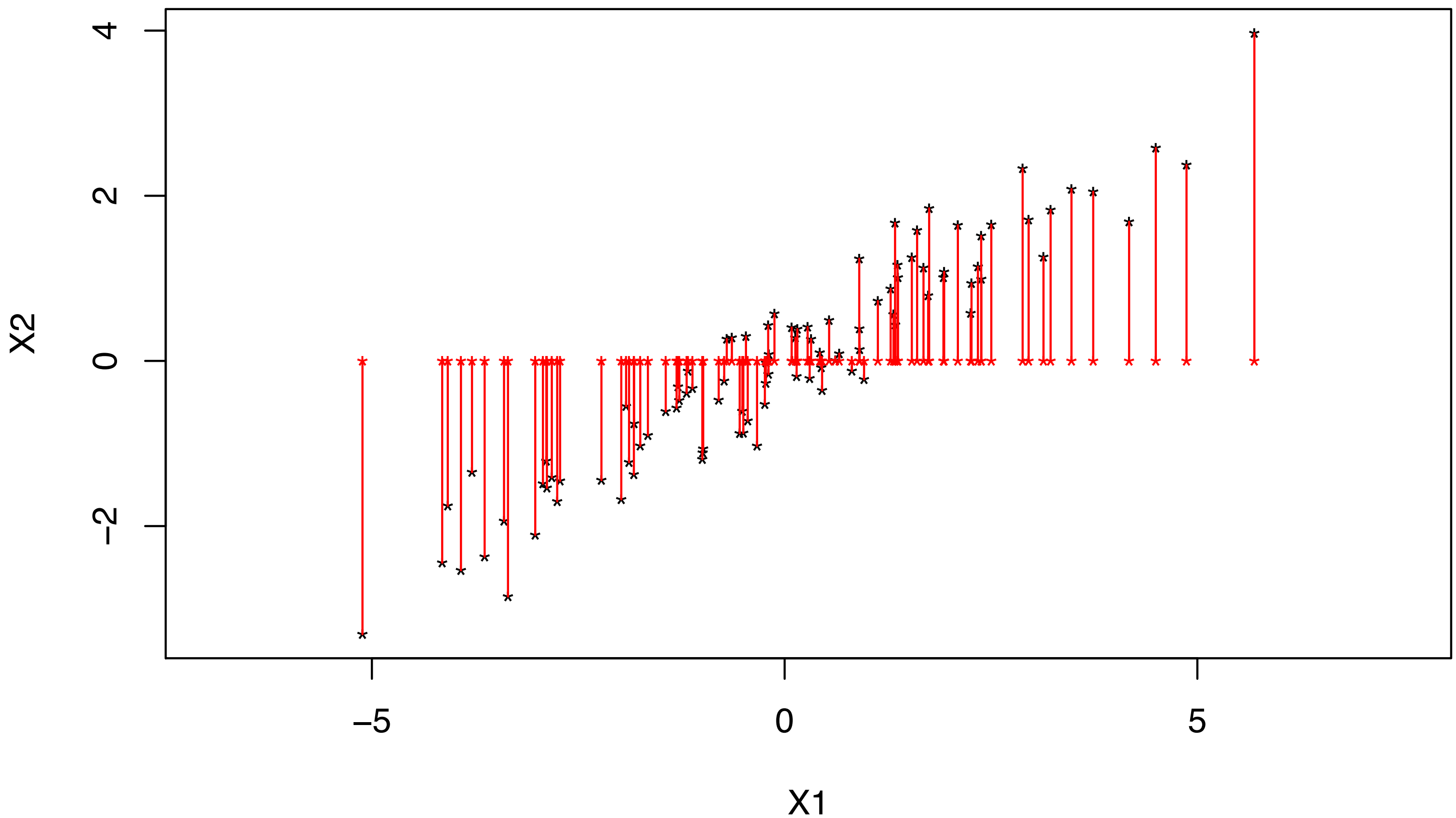


The first thing usually done in these problems is to center the data (easier to understand the geometry for centered data). For for $i = 1, \ldots, n$, we replace $(X_{i1}, X_{i2})^T$ by $(X_{i1} - \bar{X}_1, X_{i2} - \bar{X}_2)^T$:

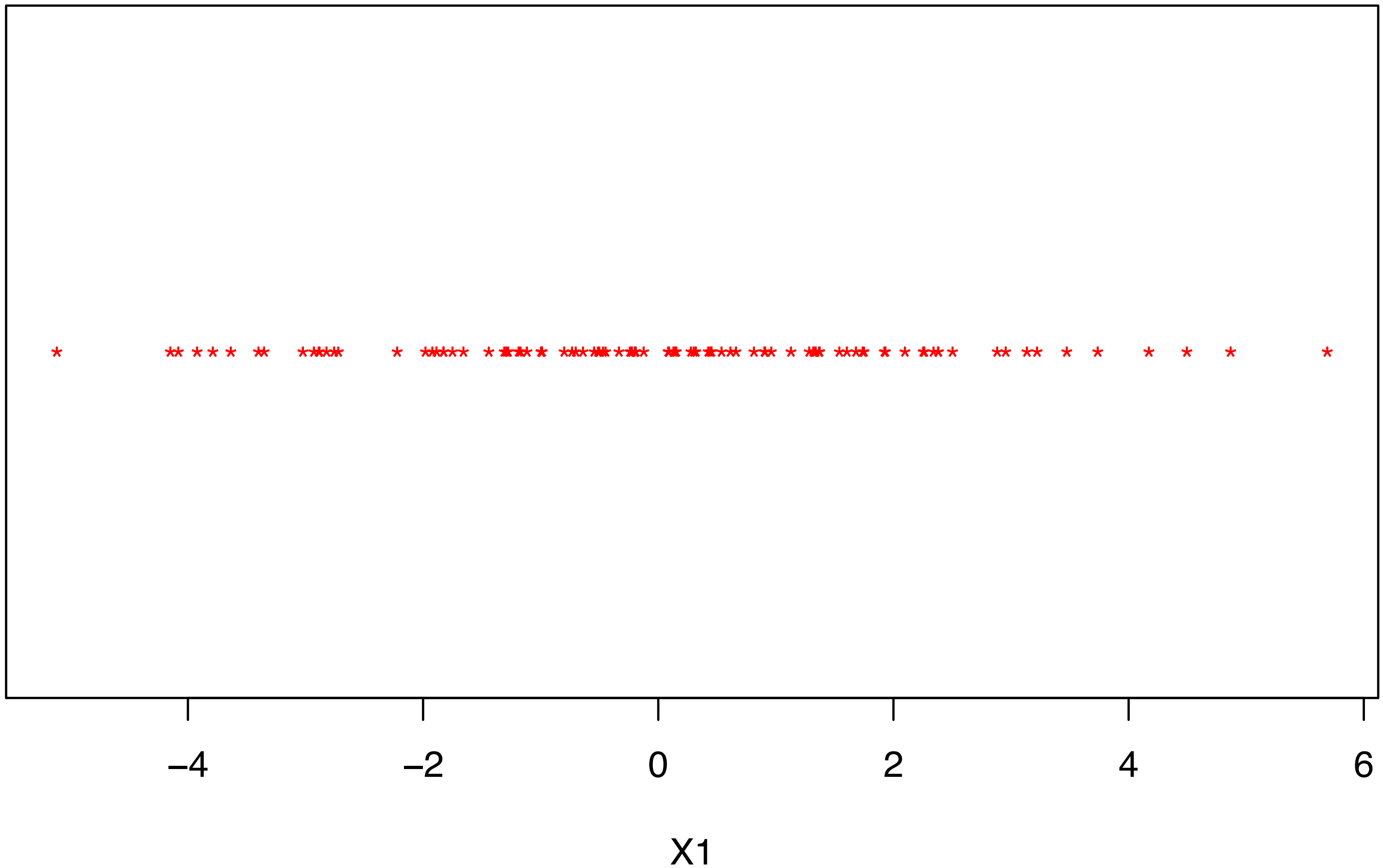


Until further notice, from now on in this chapter, to avoid heavy notations, when we refer to X_{ij} we mean $X_{ij} - \bar{X}_j$.

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



Keeping only the first component



☛ This is not very interesting because we lose all the information about the second component X_2.

☛ Suppose for example that the data contain the age (X_1) and the height (X_2) of $n = 100$ individuals.

☛ Then this amount to keeping only the age and drop completely the data about height.

☛ This does not sound like a very good idea.

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

A simple approach is to take a linear combination of the age and the height.

