

02409 Multivariate Statistics

Lecture C, September 15 2025

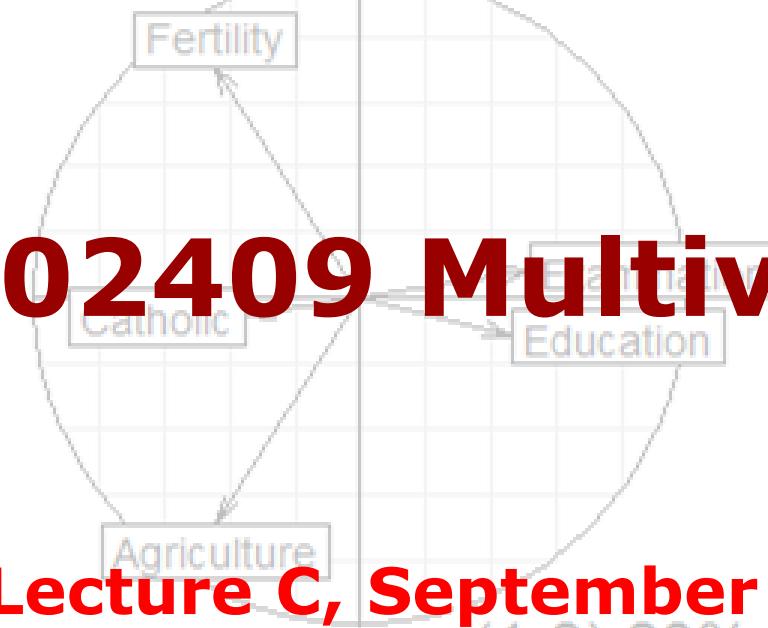
Anders Stockmarr

anst@dtu.dk

Course developers:

Anders Stockmarr

Anders Nymark Christensen



(1-3) 60%

V. De Geneve

Factor 1 [41%]

Factor 3 [19%]

Clustering 4 groups

Groups

28

16

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

Agenda

- Some examples
- A brief recap on estimation
- Partial correlation - examples
- Multiple correlation - examples
- PCA

Example - Exercise 2.4

In an investigation on the relationship between weather and crops the following variables were measured:

- X_1 : Yield of hay in cwt per acre
- X_2 : Spring rainfall in inches
- X_3 : Accumulated temperature above 42°F in the spring

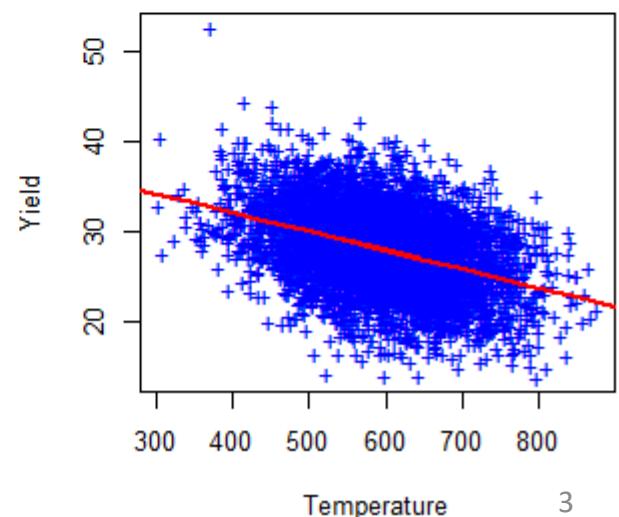
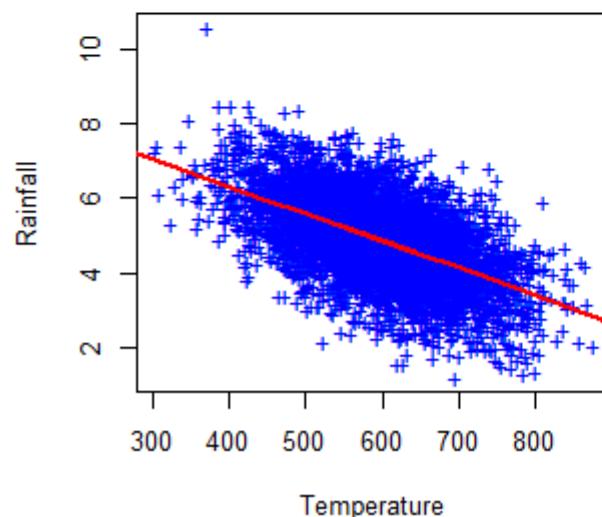
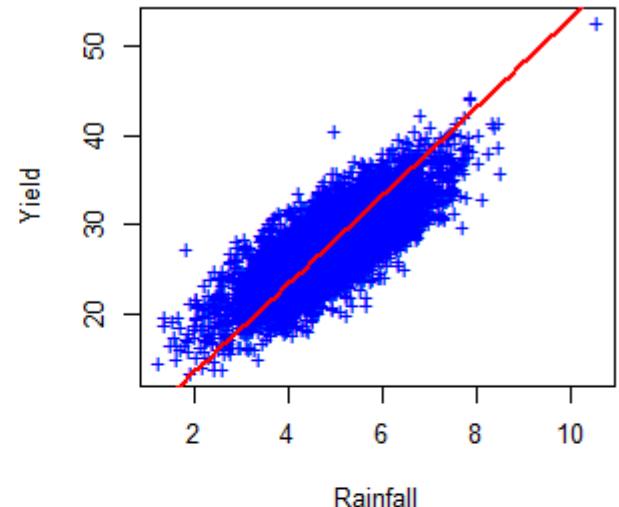
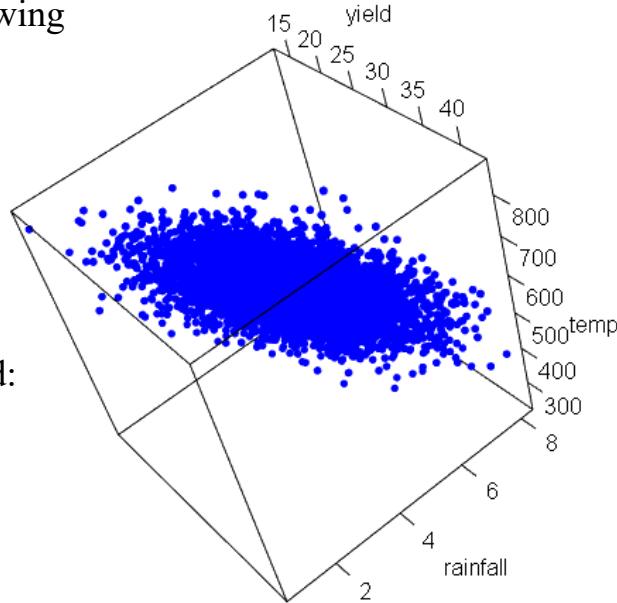
Based on observations from 20 years the following estimates were obtained:

- $\hat{\mu}_1 = 28.02 \quad \hat{\sigma}_1 = 4.42 \quad \hat{\rho}_{12} = 0.80$
- $\hat{\mu}_2 = 4.91 \quad \hat{\sigma}_2 = 1.10 \quad \hat{\rho}_{13} = -0.40$
- $\hat{\mu}_3 = 594 \quad \hat{\sigma}_3 = 85 \quad \hat{\rho}_{23} = -0.56$

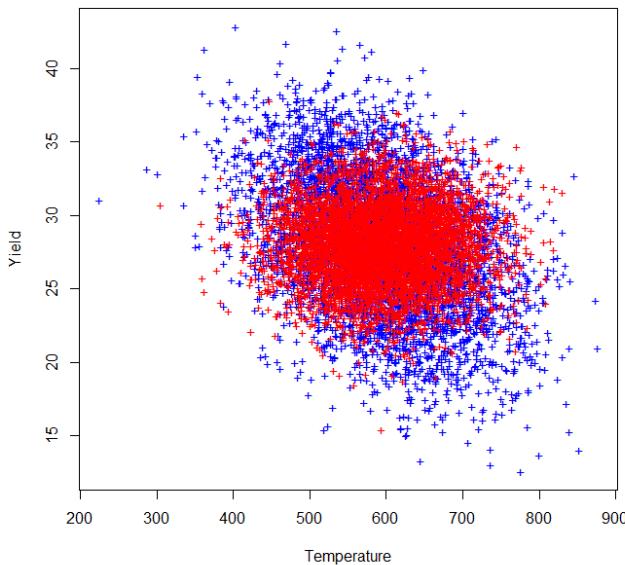
Compute the partial correlation between yield and temperature given the rainfall.

Another way to consider conditional distributions, is to fit a regression line, subtract it, and look at the residuals. I.e., remove the information that is associated with a specific value of the other variable.

Pictures are simulated data, based on the estimates above.



Example - Exercise 2.4



Red dots are simulations from the conditional distributions given a rainfall of 4.91; **far less variable and with less correlation**; but that is because the rain doesn't vary....

Partial covariance:

$$\Sigma = \begin{pmatrix} 4.42^2 & 0.8 \cdot 4.42 \cdot 1.1 & -0.40 \cdot 4.42 \cdot 85 \\ * & 1.1^2 & -0.56 \cdot 1.1 \cdot 85 \\ * & * & 85^2 \end{pmatrix}$$

$$\Sigma_{yt,yt} = \begin{pmatrix} 4.42^2 & -0.40 \cdot 4.42 \cdot 85 \\ * & 85^2 \end{pmatrix}$$

$$\Sigma_{yt,r} = \begin{pmatrix} 0.8 \cdot 4.42 \cdot 1.1 \\ -0.56 \cdot 1.1 \cdot 85 \end{pmatrix}, \Sigma_{r,yt} = \Sigma_{yt,r}^T$$

$$\Sigma_{r,r} = 1.1^2$$

$$\Sigma_{yt,yt|r} = \Sigma_{yt,yt} - \Sigma_{yt,r} \Sigma_{r,r}^{-1} \Sigma_{r,yt}$$

$$= \begin{pmatrix} 4.42^2 & -0.40 \cdot 4.42 \cdot 85 \\ * & 85^2 \end{pmatrix} -$$

$$\begin{pmatrix} 0.8 \cdot 4.42 \cdot 1.1 \\ -0.56 \cdot 1.1 \cdot 85 \end{pmatrix} \left(\frac{1}{1.1^2} \right) (0.8 \cdot 4.42 \cdot 1.1 \quad -0.56 \cdot 1.1 \cdot 85)$$

$$= \begin{pmatrix} 7.03 & 18.03 \\ * & 4959.24 \end{pmatrix}$$

Partial correlation:

$$\frac{18.03}{\sqrt{7.03 \cdot 4959.24}} = 0.097$$

Compare with the initial correlation: -0.40!

Estimation of parameters I

The i'th
observation

$$X_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{bmatrix}$$

The mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

The empirical
variance- covariance
or dispersion matrix

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

Observations collected
in data matrix

$$X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

Estimation of parameters II

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

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

|||| Theorem 1.32

Let the situation be as stated above. Then the $100(1 - \alpha)\%$ confidence ellipsoid for the unknown mean μ is

$$\{\mu | (\mu - \bar{\mathbf{x}})^T \mathbf{s}^{-1} (\mu - \bar{\mathbf{x}}) \leq \frac{p(n-1)}{(n-p)n} F(p, n-p)_{1-\alpha}\}$$

and the $100(1 - \alpha)\%$ prediction ellipsoid for a coming observation \mathbf{x} is

$$\{\mathbf{x} | (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{s}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \leq \frac{p(n-1)(n+1)}{(n-p)n} F(p, n-p)_{1-\alpha}\}$$

An example from fitness data

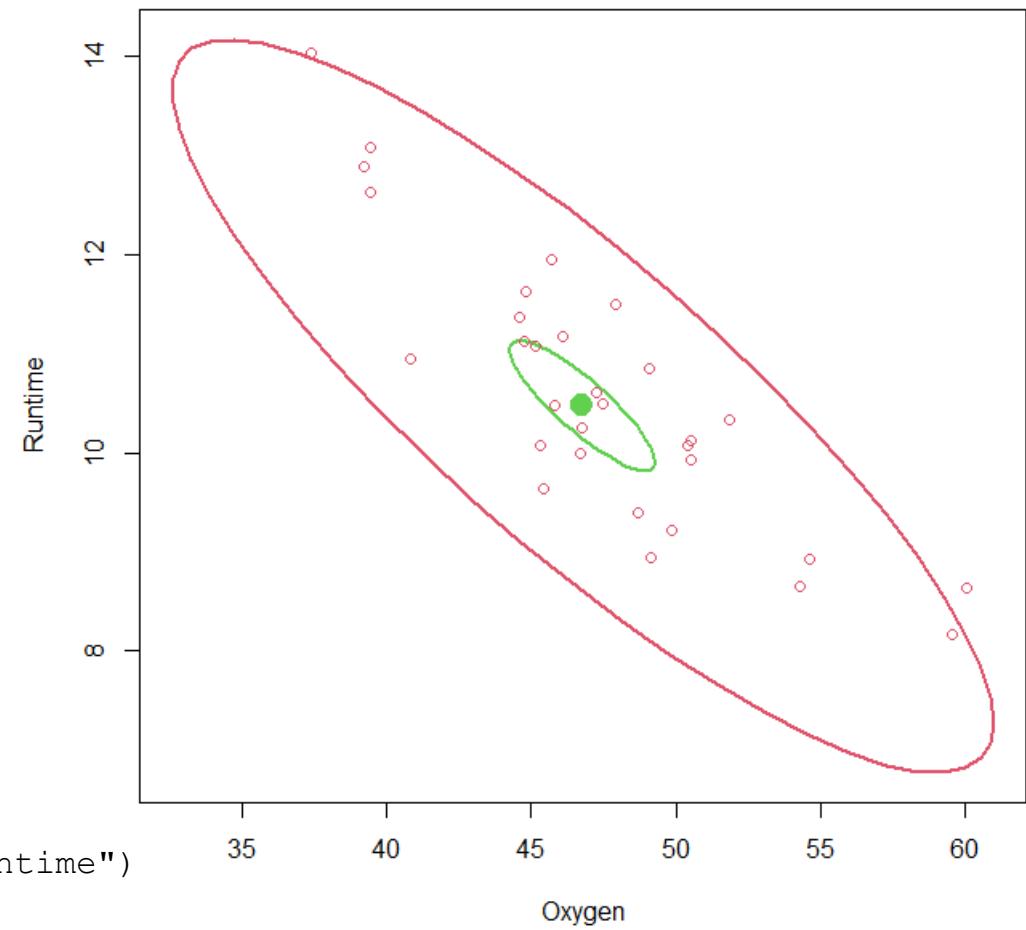
```
Fit<-read.table("Data/sundhed.txt") [,3:4]
names(Fit)<-c("Oxygen", "Runtime")

n <- length(Fit$Oxygen)
P <-2
shape <- var(Fit)
center <- sapply(Fit,mean,1)

conf95 <-
  sqrt(p * (n-1) * qf(0.95, p, n-p)/(n*(n-p)))
pred95 <- sqrt(n+1)*rconf95

library(car)
pred.elip95 <-
car::ellipse(center, shape, pred95, draw = FALSE)
conf.elip95 <-
car::ellipse(center, shape, conf95, draw = FALSE)

plot(pred.elip95, type='l', col = 2,lwd=2,xlab="Oxygen",ylab="Runtime")
lines(conf.elip95, type='l', col = 3,lwd=2)
points(Fit$Oxygen, Fit$Runtime,col=2)
points(center[1],center[2],cex=2,pch=19,col=3)
```



Partial correlation coefficient recap

The partial correlations between some variables given other variables are simply correlations in the conditional distribution of the 'some' variables given the 'other' variables. It follows that

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1 - \rho_{ik}^2)(1 - \rho_{jk}^2)}}$$

$$\rho_{ij|kl} = \frac{\rho_{ij|k} - \rho_{il|k} \cdot \rho_{jl|k}}{\sqrt{(1 - \rho_{il|k}^2) \cdot (1 - \rho_{jl|k}^2)}}$$

Let us look at some examples! ☺

Exam 2016 - Question 2 What is the partial correlation?

