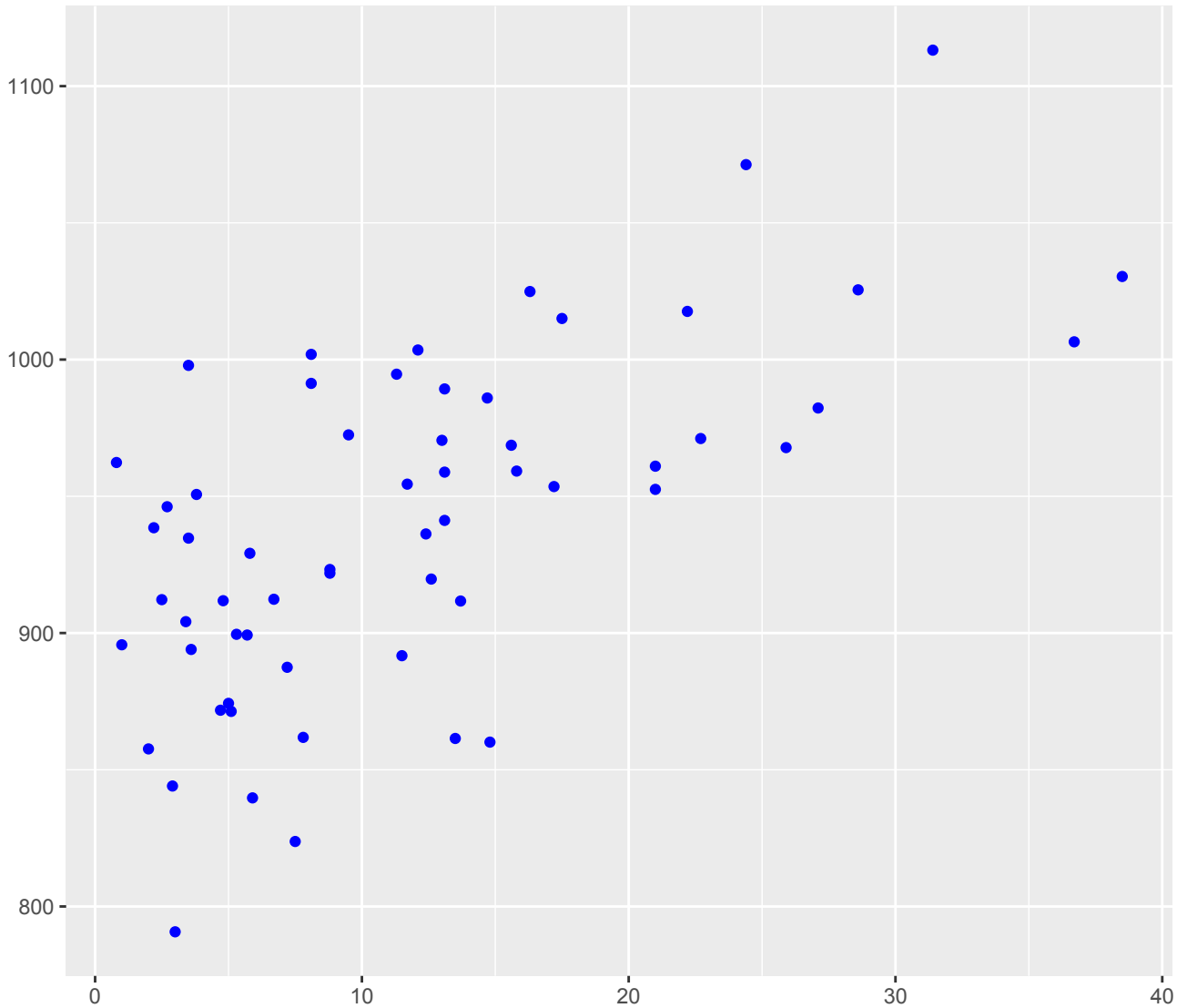
Unsupervised Learning I

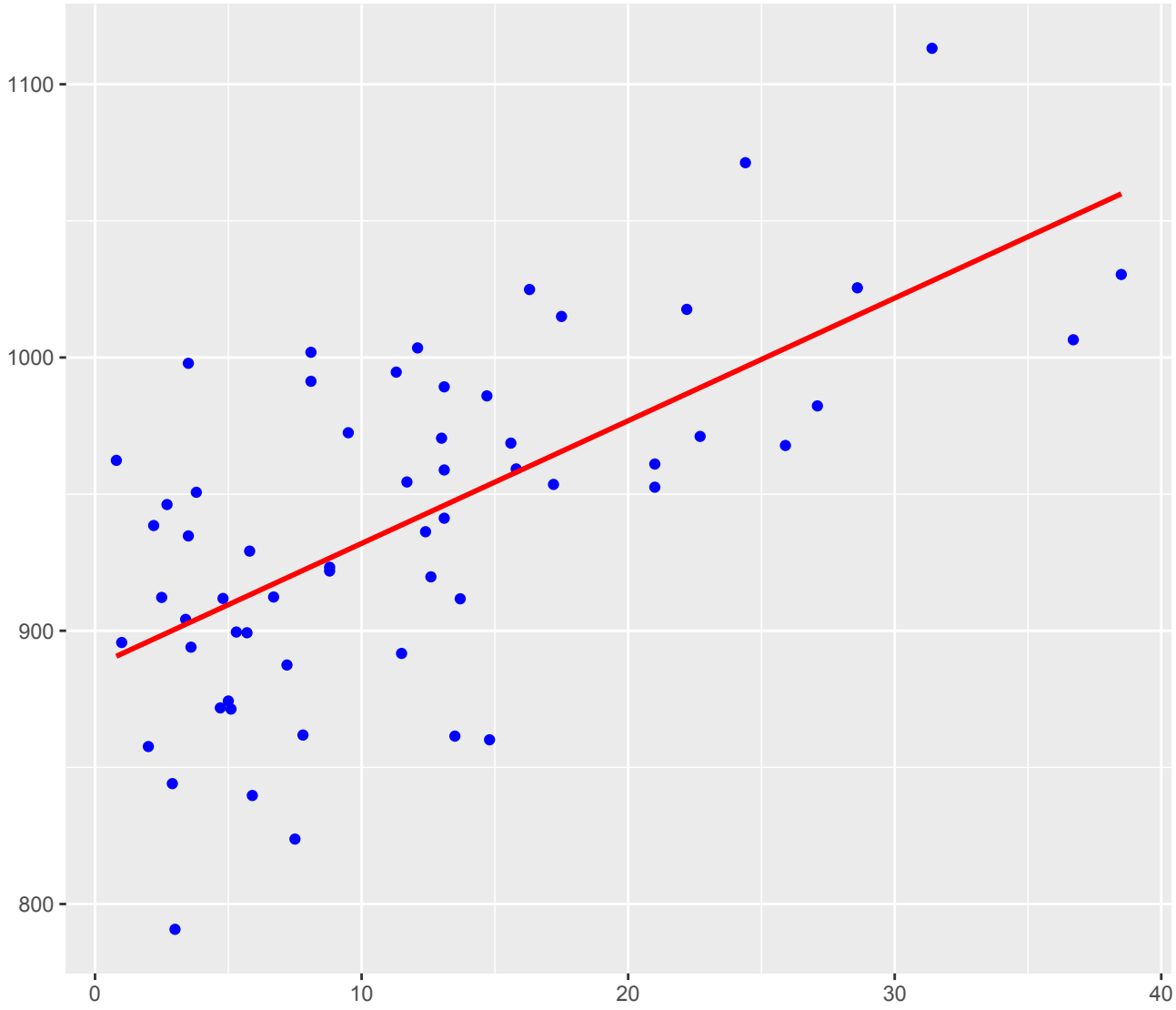
Roberta De Vito

BROWN
Public Health

Until now...