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

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

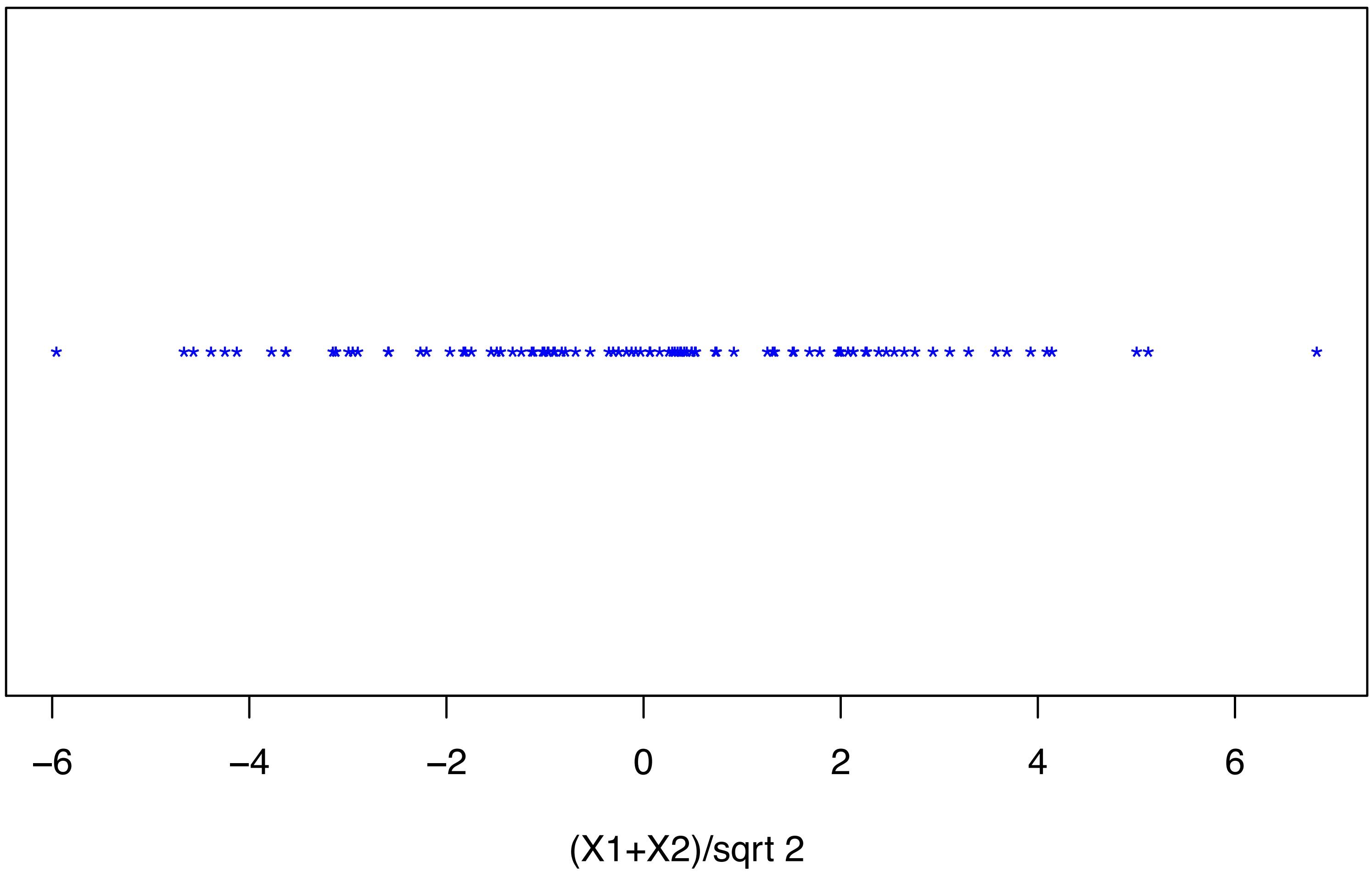
which would just take the average of the age and the height of each individual. We refer to $1/2$ and $1/2$ as the weight of age and height, respectively.

☛ Often in these problems we prefer to rescale linear combinations so that the sum of the square of the weights equals 1, which in this case would give instead

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

The values

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



We can interpret this linear combination as an orthogonal projection:

☛ We have

$$Y_i = X_{i1}/\sqrt{2} + X_{i2}/\sqrt{2} = X_i^T a\,,$$

where

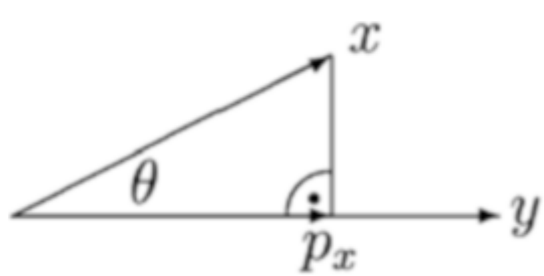
$$X_i = (X_{i1}, X_{i2})^T, \quad a = (1/\sqrt{2}, 1/\sqrt{2})^T.$$

☛ Since, $\|a\| = 1$, we can also write

$$Y_i = \frac{X_i^T a}{\|a\|}\,.$$

☛ We have seen before that the orthogonal projection p_x of a vector x onto a vector y was obtained by

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



☛ Thus the coordinates of the orthogonal projection p_x of a point x onto the line passing through the origin and a point y are given by

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

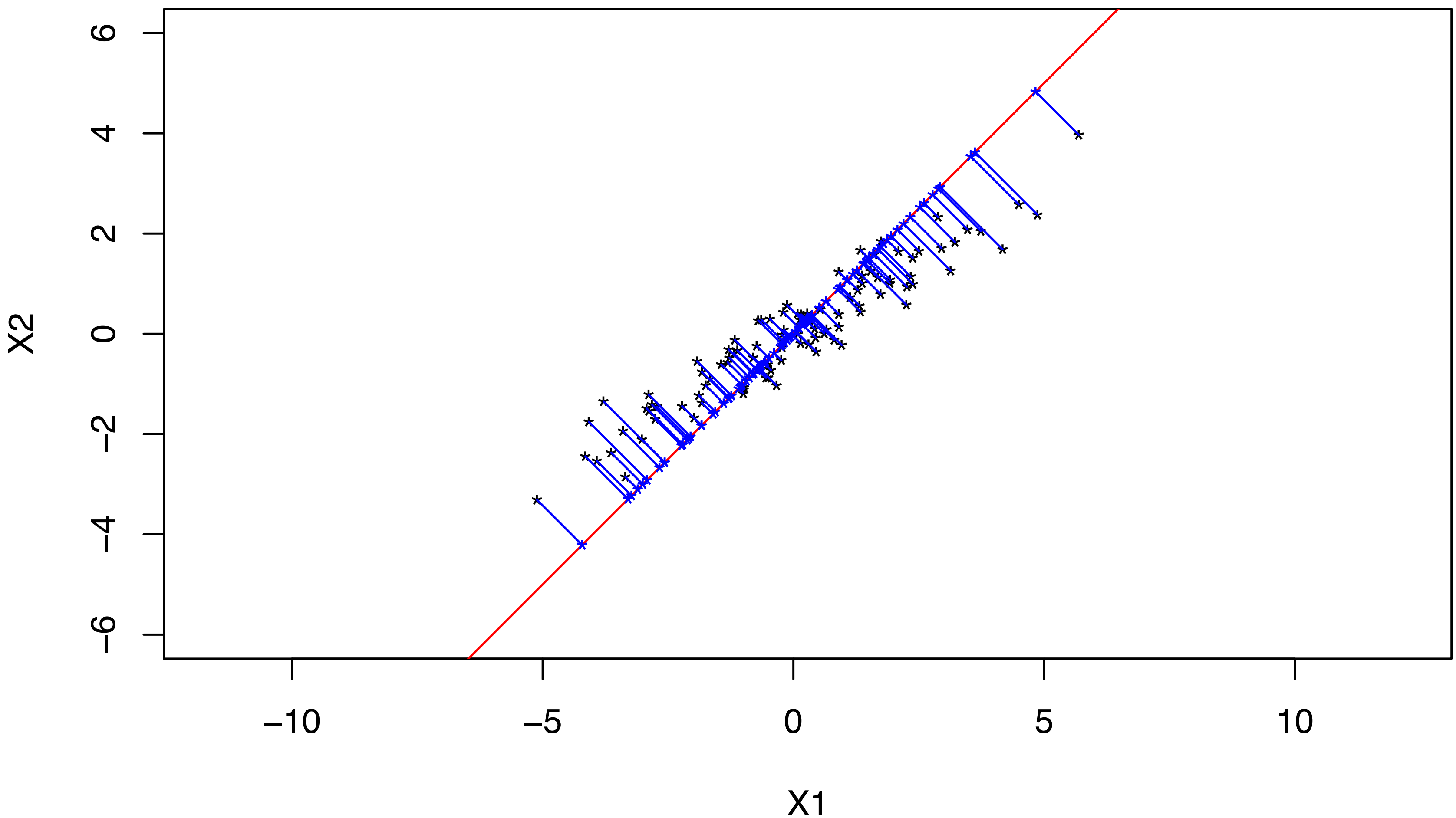
($y/\|y\|$ just gives the direction of the line on which we project x).