A: 0

0%

B: ρ

0%

C: ρ^2

0%

D: $1 - \rho^2$

0%

E: $\rho / \sqrt{1 - \rho^2}$

0%

F: Don't know

0%

Partial correlation coefficient Exam 2016 – problem 2

Consider a random variable

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1 - \rho_{ik}^2)(1 - \rho_{jk}^2)}}$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

with dispersion (variance-covariance) matrix

Question 2.1.

The partial correlation $\rho_{24|3}$ is

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

Note that the variances are 1; covariances and correlations are the same...

$$\rho_{24|3} = \frac{\boxed{\rho_{24}} + \boxed{\rho_{23}} \cdot \boxed{\rho_{43}}}{\sqrt{(1 - \rho_{23})(1 - \rho_{43})}} \\ \frac{\rho^2 - \rho \cdot \rho}{\sqrt{(1 - \rho)(1 - \rho)}} = 0$$

Option A

Example - Cement strength

Correlation Matrix					
	C3S	C3A	BLAINE	Strgth3	Strgth28
C3S	1.000
C3A	-0.309	1.000	.	.	.
BLAINE	0.091	0.192	1.000	.	.
Strgth3	0.158	0.120	0.745	1.000	.
Strgth28	0.344	-0.166	0.320	0.464	1.000

Partial Correlation Matrix					
	C3S	C3A	BLAINE	Strgth3	Strgth28
C3S	1.0000				
C3A	-.3340	1.0000			
BLAINE					
Strgth3	0.1358	-.0352		1.0000	
Strgth28	0.3337	-.2446		0.3570	1.0000

Partial correlation coefficient III

||| Theorem 1.37

Let $R = R_{ij|m+1\dots p}$ be the empirical partial correlation coefficient between Z_i and Z_j conditioned on (or: for given) Z_{m+1}, \dots, Z_p . It is assumed to be computed from the unbiased estimates of the variance-covariance matrix and from n observations. Then

$$\frac{R}{\sqrt{1 - R^2}} \sqrt{n - 2 - (p - m)} \sim t(n - 2 - (p - m)),$$

if $\rho_{ij|m+1,\dots,p} = 0$.

||| Example 1.39

Let us investigate whether the value of $r_{24|3}$ is significantly different from 0. We find with $r_{24|3} = R$:

$$\begin{aligned} \frac{R}{\sqrt{1 - R^2}} \sqrt{n - 2 - (p - m)} &= \frac{-0.035}{\sqrt{1 - 0.035^2}} \cdot \sqrt{51 - 2 - (5 - 4)} \\ &= -0.243 = t(48)_{40\%}. \end{aligned}$$

A hypothesis that $\rho_{24|3}$ is 0 will therefore be accepted using a test at level α for $\alpha < 80\%$. (Note: this is by nature a two-sided test.)

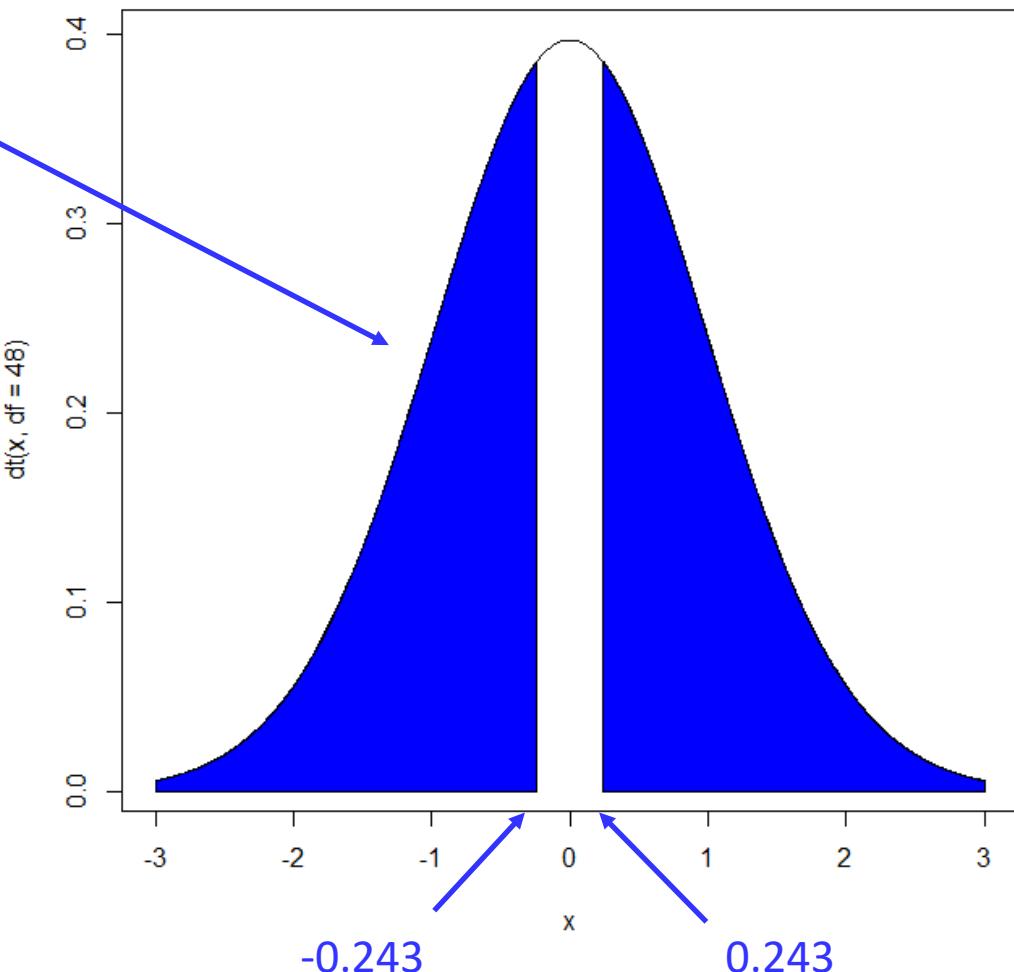
Note that from the example text one can deduce (nearly) that $2 \cdot p(t_{48} < -0.243) = 0.81$; the p-value is 0.81.

Test-values, example 1.37

```
p<-2*pt (-.243, df=48)  
P  
[1] 0.8090408
```

There is therefore no evidence in the data
that supports that C3A and Strength3 are correlated,
once the BLAINE level is known.

Area: 0.81



Multiple correlation coefficient - example

We now define the multiple correlation coefficient between Y_i , $i = 1, \dots, m$ and X as the maximal correlation between Y_i and a linear combination of X 's elements. It is denoted $\rho_{y_i|x}$.

|||| Theorem 1.42

We consider the situation above. Let σ_i be the i 'th column in Σ_{xy} , i.e. σ_i^T is the i 'th row in Σ_{yx} . Further, let σ_{ii} denote the i 'th diagonal element, i.e. the variance of Y_i .

Then

$$\rho_{y_i|x} = \frac{\sqrt{\sigma_i^T \Sigma_{xx}^{-1} \sigma_i}}{\sqrt{\sigma_{ii}}}.$$

If we let

$$\Sigma_i = \begin{bmatrix} \sigma_{ii} & \sigma_i^T \\ \sigma_i & \Sigma_{xx} \end{bmatrix},$$

then

$$1 - \rho_{y_i|x}^2 = \frac{\det \Sigma_i}{\sigma_{ii} \det \Sigma_{xx}} = \frac{V(Y_i|X)}{V(Y_i)},$$

Exam 2016 problem 2.2 What is the multiple correlation?

A: 0

0%

B: ρ

0%

C: ρ^2

0%

D: $1 - \rho^2$

0%

E: $\rho / \sqrt{1 - \rho^2}$

0%

F: Don't Know

0%

Exam 2016 – Problem 2.2

Consider a random variable

$$\Sigma_i = \begin{bmatrix} \sigma_{ii} & \sigma_i^T \\ \sigma_i & \Sigma_{xx} \end{bmatrix},$$

$$1 - \rho_{y_i|X}^2 = \frac{\det \Sigma_i}{\sigma_{ii} \det \Sigma_{xx}} = \frac{V(Y_i|X)}{V(Y_i)},$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

with dispersion (variance-covariance) matrix

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

Question 2.2.

The squared multiple correlation $\rho_{1|23}^2$ is

- 1 0
- 2 ρ
- 3 ρ^2
- 4 $1 - \rho^2$
- 5 $\rho/\sqrt{1 + \rho^2}$
- 6 Don't know.

$$\begin{aligned} \rho_{1|23}^2 &= 1 - \frac{\det \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}}{1 \cdot \det \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}} = \\ &= 1 - \frac{1 \cdot 1 \cdot 1 + \rho \cdot \rho \cdot \rho^2 + \rho^2 \cdot \rho \cdot \rho - \rho^2 \cdot 1 \cdot \rho^2 - \rho \cdot \rho \cdot 1 - 1 \cdot \rho \cdot \rho}{1 - \rho^2} = \\ &= 1 - \frac{1 + \rho^4 + \rho^4 - \rho^4 - \rho^2 - \rho^2}{1 - \rho^2} = \\ &= 1 - \frac{\rho^4 - 2\rho^2 + 1}{1 - \rho^2} = 1 - \frac{(1 - \rho^2)(1 - \rho^2)}{1 - \rho^2} = \rho^2 \end{aligned}$$

Alternative:

$$\begin{aligned} \rho_{1|23}^2 &= \Sigma_{1,23} \Sigma_{23,23}^{-1} \Sigma_{23,1} = (\rho \quad \rho^2) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho \\ \rho^2 \end{pmatrix} \\ &= \frac{\rho^2}{1 - \rho^2} (1 \quad \rho) \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \rho \end{pmatrix} = \frac{\rho^2}{1 - \rho^2} (1 - \rho^2 \quad 0) \begin{pmatrix} 1 \\ \rho \end{pmatrix} = \rho^2 \quad \text{Option C} \end{aligned}$$

Example - Cement strength

Correlation Matrix					
	C3S	C3A	BLAINE	Strgth3	Strgth28
C3S	1.000
C3A	-0.309	1.000	.	.	.
BLAINE	0.091	0.192	1.000	.	.
Strgth3	0.158	0.120	0.745	1.000	.
Strgth28	0.344	-0.166	0.320	0.464	1.000

Partial Correlation Matrix					
	C3S	C3A	BLAINE	Strgth3	Strgth28
C3S	1.0000				
C3A	-.3340	1.0000			
BLAINE					
Strgth3	0.1358	-.0352		1.0000	
Strgth28	0.3337	-.2446		0.3570	1.0000

Multiple correlation coefficient II

|||| Theorem 1.45

Let $R = \hat{\rho}_{y_i|x}$ be the empirical multiple correlation coefficient between Y_i and $\mathbf{X} = (Z_{m+1}, \dots, Z_p)$ based upon n observations. Then

$$\frac{R^2}{1 - R^2} \cdot \frac{n - (p - m) - 1}{p - m} \sim F(p - m, n - (p - m) - 1),$$

if $\rho_{y_i|x} = \rho_{y_i|z_{m+1}, \dots, z_p} = 0$.

|||| Example 1.44

To get an impression of to which degree the content of C_3A and C_3S in example 1.39 can explain the variation in e.g. 3-day strength we can compute the multiple correlation coefficient between strength day 3 and (C_3S , and C_3A). We find

$$1 - \hat{\rho}_{4|12}^2 = \frac{\det \begin{bmatrix} 1 & 0.158 & 0.120 \\ 0.158 & 1 & -0.309 \\ 0.120 & -0.309 & 1 \end{bmatrix}}{1 \cdot \det \begin{bmatrix} 1 & -0.309 \\ -0.309 & 1 \end{bmatrix}}$$

where the indices of the variables correspond to those used in example 1.35. We find

$$\hat{\rho}_{4|12}^2 = 1 - 0.9435 = 0.0565.$$

The data therefore indicate that only about 6% of the variation in the strength of the cement (from samples which have been collected the way these data have been collected) can be explained by variations in C_3S - and C_3A - content alone.

|||| Example 1.47

Consider the situation in example 1.44. We now want to examine if it can be assumed that the multiple correlation between X_4 and (X_1, X_2) is 0. (Note that $p = 3$ and $m = 1$.) We find the statistic

$$\frac{R^2}{1 - R^2} \frac{51 - (3 - 1) - 1}{3 - 1} = \frac{0.0565}{0.9435} \cdot \frac{48}{2} = 1.44.$$

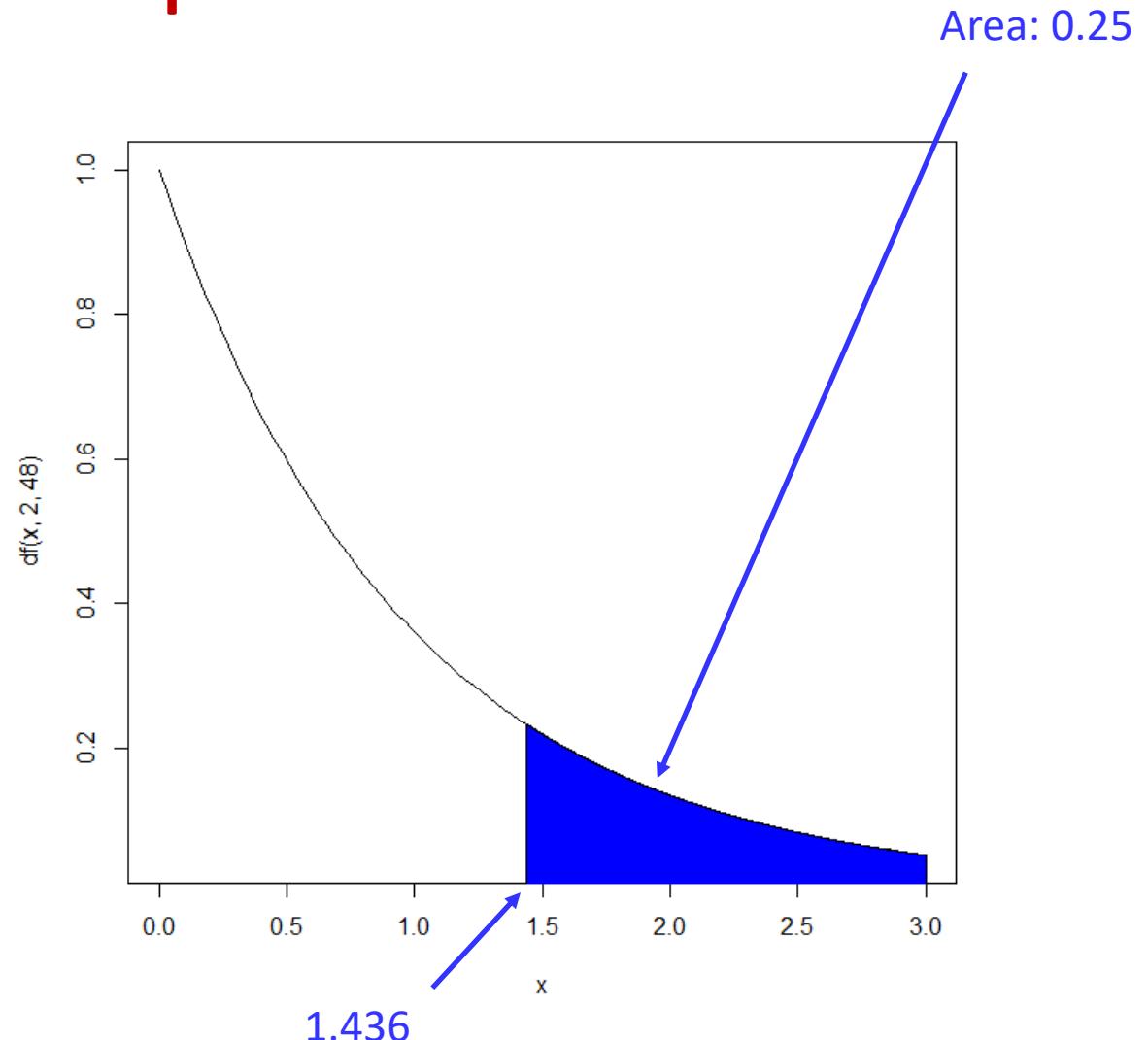
Since

$$F(2, 48)_{0.90} = 2.42,$$

we will at least accept a hypothesis that $\rho_{4|12} = 0$ for any level $\alpha < 10\%$. With the available data it cannot be rejected that $\rho_{4|12} = 0$. This does not mean that it is not different from 0 (which it probably is), only that we cannot be sure using the available data because the true (but unknown) value of $\rho_{4|12}$ is probably rather small.