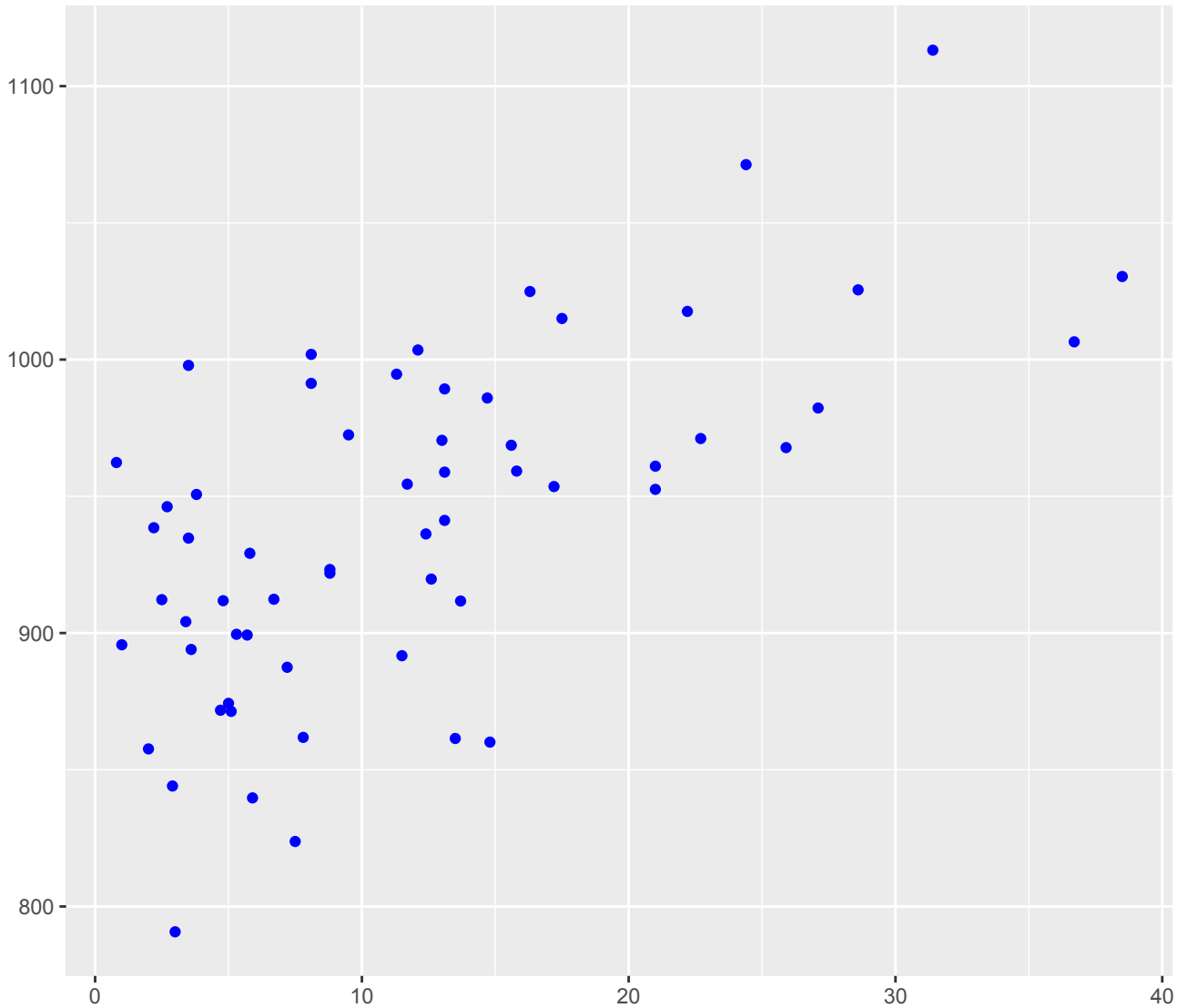


Until now...



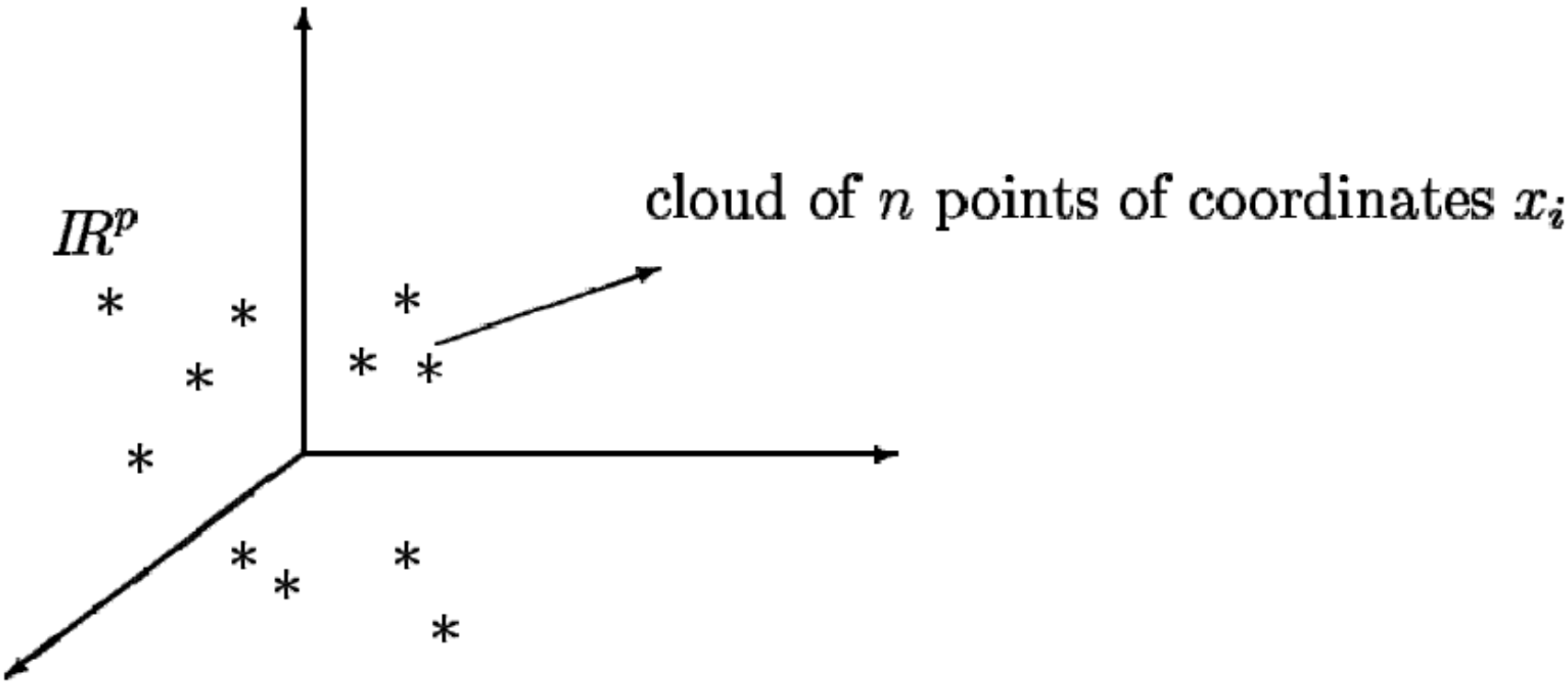
Q1 on prismia

Until now...