☛ Since we saw that

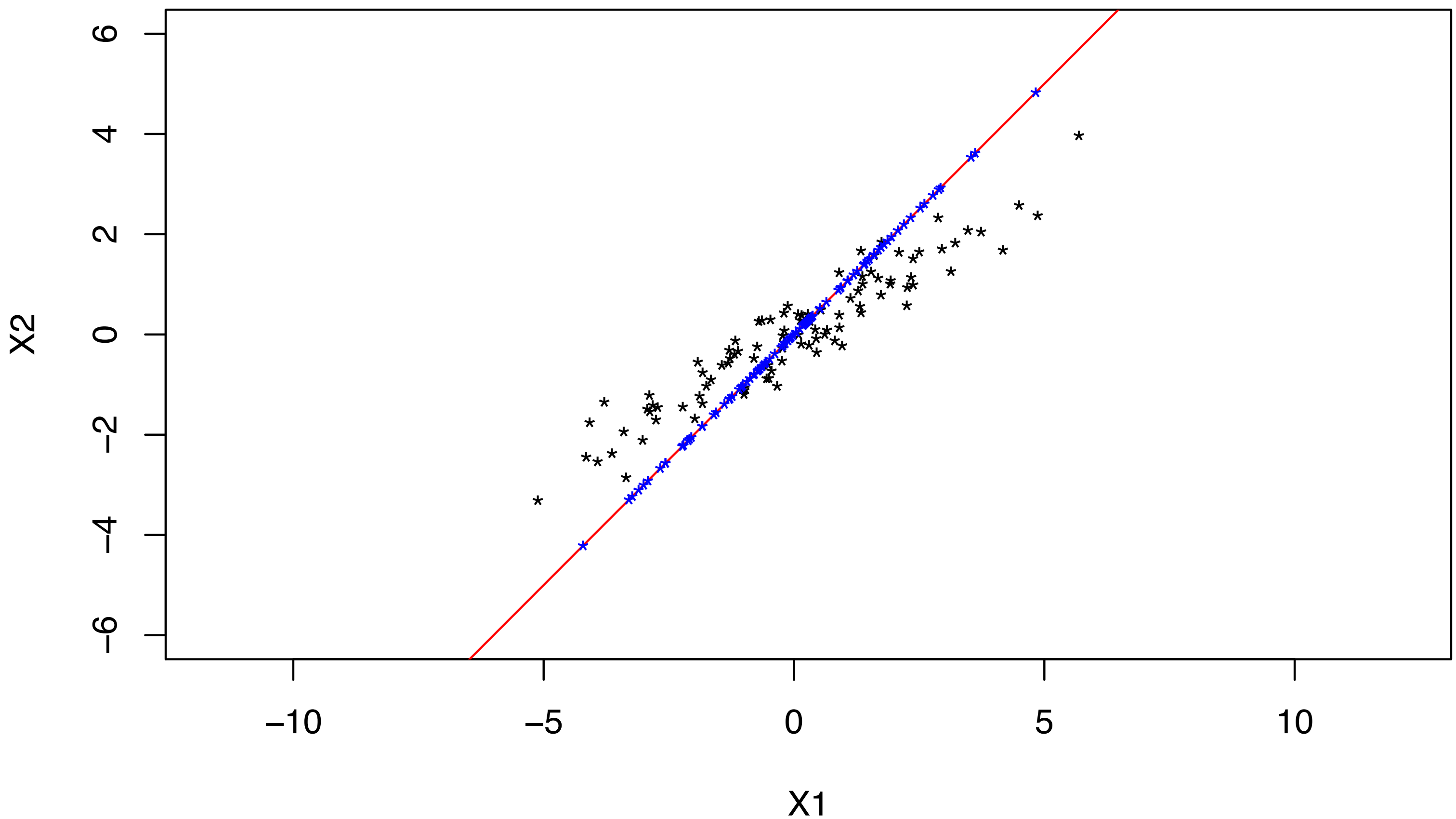
$$Y_i = \frac{X_i^T a}{\|a\|}\,,$$

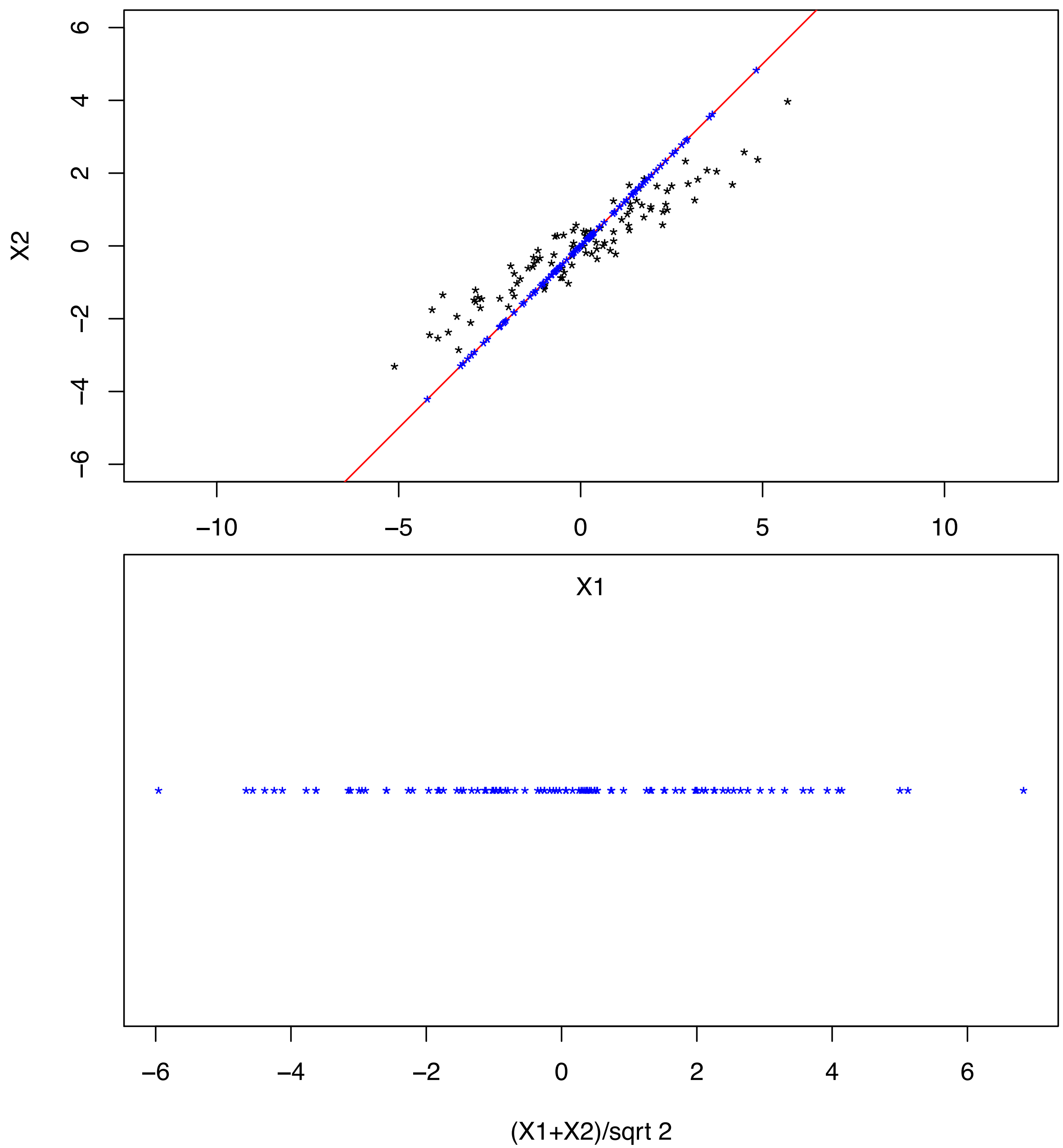
taking $X_i = x$ and $y = a$, we conclude that Y_i are the coordinates of the orthogonal projection of X_i onto the line passing through the origin and $a = (1/\sqrt{2}, 1/\sqrt{2})^T$.

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



Taking a scaled average of the two components corresponds to projecting the data on the red line and keeping only the projected values (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

☛ However 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.