Test-values, example 1.47

```
R2<-1-det(matrix(c(1,0.158,0.12,0.158,1,  
-0.309,0.12,-0.309,1),nrow=3))/  
det(matrix(c(1,-0.309,-0.309,1),nrow=2))  
R2  
  
# 1.436485  
  
(R2/(1-R2))*48/2  
  
# 1.436  
  
1-pf((R2/(1-R2))*48/2,2,48)  
  
# 0.2477989
```



PCA: Principle Components Analysis

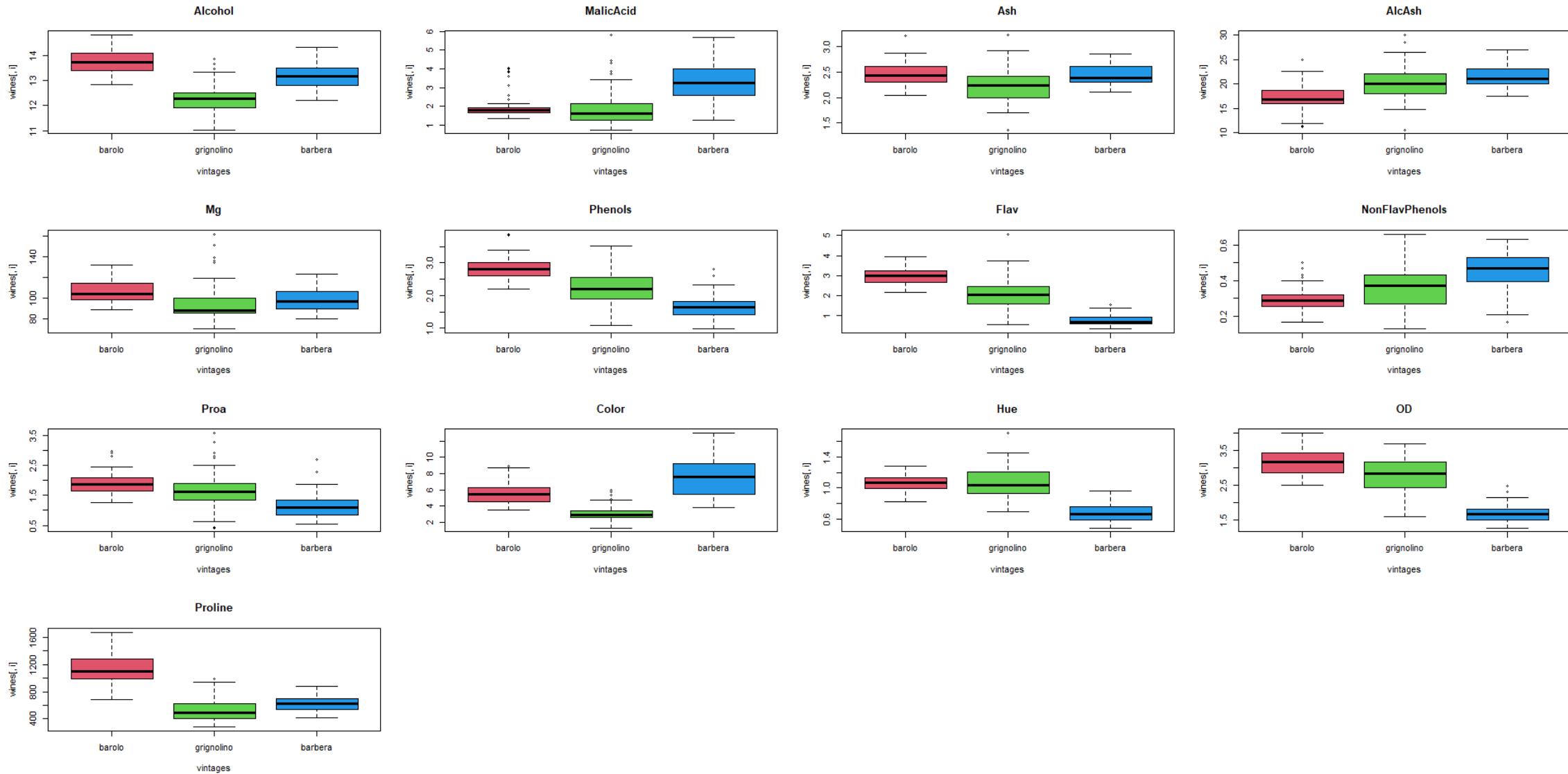
Example: The Wine data

```
load("Data/Winedata.Rdata")
```

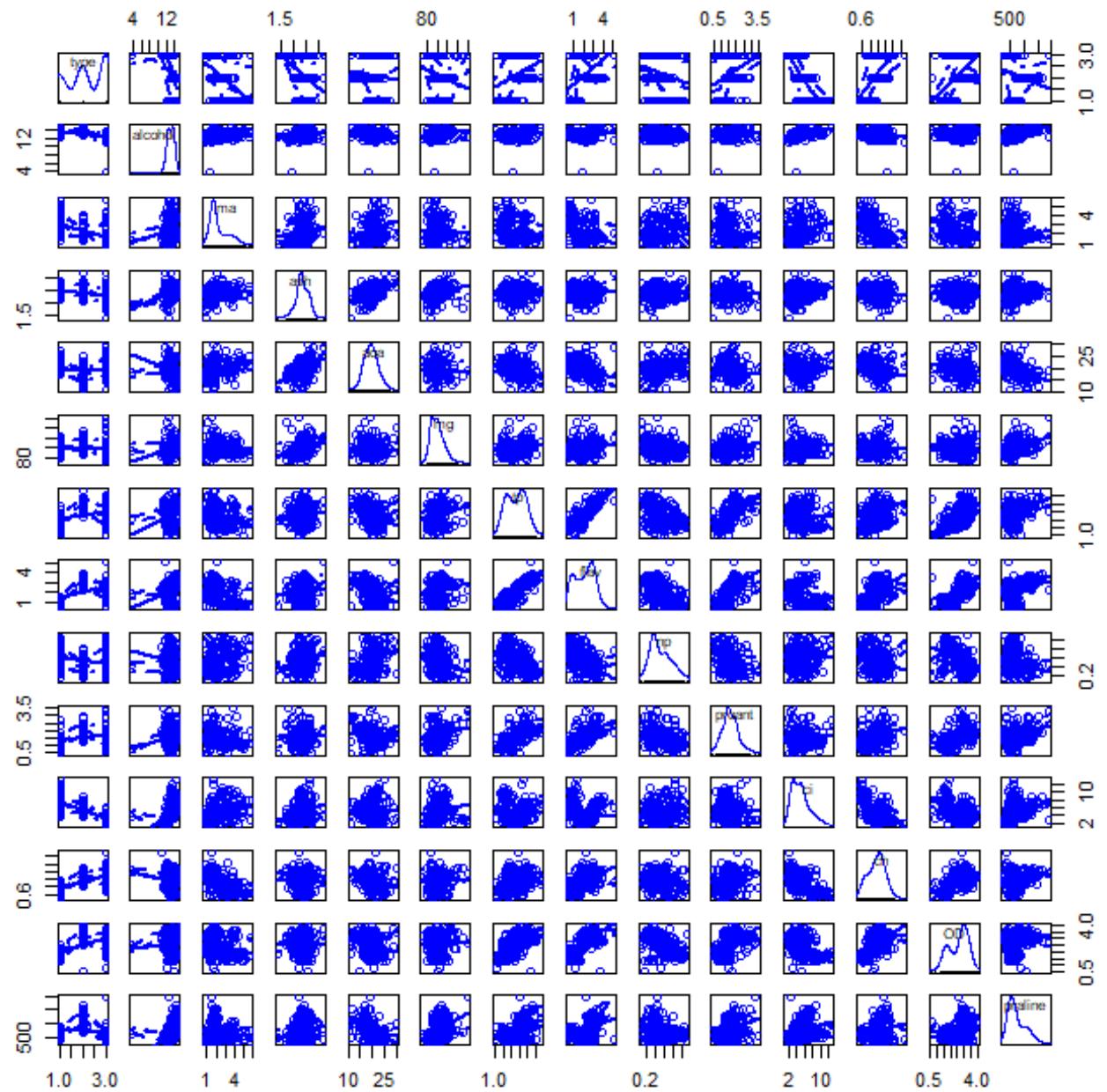
- A dataset consisting of n=178 Italian wines. Of these, 59 are Barolo wines, 71 are Grignolino wines, and 48 are Barbera wines. Two sub-elements: `wines` and `vintages` (Barolo, Grignolino or Barbera).
- 13 characteristics of the wines:

1) Alcohol	8) Nonflavanoid phenols: NonFlavPhenols
2) MalicAcid	9) Proanthocyanins: Proa
3) Ash	10) Color intensity: Color
4) Alkalinity of ash: AlcAsh	11) Color hue: Hue
5) Magnesium: Mg	12) OD280/OD315 protein measurement: OD
6) Total Phenols: Phenols	13) Proline (amino acid): Proline
7) Flavanoids: Flavor	

Variation of Vintages



The Wine Data



Separation of Wines

- What causes wines to be different?
- With the 13 characteristics, we can distinguish wines through differences in the characteristics. But are all 13 characteristics necessary? Some may be redundant.
- If we can identify scales (linear combinations of the characteristics) where the characteristics vary the most, we can also (hope to) find a scale that differentiates optimally between the wines.

Separation of Wines

- The variance covariance matrix of the wine characteristics:

	Alcohol	MalicAcid	Ash	AlcAsh	Mg	Phenols	Flav	NonFlavPhenols	Proa	Color	Hue	OD	Proline
Alcohol	0.66	0.09	0.05	-0.84	3.14	0.15	0.19		-0.02	0.06	1.03	-0.01	0.04
MalicAcid	0.09	1.25	0.05	1.08	-0.87	-0.23	-0.46		0.04	-0.14	0.64	-0.14	-0.29
Ash	0.05	0.05	0.08	0.41	1.12	0.02	0.03		0.01	0.00	0.16	0.00	0.00
AlcAsh	-0.84	1.08	0.41	11.15	-3.97	-0.67	-1.17		0.15	-0.38	0.15	-0.21	-0.66
Mg	3.14	-0.87	1.12	-3.97	203.99	1.92	2.79		-0.46	1.93	6.62	0.18	0.67
Phenols	0.15	-0.23	0.02	-0.67	1.92	0.39	0.54		-0.04	0.22	-0.08	0.06	0.31
Flav	0.19	-0.46	0.03	-1.17	2.79	0.54	1.00		-0.07	0.37	-0.40	0.12	0.56
NonFlavPhenols	-0.02	0.04	0.01	0.15	-0.46	-0.04	-0.07		0.02	-0.03	0.04	-0.01	-0.04
Proa	0.06	-0.14	0.00	-0.38	1.93	0.22	0.37		-0.03	0.33	-0.03	0.04	0.21
Color	1.03	0.64	0.16	0.15	6.62	-0.08	-0.40		0.04	-0.03	5.37	-0.28	-0.71
Hue	-0.01	-0.14	0.00	-0.21	0.18	0.06	0.12		-0.01	0.04	-0.28	0.05	0.09
OD	0.04	-0.29	0.00	-0.66	0.67	0.31	0.56		-0.04	0.21	-0.71	0.09	0.50
Proline	164.57	-67.55	19.32	-463.36	1769.16	98.17	155.45		-12.20	59.55	230.77	17.00	69.93

99166.72

Separation of Wines

- To avoid scaling problems, we must scale the data to the same scale. Variances: Look in correlations instead-
- The `scale` function in R subtracts the mean and divide by the sd:

$$X^{scaled} = \frac{X - \text{mean}(X)}{\text{sd}(X)}$$

	Alcohol	MalicAcid	Ash	AlcAsh	Mg	Phenols	Flav	NonFlavPhenols	Proa	Color	Hue	OD	Proline	
Alcohol	1.00	0.09	0.21	-0.31	0.27	0.29	0.24		-0.16	0.14	0.55	-0.07	0.07	0.64
MalicAcid	0.09	1.00	0.16	0.29	-0.05	-0.34	-0.41		0.29	-0.22	0.25	-0.56	-0.37	-0.19
Ash	0.21	0.16	1.00	0.44	0.29	0.13	0.12		0.19	0.01	0.26	-0.07	0.00	0.22
AlcAsh	-0.31	0.29	0.44	1.00	-0.08	-0.32	-0.35		0.36	-0.20	0.02	-0.27	-0.28	-0.44
Mg	0.27	-0.05	0.29	-0.08	1.00	0.21	0.20		-0.26	0.24	0.20	0.06	0.07	0.39
Phenols	0.29	-0.34	0.13	-0.32	0.21	1.00	0.86		-0.45	0.61	-0.06	0.43	0.70	0.50
Flav	0.24	-0.41	0.12	-0.35	0.20	0.86	1.00		-0.54	0.65	-0.17	0.54	0.79	0.49
NonFlavPhenols	-0.16	0.29	0.19	0.36	-0.26	-0.45	-0.54		1.00	-0.37	0.14	-0.26	-0.50	-0.31
Proa	0.14	-0.22	0.01	-0.20	0.24	0.61	0.65		-0.37	1.00	-0.03	0.30	0.52	0.33
Color	0.55	0.25	0.26	0.02	0.20	-0.06	-0.17		0.14	-0.03	1.00	-0.52	-0.43	0.32
Hue	-0.07	-0.56	-0.07	-0.27	0.06	0.43	0.54		-0.26	0.30	-0.52	1.00	0.57	0.24
OD	0.07	-0.37	0.00	-0.28	0.07	0.70	0.79		-0.50	0.52	-0.43	0.57	1.00	0.31
Proline	0.64	-0.19	0.22	-0.44	0.39	0.50	0.49		-0.31	0.33	0.32	0.24	0.31	1.00

Separation of Wines

- Let us consider the correlation matrix:

```
Sigma<-var(scale(wines))
```

The sum of the standardized variances:

```
sum(diag(Sigma))  
[1] 13
```

Of course – there are 13 variables.

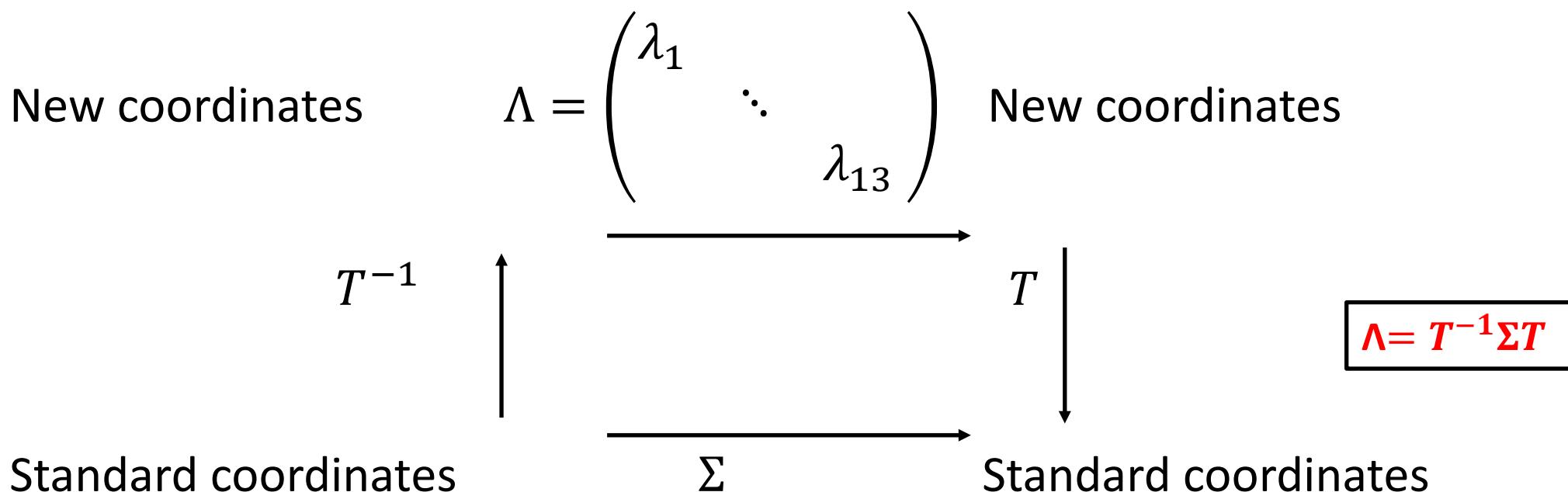
Original question: In which direction (scale) do wines data vary the most?

