Multivariate Statistics

Dr. Alan Dabney
Associate Professor
Department of Statistics
Texas A&M University

Topic One:
Introductory Material

What does "multivariate" mean?

- **Univariate**: Record single variable for each individual.
  - Examples:
    - Survival time after breast tumor removal and chemo.
    - Proportion of devices with error in assembly line.
    - SAT exam score after participating in tutoring program.
- **Multivariate**: Record $p \geq 1$ variables for each individual.
  - Examples:
    - Total blood count, LDL, and HDL after blood test.
    - Trading volume, % change, and maximum price of a stock.
    - Hardness, pressure resistance, electrical conductivity, and color of a gem.

Common applications

- Dimension reduction.
- Formal inference.
- Correlation.
- Discriminant analysis.
- Clustering.

Dimension reduction

- **Principal components**:
  - Explain most of the total variability among the p variables using $k < p$ linear combinations of the original variables.
- **Factor analysis**:
  - Construct random quantities (*factors*) that describe variable groupings such that within-group correlations are maximized and between-group correlations are minimized.

Dimension reduction

- **Principal components**:
  - Explain most of the total variability among the p variables using $k < p$ linear combinations of the original variables.
- **Factor analysis**:
  - Construct random quantities (*factors*) that describe variable groupings such that within-group correlations are maximized and between-group correlations are minimized.

Example: Census data

- Tract-level data on five socioeconomic variables in the Madison, WI area.
- Variables:
  - Total population (thousands).
  - Percent of population with professional degree.
  - Percent aged 16 or over who are employed.
  - Percent employed by government.
  - Median home value (in $100,000s).

Data from Table 8.5 in textbook.

Example: Census data

Variable	PC_1	PC_2	PC_3	PC_4	PC_5
Population			-0.188	0.977	
Professional	-0.105	-0.130	0.961	0.171	-0.139
Employment	0.492	-0.864			
Government	-0.863	-0.480	-0.153		
Home			0.125		0.989
Cum. Var.	67.70	92.79	98.09	99.90	1.000

Example: Census data

Variable	PC_1	PC_2	PC_3	PC_4	PC_5
Population			-0.188	0.977	
Professional	-0.105	-0.130	0.961	0.171	-0.139
Employment	0.492	-0.864			
Government	-0.863	-0.480	-0.153		
Home			0.125		0.989
Cum. Var.	67.70	92.79	98.09	99.90	1.000

Empty cells are just small numbers.

Example: Census data

Variable	PC_1	PC_2	PC_3	PC_4	PC_5
Population			-0.188	0.977	
Professional	-0.105	-0.130	0.961	0.171	-0.139
Employment	0.492	-0.864			
Government	-0.863	-0.480	-0.153		
Home			0.125		0.989
Cum. Var.	67.70	92.79	98.09	99.90	1.000

Roughly, linear combination:

$$0.5 \times \text{Employment} - 0.9 \times \text{Government}$$

explains 70% of variance from all five variables. This linear combination is a weighted difference between the two employment variables.

Example: Census data

Variable	PC_1	PC_2	PC_3	PC_4	PC_5
Population			-0.188	0.977	
Professional	-0.105	-0.130	0.961	0.171	-0.139
Employment	0.492	-0.864			
Government	-0.863	-0.480	-0.153		
Home			0.125		0.989
Cum. Var.	67.70	92.79	98.09	99.90	1.000

Roughly, linear combination:

$$-0.9 \times \text{Employment} - 0.5 \times \text{Government}$$

explains an *additional* 25% of variance. This linear combination is a weighted sum of the two employment variables.