☛ This depends on how we define "lose information".

reduce the dimension so it try to reduce the variane of projected data

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

do orthognal projection on directions which are orthognal to each other

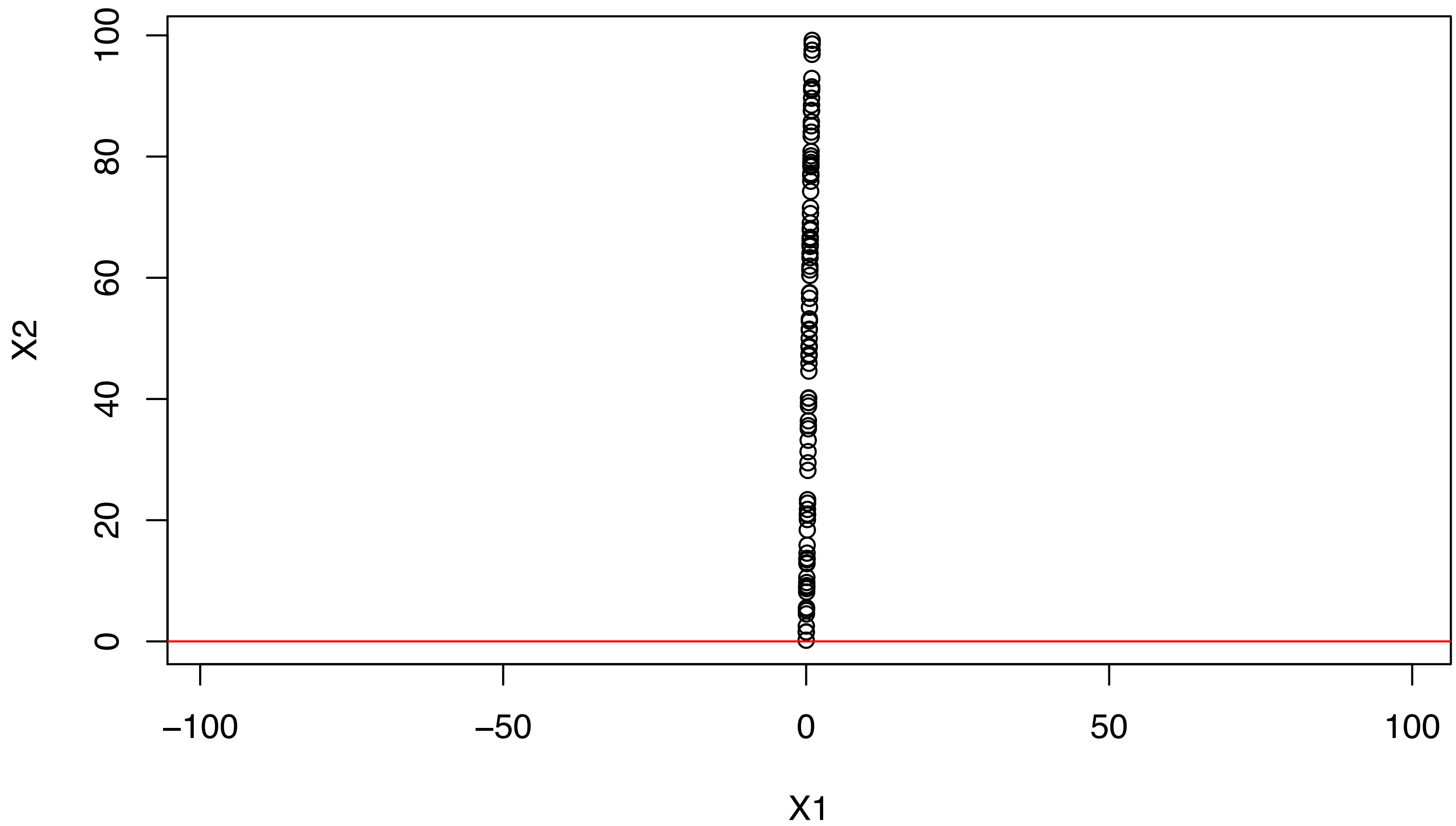
☛ Moreover, in PCA, "lose as little information as possible" is defined as "keep as much of the variability of the original data as possible".

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

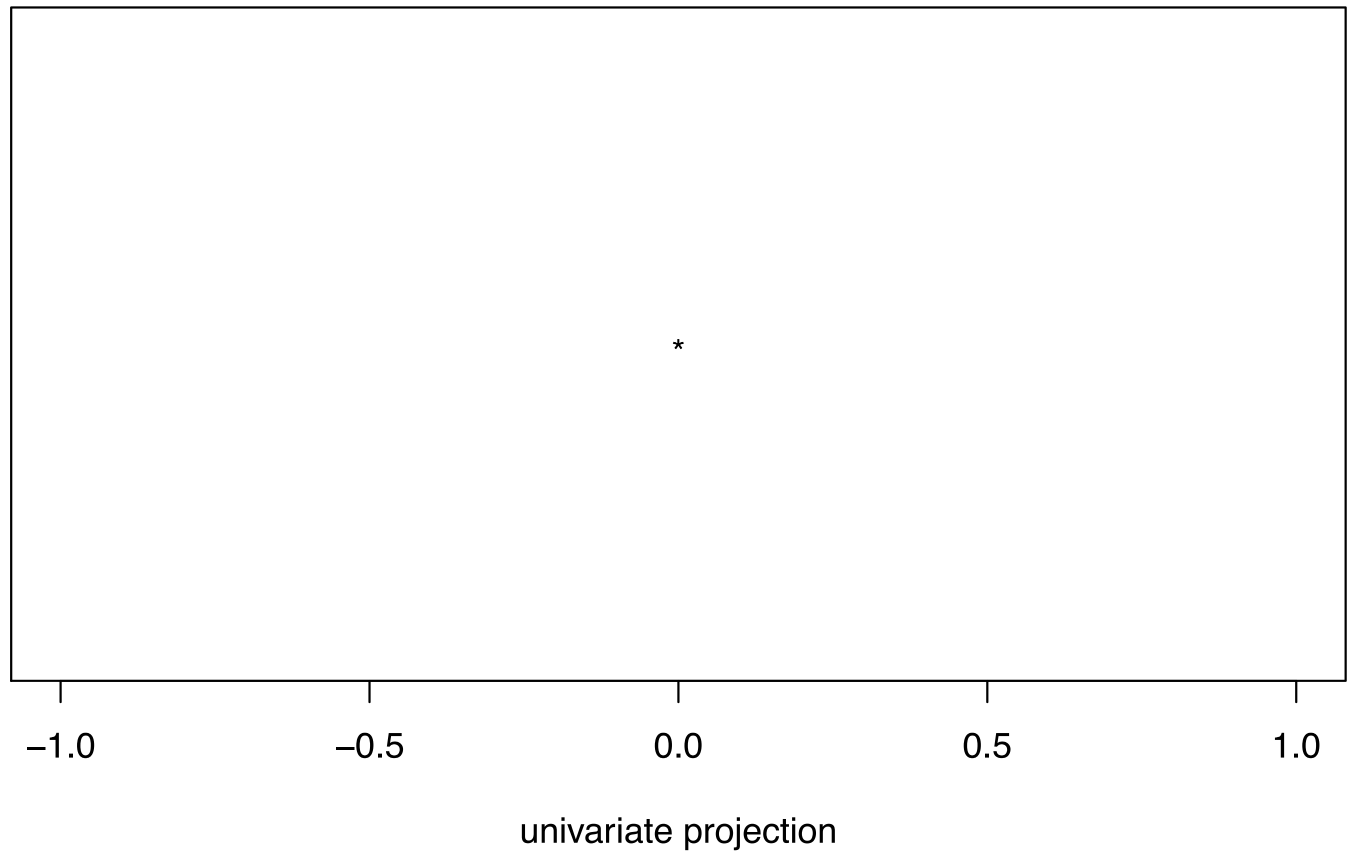
$$\text{var}(Y_i)$$

is as large as possible.