Let us start by representing the data in a set of coordinates where no correlation is present, to get an overview not disturbed by correlations.

Separation of Wines

- No correlations means that Σ is represented by a diagonal matrix in these directions;
- in other words that the new coordinates T consists of eigenvectors for Σ ; solutions to the equation

$$\Sigma v = \lambda v$$



Separation of Wines

- Eigenvectors in R:

```
T<-eigen(Sigma)$vectors
```

The inverse of T is equal to the matrix transpose $t(T)$:

```
Lambda<-t (T) %*% Sigma %*% T
```

```
round(Lambda, digits=2)
```

Separation of Wines

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
[1,]	4.71	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[2,]	0.00	2.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[3,]	0.00	0.0	1.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[4,]	0.00	0.0	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[5,]	0.00	0.0	0.00	0.00	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[6,]	0.00	0.0	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[7,]	0.00	0.0	0.00	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.00	0.0
[8,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.0
[9,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.0
[10,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.0
[11,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.0
[12,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.0
[13,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.1

- It turns out that this matrix **provides us with the answer to our problem**:
- Any (normed) linear combination of the 13 standardized characteristics will also be a (normed) linear combination of the 13 uncorrelated eigenvalues;
- A little consideration shows that because of this, the variance of any normed linear combination can never exceed the maximum variance of the eigenvectors – **4.71**.
- The solution is thus **the first eigenvector**: $T[, 1]$.

Separation of Wines

- The combination of the scaled data that varies the most:

```
round(T[,1], digits=2)
[1] -0.14  0.25  0.00  0.24 -0.14 -0.39 -0.42  0.30 -0.31  0.09 -0.30
[12] -0.38 -0.29
```

- Thus the most varying combination of the scaled data is

$$-0.14 * \widetilde{\text{Alcohol}} + 0.25 * \widetilde{\text{MalicAcid}} + 0 * \widetilde{\text{Ash}} + 0.24 * \widetilde{\text{AlcAsh}} - 0.14 * \widetilde{\text{Mg}} - 0.39 * \widetilde{\text{Phenols}} - 0.42 * \widetilde{\text{PhenolsFlav}} \\ + 0.30 * \widetilde{\text{NonFlavPhenols}} - 0.31 * \widetilde{\text{Proa}} + 0.09 * \widetilde{\text{Color}} - 0.30 * \widetilde{\text{Hue}} - 0.38 * \widetilde{\text{OD}} - 0.29 * \widetilde{\text{Proline}}$$

Where the \sim versions are the scaled variables, with the mean subtracted and divided by the standard deviation.

- This is the scale that we want to look at, when we seek to maximize the separation of wines.

Separation of Wines

- The combination of the original data that varies the most:

```
round(T[,1]*sqrt(diag(var(wines))), digits=2)
[1] -0.12  0.27  0.00  0.80 -2.03 -0.25 -0.42  0.04 -0.18  0.21 -
0.07
[12] -0.27 -90.30
```

- Thus the most varying combination of the original data is

$$-0.12 * \text{Alcohol} + 0.27 * \text{MalicAcid} + 0 * \text{Ash} + 0.8 * \text{AlcAsh} - 2.03 * \text{Mg} - 0.25 * \text{Phenols} - 0.42 * \text{PhenolsFlav} + 0.04 * \text{NonFlavPhenols} - 0.18 * \text{Proa} + 0.21 * \text{Color} - 0.07 * \text{Hue} - 0.27 * \text{OD} - 90.30 * \text{Proline}$$

Separation of Wines

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
[1,]	4.71	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[2,]	0.00	2.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[3,]	0.00	0.0	1.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[4,]	0.00	0.0	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[5,]	0.00	0.0	0.00	0.00	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[6,]	0.00	0.0	0.00	0.00	0.00	0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.0
[7,]	0.00	0.0	0.00	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.00	0.0
[8,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.0
[9,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.0
[10,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.0
[11,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.0
[12,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.0
[13,]	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.1

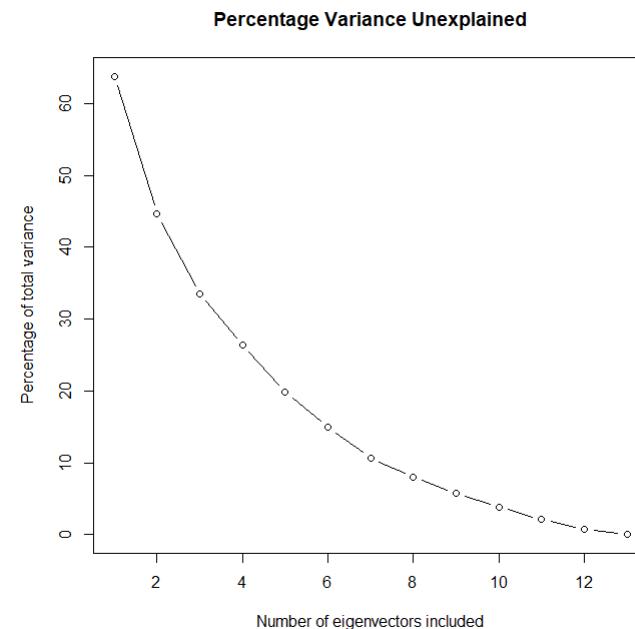
- **Further conclusions:**
- The scale, uncorrelated with the first eigenvector, that varies the most, is exactly the 2nd eigenvector $T[, 2]$, with variance **2.5**. And so on...
- The total variation after the coordinate shift is unchanged:

```
sum(diag(Lambda))  
[1] 13
```

- Also note that the contribution from the 13th eigenvector is only $0.1/13$, **0.7%**

Separation of Wines

```
plot(100*(13-cumsum(diag(Lambda)))/13,type="b",
  main="Percentage Variance Unexplained",
  xlab='Number of eigenvectors included',
  ylab='Percentage of total variance')
```



Eigenvectors	% variance explained
1	36
2	55
3	67
4	74
5	80
6	85
7	89
8	92
9	94
10	96
11	98
12	99
13	100

Principal Component Analysis

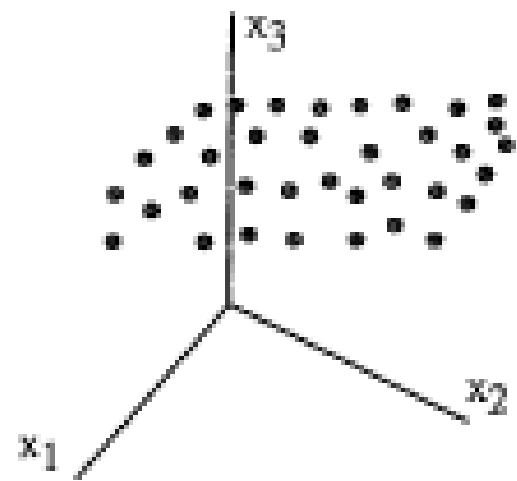
- PCA is a method to handle many variables, which are mutually correlated.
- PCA seeks to identify underlying dimensions in a data material, and to estimate the relationship between these.
- PCA may be used as a data reducing method, often in relation to multiple regression.
- PCA reduces the number of explanatory variables to a lesser number of "principal components", with (we hope) nearly as much of the variation as the initial variables.

What PCA does is essentially the contents of the preceding slides!

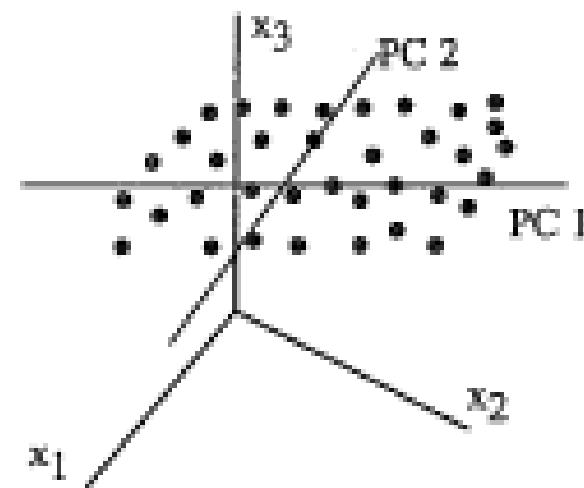
- The eigenvectors on the previous slides are **exactly** what is known as the **Principal Components**.

Identification of Lower- Dimensional Spaces

*Figure 3.10 Data swarm
(quasi-planar)*



*Figure 3.11 Data swarm with
2 PCs*



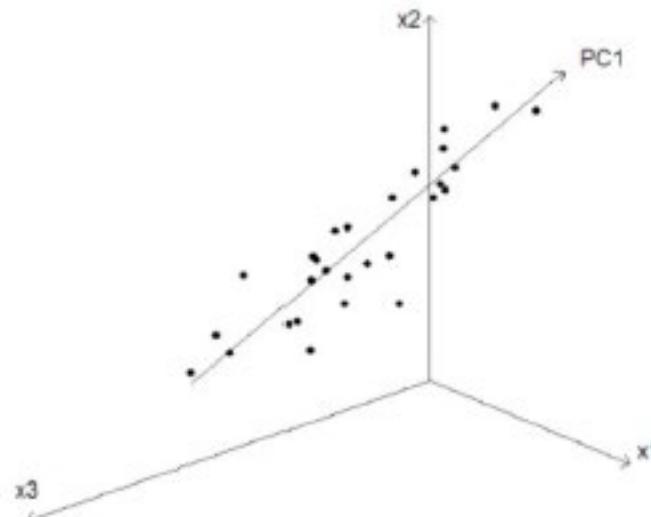
The First Principal Component – PC1

We look for the direction that **explains as much as possible of the variation in the data**. We assume here 3 variables and n=28:

$$p_1 = t_{11}X_1$$

$$\text{where } \sum_{j=1}^3 t_{1j}^2 = 1.$$

- p are the "scores";
- t are the "loadings".



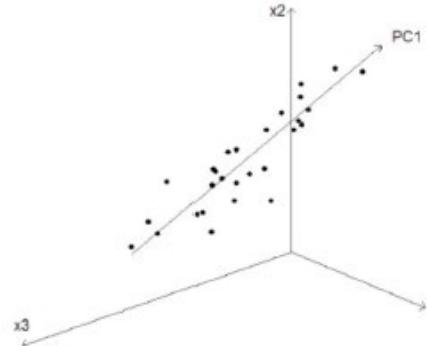
The First Principal Component – PC1

We look for the direction that **explains as much as possible of the variation in the data**. We assume here 3 variables and n=28:

$$p_1 = t_{11}X_1 + t_{21}X_2 + t_{31}X_3,$$

where $\sum_{j=1}^3 t_{1j}^2 = 1$.

- p are the "scores";
- t are the "loadings".



- Important distinction from the book:
 - In the book, **the scores ARE the principal components**, not the direction!
 - The turning (coordinate change) matrix T is labeled P , and the principal components are labeled Y :

$$Y = P^T X$$

Here:

$$Scores = T^T X$$

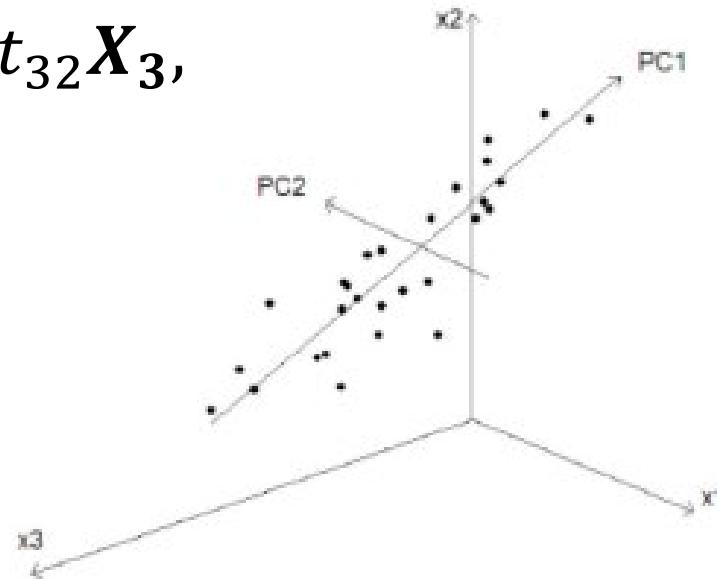
The Second Principal Component – PC2

We consider the plane perpendicular to **PC1**, and find the linear combination that explains the 2nd most variation:

$$p_2 = t_{12}X_1 + t_{22}X_2 + t_{32}X_3,$$

$$\text{where } \sum_{j=1}^3 t_{2j}^2 = 1$$

PC1 and PC2 are orthogonal.



A few theorems:

||| Theorem 6.3

The principal components are uncorrelated and the variance of the i 'th component is λ_i i.e. the i 'th largest eigenvalue.

||| Theorem 6.5

The total variance i.e. the sum of variance of the original variables is equal to the sum of the variance of the principal components i.e.

$$\sum_i V(X_i) = \sum_i V(Y_i)$$

Important note

- Theorems on the previous slide only works **if the variance matrix Σ is essentially known**, ie. if the Principal Components are known.
- Very often, the variance matrix is not known and will be estimated from the data (like the wines data);

**In that case, the ‘theorems’ one the previous slide only holds approximately,
and only for large samples!**

- Reason: When estimated from data, the principal components in terms of directions are themselves stochastic variables; converting observations to scores is done by **multiplying with a RANDOM matrix**, not a fixed matrix.
- If the uncertainty of the the Principal Components may be neglected (large samples), one may assume the theorems to hold even for estimated variances.

An observation

- If X is multivariate normal with unknown variance, are the scores then multivariate normal?

NO!

Reason as on the previous slide.

- However, such an assumption is often not needed: Often PCA is used as an explorative data technique that bypasses statistical modeling (for a start).