The Geometric Point of View

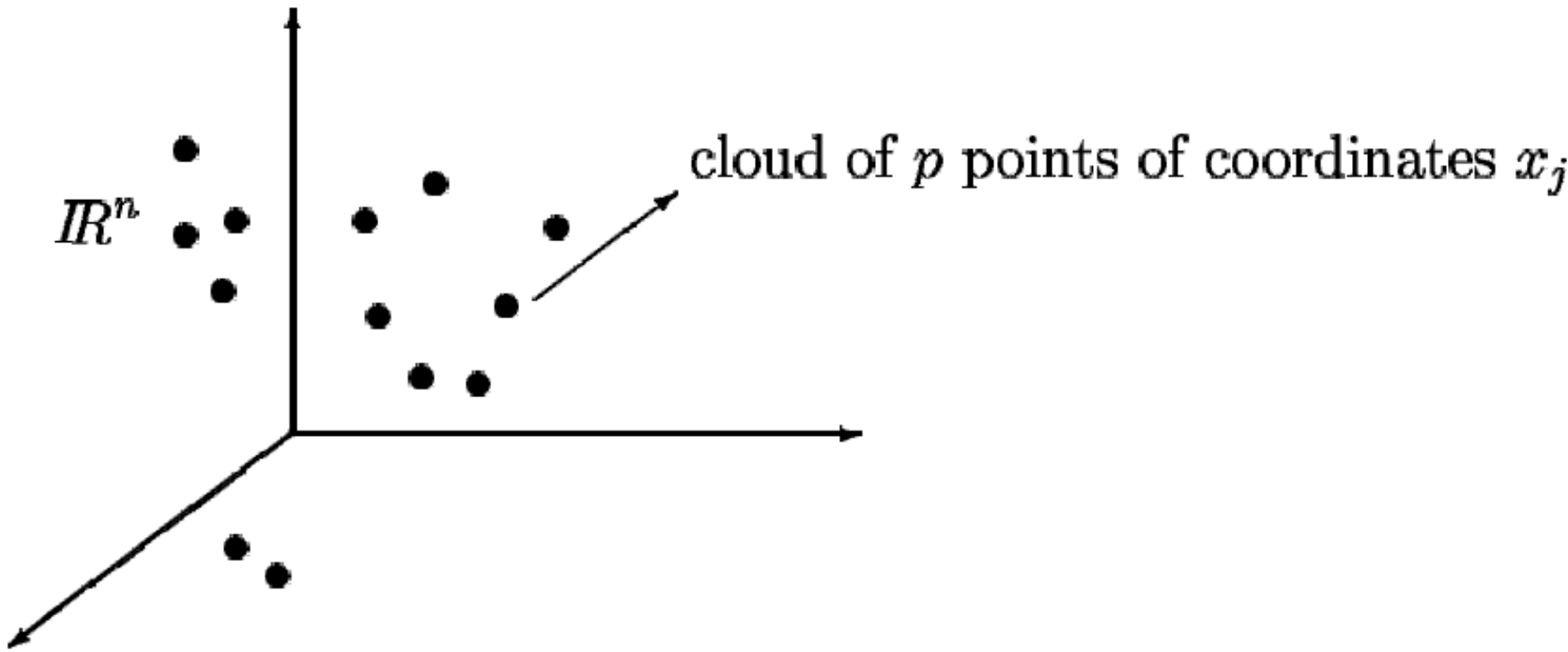
The Geometric Point of View



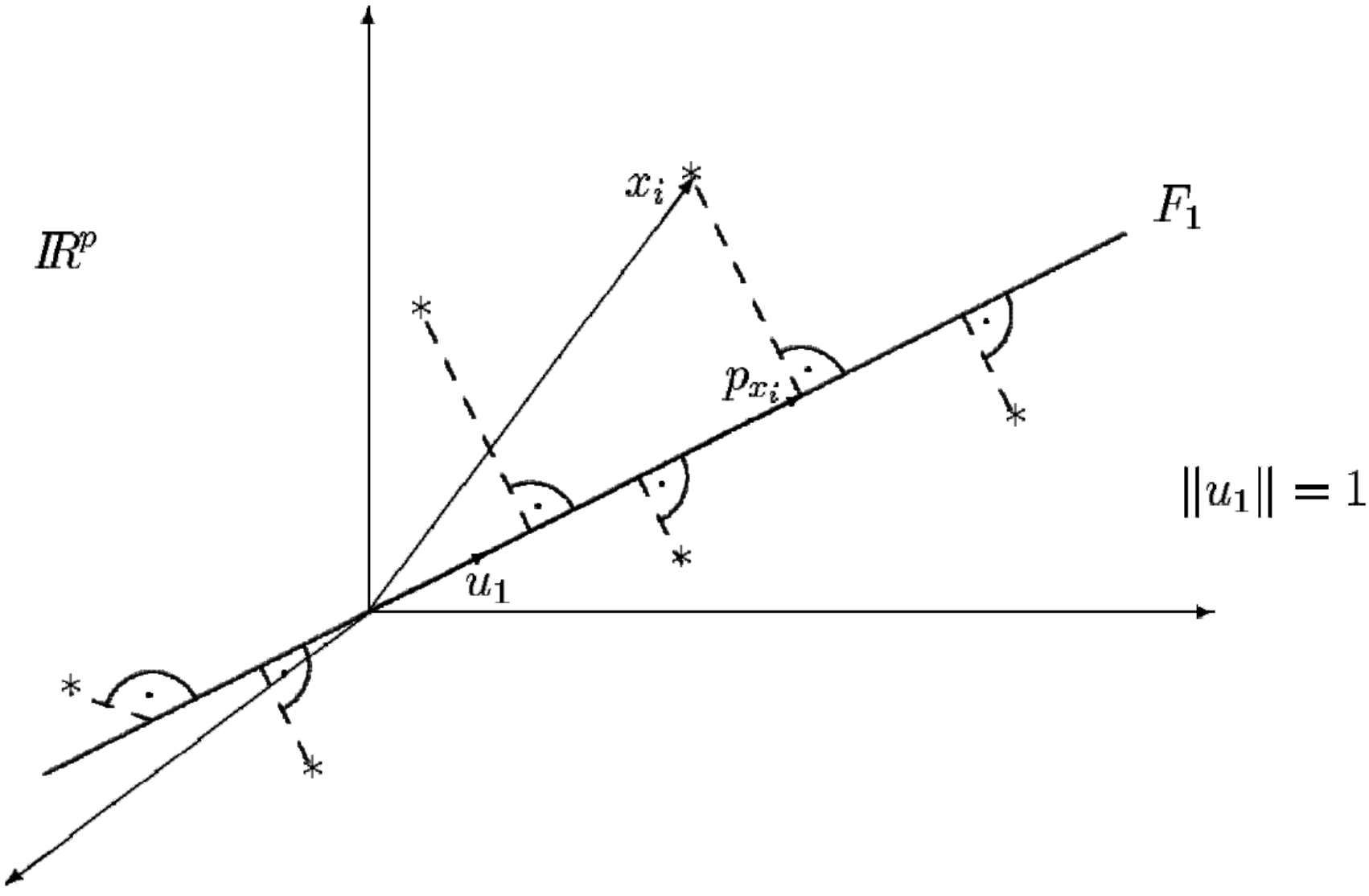
Q2 on prismia

The Geometric Point of View

The Geometric Point of View



Fitting the p-dimensional Point Cloud



Fitting the p-dimensional Point Cloud

Fitting the p-dimensional Point Cloud

What is the measure to represent the association between two variables?

The vector u_1 which minimizes the LS is the eigenvector of $XX^\top$ associated with the largest eigenvalue λ_1 of $XX^\top$.