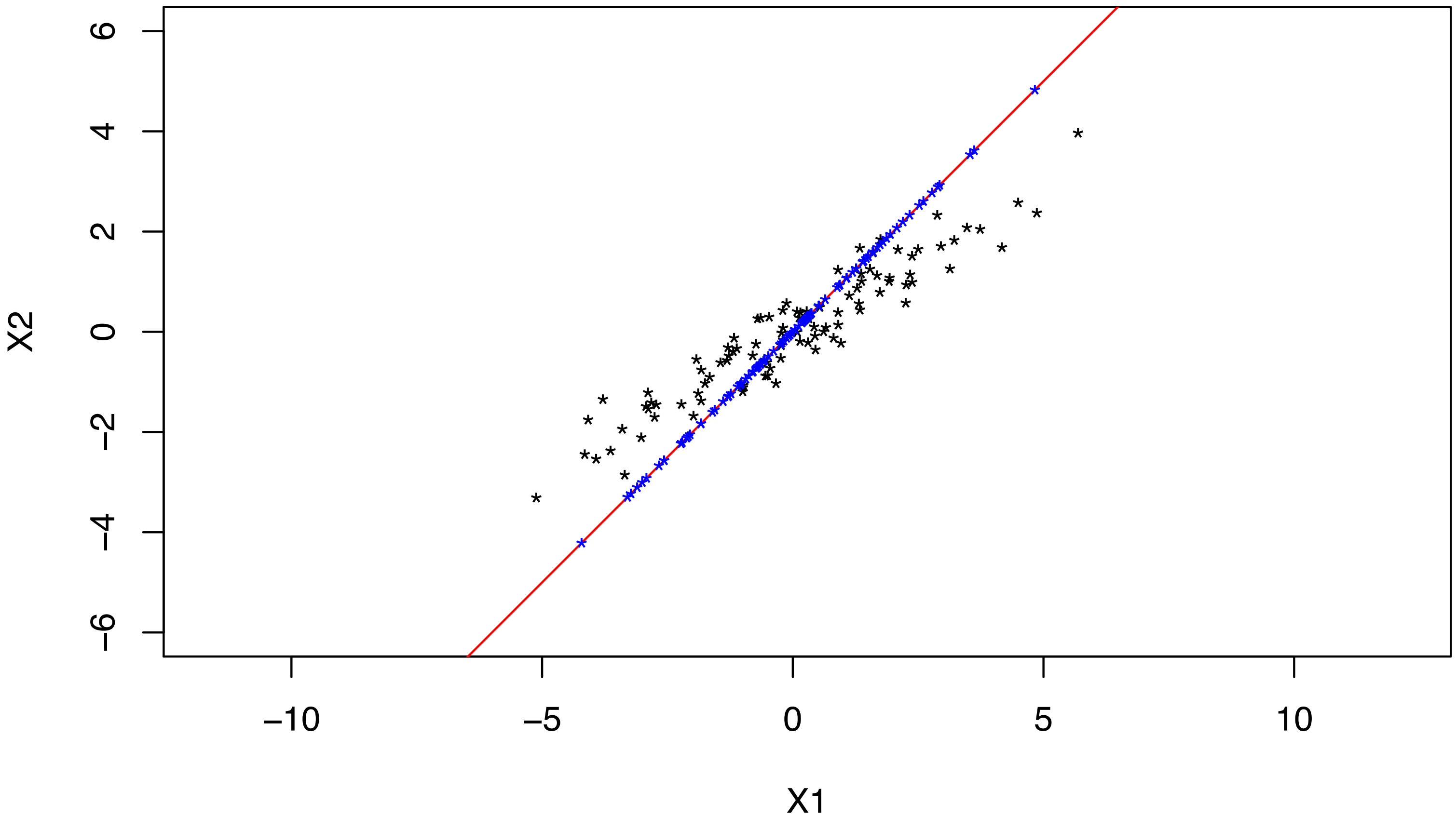
Why do we want to maximise variance? Here is an example where the projected data are not variable: project the data on the red line



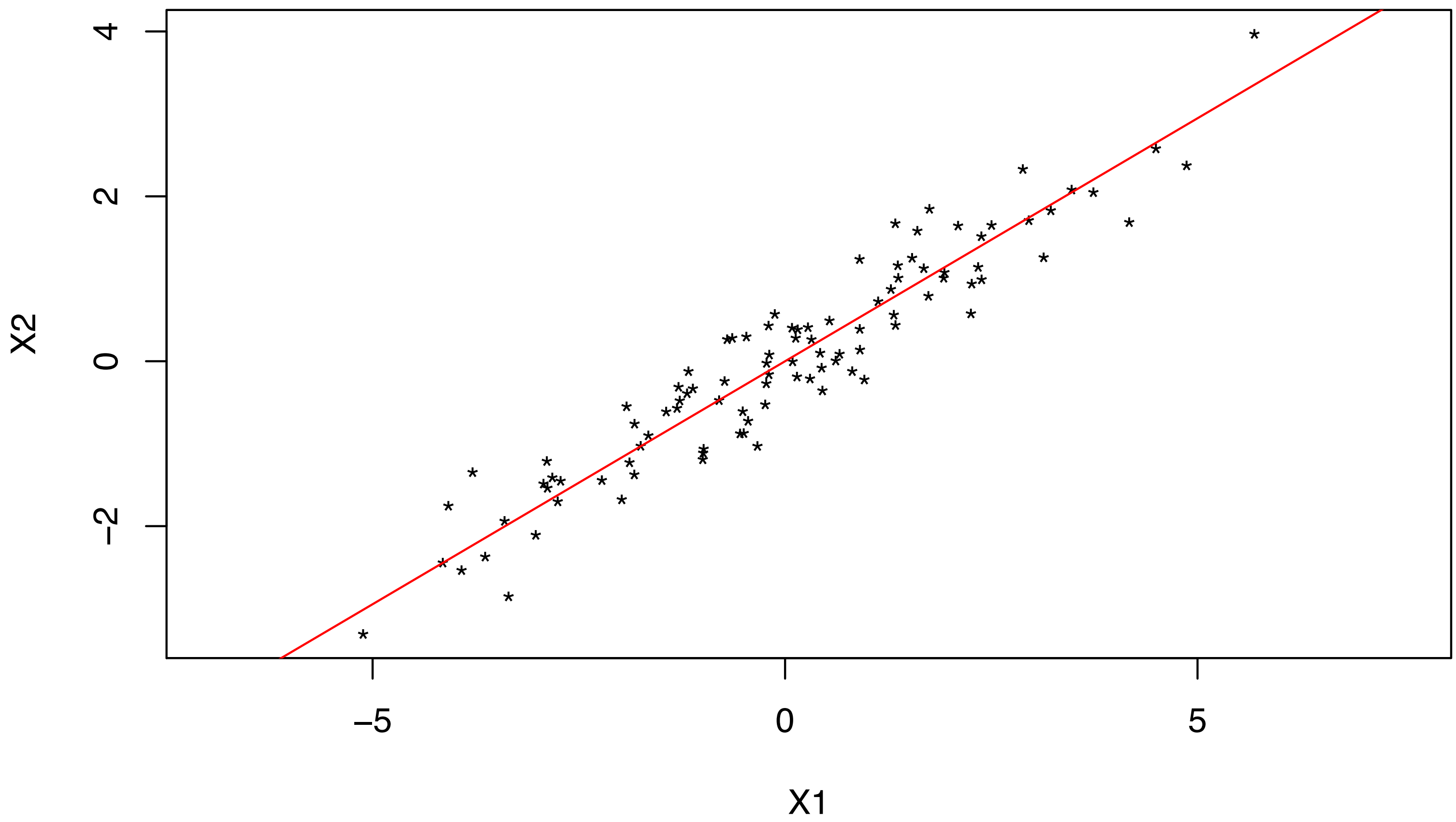
The data are all projected on the same point: the projected data have zero variance and we don't learn anything about the data.