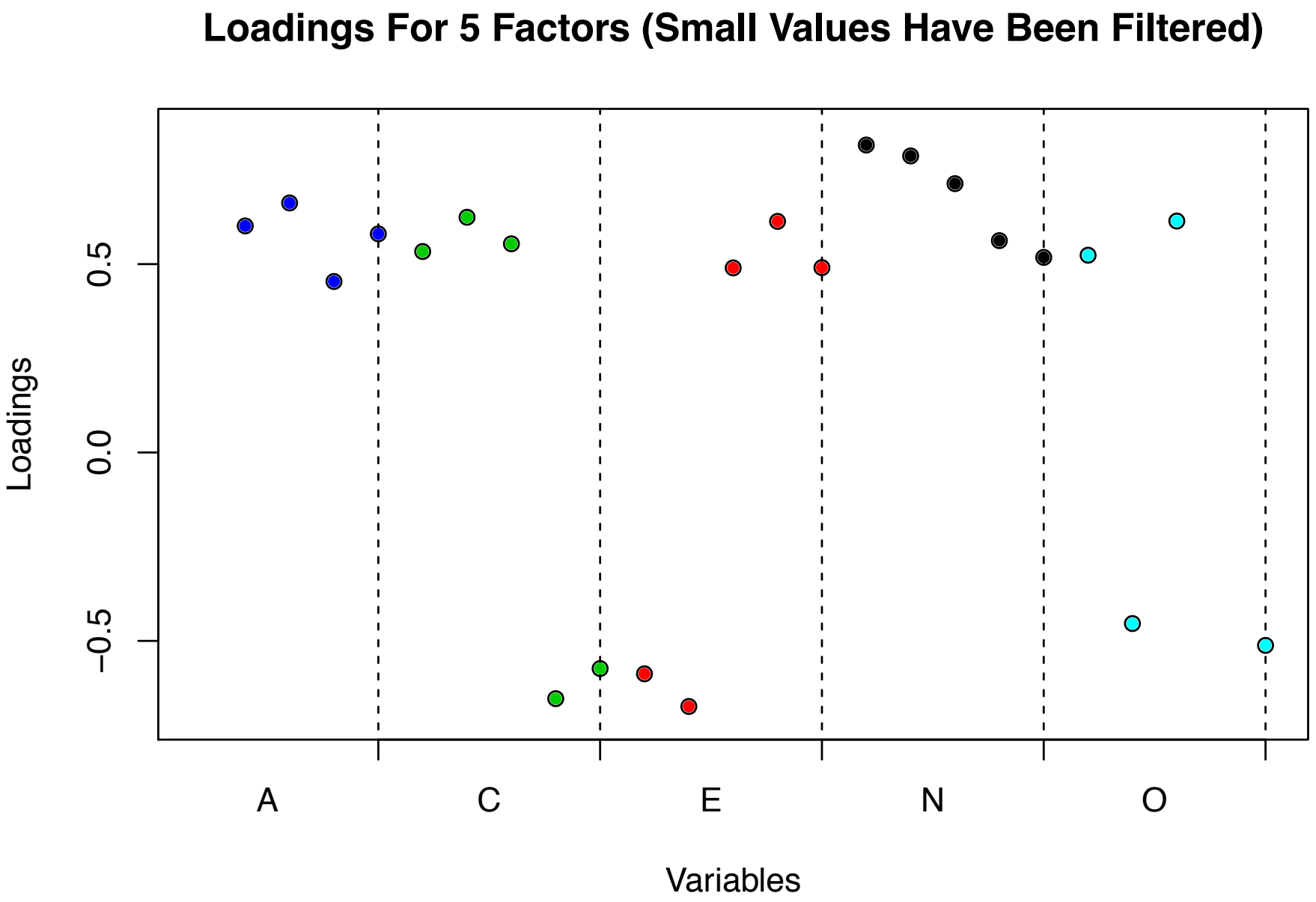
Dimension reduction

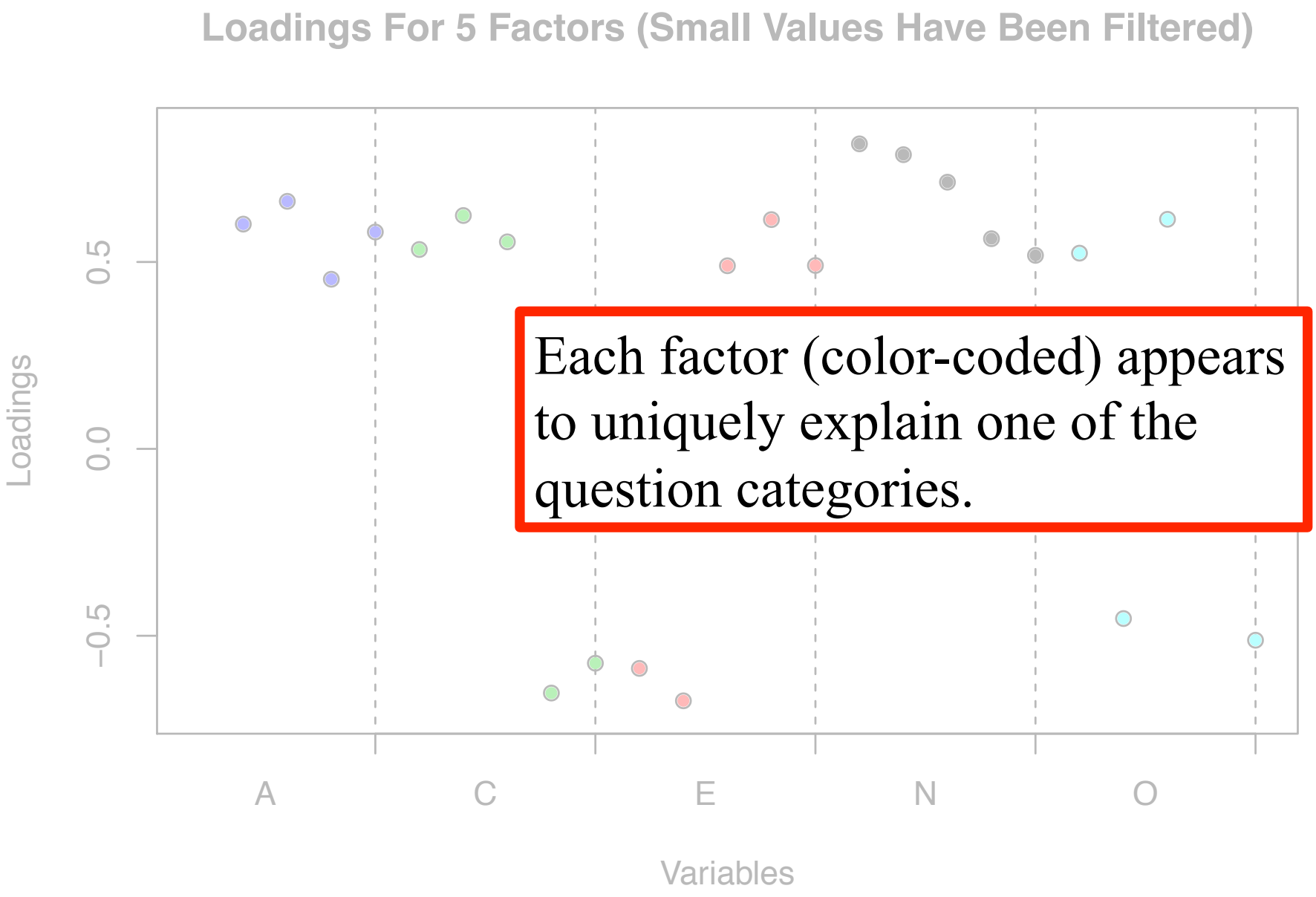
- **Principal components**:
  - Explain most of the total variability among the p variables using $k < p$ linear combinations of the original variables.
- **Factor analysis**:
  - Construct random quantities (*factors*) that describe variable groupings such that within-group correlations are maximized and between-group correlations are minimized.

Example: Personality assessment

- Subset of Synthetic Aperture Personality Assessment (SAPA) data.
  - For each of 1,000 individuals, 25 self-report personality variables. Variables organized into 5 categories:
    - Agreeableness (A).
    - Conscientiousness (C).
    - Extraversion (E).
    - Neuroticism (N).
    - Openness (O).

Data from ‘bfi’ dataset in R’s ‘psych’ package.


Loadings For 5 Factors (Small Values Have Been Filtered)
Loadings
0.5
0.0
−0.5
A
C
E
N
O
Variables

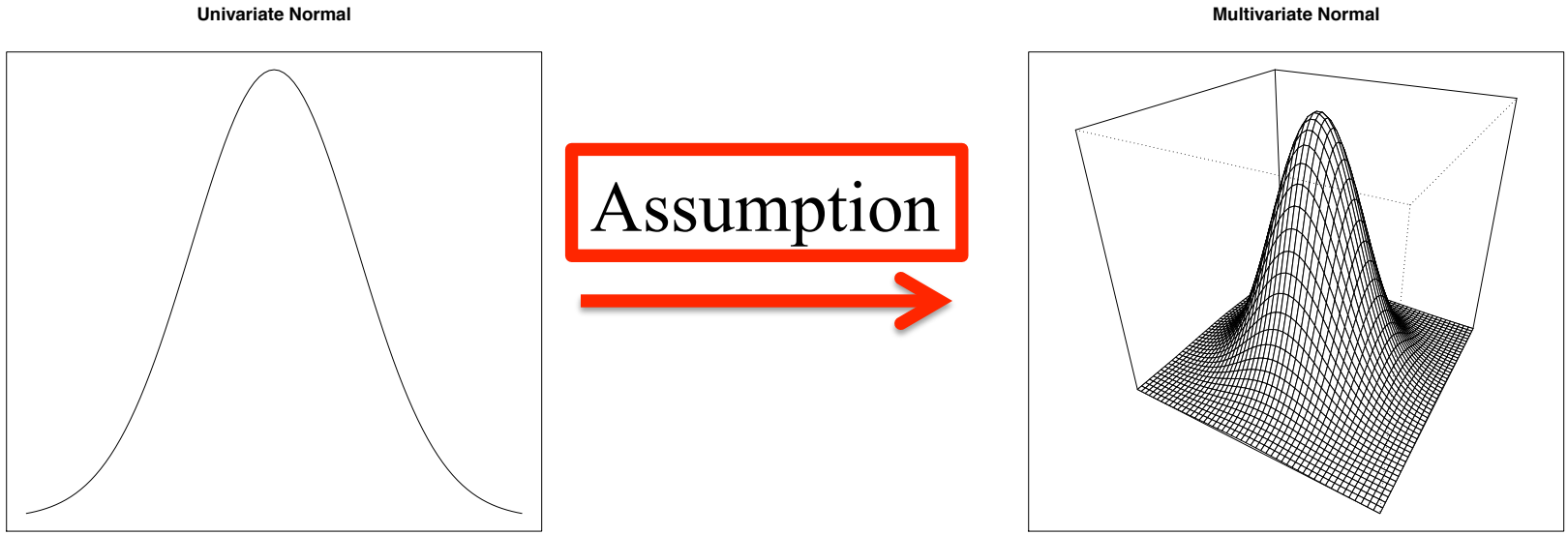

Loadings For 5 Factors (Small Values Have Been Filtered)
Loadings
0.5
0.0
−0.5
A
C
E
N
O
Variables
Each factor (color-coded) appears to uniquely explain one of the question categories.

Inference on mean vectors

- Multivariate ANOVA:
  - A generalization of univariate ANOVA.
- Multivariate linear regression:
  - A generalization of univariate linear regression.

Inference on mean vectors

- Multivariate ANOVA:
  - A generalization of univariate ANOVA.
- Multivariate linear regression:
  - A generalization of univariate linear regression.



Example: Nursing home data

- Data on 516 homes.
- Response vector:
  - Cost of nursing labor.
  - Cost of dietary labor.
  - Cost of maintenance labor.
  - Cost of housekeeping labor.
- Comparison group = home type:
  - Private.
  - Nonprofit.
  - Government.

Data from Example 6.10 in textbook.

Example: Nursing home data

- Data on 516 homes.
- Response vector:
  - Cost of nursing labor.
  - Cost of dietary labor.
  - Cost of maintenance lab
  - Cost of housekeeping la
- Comparison group = home type.
  - Private.
  - Nonprofit.
  - Government.

H_0: No difference in mean expense vectors between home types.

P-value < 0.0001.

Data from Example 6.10 in textbook.

Canonical correlation

- Assessment of correlation between two random vectors.
- Generalization of univariate correlation between two univariate variables.
- Canonical variables = linear combinations of variable components.
- Canonical correlations = univariate correlations between those linear combinations.

Example: Savings data

- Intercountry life-cycle savings data:
  - For each of 50 countries, 5 variables:
    - Savings ratio (*sr*).
    - % population < 15 (p_{15}).
    - % population > 75 (p_{75}).
    - Per-capita disposable income (*pdi*).
    - % growth rate in disposable income (*gpdi*).
- Question: How correlated are the population variables with the other variables?

Data from 'stats' package in R.

Example: Savings data

- Canonical variables:
  - First:
    - Population: $-0.009 \times p_{15} + 0.049 \times p_{75}$
    - Other: $0.008 \times sr + 0.000 \times dpi + 0.004 \times ddpi$
    - Canonical correlation: 0.825
  - Second:
    - Population: $-0.036 \times p_{15} - 0.260 \times p_{75}$
    - Other: $0.033 \times sr + 0.000 \times dpi - 0.012 \times ddpi$
    - Canonical correlation: 0.365

Example: Savings data

- Canonical variables:
  - First:
    - Population: $-0.009 \times p_{15} + 0.049 \times p_{75}$
    - Other: $0.008 \times sr + 0.000 \times dpi + 0.004 \times ddpi$
    - Canonical correlation: 0.825
  - Second:
    - Popula
    - Other:
    - Canonical correlation: 0.565

Percent population above 75 moderately positively correlated with savings ratio and percent growth in disposable income.