The first factorial axis

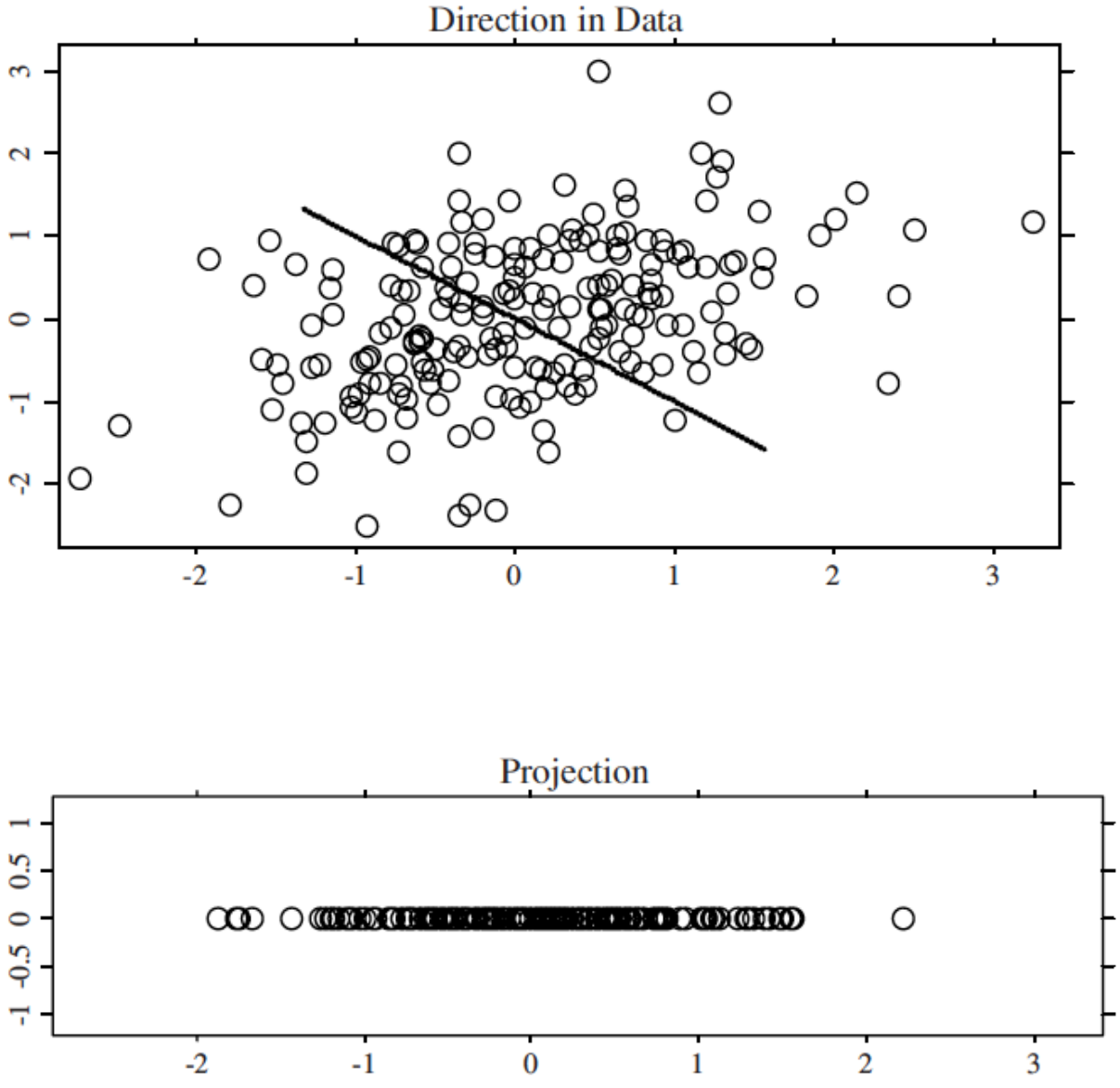
The first factorial axis

What about the second? The second factorial axis u_2 is the

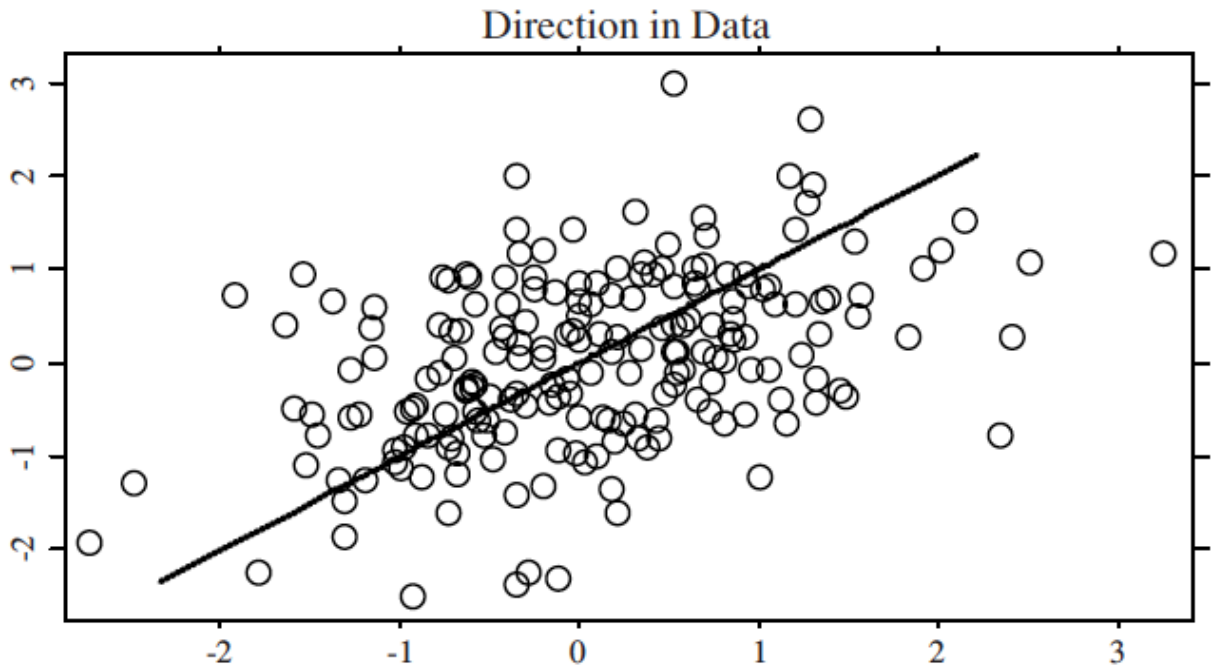
eigenvector of $XX^\top$ corresponding to the second largest eigenvalue λ_2 of $XX^\top$.

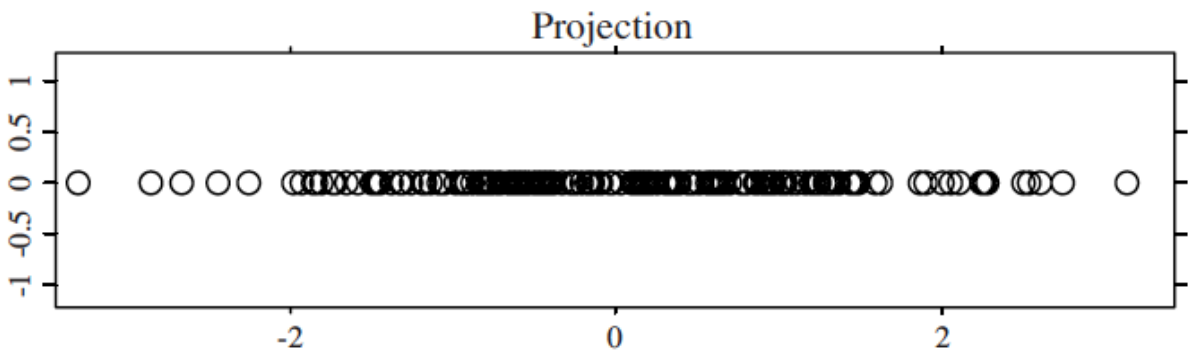
Principal Component Analysis: the direction

Principal Component Analysis: the direction



Principal Component Analysis: the direction





Principal Component Analysis: the theorem

For a given $X \sim (\mu, \Sigma)$ let $Y = \Gamma(X - \mu)$ be the PC transformation then

1. $E[Y_j] = 0$
2. $Var[Y_j] = \lambda_j$
3. $Cov[Y_i, Y_j] = 0$
4. $Var(Y_1) \geq Var[Y_2] \geq \cdots \geq Var[Y_p] \geq 0$
5. $\sum_{j=1}^{p} Var(Y_j) = tr(\Sigma)$

In practice: Swiss Bank Notes

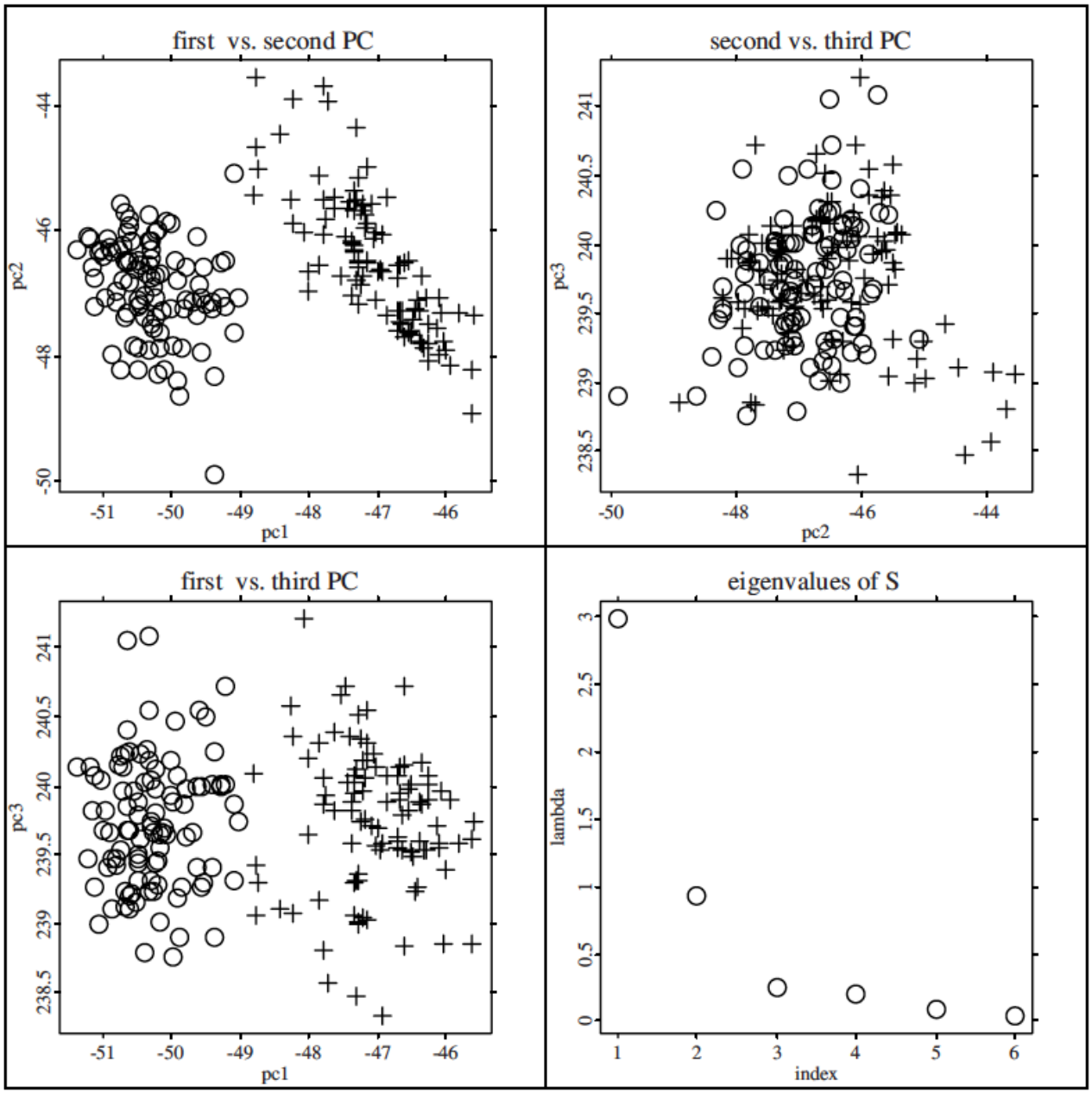
- 200 observations: 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes.
- The variables