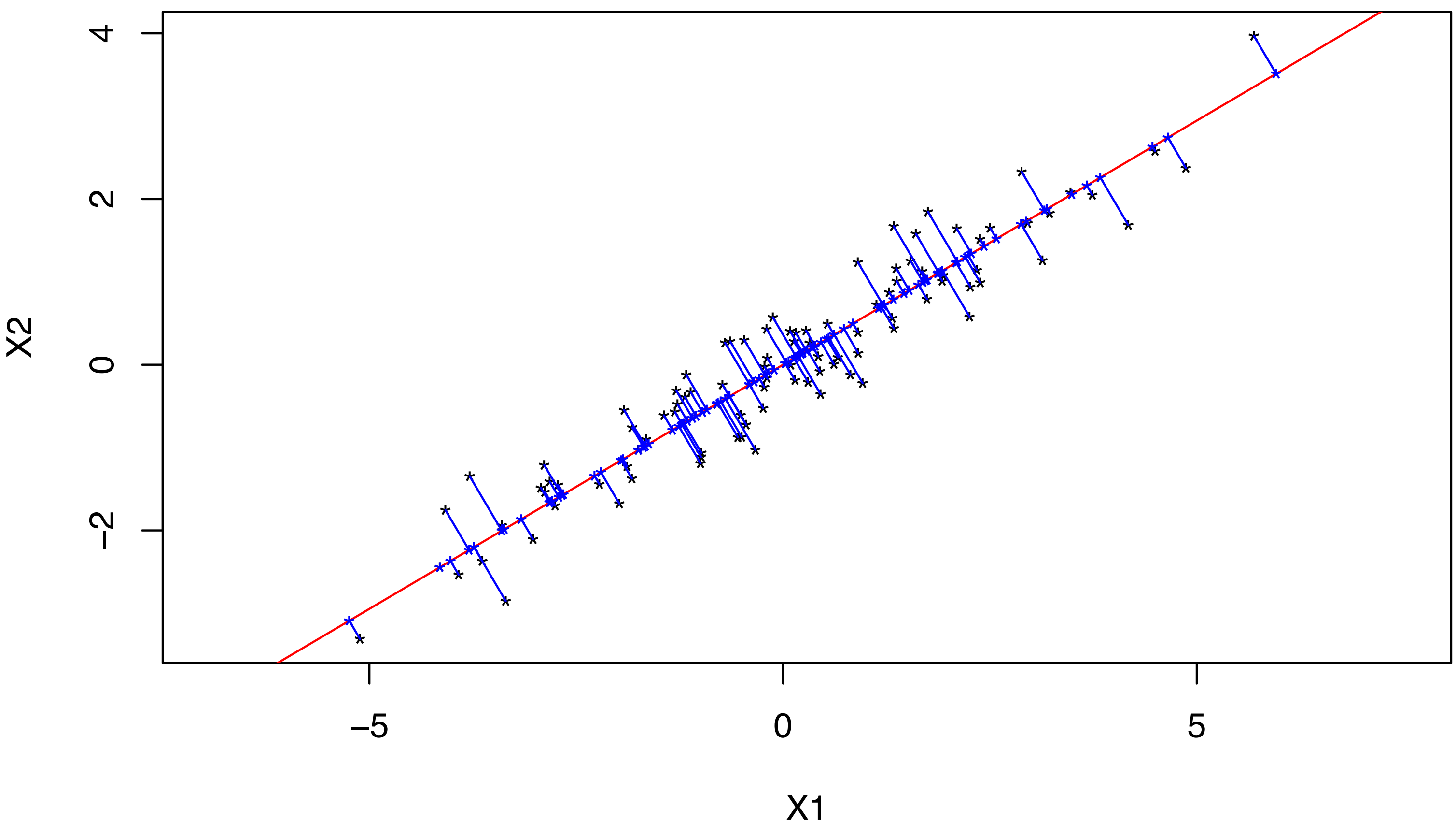


Getting back to our example, remember that we projected the data on this line:



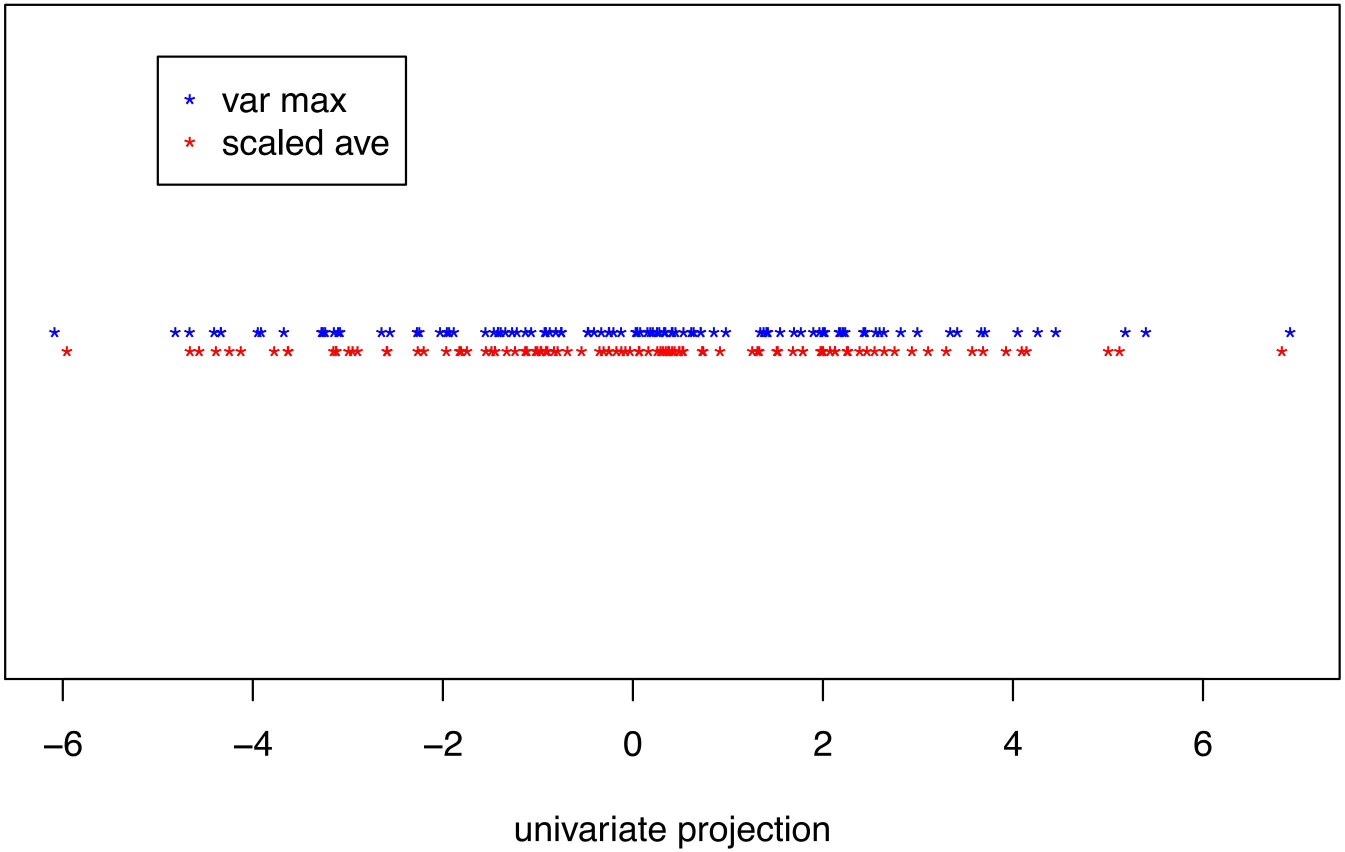
However we would have kept more information if we had instead projected the data on the following line:





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 projection:



The projection in blue is in fact the one that maximises the variance of the projected data.

观测值：行

PCA后会得到新观测值在新坐标轴上的得分

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

第一个维度得分

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

第i行在第一个坐标轴上的投影

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

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

$$\begin{array}{c|c} & \text{语数英} \\ \hline \begin{array}{c} \text{甲} \\ \text{乙} \\ \text{丙} \end{array} & \left[\begin{array}{ccc} & & \\ & & \\ & & \end{array}\right] \end{array} \qquad \begin{array}{c|c} & \begin{array}{cc} PC1 & PC2 \end{array} \\ \hline \begin{array}{c} \text{甲} \\ \text{乙} \\ \text{丙} \end{array} & \left[\begin{array}{cc} y_{11} & y_{12} \\ y_{21} & y_{22} \\ & \end{array}\right] \end{array}$$

and

$$\text{var}(Y_{i1})$$

is as large as possible.

每个观测值会有每行

- ☛ We use Y_{i1} instead of Y_i as there will be more than one projection.
- ☛ Constraint $\|a\| = 1$ makes problem well-defined, otherwise $\text{var}(Y_{i1})$ can be made as large as we want by multiplying a by arbitrary large scalar.

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

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

☛ Remember that the γ_j's are only defined up to a change of sign, so each γ_j can be replaced by $-\gamma_j$.

☛ It can be shown that the a that maximises $\mathrm{var}(Y_{i1})$ is equal to

$$a = \gamma_1,$$

the eigenvector (column vector) of Σ with largest eigenvalue.

☛ For $a = \gamma_1$, the variable

$\gamma_1^T X_i$

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

is called the first principal component of X_i. It is the projected value of X_i in the direction of γ_1.

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

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

is called the first principal component of X_i. It is a number (scalar).