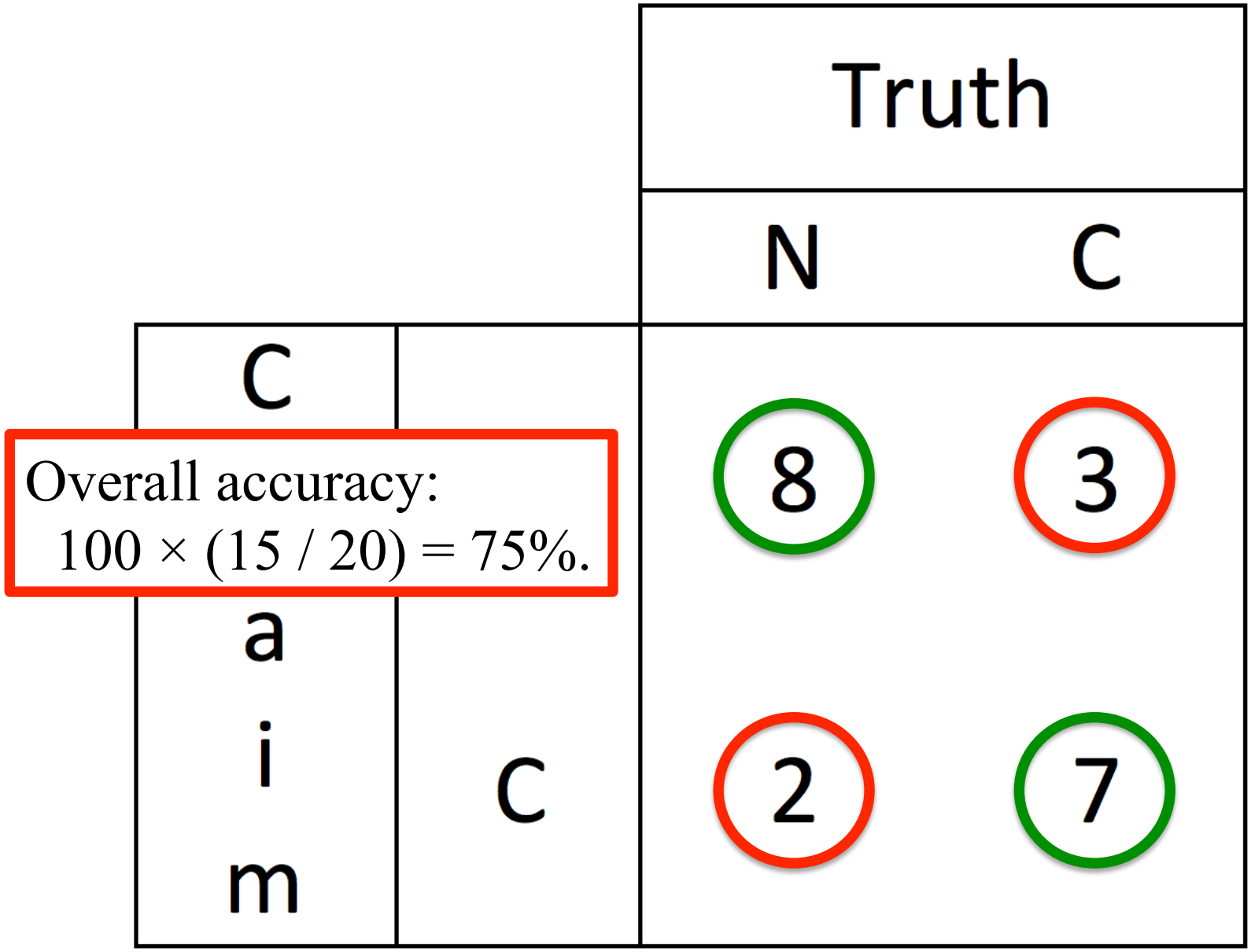
Discriminant analysis

- Given response vectors from 2 or more classes:
  - Define *discriminants* whose values separate the classes as much as possible.
  - Define a rule for *assigning* new observations to one of the classes.
  - The two tasks are often combined.
- We typically require *training* data (data for which class membership is known).
  - But, when assigning to classes, our goal is to assign *new* observations.

Example: Cervical cancer

- "Expression" values (continuous numbers) for 714 micro RNA in 58 humans, half of whom are healthy, half of whom have cervical cancer.
  - Set aside 19 individuals in each class for *training*, with the remaining used for *testing*.
- Linear Discriminant Analysis (LDA):
  - One simple discrimination method.
  - Assumption:
    - Each class has own multivariate (p = 714!) normal distribution mean.
    - The classes share the same *covariance matrix*.

		Truth	
		N	C
Claim	N	8	3
	C	2	7


Truth
N
C
C
a
i
m
C
8
3
2
7
Overall accuracy:
100 × (15 / 20) = 75%.

		Truth	
		N	C
C		8	3
a			
i	C	2	7
m			

Sensitivity:
$100 \times (7 / 10) = 70\%$.

		Truth	
		N	C
C		8	3
a			
i	C	2	7
m			

Specificity:
$100 \times (8 / 10) = 80\%$.