X_1: Length of the bank note
X_2: Height of the bank note, measured on the left
X_3: Height of the bank note, measured on the right
X_4: Distance of inner frame to the lower border
X_5: Distance of inner frame to the upper border
X_6: Length of the diagonal

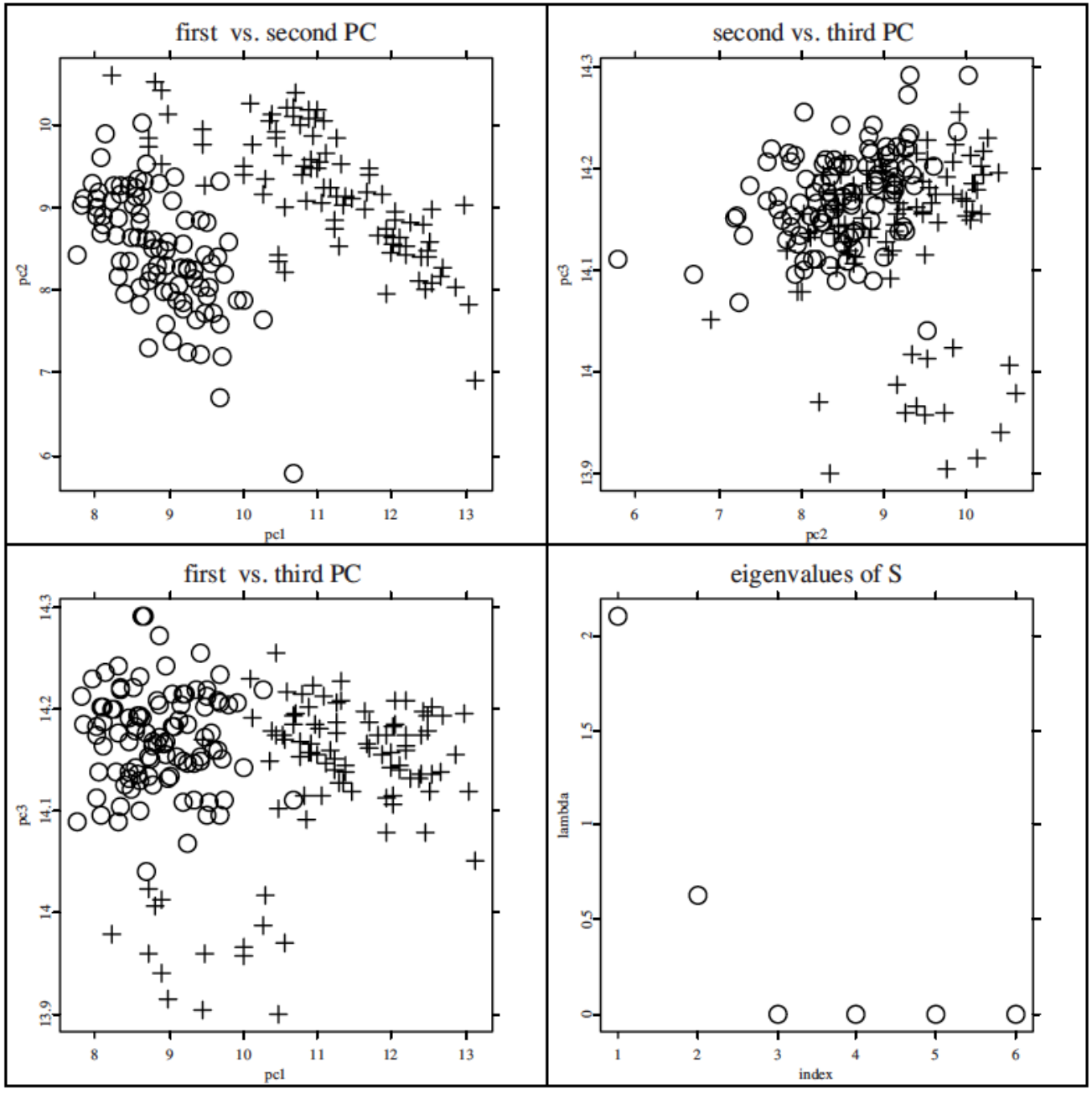
In practice: Swiss Bank Notes

Length	Height (left)	Height (right)	Inner Frame (lower)	Inner Frame (upper)	Diagonal
214.8	131.0	131.1	9.0	9.7	141.0
214.6	129.7	129.7	8.1	9.5	141.7
214.8	129.7	129.7	8.7	9.6	142.2
214.8	129.7	129.6	7.5	10.4	142.0
215.0	129.6	129.7	10.4	7.7	141.8
215.7	130.8	130.5	9.0	10.1	141.4