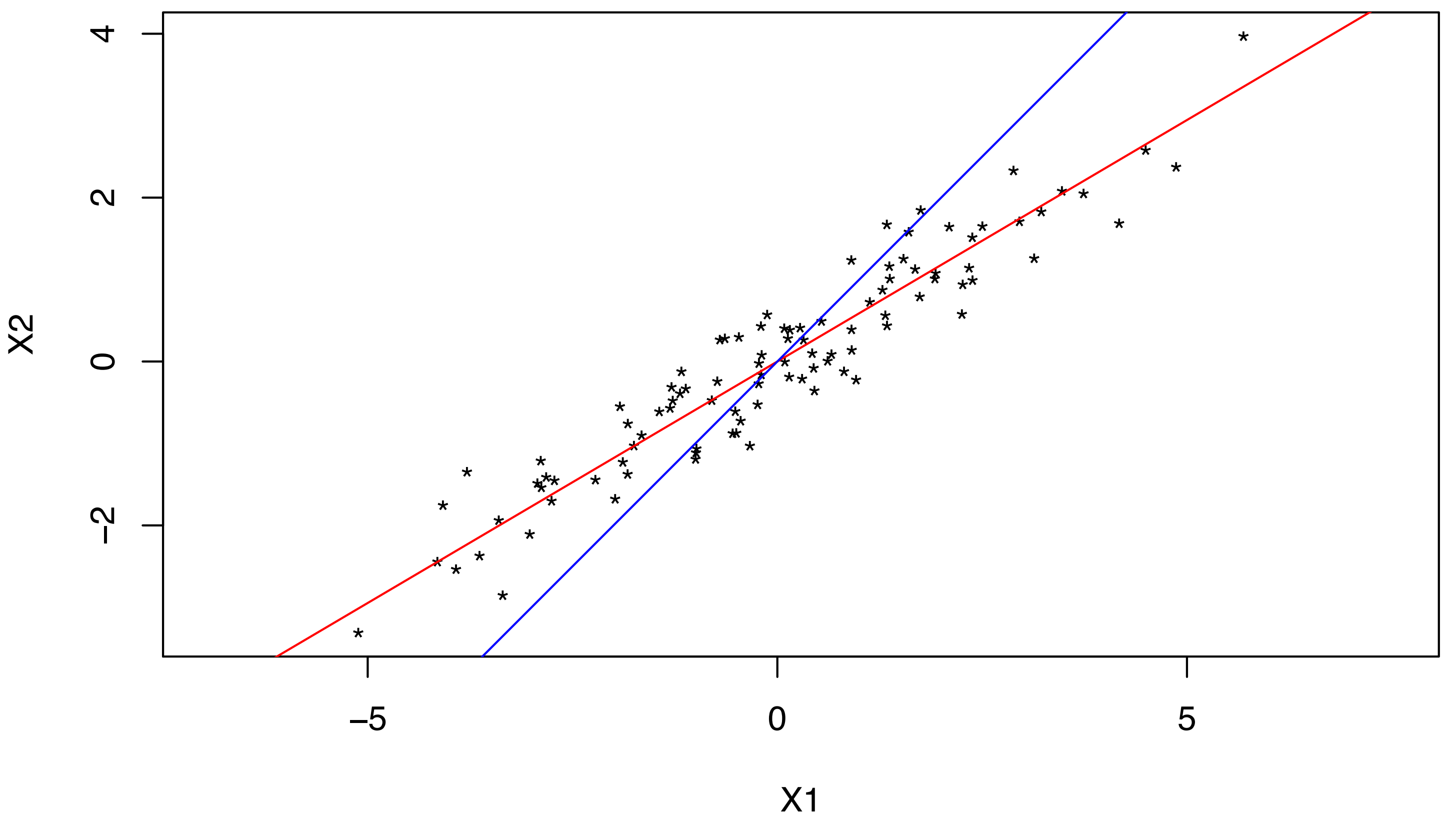
☛ It is the linear projection of the data that has maximum variance.

☛ We always center the data before projecting.

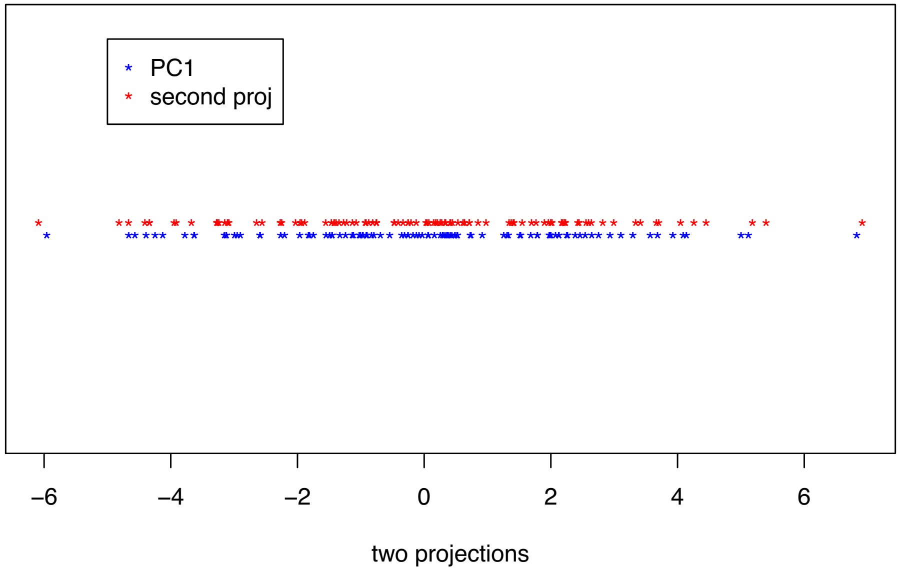
$[\quad]_{p \times p}$ $\qquad (X_i - \mu)$

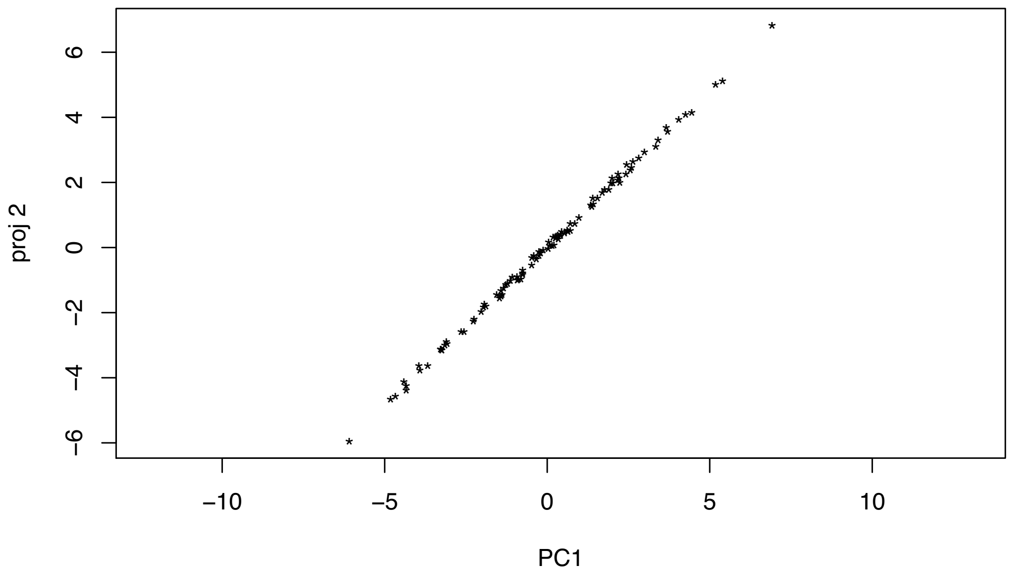
☛ In PCA, once we have found a univariate projection, how do we add a second projection?

☛ Should not just project the data on any other line. Example: could project next on blue line.