PCA of the Wines Data

```
wines.PC<- PCA(scale(wines))
names(wines.PC)
[1] "scores"      "loadings"     "var"          "totalvar"
[5] "centered.data"

summary(wines.PC)

PCA model of a mean-centered matrix of 178 by 13
Number of PCs to cover 90 percent of the variance: 8

      Var Cumul. var.
PC 1 36.198848   36.19885
PC 2 19.207490   55.40634
PC 3 11.123631   66.52997
PC 4  7.069030   73.59900
PC 5  6.563294   80.16229
PC 10 1.930019   96.16972
```

- Let's take a look at the components

Remember this:

Eigenvectors	% variance explained
1	36
2	55
3	67
4	74
5	80
6	85
7	89
8	92
9	94
10	96
11	98
12	99
13	100

PCA of the Wines Data – the Loadings

- The **loadings** are the *coordinates of the principal components*:

```
> head(wines$loadings, n=3)
   PC 1      PC 2      PC 3      PC 4      PC 5      PC 6
Alcohol -0.144329395 -0.4836515 -0.20738262 -0.0178563 0.26566365 -0.2135386
MalicAcid 0.245187580 -0.2249309 0.08901289 0.5368903 -0.03521363 0.5368138
Ash     0.002051061 -0.3160688 0.62622390 -0.2141756 0.14302547 -0.1544747
                    PC 7      PC 8      PC 9      PC 10     PC 11
Alcohol -0.05639636 -0.39613926 -0.50861912 -0.21160473 0.22591696
MalicAcid 0.42052391 -0.06582674 0.07528304 0.30907994 -0.07648554
Ash     -0.14917061 0.17026002 0.30769445 0.02712539 0.49869142
                    PC 12     PC 13
Alcohol 0.26628645 -0.01496997
MalicAcid -0.12169604 -0.02596375
Ash     0.04962237 0.14121803
> head(T, n=3)
     [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
[1,] -0.144329395 -0.4836515 0.20738262 -0.0178563 0.26566365 0.2135386
[2,]  0.245187580 -0.2249309 0.08901289 0.5368903 -0.03521363 0.5368138
[3,]  0.002051061 -0.3160688 0.62622390 -0.2141756 0.14302547 0.1544747
     [,7]     [,8]     [,9]     [,10]    [,11]    [,12]
[1,]  0.05639636  0.39613926 0.50861912 0.21160473 -0.22591696 -0.26628645
[2,] -0.42052391  0.06582674 -0.07528304 -0.30907994  0.07648554  0.12169604
[3,]  0.14917061 -0.17026002 -0.30769445 -0.02712539 -0.49869142 -0.04962237
     [,13]
[1,]  0.01496997
[2,]  0.02596375
[3,] -0.14121803
>
```

- The principal components are only identified up to a sign change.

PCA of the Wines Data – the Scores

- The **scores** are the *new coordinates* of the (scaled) wines data *relative to the principal components*:

```
> head(wines$PC$scores, n=3)
      PC 1       PC 2       PC 3       PC 4       PC 5       PC 6
[1,] -3.307421 -1.4394023 -0.1652728 -0.2150246 -0.6910933 -0.2232504
[2,] -2.203250  0.3324551 -2.0207571 -0.2905387  0.2569299 -0.9245123
[3,] -2.509661 -1.0282507  0.9800541  0.7228632  0.2503270  0.5477310
      PC 7       PC 8       PC 9       PC 10      PC 11      PC 12
[1,]  0.59474883  0.06495586 -0.6396384 -1.0180840  0.4502932 -0.5392891439
[2,]  0.05362434  1.02153432  0.3079780 -0.1592521  0.1422560 -0.3871456499
[3,]  0.42301218 -0.34324787  1.1745213 -0.1130420  0.2858665 -0.0005819316
      PC 13
[1,]  0.066052305
[2,] -0.003626273
[3,] -0.021655423

> head(scale(wines) %*% T, n=3)
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] -3.307421 -1.4394023 -0.1652728 -0.2150246 -0.6910933  0.2232504
[2,] -2.203250  0.3324551 -2.0207571 -0.2905387  0.2569299  0.9245123
[3,] -2.509661 -1.0282507  0.9800541  0.7228632  0.2503270 -0.5477310
      [,7]      [,8]      [,9]      [,10]     [,11]     [,12]
[1,] -0.59474883 -0.06495586  0.6396384  1.0180840 -0.4502932  0.5392891439
[2,] -0.05362434 -1.02153432 -0.3079780  0.1592521 -0.1422560  0.3871456499
[3,] -0.42301218  0.34324787 -1.1745213  0.1130420 -0.2858665  0.0005819316
      [,13]
[1,] -0.066052305
[2,]  0.003626273
[3,]  0.021655423
>
```

- Note the same sign changes

PCA of the Wines data

var, totalvar and centered.data

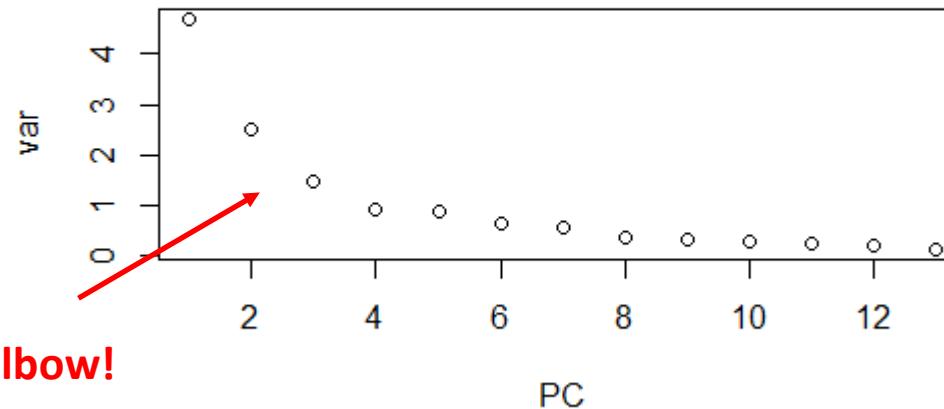
```
wines.PC$var  
PC 1      PC 2      PC 3      PC 4      PC 5      PC 6      PC 7      PC 8  
4.7058503 2.4969737 1.4460720 0.9189739 0.8532282 0.6416570 0.5510283 0.3484974  
PC 9      PC 10     PC 11     PC 12     PC 13  
0.2888799 0.2509025 0.2257886 0.1687702 0.1033779
```

```
wines.PC$totalvar  
[1] 13  
wines.PC$centered.data  
[1] TRUE
```

We recognize the eigenvalues of the matrix Λ , and the sum of these. Lastly, an indicator that we have ‘done the right thing’ (in this case).

Selecting the Number of Principal Components – the Skree Plot

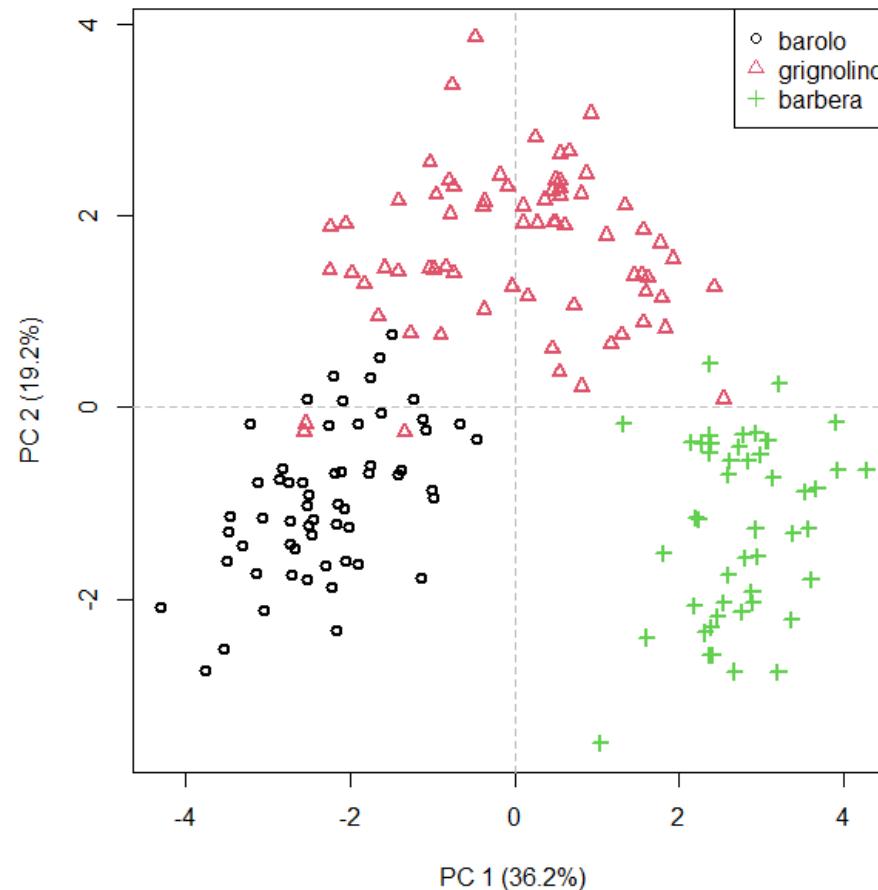
```
plot(1:13, wines.PC$var, xlab="PC", ylab="var")
```



- Rules of thumb:
 - i. You select a number of principal components where the 'elbow' of the graph is.
 - ii. You usually (but not always) only select principal components with a variance greater than 1 – if the value is lower, the PC explain less than one average ordinary observation.
 - iii. You fix the amount of total variation that you need explained – t.ex. 80%.

The Score plot

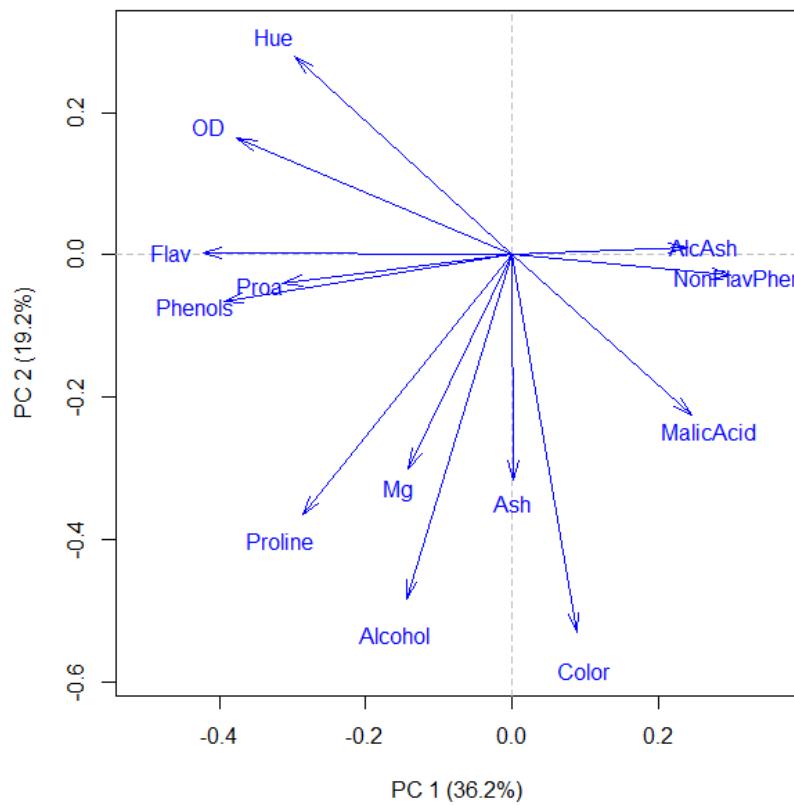
```
scoreplot(wines.PC, col = vintages, pch= as.numeric(vintages), lwd=2)  
legend("topright",levels(vintages), col=1:3,pch=1:3)
```



Apparently, we can
more or less separate
the vintages from just
TWO dimensions!

The Loading plot

```
loadingplot(wines.PC, show.names= TRUE)
```



The **Loading plot** shows the points

$$\begin{bmatrix} \text{Corr}(X_i, PC1) \\ \text{Corr}(X_i, PC2) \end{bmatrix}$$

For the 13 variables.

Some considerations (not trivial):

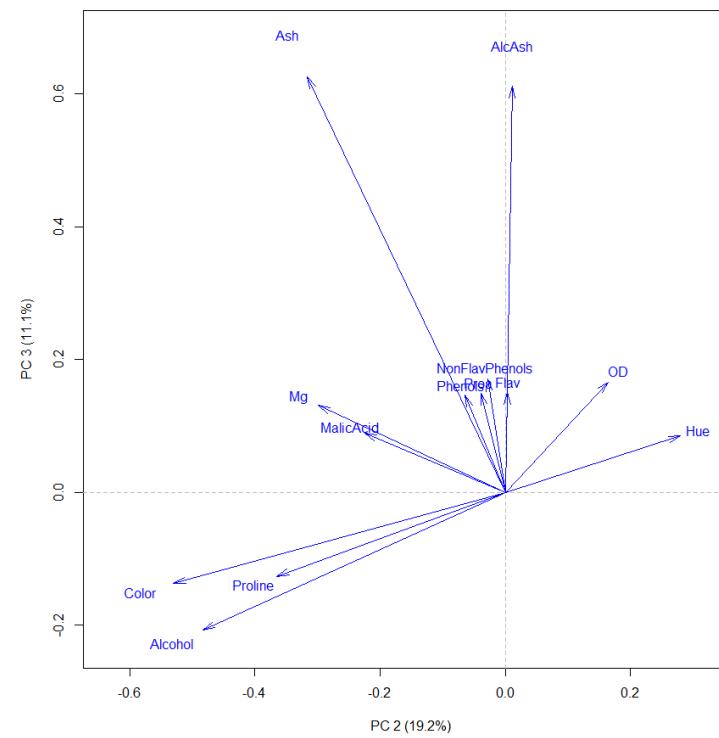
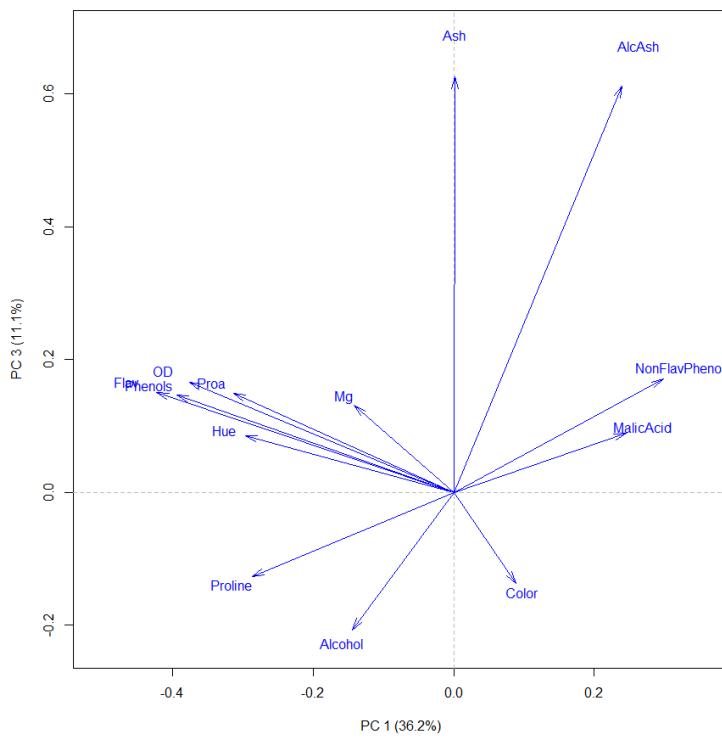
Barolo: Sweet taste, alcoholic, low nutritional value, saturated color.

Grignolino: Sweet taste, low nutritional value, bright color.

Barbera: Bitter taste, high nutritional value, deep color.