In practice: Swiss Bank Notes



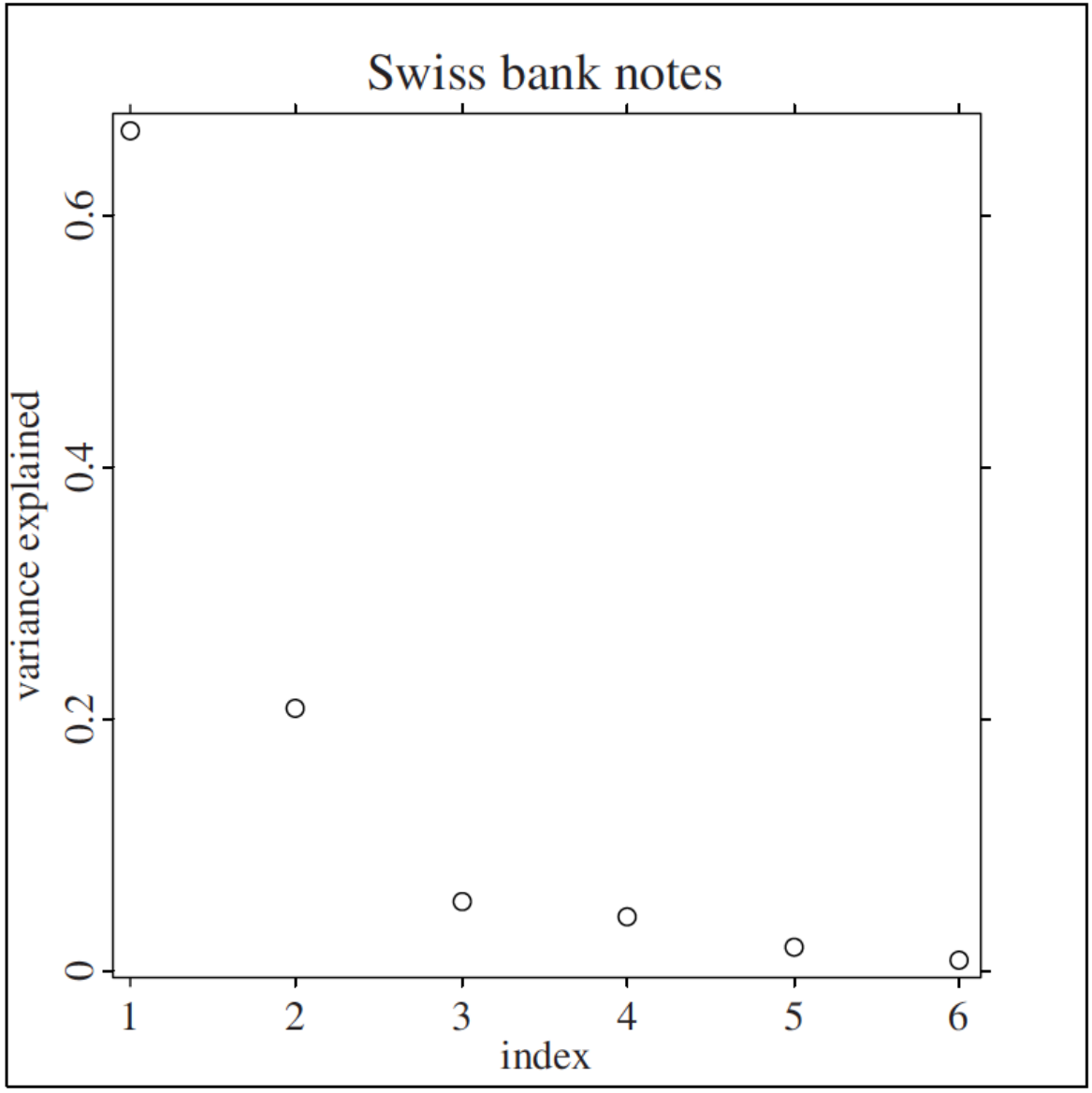
In practice: Swiss Bank Notes



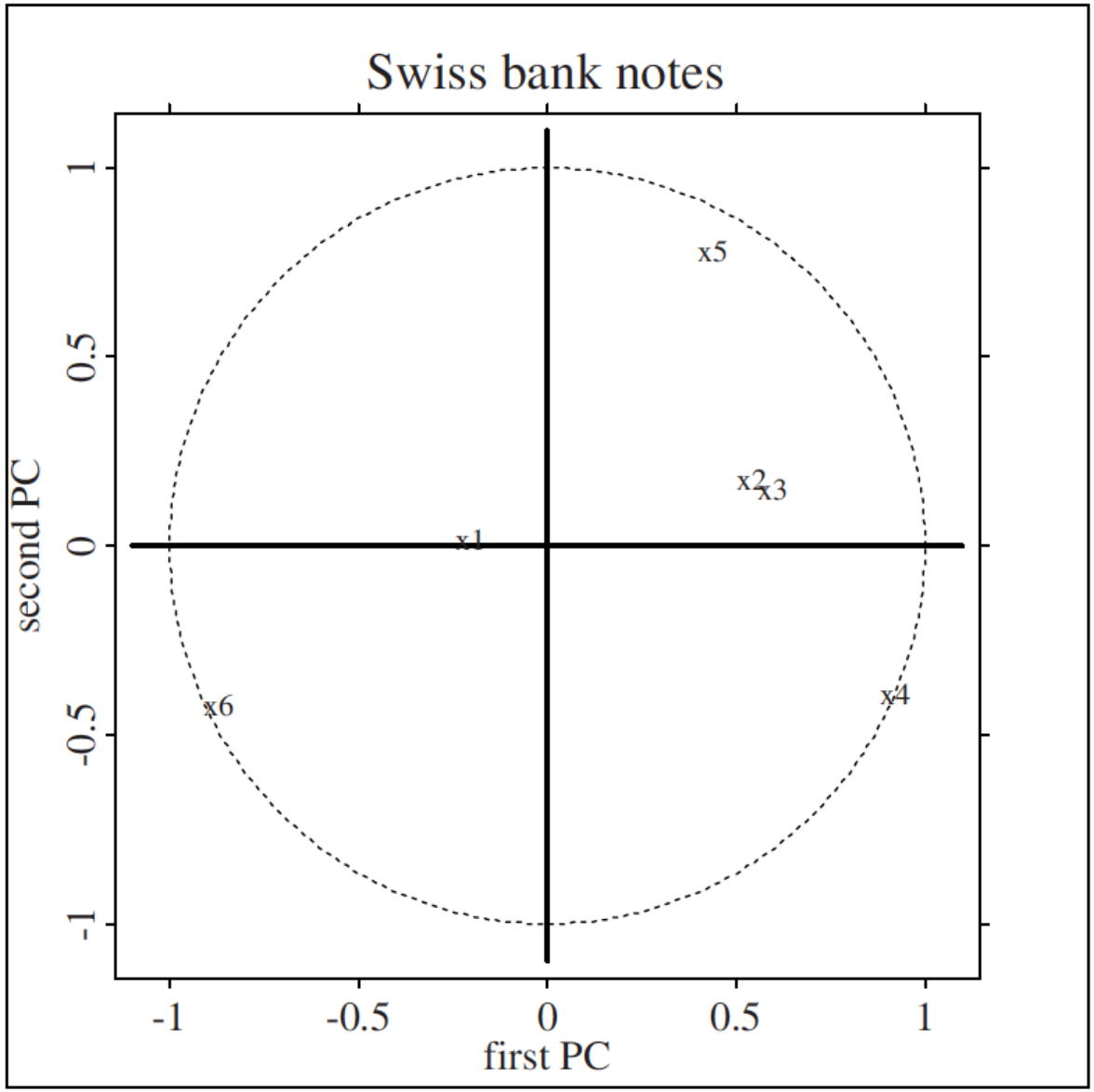
In practice: Swiss Bank Notes

$$
\begin{aligned}
y_1 &= -0.044x_1 + 0.112x_2 + 0.139x_3 + 0.768x_4 + 0.202x_5 - 0.579x_6 \\
y_2 &= 0.011x_1 + 0.071x_2 + 0.066x_3 - 0.563x_4 + 0.659x_5 - 0.489x_6
\end{aligned}
$$