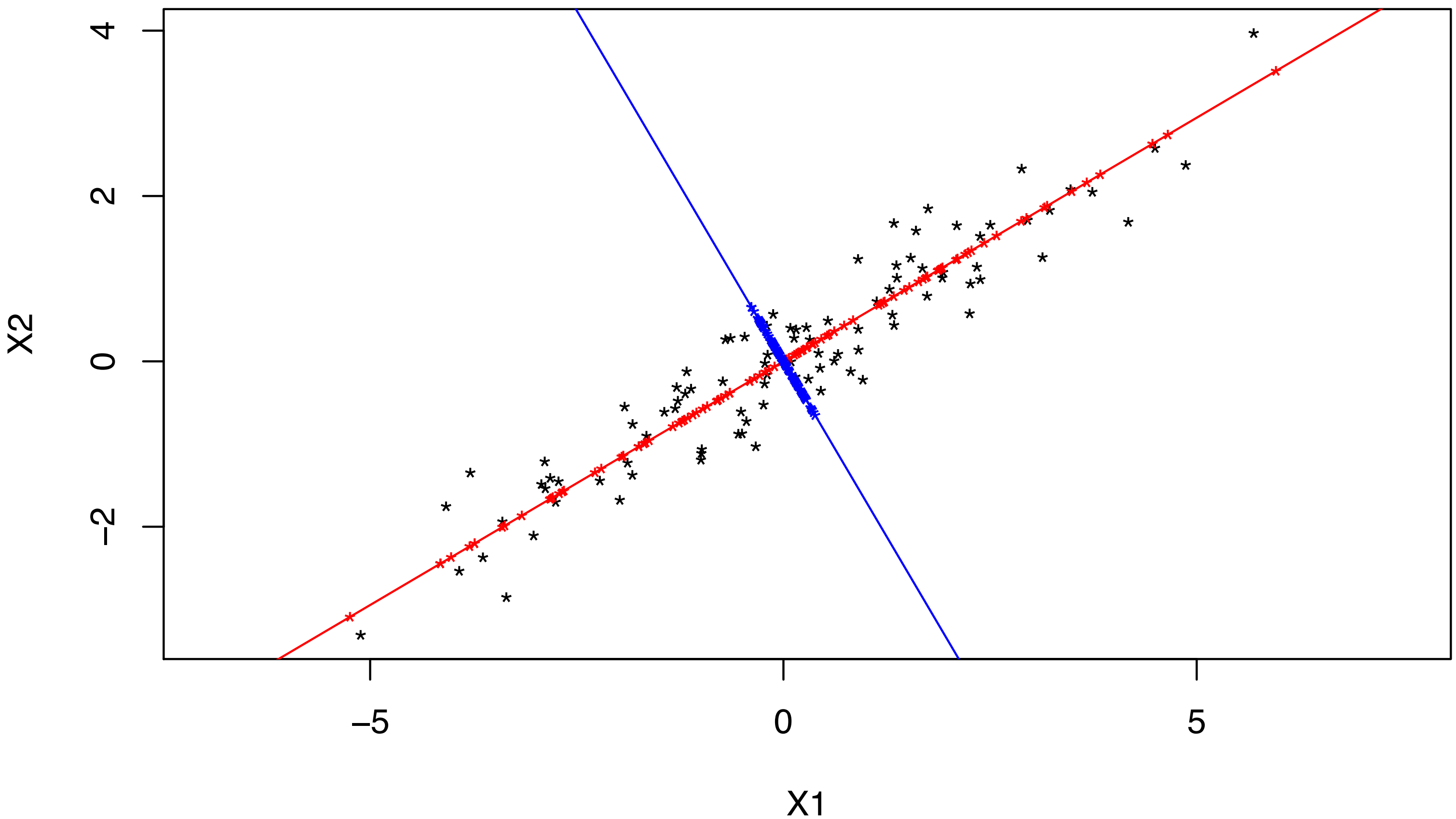
Those two projections are essentially redundant, we don't learn much more about the data by adding the second projection:



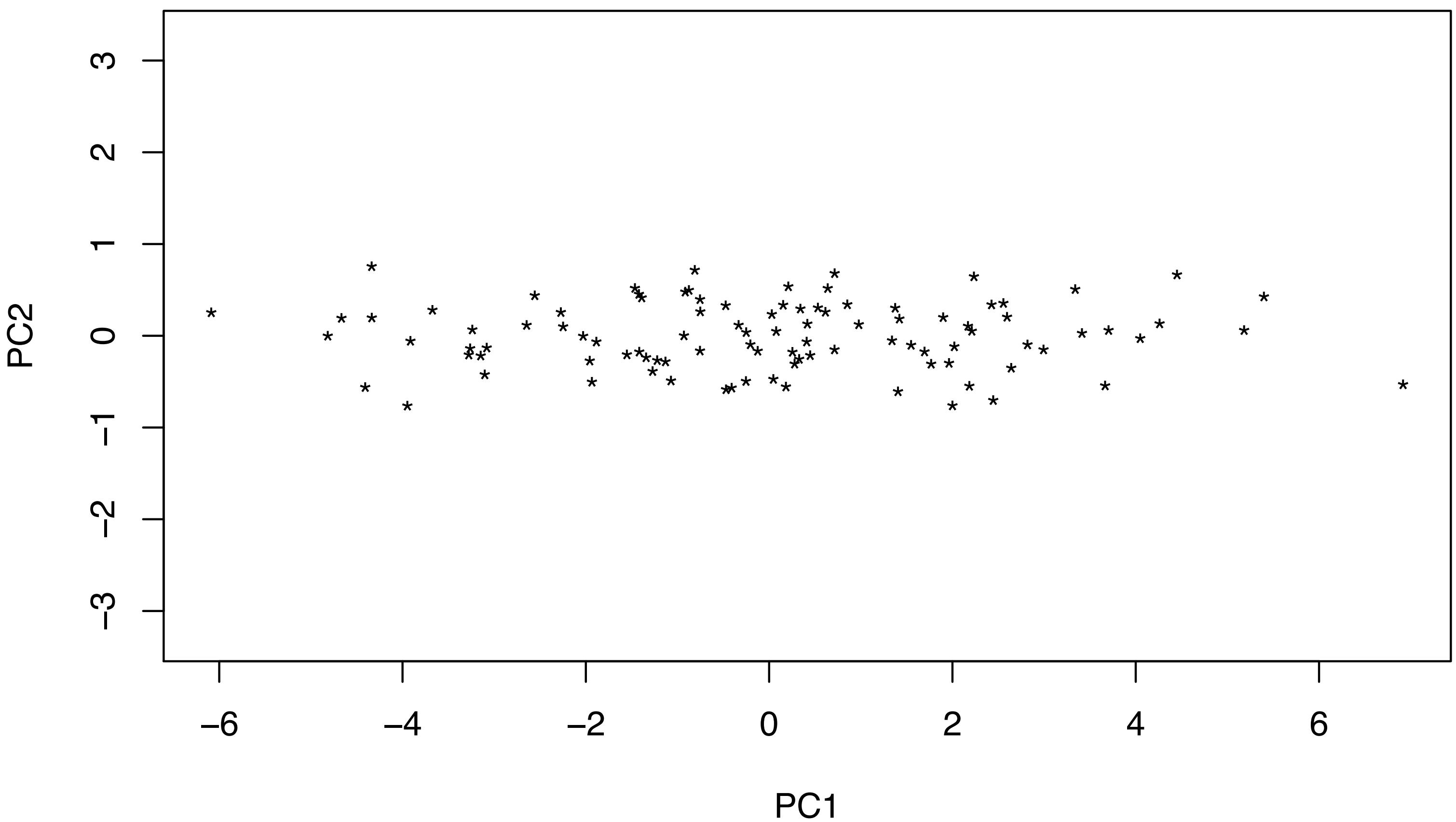


☛ We should rather project the data onto a line as different as possible from the previous one, to learn complementary information. How?

☛ Project on line perpendicular to the direction used for PC1. Variable obtained by this second projection is called second principal component.



In this example, since $p = 2$, the data projected on the two lines are just the same as the original data, but where the axes have been rotated to match the blue and the red lines.



More generally, in PCA, we project p-dimensional data $X_i \sim (\mu, \Sigma)$ onto $q \leq p$ dimensions defined by the first q orthonormal eigenvectors $\gamma_1, \ldots, \gamma_q$ of Σ corresponding the q largest e.vals $\boxed{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p}$ of Σ, as follows:

第1个最大 ⇒ decreasing order

- We start by taking the first principal component of X_i

X_i的第一个主成分 → 第i列

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

特征向量

with γ_1 the eigenvec of Σ corresponding to its largest eigenval, λ_1.

特

- 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 evec of Σ corresponding to its kth largest eval, λ_k.

- The γ_j's are of norm 1 and orthogonal to each other. Thus the projection directions are orthogonal to each other.

X_i的主成分

- In matrix notation, letting $Y_i = (Y_{i1}, \ldots, Y_{ip})^T$, we have

特征向量矩阵(X_i的)

$$Y_i = \underbrace{\Gamma^T}_{p\times p}(X_i - \underbrace{\mu)}_{p},$$

如果他是一个向量

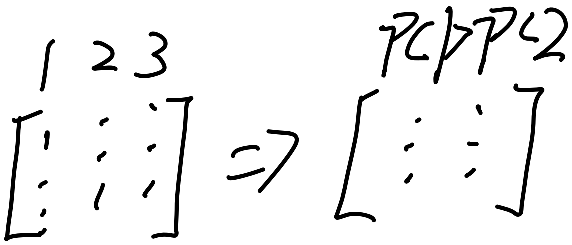
where the kth column of the $p \times p$ matrix Γ is the column vector γ_k.

- Suppose we construct $Y_{i1}, \ldots, Y_{ip}$ as described above. Then we have (see written notes)

pxp矩阵的第k列是特征向量k

压缩成两个主成分

不同观测在新主成分上的映射的均值都是0



$$E(Y_{ij}) = 0, \quad \text{for } j = 1, \ldots, p$$ (X_i的第j个主成分)均值=0

$$\text{var}(Y_{ij}) = \lambda_j, \quad \text{for } j = 1, \ldots, p$$ (X_i的第j个主成分)方差是特征值λ_j

$$\text{cov}(Y_{ik}, Y_{ij}) = 0, \quad k \neq j$$ (每个观测在不同主成分上的映射不相关)

第一列

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

(X_i在第1个主成分的映射 > 第2个 > 第3个 > 第p个 ……

$$\sum_{j=1}^{p} \text{var}(Y_{ij}) = \text{tr}(\Sigma)$$ 对角线相加

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