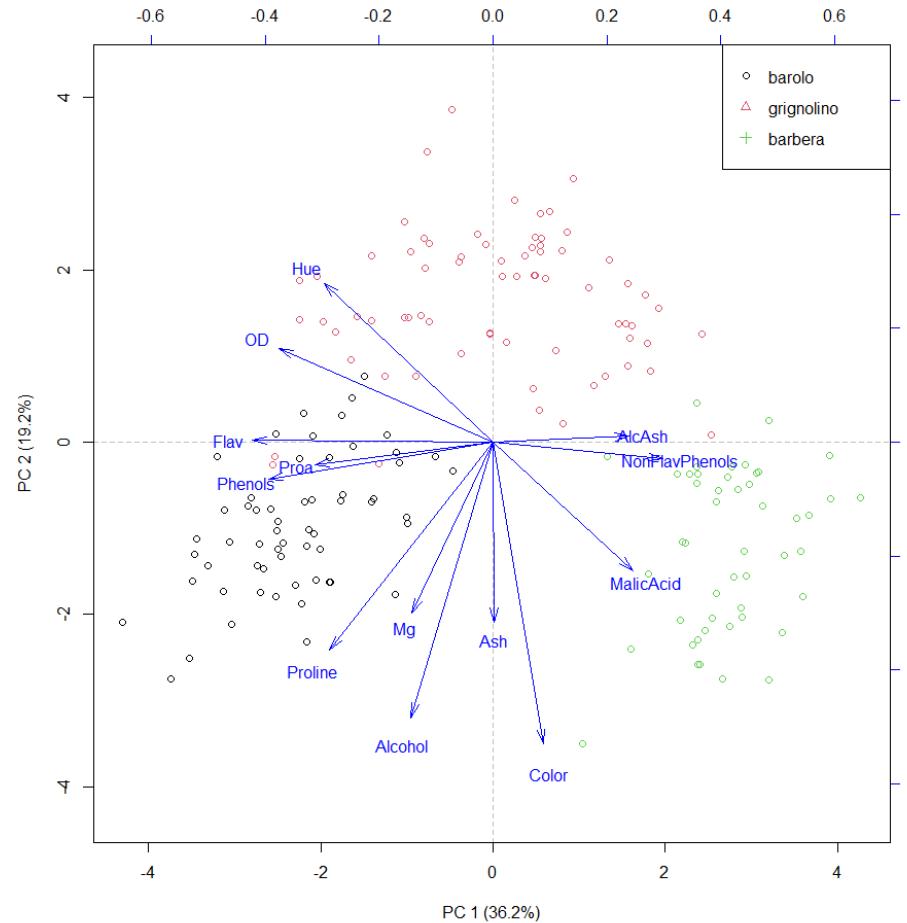
Higher Order Loading Plots

```
par(mfrow=c(1,2))
loadingplot(wines.PC, pc=c(1,3), show.names= TRUE)
loadingplot(wines.PC, pc=c(2,3), show.names= TRUE)
```



The Biplot

```
biplot(wines.PC, score.col = vintages, show.names = "loadings")
legend("bottomright", levels(vintages), col=1:3, pch=1:3)
```

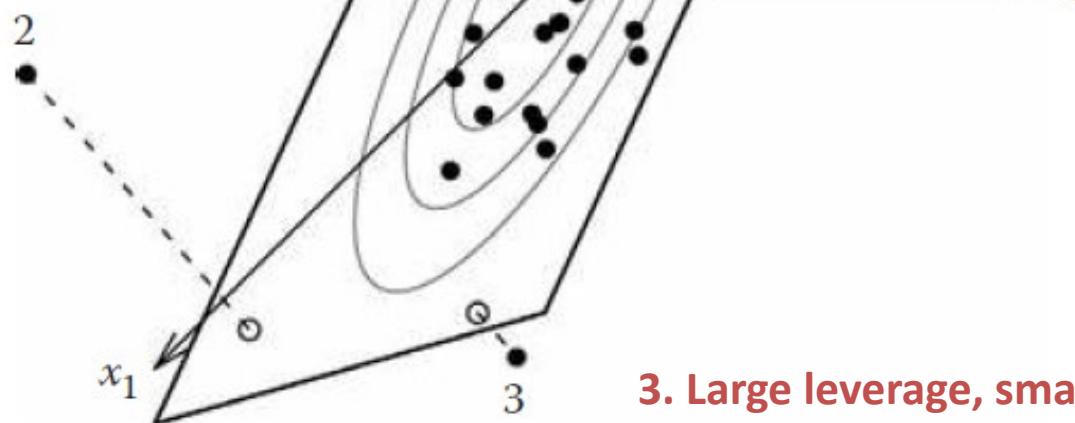


Diagnostic Plots – Residuals and Leverage

1. Large residual, far from the PCA space but central



2. Large residual, far from the PCA space, large leverage.

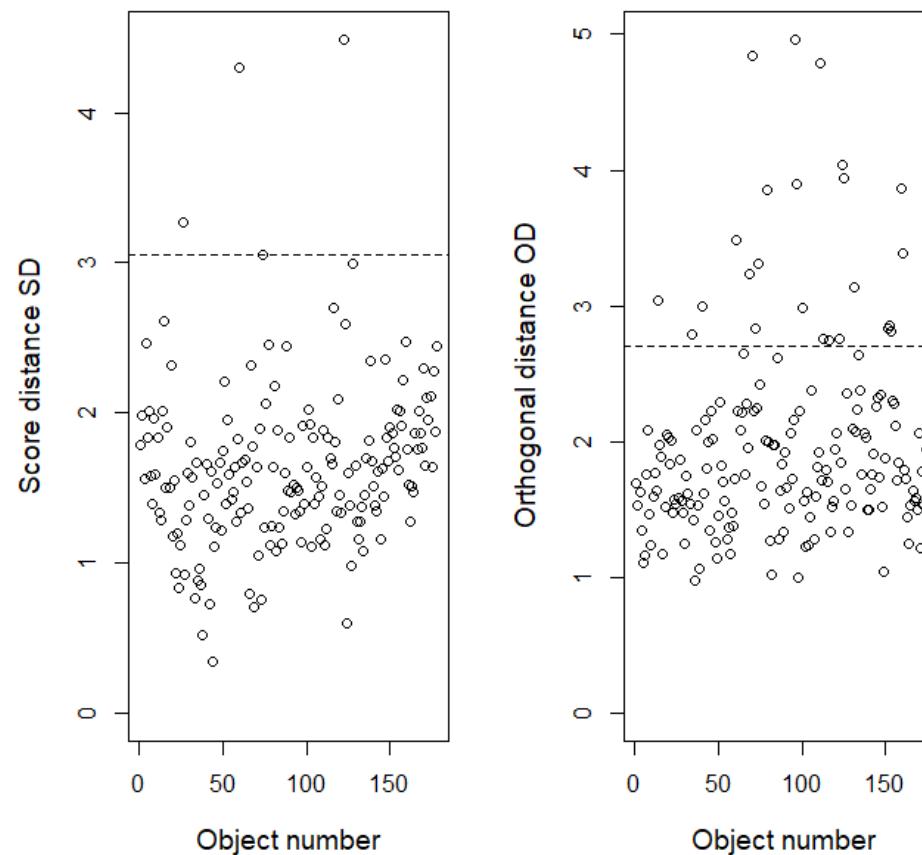


3. Large leverage, small residual

Diagnostic Plots

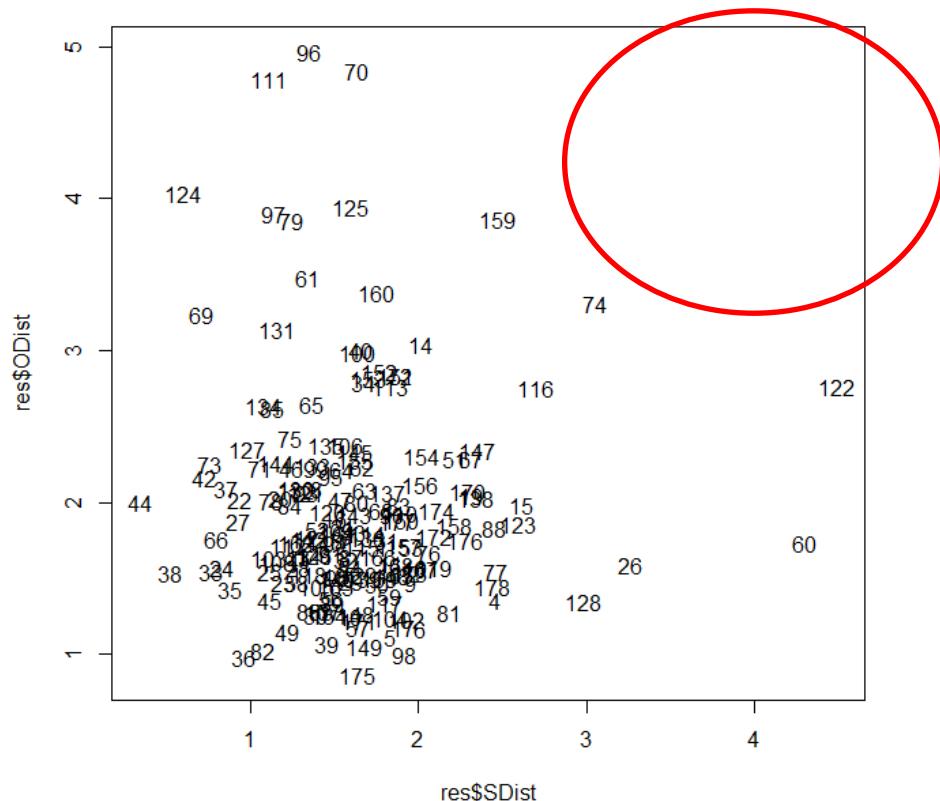
Plotting distances for 3 principal components, leverage (left) and residual (right):

```
wines.PCA<- princomp(wines, cor = TRUE)  
res<-pcaDiagplot(wines, wines.PCA, a=3)
```



Diagnostic Plots

```
par(mfrow=c(1, 1))
plot(res$SDist, res$ODist, type="n")
text(res$SDist, res$ODist, labels=as.character(1:178))
```



No points have high leverage and also high residual

Complex data

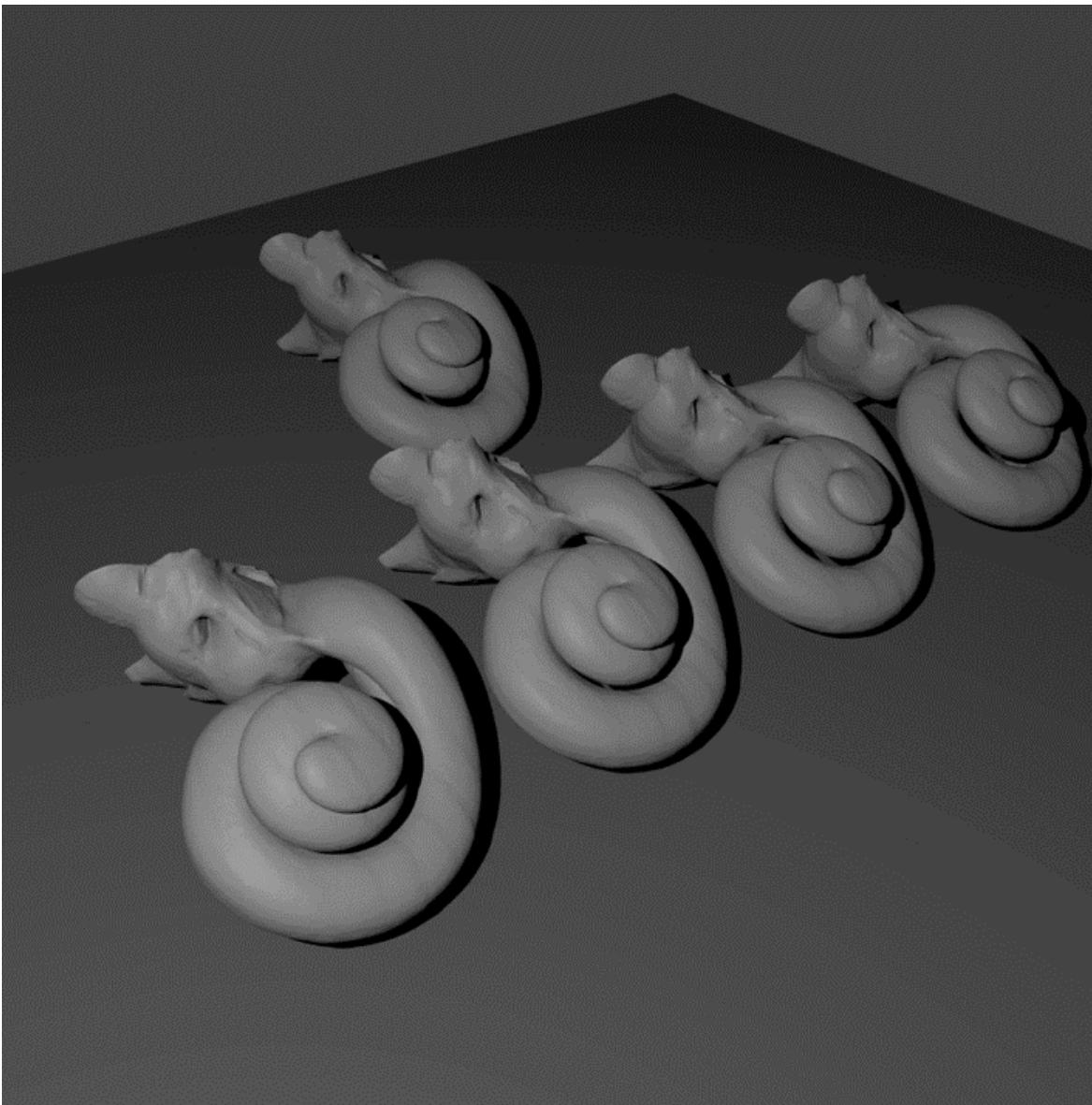
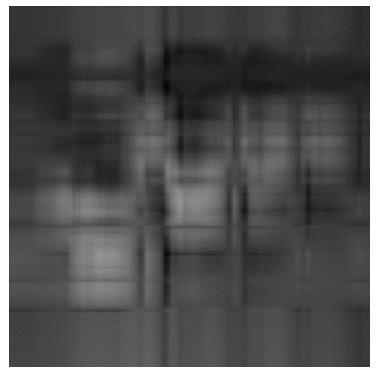


Image reconstruction from principal components



3 PCs



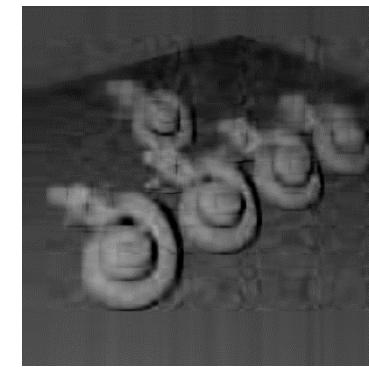
6 PCs



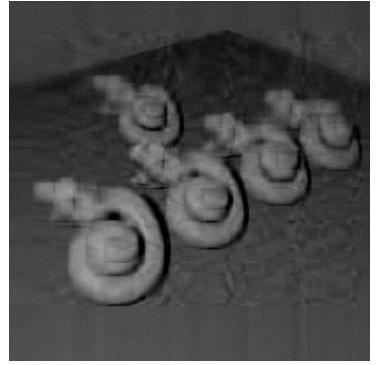
9 PCs



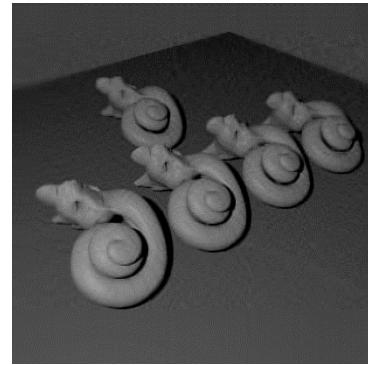
12 PCs



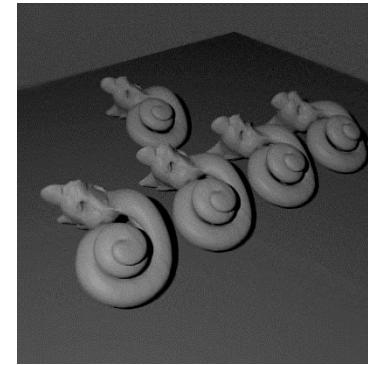
15 PCs



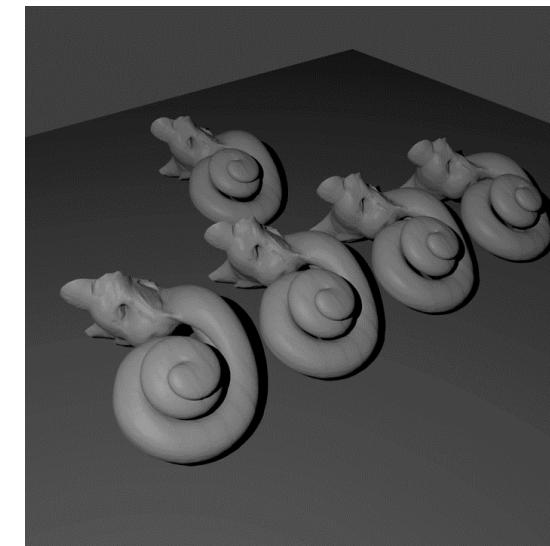
18 PCs



50 PCs



100 PCs

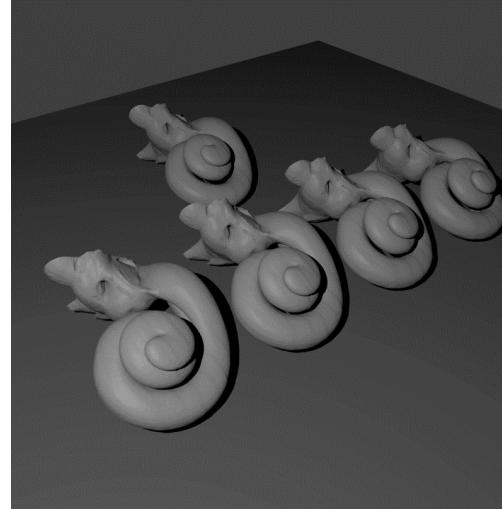


Original: 719x719 pixels

HOW?