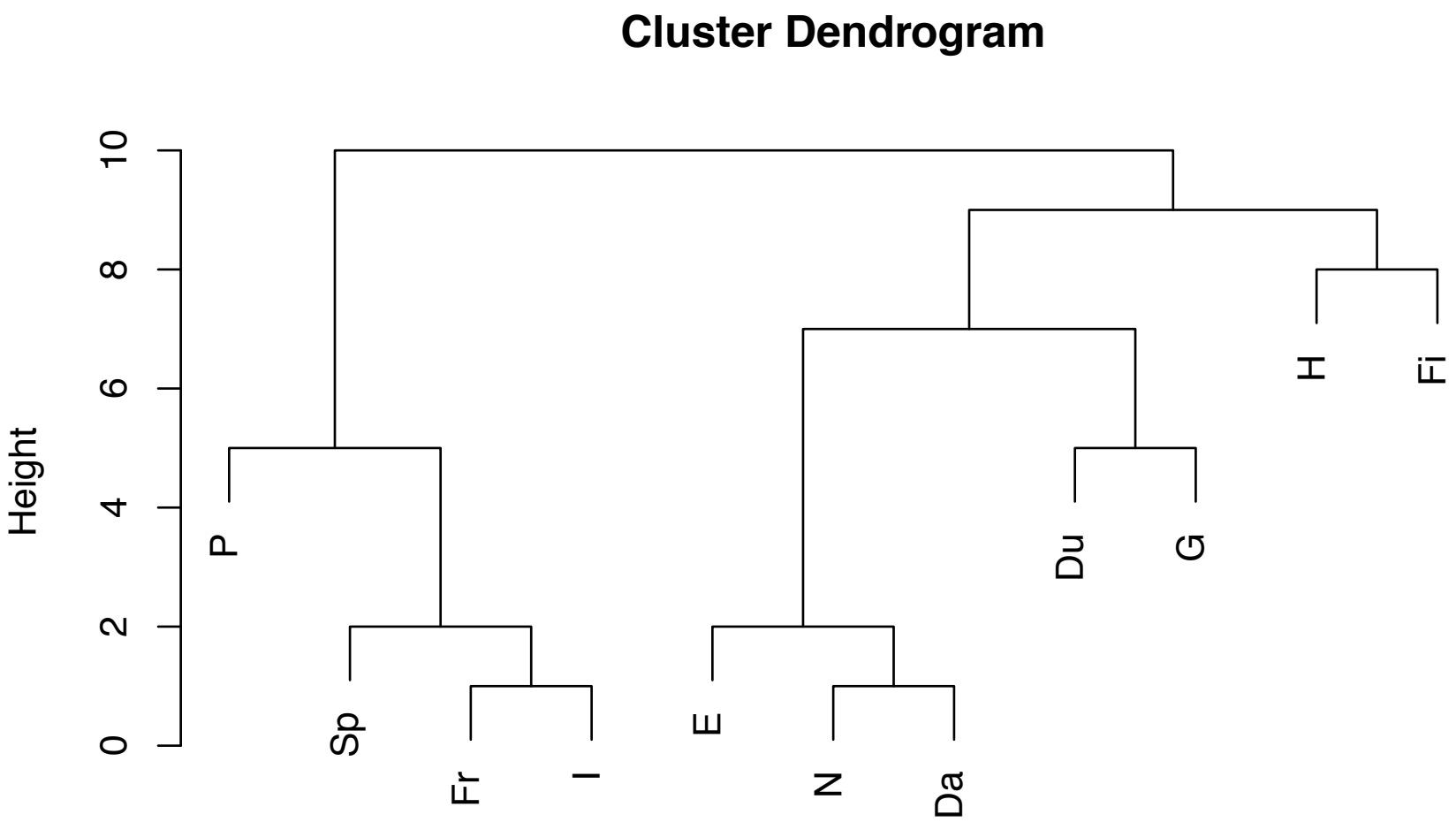
Cluster analysis

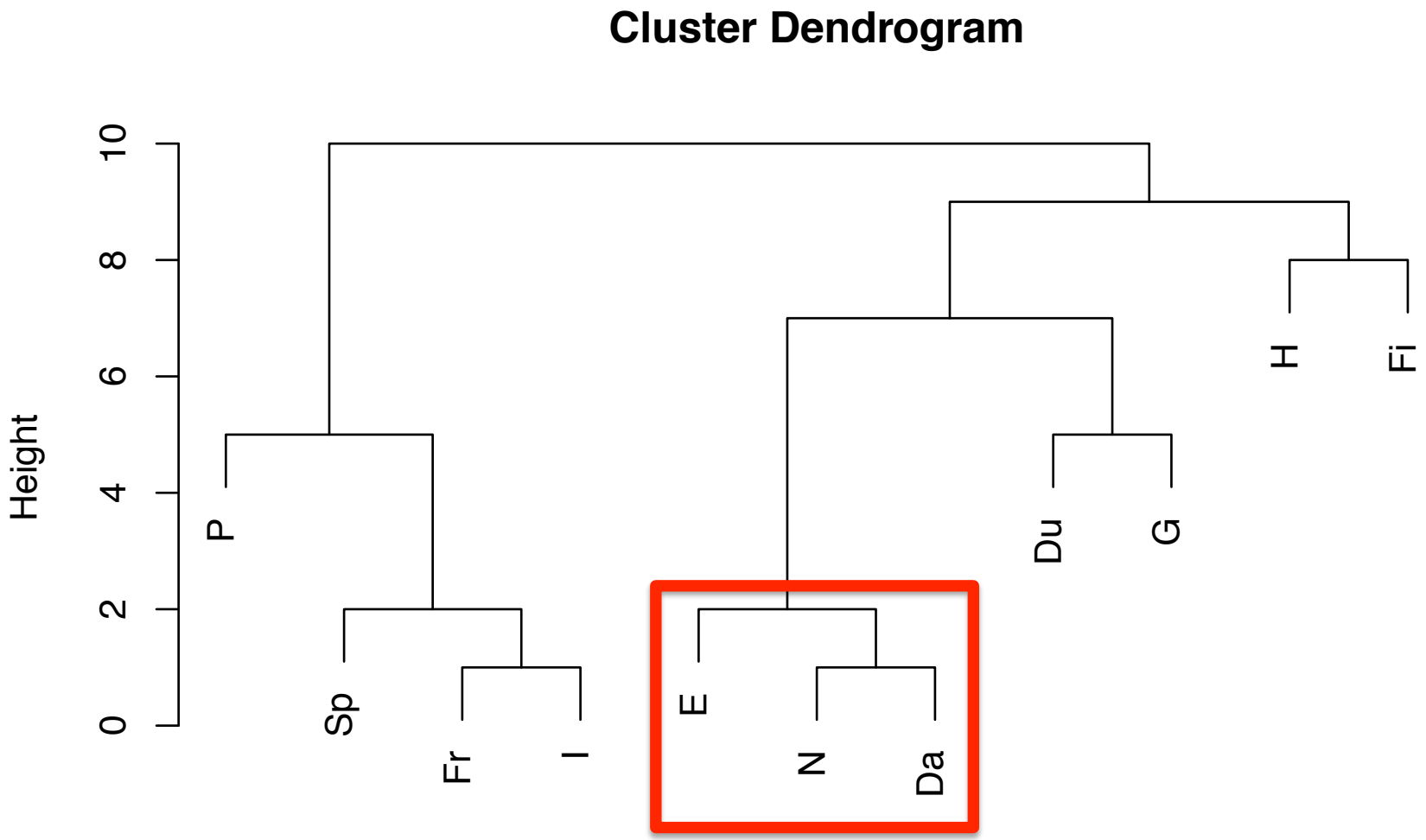
- Clustering refers to joining individuals or variables into groups based on similarities.
  - Joining individuals: Which individuals are most alike, in terms of their vector observations?
  - Joining variables: Which variables are most alike, in terms of their values across individuals?
- Discrimination is *supervised* (known groupings), while clustering is *unsupervised* (unknown groupings).

Example: Language similarity

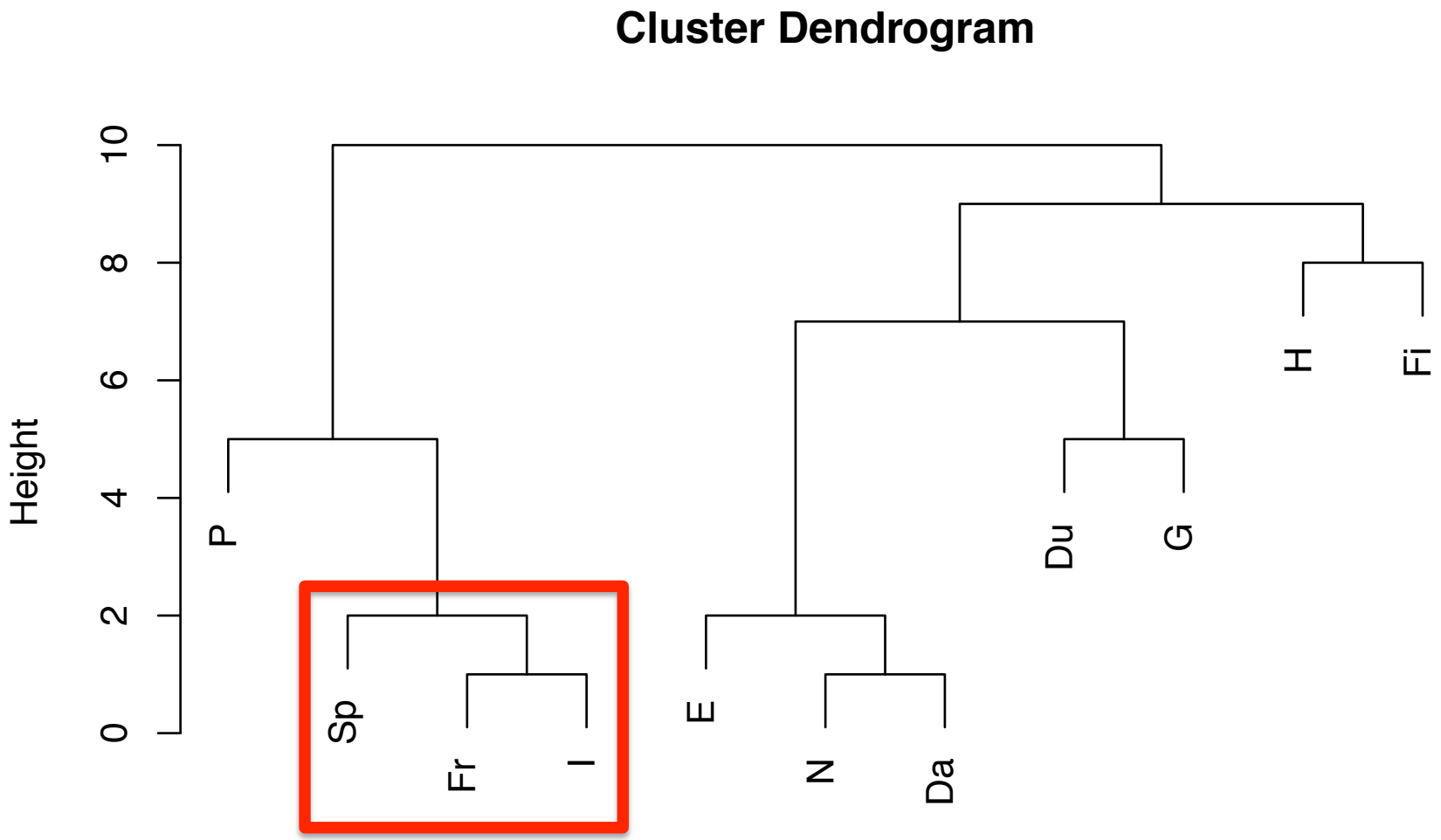
- Data on spelling of numbers one through ten, in each of 11 languages.
  - Languages: English (E), Norwegian (N), Danish (Da), Dutch (Du), German (G), French (Fr), Spanish (Sp), Italian (I), Polish (P), Hungarian (H), Finnish (Fi).
- Converted to be concordance of first letters.
- Which languages are most alike, and how do the languages group together?

Data from Example 12.2 in textbook.

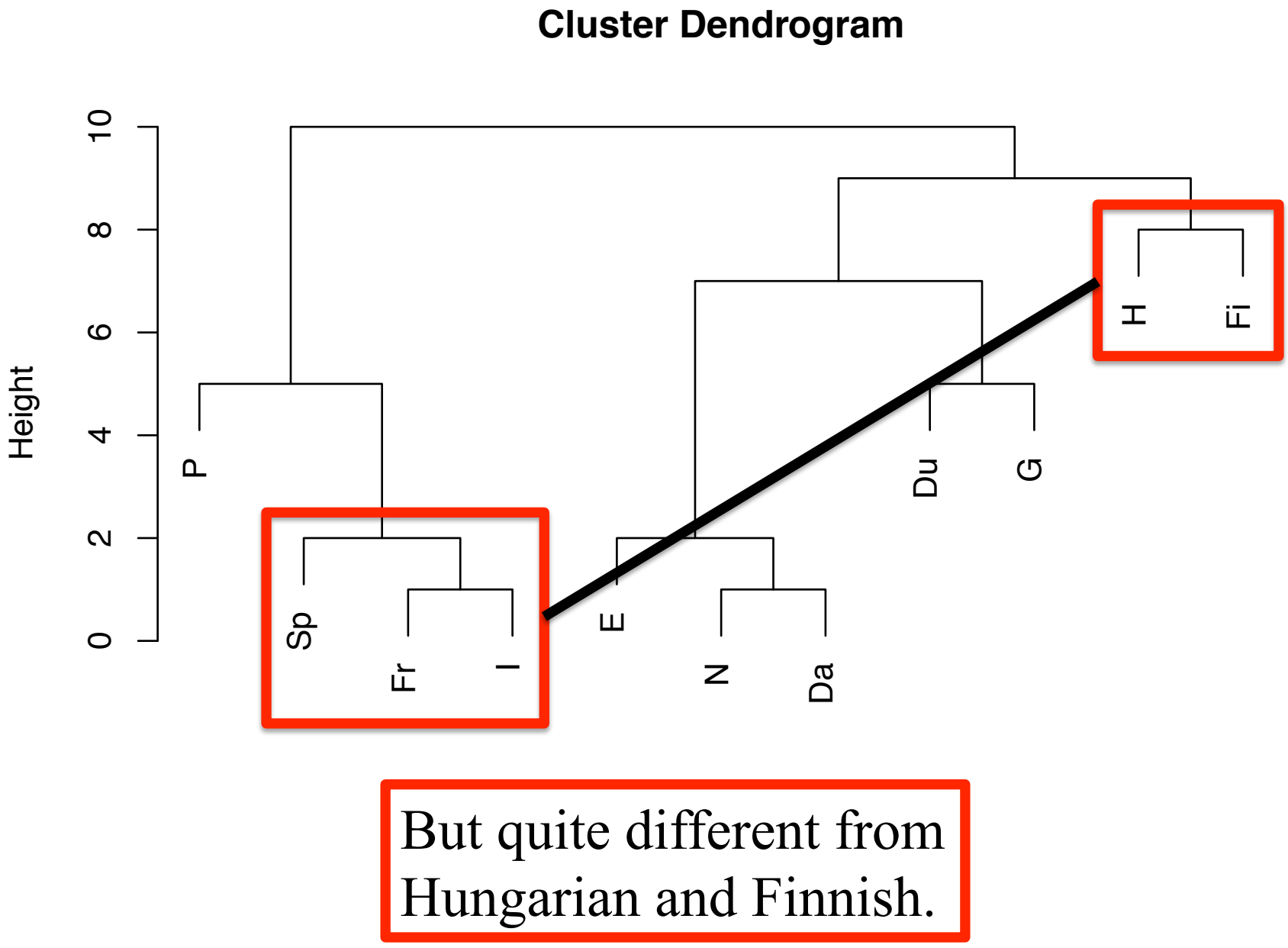

Cluster Dendrogram
Height
0
2
4
6
8
10
P
Sp
Fr
I
E
N
Da
Du
G
H
Fi



English, Norwegian, Danish are similar.