In practice: Swiss Bank Notes



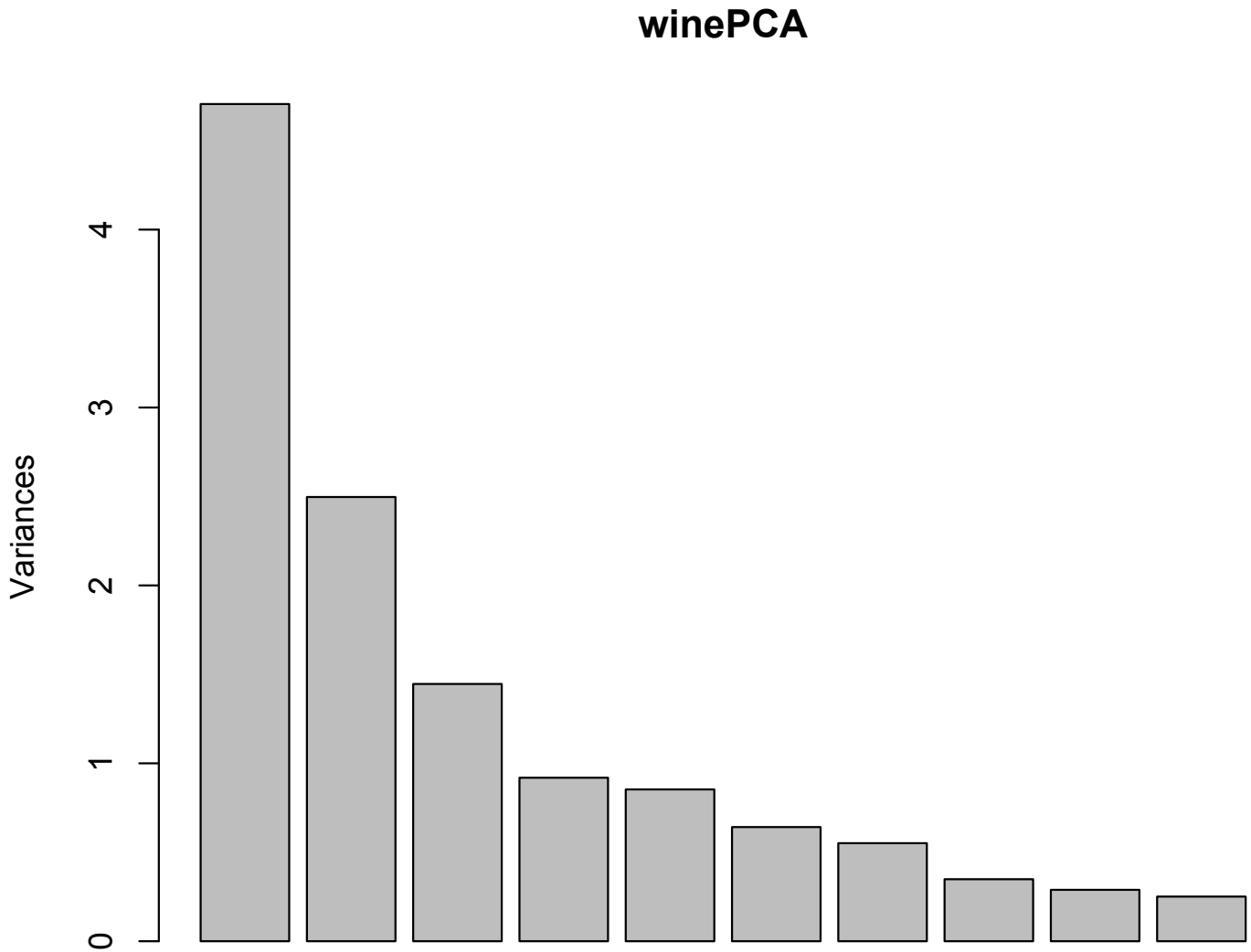
In practice: Swiss Bank Notes



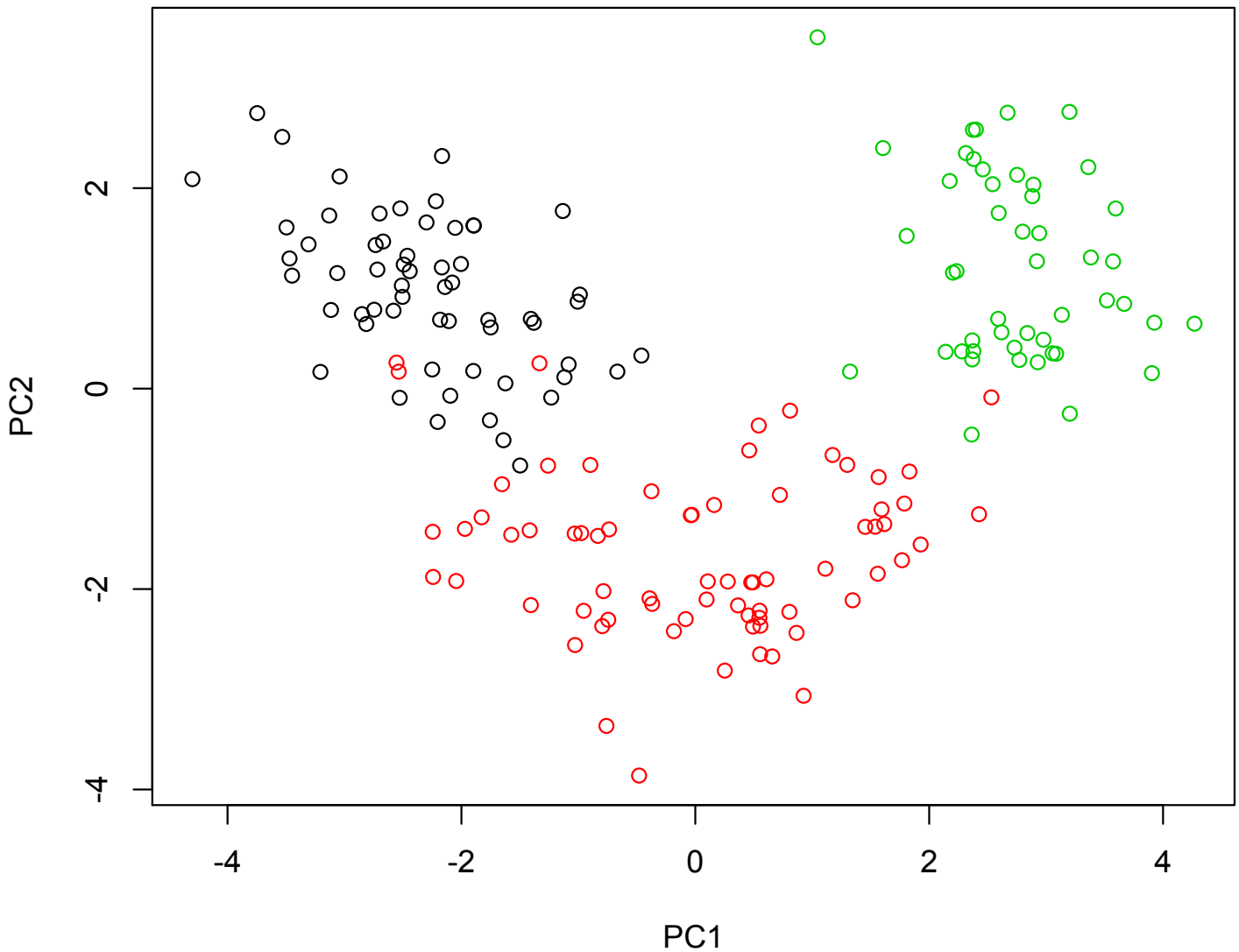
Another example: The wine

Type	The type of wine, into one of three classes 1 (59 obs), 2(71 obs), and 3 (48 obs)
Alcohol	Alcohol
Malic	Malic acid
Ash	Ash
Alcalinity	Alcalinity of ash
Magnesium	Magnesium
Phenols	Total phenols
Flavanoids	Flavanoids
Nonflavanoids	Nonflavanoid phenols
Proanthocyanins	Proanthocyanins
Color	Color intensity
Hue	Hue
Dilution	D280/OD315 of diluted wines
Proline	Proline

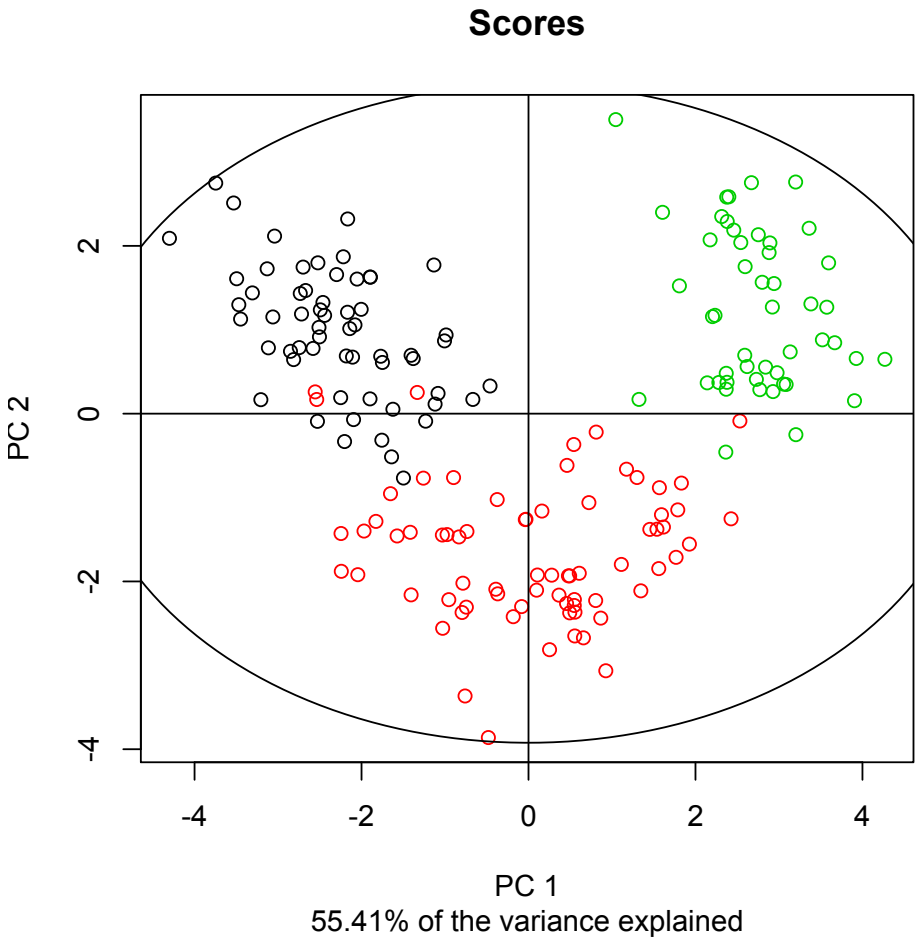
How many PCA will you consider?