The image is **RGBA** coded:

- **Red**
- **Green**
- **Blue**
- Alpha - opaqueness



Red, **green** and **blue** is all the colors that the human eye can perceive -
Thus each of the 719x719 pixels is represented by a 4-dimensional vector – 4 layers.

Reconstruct the information in each layer from 3, 6, 9, 12, 15, 18, 50 and 100 principal components.

Pixel columns within layers are variables, rows observations.

- 719 principal components capture all information
- But the data is **structured**/essentially contained in a subspace, so much less will extract nearly all the information. More on this later.

Example 6.9: Boxes with randomly generated sides

X_1 : longest side

X_2 : second longest side

X_3 : smallest side

X_4 : longest diagonal

X_5 : radius in the circumscribed sphere divided by radius in the inscribed sphere

X_6 : longest side + second longest side)/shortest side

X_7 : surface area/volume.

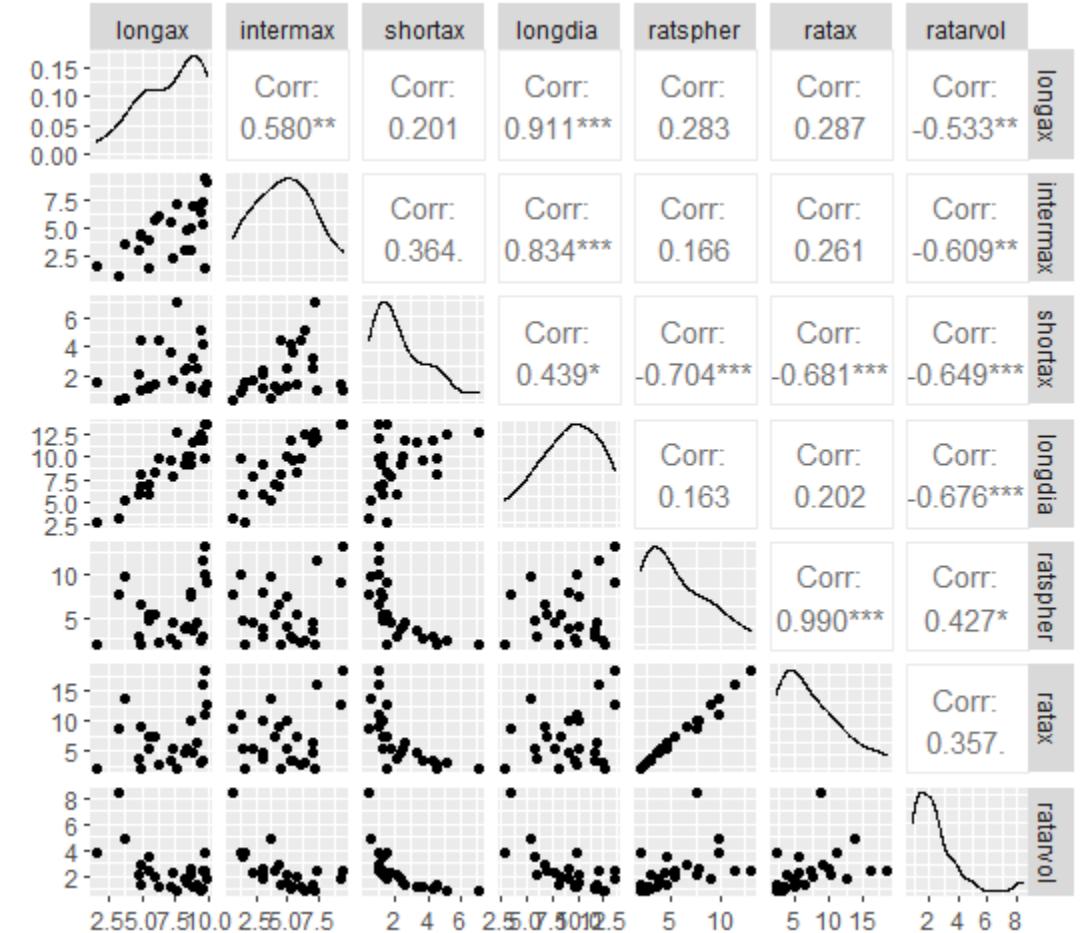
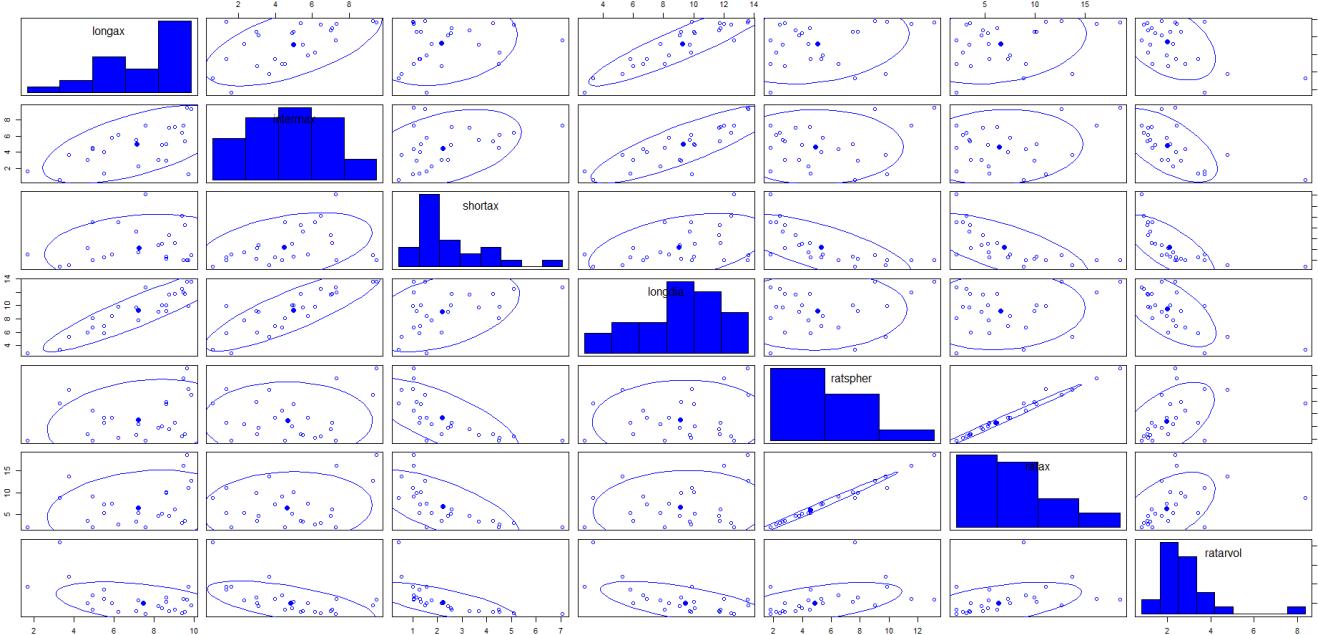
Box	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	3.760	3.660	0.540	5.275	9.768	13.741	4.782
2	8.590	4.990	1.340	10.022	7.500	10.162	2.130
:	:	:	:	:	:	:	:
24	8.210	3.080	2.420	9.097	3.753	4.657	1.719
25	9.410	6.440	5.110	12.495	2.446	3.103	0.914

From J.C. Davis. *Statistics and data analysis in geology*. John Wiley, New York 1973

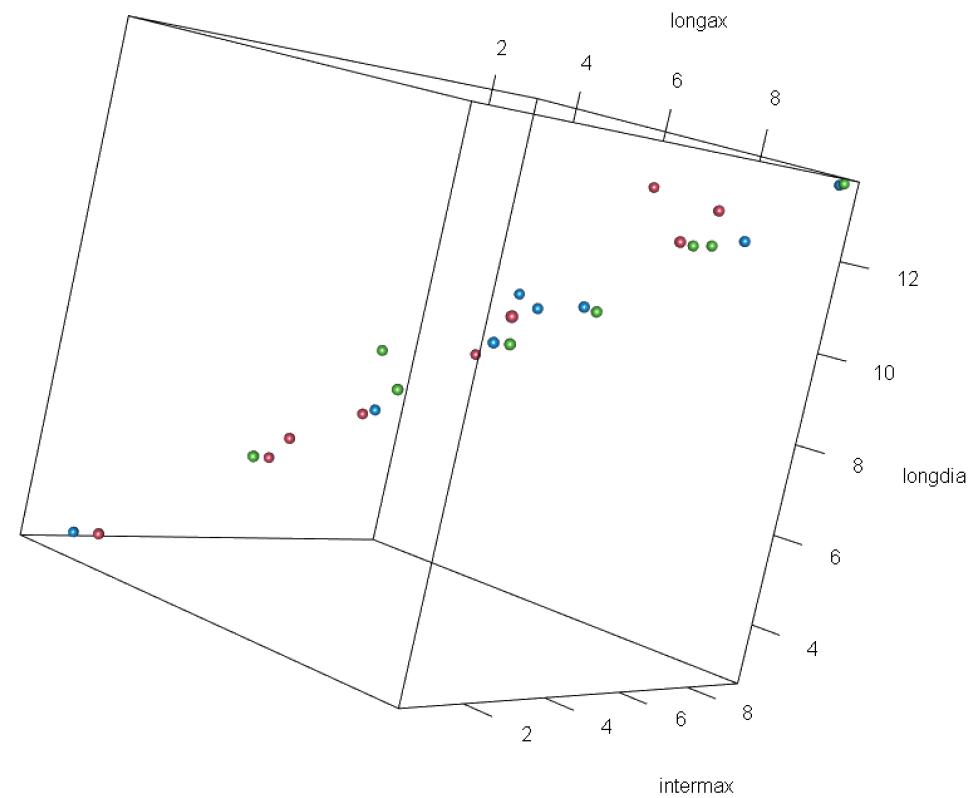
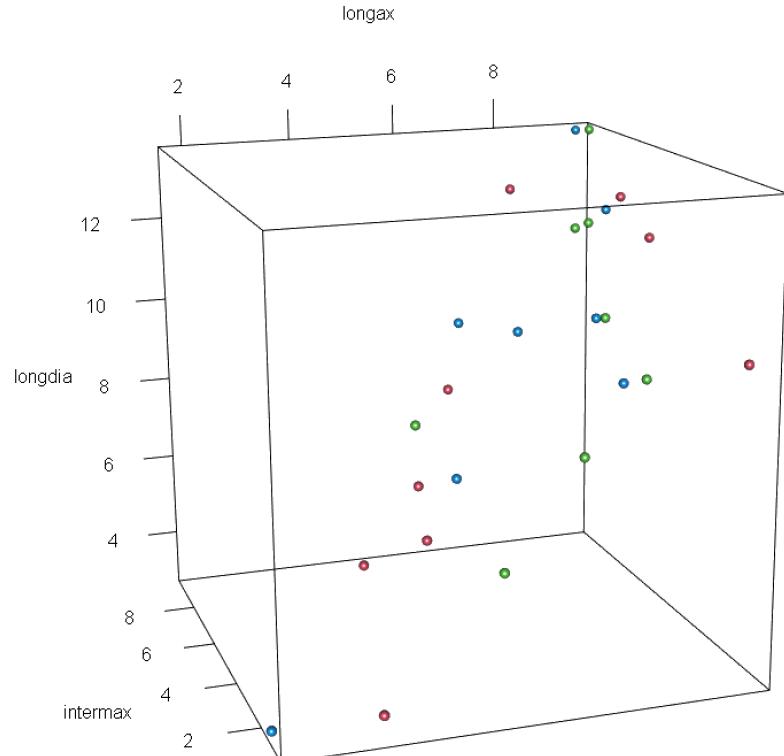
Pairwise distribution of the 7 variables

```
library(car)
scatterplotMatrix(BoxData, regLine=F, smooth=F,
  ellipse=list(levels=0.9, robust=TRUE, fill=F),
  diagonal=list(method="histogram", breaks="FD"))

library(GGally)
ggpairs(BoxData)
```



Distribution of longax, intermax, and longdia



Seen from another angle

Estimated variance and correlation matrices

Boxes with randomly generated sides

```
> round(var(BoxData), digits=2)
      longax intermax shortax longdia ratspher ratax ratarvol
longax    5.40     3.26    0.78    6.39     2.16   3.04    -2.00
intermax  3.26     5.85    1.46    6.08     1.31   2.88    -2.37
shortax   0.78     1.46    2.77    2.20    -3.84  -5.17    -1.74
longdia   6.39     6.08    2.20    9.11     1.61   2.78    -3.28
ratspher  2.16     1.31   -3.84    1.61    10.71  14.77    2.25
ratax    3.04     2.88   -5.17    2.78    14.77  20.78    2.62
ratarvol -2.00    -2.37   -1.74   -3.28     2.25   2.62    2.59
```

```
> round(cor(BoxData), digits=2)
      longax intermax shortax longdia ratspher ratax ratarvol
longax   1.00     0.58    0.20    0.91     0.28   0.29    -0.53
intermax  0.58     1.00    0.36    0.83     0.17   0.26    -0.61
shortax   0.20     0.36    1.00    0.44    -0.70  -0.68    -0.65
longdia   0.91     0.83    0.44    1.00     0.16   0.20    -0.68
ratspher  0.28     0.17   -0.70    0.16     1.00   0.99    0.43
ratax    0.29     0.26   -0.68    0.20     0.99   1.00    0.36
ratarvol -0.53    -0.61   -0.65   -0.68     0.43   0.36    1.00
```

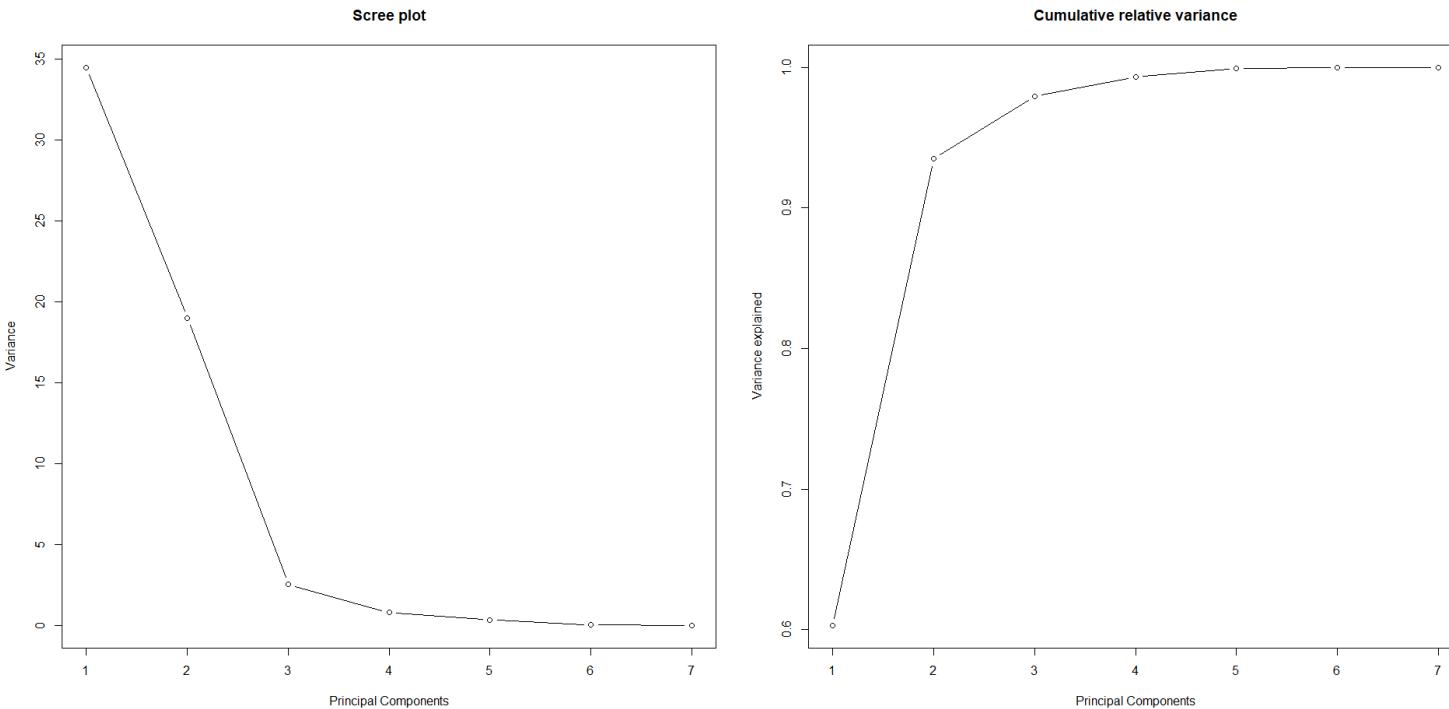
Eigenvalues and eigenvectors of dispersion matrix