Spanish, French and Italian are very similar, too.



But quite different from Hungarian and Finnish.

Notation

Our data: $p \geq 1$ variables measured on each of n items / individuals / trials:

$$x_{jk} = \text{measurement of } k\text{th variable on } j\text{th item}$$

Can represent with an array:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Notation

Our data: $p \geq 1$ variables measured on each of n items / individuals / trials:

$$x_{jk} = \text{measurement of } k\text{th variable on } j\text{th item}$$

Can represent with an array:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Variable 1

Notation

Our data: $p \geq 1$ variables measured on each of n items / individuals / trials:

$$x_{jk} = \text{measurement of } k\text{th variable on } j\text{th item}$$

Can represent with an array:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Item 1

Example: Pottery chemical composition

- Twenty-six samples found at four kiln sites. Each sample assessed for percent composition of five metal oxides: Al, Fe, Mg, Ca, Na.

Data from DASL data repository.

Example: Pottery chemical composition

- Twenty-six samples found at four kiln sites. Each sample assessed for percent composition of five metal oxides: Al, Fe, Mg, Ca, Na.

$$\mathbf{X} = \begin{bmatrix} 0.15 & 0.51 \\ 0.12 & 0.17 \\ 0.13 & 0.20 \\ 0.16 & 0.14 \end{bmatrix}$$

Ca Na

Item 1
Item 2
Item 3
Item 4

Here's a toy subset: Four items, two variables.

Data from DASL data repository.

Descriptive statistics

- Key quantities with multivariate data:
  - Location: e.g., mean.
  - Spread: e.g., variance.
  - *Linear* association: e.g., correlation.

Location and spread

For variable k, $k = 1, 2, \ldots, p$, the sample mean is

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk}$$

Similarly, the sample variance for variable k is

$$s_k^2 = s_{kk} = \frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2$$

The sample standard deviation for variable k is $\sqrt{s_{kk}}$.

Location and spread

For variable k, $k = 1, 2, \ldots, p$, the sample mean is

$$\bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk}$$

Similarly, the sample variance for variable k is

$$s_k^2 = s_{kk} = \frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2$$

The sample standard deviation for variable k is $\sqrt{s_{kk}}$.

NOTE: You will often see $n - 1$ in the denominator instead of n. We will discuss in class later in the semester (also see textbook).

Covariance and correlation

Consider variables i and k, $i = 1, 2, \ldots, p$, $k = 1, 2, \ldots, p$. The sample covariance between variables i and k is

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

Similarly, the sample correlation is

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2}}$$

If $i = k$, the covariance reduces to the variance. Also, the correlation is one. Note also that $s_{ik} = s_{ki}$ and $r_{ik} = r_{ki}$.

Covariance and correlation

- Magnitude and sign of covariance and correlation indicate strength and direction of association.
  - Zero: No linear association.
  - Positive sign: Large values of the two variables tend to occur together, and small values of the two variables tend to occur together.
  - Negative sign: Large values of one variable tend to occur with small values of the other.
- Covariance is unbounded, while correlation is bounded to be between -1 and 1.
- Covariance changes if you change the units of measurement. Correlation does not.
- Limited to *linear* associations.
- Both are susceptible to outlying values.

The sample mean can be written as $\bar{\mathbf{x}} = [\bar{x}_1 \bar{x}_2 \ldots \bar{x}_p]'$. This is column vector of length p (equivalently, a $p \times 1$ matrix). The $p \times p$ sample variance / covariance matrix can be written as

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

The $p \times p$ sample correlation matrix can be written as

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

The sample mean can be written as $\bar{\mathbf{x}} = [\bar{x}_1 \bar{x}_2 \ldots \bar{x}_p]'$. This is column vector of length p (equivalently, a $p \times 1$ matrix). The $p \times p$ sample variance / covariance matrix can be written as

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Symmetric

The $p \times p$ sample correlation matrix can be written as

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Symmetric

The sample mean can be written as $\bar{\mathbf{x}} = [\bar{x}_1 \bar{x}_2 \ldots \bar{x}_p]'$. This is column vector of length p (equivalently, a $p \times 1$ matrix). The $p \times p$ sample variance / covariance matrix can be written as

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

To remind us that we are dividing by n instead of $n - 1$.

The $p \times p$ sample correlation matrix can be written as

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Example: Pottery chemical composition

The sample means are

$$\bar{x}_1 = \frac{1}{4}\sum_{j=1}^{4} x_{j1} = \frac{1}{4}(0.15 + 0.12 + 0.13 + 0.16) = 0.140$$

$$\bar{x}_2 = \frac{1}{4}\sum_{j=1}^{4} x_{j2} = \frac{1}{4}(0.51 + 0.17 + 0.20 + 0.14) = 0.255$$

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 0.140 \\ 0.255 \end{bmatrix}$$

The sample variances and covariances are

$$
\begin{aligned}
s_{11} &= \frac{1}{4}\sum_{j=1}^{4}\left(x_{j1}-\bar{x}_1\right)^2 \\
&= \frac{1}{4}((0.150-0.140)^2+(0.120-0.140)^2+ \\
&\qquad (0.130-0.140)^2+(0.160-0.140)^2) = 0.000250 \\
s_{22} &= \frac{1}{4}\sum_{j=1}^{4}\left(x_{j2}-\bar{x}_2\right)^2 \\
&= \frac{1}{4}((0.510-0.255)^2+(0.170-0.255)^2+ \\
&\qquad (0.200-0.255)^2+(0.140-0.255)^2) = 0.022125 \\
s_{12} &= \frac{1}{4}\sum_{j=1}^{4}\left(x_{j1}-\bar{x}_1\right)\left(x_{j2}-\bar{x}_2\right) \\
&= \frac{1}{4}((0.150-0.140)(0.510-0.255)+(0.120-0.140)(0.170-0.255)+ \\
&\qquad (0.130-0.140)(0.200-0.255)+(0.160-0.140)(0.140-0.255)) \\
&\qquad = 0.000625
\end{aligned}
$$

In matrix form:

$$
\mathbf{S}_n = \begin{bmatrix} 0.000250 & 0.000625 \\ 0.000625 & 0.022125 \end{bmatrix}
$$

The sample variances and covariances are

$$s_{11} = \frac{1}{4}\sum_{j=1}^{4}(x_{j1} - \bar{x}_1)^2$$

$$= \frac{1}{4}((0.150 - 0.140)^2 + (0.120 - 0.140)^2 +$$

$$(0.130 - 0.140)^2 + (0.160 - 0.140)^2) = 0.000250$$

$$s_{22} = \frac{1}{4}\sum_{j=1}^{4}(x_{j2} - \bar{x}_2)^2$$

The sample correlation is

$$r_{12} = r_{21} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{0.000625}{\sqrt{0.000250}\sqrt{0.022125}} = 0.265747$$

In matrix form:

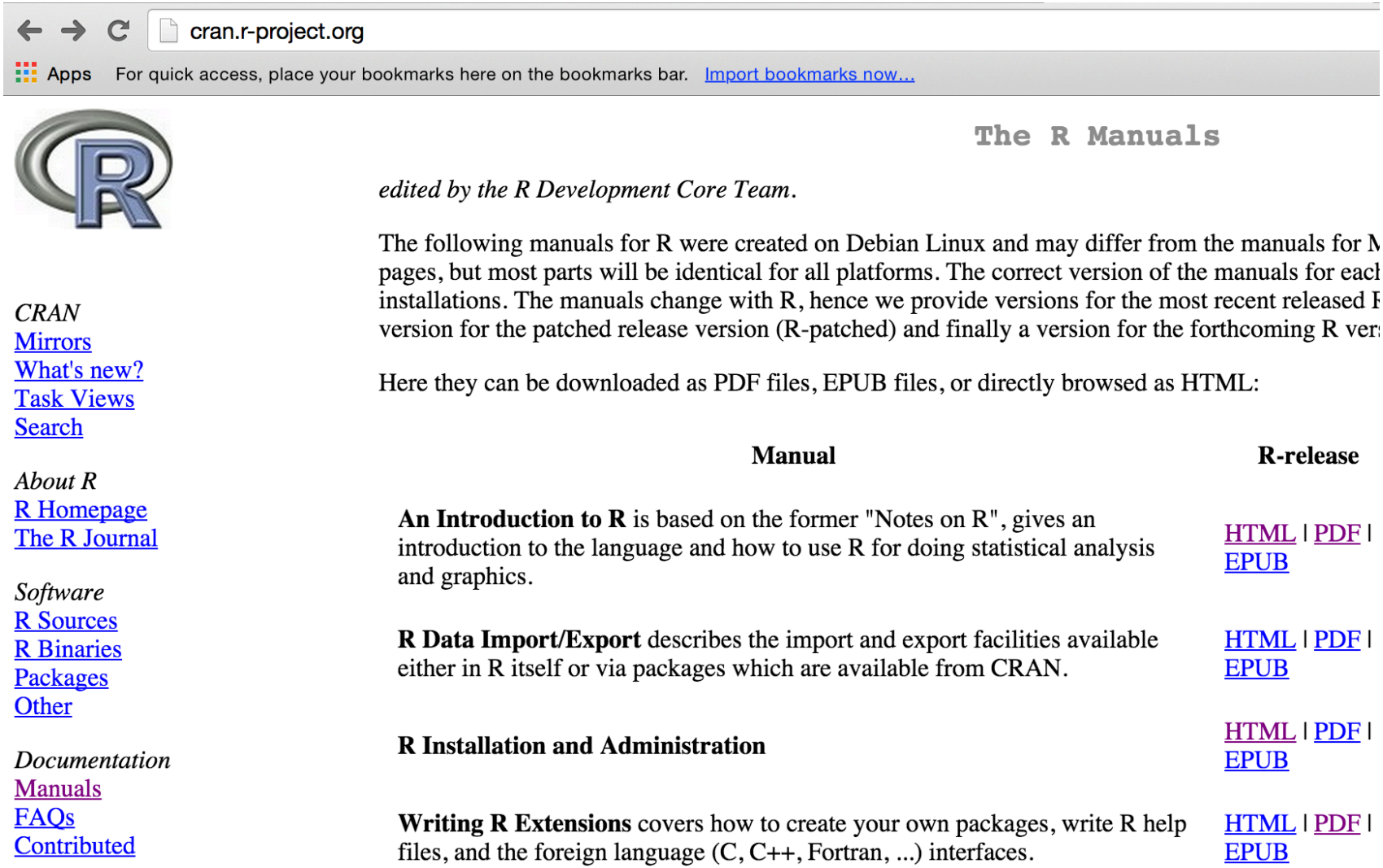
$$\mathbf{R} = \begin{bmatrix} 1.000000 & 0.265747 \\ 0.265747 & 1.000000 \end{bmatrix}$$

$$= 0.000625$$

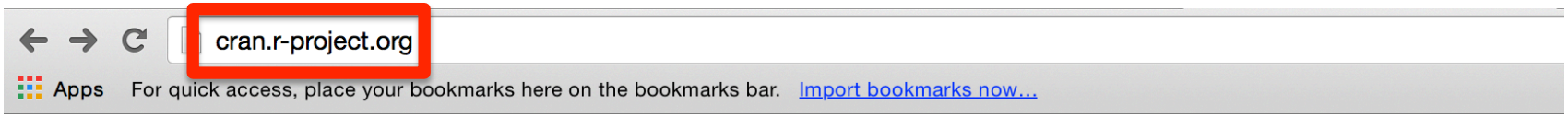
In matrix form:

$$\mathbf{S}_n = \begin{bmatrix} 0.000250 & 0.000625 \\ 0.000625 & 0.022125 \end{bmatrix}$$

R: Open source statistical software



R: Open source statistical software



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The R Manuals

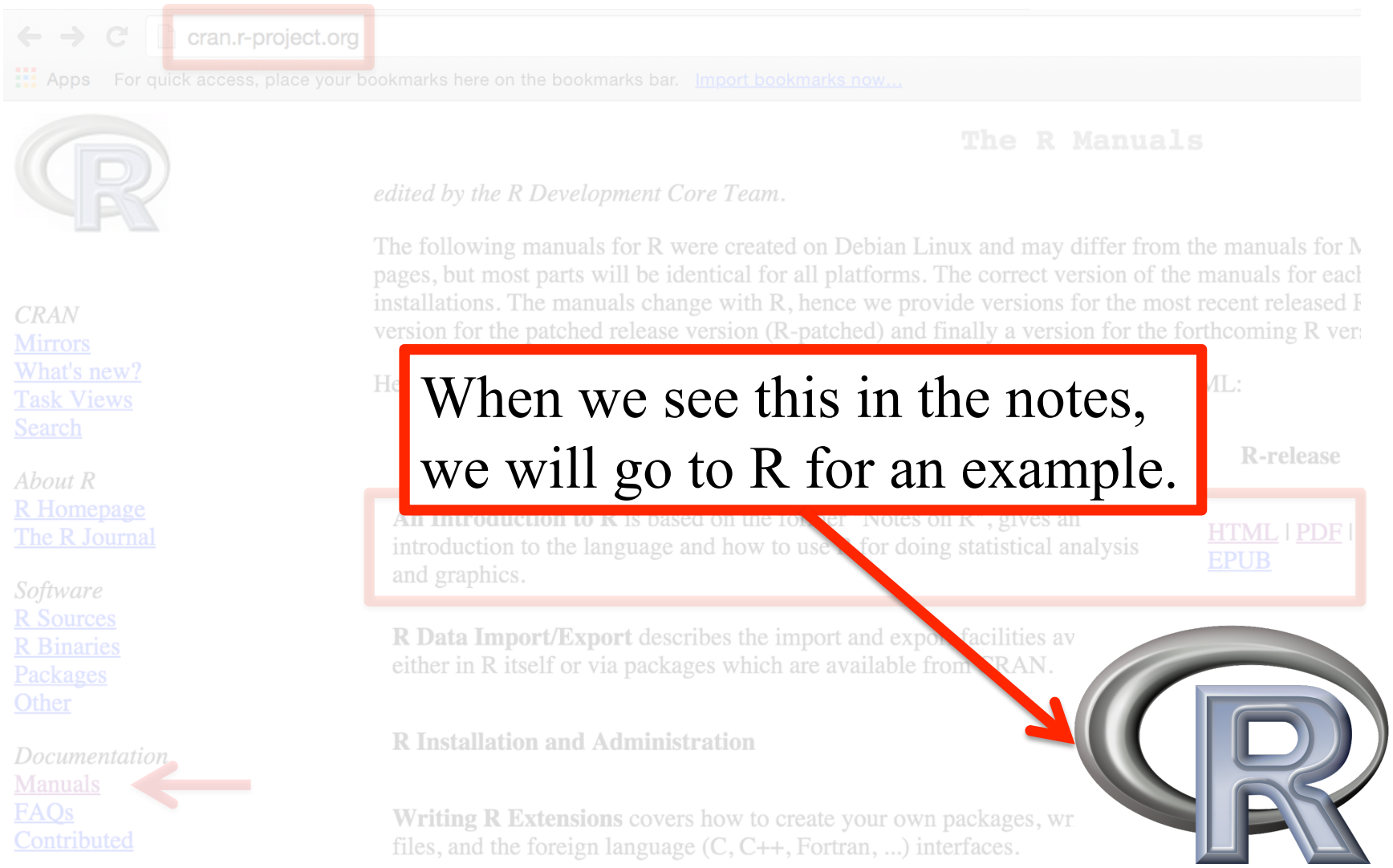
edited by the R Development Core Team.

The following manuals for R were created on Debian Linux and may differ from the manuals for M
pages, but most parts will be identical for all platforms. The correct version of the manuals for eacl
installations. The manuals change with R, hence we provide versions for the most recent released F
version for the patched release version (R-patched) and finally a version for the forthcoming R ver

Here they can be downloaded as PDF files, EPUB files, or directly browsed as HTML:

Manual	R-release
An Introduction to R is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics.	HTML | PDF | EPUB
R Data Import/Export describes the import and export facilities available either in R itself or via packages which are available from CRAN.	HTML | PDF | EPUB
R Installation and Administration	HTML | PDF | EPUB
Writing R Extensions covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces.	HTML | PDF | EPUB