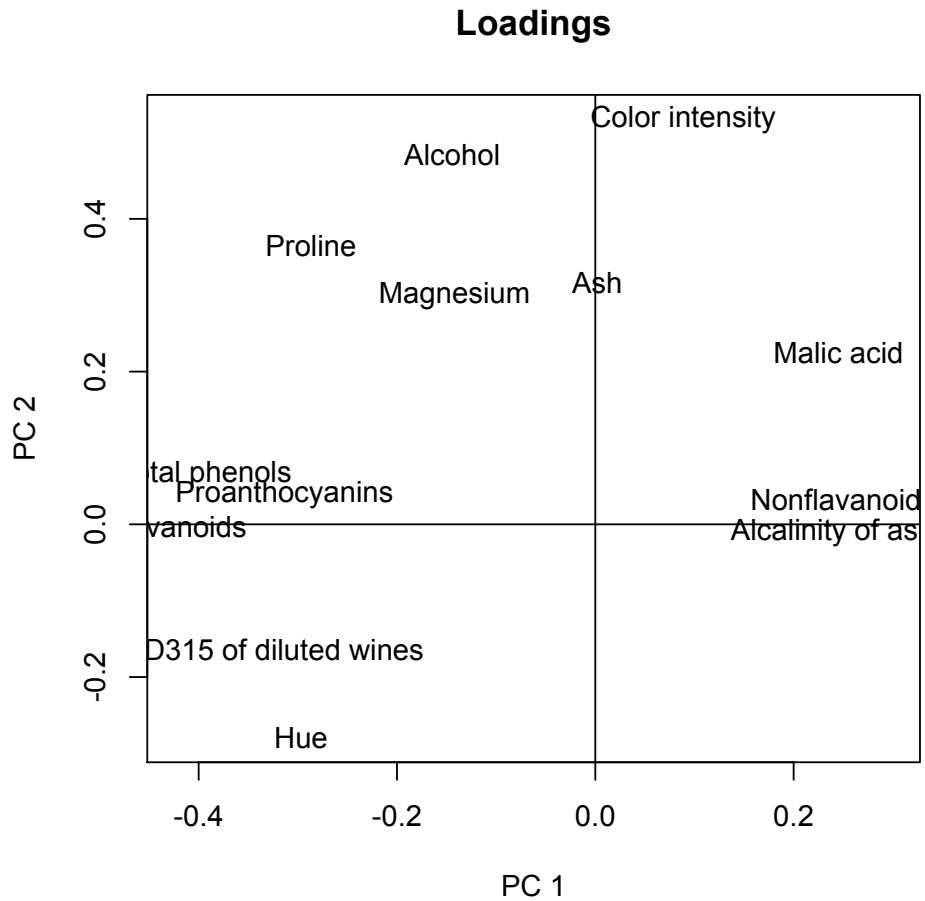


Cv1 (black), Cv2 (red), Cv3(green)



Cv1 (black), Cv2 (red), Cv3(green)





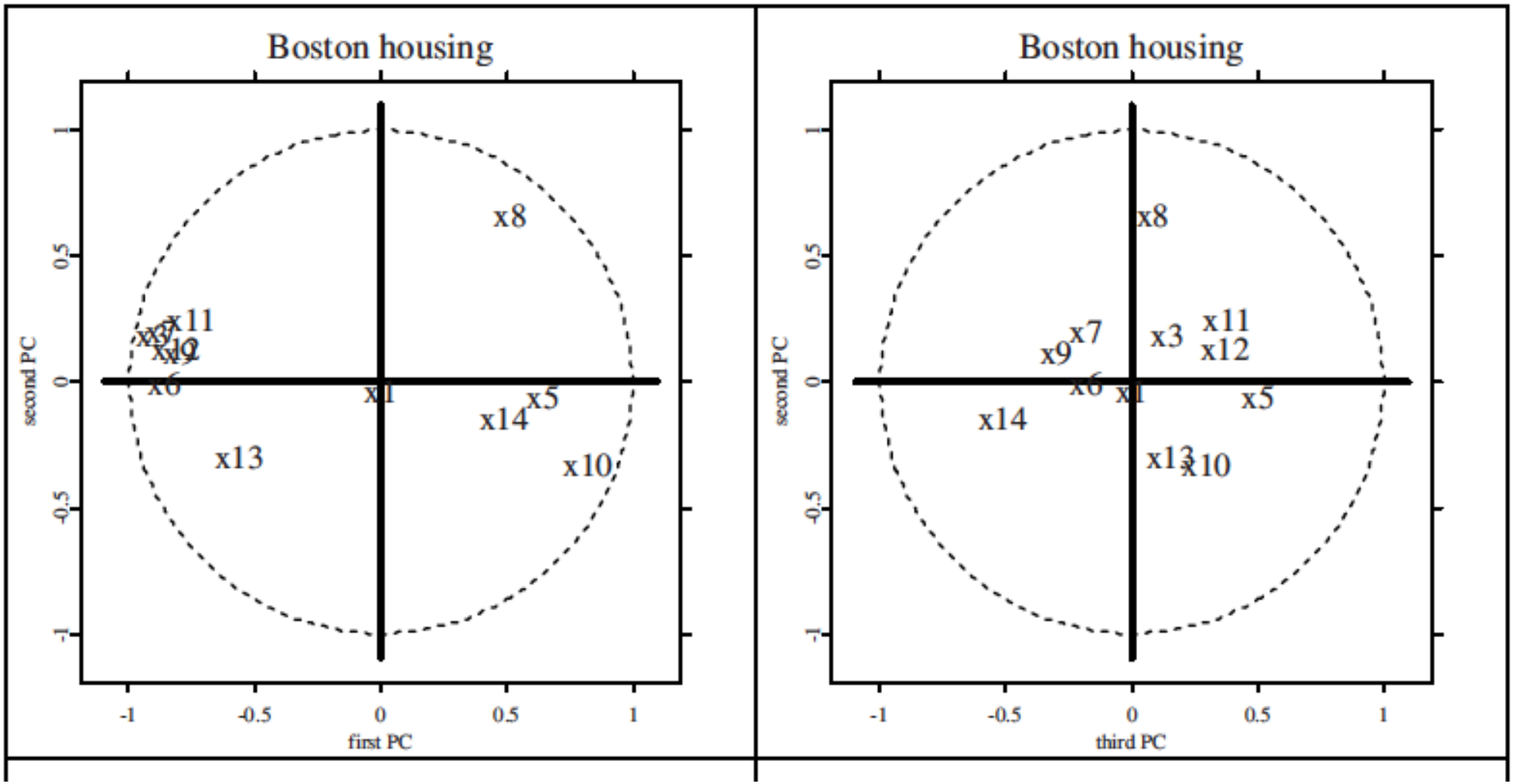
The Boston housing Example

X_1: per capita crime rate,
X_2: proportion of residential land zoned for large lots,
X_3: proportion of nonretail business acres,
X_4: Charles River (1 if tract bounds river, 0 otherwise),
X_5: nitric oxides concentration,
X_6: average number of rooms per dwelling,
X_7: proportion of owner-occupied units built prior to 1940,
X_8: weighted distances to five Boston employment centers,
X_9: index of accessibility to radial highways,
X_{10}: full-value property tax rate per \$10,000,
X_{11}: pupil/teacher ratio ,
X_{12}: $1000(B-0.63)^2\, \boldsymbol{I}(B<0.63)$ where B is the proportion of African American,
X_{13}: % lower status of the population,
X_{14}: median value of owner-occupied homes in \$1000.

The Boston housing Example

eigenvalue	percentages	cumulated percentages
7.2852	0.5604	0.5604
1.3517	0.1040	0.6644
1.1266	0.0867	0.7510
0.7802	0.0600	0.8111
0.6359	0.0489	0.8600
0.5290	0.0407	0.9007
0.3397	0.0261	0.9268
0.2628	0.0202	0.9470
0.1936	0.0149	0.9619
0.1547	0.0119	0.9738
0.1405	0.0108	0.9846
0.1100	0.0085	0.9931
0.0900	0.0069	1.0000

The Boston housing Example



The Boston housing Example