Eigenvalues of the Corelation Matrix			
	Eigenvalue	Proportion	Cumulated
1	34.49	0.60	0.60
2	19.00	0.33	0.93
3	2.54	0.04	0.98
4	0.81	0.01	0.99
5	0.34	0.006	0.999
6	0.03	0.0006	0.99996
7	0.003	0.00004	1.0000

X_1 : longest side
 X_2 : second longest side
 X_3 : smallest side
 X_4 : longest diagonal
 X_5 : radius in circumscribed sphere divided by radius in inscribed sphere
 X_6 : (longest side + second longest side)/shortest side
 X_7 : surface area/volume

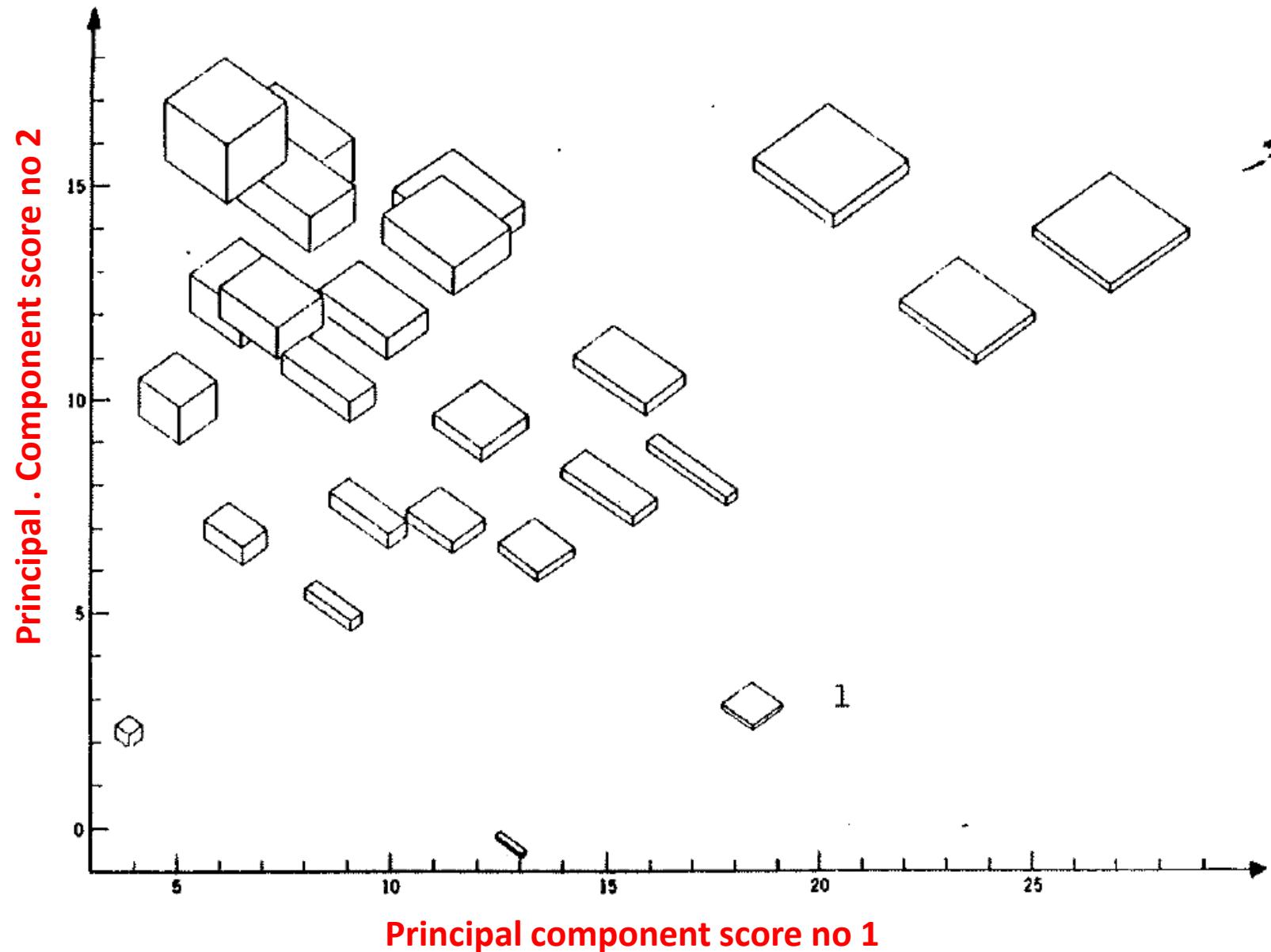
	Eigenvectors						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
longax	0.164247	0.422023	-0.644914	0.090057	0.225052	-0.414957	0.384782
intermax	0.142092	0.447420	0.712955	0.049719	0.394949	-0.066404	0.329324
shortax	-0.172881	0.256798	0.130098	-0.629178	-0.607154	-0.279860	0.210758
longdia	0.169853	0.649513	-0.146285	-0.212135	0.032942	0.402806	-0.565282
ratspher	0.546410	-0.134786	-0.105316	-0.164491	-0.161154	0.596078	0.513655
ratax	0.768262	-0.133249	0.148715	0.062070	-0.206930	-0.464846	-0.327335
ratarvol	0.072837	-0.313012	-0.065187	-0.719473	0.595697	-0.106941	-0.092389

The Scree and Variance Explained Plots



The **Scree and Variance Explained Plots** show the eigenvalues $\hat{\lambda}_i$, and the cumulative proportion of variance explained $\frac{\hat{\lambda}_1 + \dots + \hat{\lambda}_m}{\hat{\lambda}_1 + \dots + \hat{\lambda}_m + \hat{\lambda}_{m+1} + \dots + \hat{\lambda}_k}$

The Component Scores plot with boxes superimposed



Reconstructing the original observations from the first m principal components

Remark 6.7: Since $\mathbf{X} = \mathbf{P}\mathbf{Y}$ with the notation in the book, an obvious way of **reconstructing \mathbf{X} from the first m principal components $\mathbf{Y}^{(m)}$** is

$$\mathbf{X}^* = \mathbf{P}^{(m)}\mathbf{Y}^{(m)} = [\mathbf{p}_1, \dots, \mathbf{p}_m] \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix} = Y_1\mathbf{p}_1 + \dots + Y_m\mathbf{p}_m$$

The variance matrix of the reconstructed vector is

$$\begin{aligned} \mathbf{D}(\mathbf{X}^*) &= \mathbf{P} \mathbf{D}(\mathbf{Y}^*) \mathbf{P}^T \\ &= \lambda_1 \mathbf{p}_1 \mathbf{p}_1^T + \dots + \lambda_m \mathbf{p}_m \mathbf{p}_m^T. \end{aligned}$$

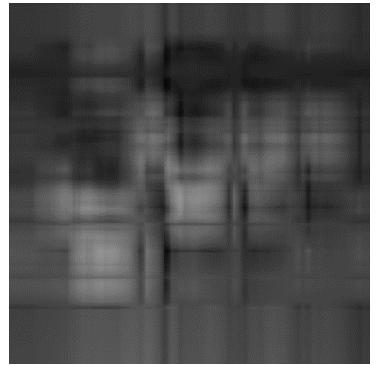
The spectral decomposition of Σ is (p. 458)

$$\Sigma = \lambda_1 \mathbf{p}_1 \mathbf{p}_1^T + \dots + \lambda_m \mathbf{p}_m \mathbf{p}_m^T + \lambda_{m+1} \mathbf{p}_{m+1} \mathbf{p}_{m+1}^T + \dots + \lambda_k \mathbf{p}_k \mathbf{p}_k^T,$$

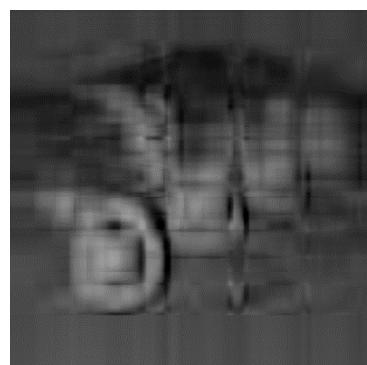
which means that

$$\Sigma - \mathbf{D}(\mathbf{X}^*) = \lambda_{m+1} \mathbf{p}_{m+1} \mathbf{p}_{m+1}^T + \dots + \lambda_k \mathbf{p}_k \mathbf{p}_k^T.$$

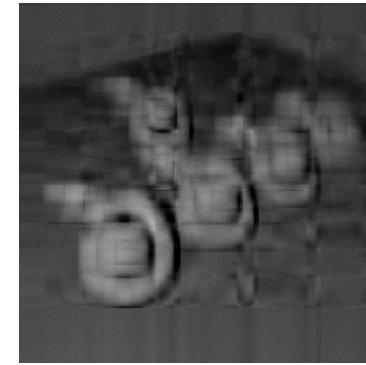
Image reconstruction from principal components – code in script.



3 PCs



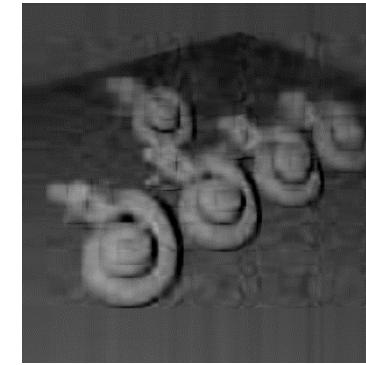
6 PCs



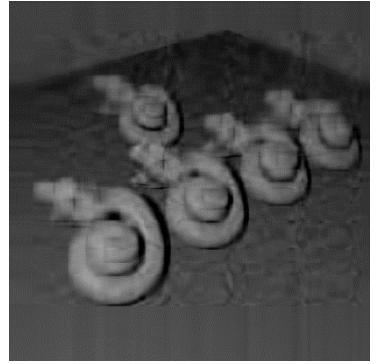
9 PCs



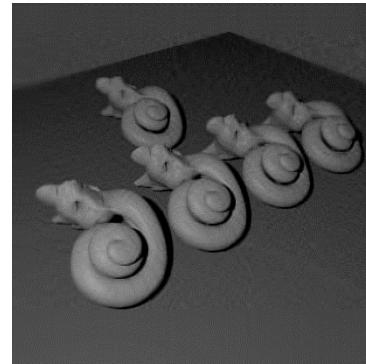
12 PCs



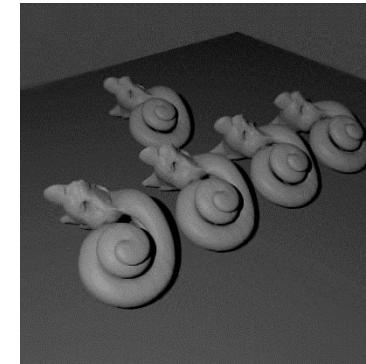
15 PCs



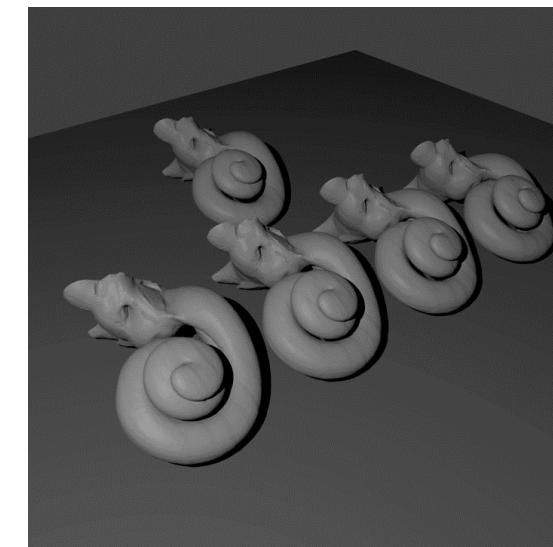
18 PCs



50 PCs



100 PCs



Original: 719x719 pixels

Test for equality of k-m smallest eigenvalues in the variance matrix

$$H_0 : \lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_k$$

$$Z_1 = -n^* \log \frac{\det \hat{\Sigma}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n^* \log \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}_*^{k-m}},$$

where

$$n^* = n - m - \frac{1}{6}(2(k - m) + 1 + \frac{2}{k - m}),$$

and

$$\hat{\lambda}_* = (\text{tr } \hat{\Sigma} - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m) / (k - m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k) / (k - m).$$

The critical region using a test at level α is approximately

$$\{(x_1, \dots, x_n) | z_1 > \chi^2(\frac{1}{2}(k - m + 2)(k - m - 1))_{1-\alpha}\}.$$

Test for equality of k-m smallest eigenvalues in the correlation matrix

$$H_0 : \lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_k$$

If we instead are using the estimated *correlation matrix* $\hat{\mathbf{R}}$ we get the criterion

$$Z_2 = -n \log \frac{\det \hat{\mathbf{R}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n \log \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}_*^{k-m}},$$

where

$$\hat{\lambda}_* = (k - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m) / (k - m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k) / (k - m).$$

The critical region for a test at level α becomes approximately equal to

$$\{x_1, \dots, x_n | z_2 > \chi^2(\frac{1}{2}(k - m + 2)(k - m - 1))_{1-\alpha}\}.$$

Test for equality of k-m smallest eigenvalues in BoxData correlation matrix

Eigenvalues of the Correlation Matrix			
	Eigenvalue	Proportion	Cumulated
1	3.39	0.48	0.48
2	2.81	0.40	0.89
3	0.44	0.06	0.95
4	0.28	0.04	0.99
5	0.08	0.01	0.999
6	0.003	0.0005	0.99996
7	0.0003	0.00004	1

Can we assume that the last two values are the same (close to 0)?

$$\prod_{i=6}^7 \hat{\lambda}_i = 8.942424e-07, \quad \hat{\lambda}_* = \frac{(\hat{\lambda}_6 + \hat{\lambda}_7)}{2} = 0.001834764, \quad n = 25, \quad -n * \log\left(\frac{\prod_{i=6}^7 \hat{\lambda}_i}{\hat{\lambda}_*^2}\right) = 33.14026$$

$$P(\chi^2(2) > 33.14) = 0.0000006$$

Conclusion: The statistical test shows that data do not support that the 2 smallest eigenvalues may be assumed to be equal.

Exercises

- Exercise 7.1
 - Principal Component Analysis on Olympic data
 - Interpretation, number of components
- Exercise 7.2
 - Principal Component Analysis on the 'Fitness' dataset
 - Interpretation, number of components, conditioning on variables
- Exercise 7.3
 - Principal Component Analysis on Beef characterization
 - Interpretation of PCA
- Exercise 7.4
 - PCA by hand and interpretation