R: Open source statistical software

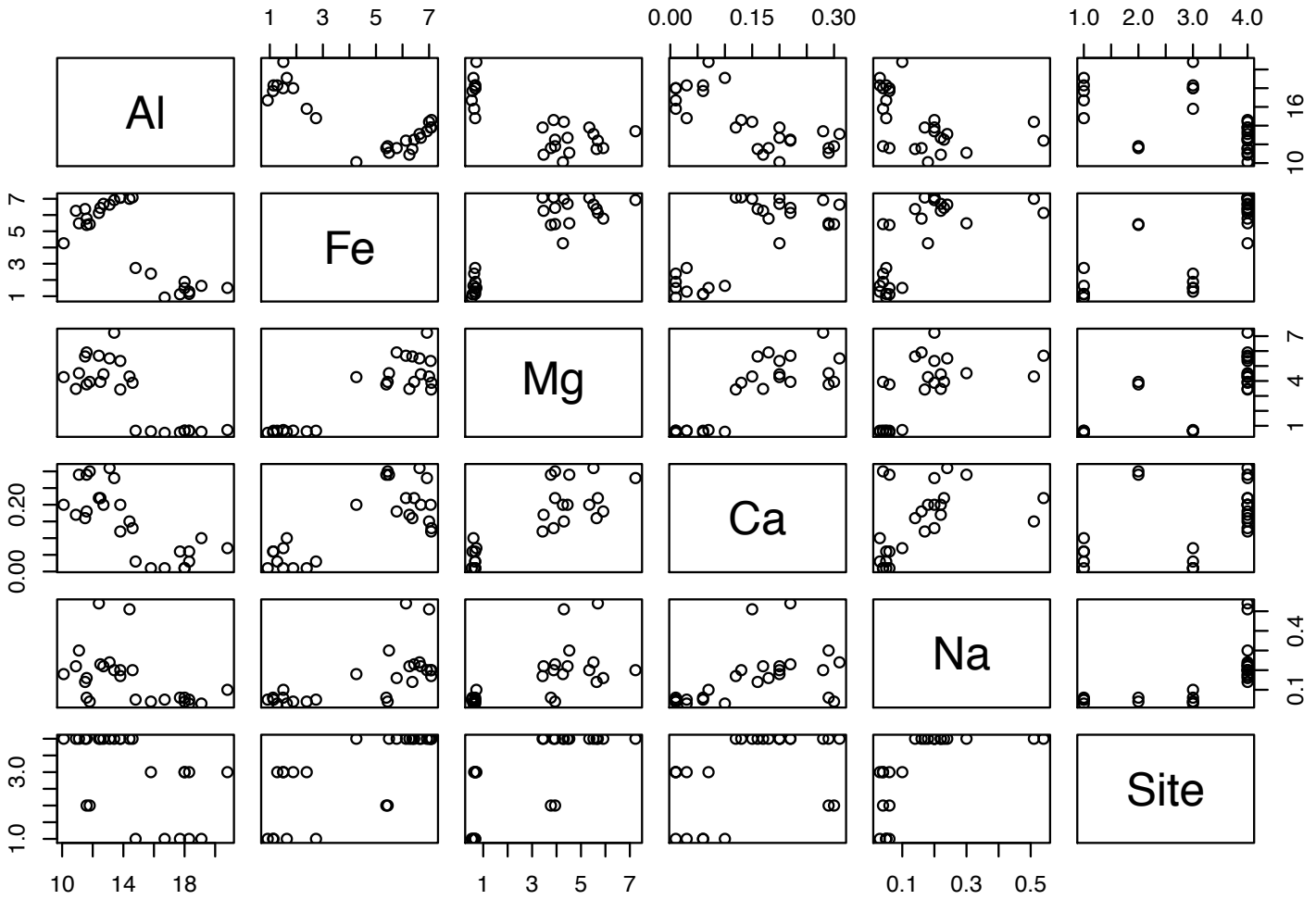


When we see this in the notes, we will go to R for an example.

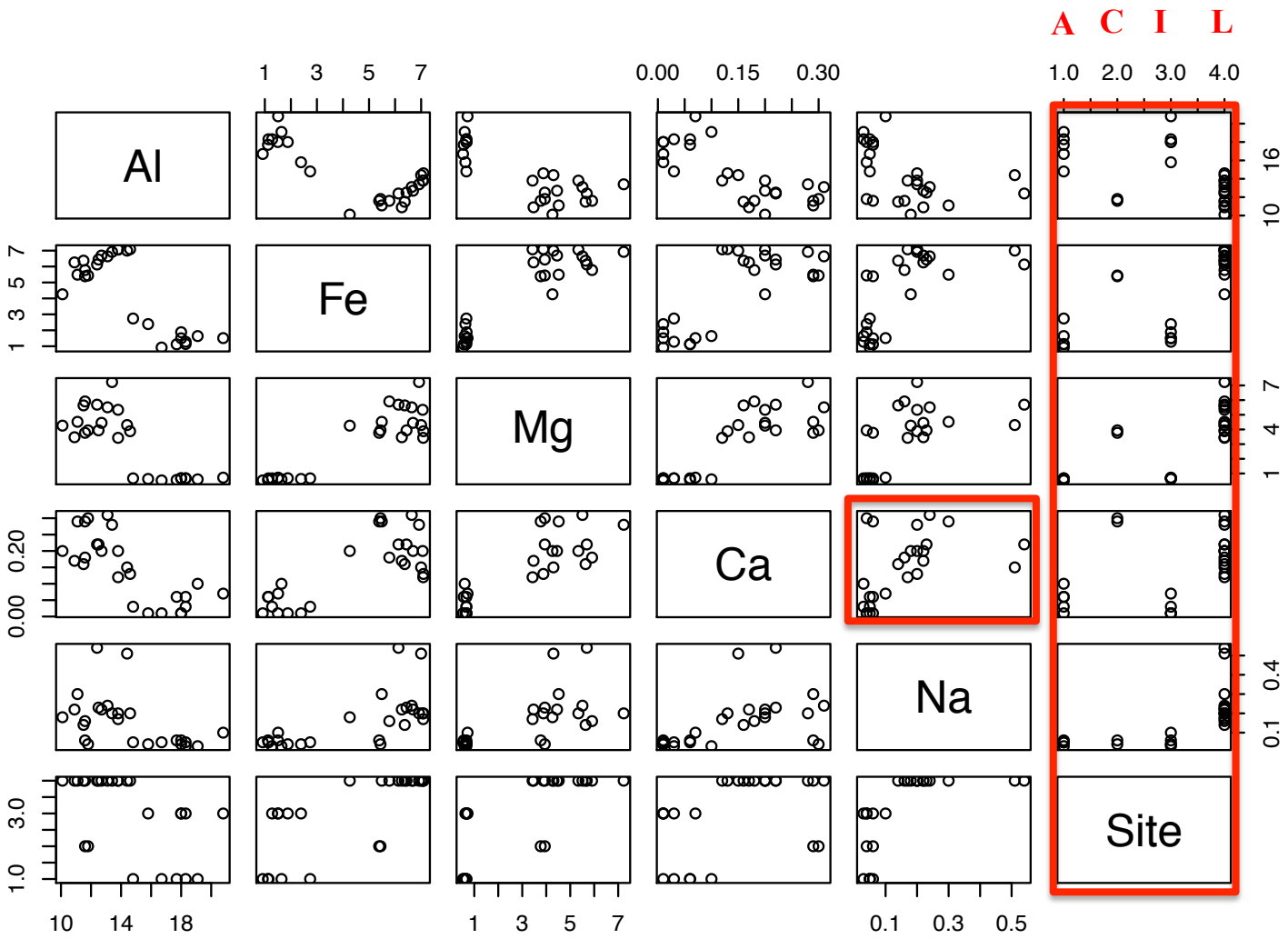
Pictures

- Pictures, together with descriptive statistics, are the main tools of *exploratory data analysis* (EDA).
- Simple scatterplots can be used to visualize relationships between two variables at a time.
- 3D scatterplots can be used to visualize relationships between three variables.
- Many others: co-plots, growth curves, star plots, Chernoff faces, …

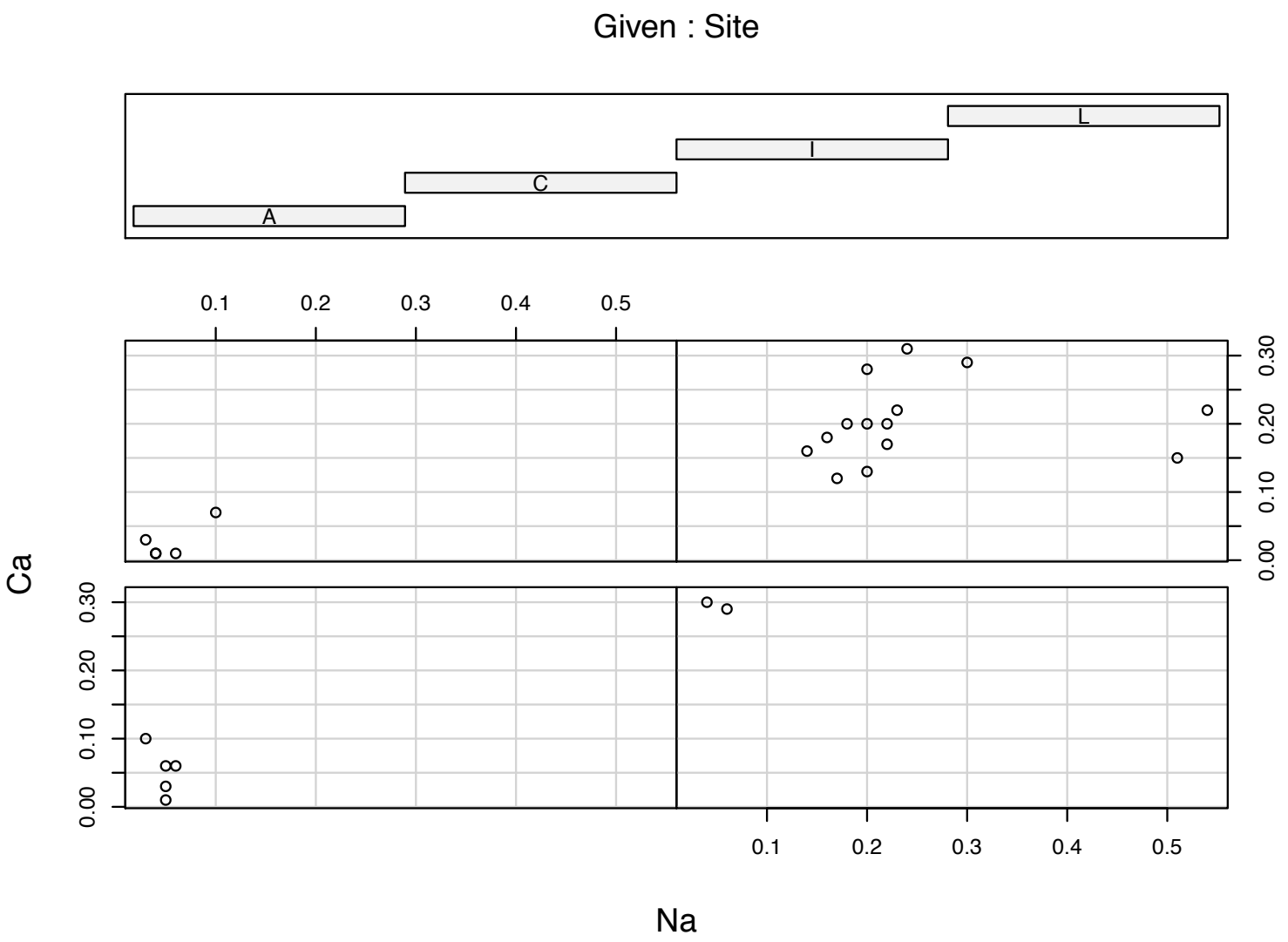
Pairwise scatterplots (pottery data)



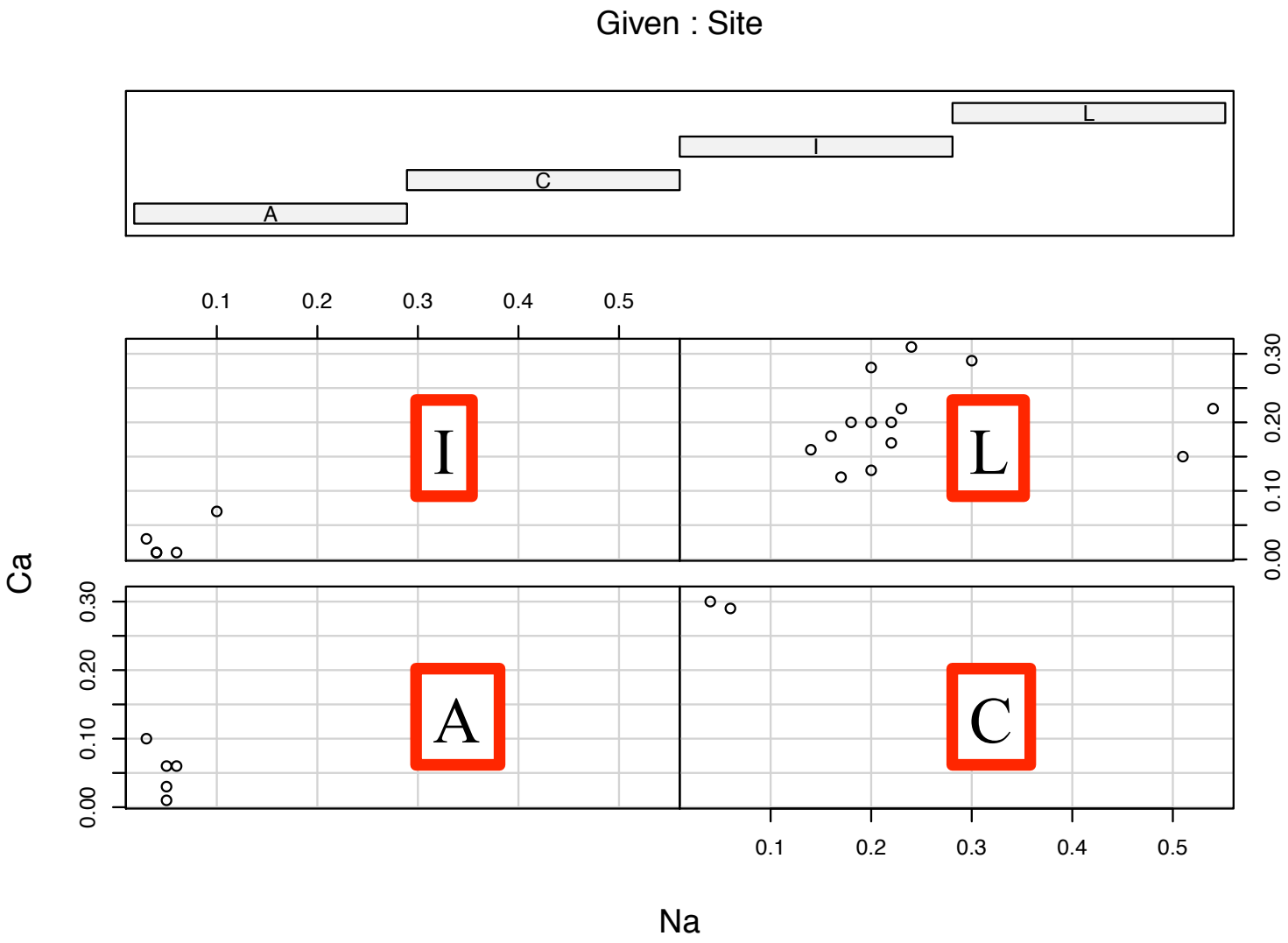
Pairwise scatterplots (pottery data)



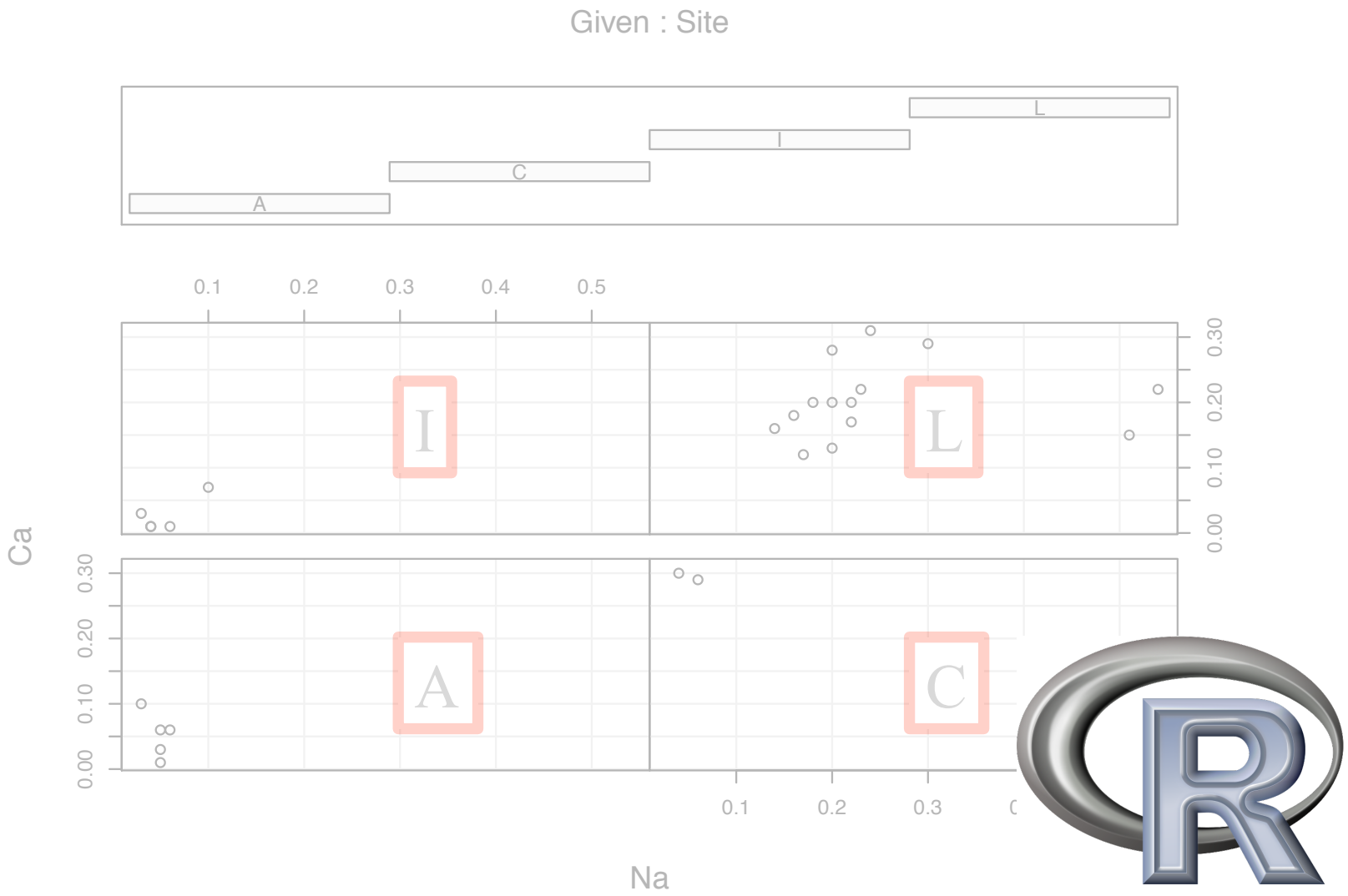
Co-plot (pottery data)



Co-plot (pottery data)



Co-plot (pottery data)

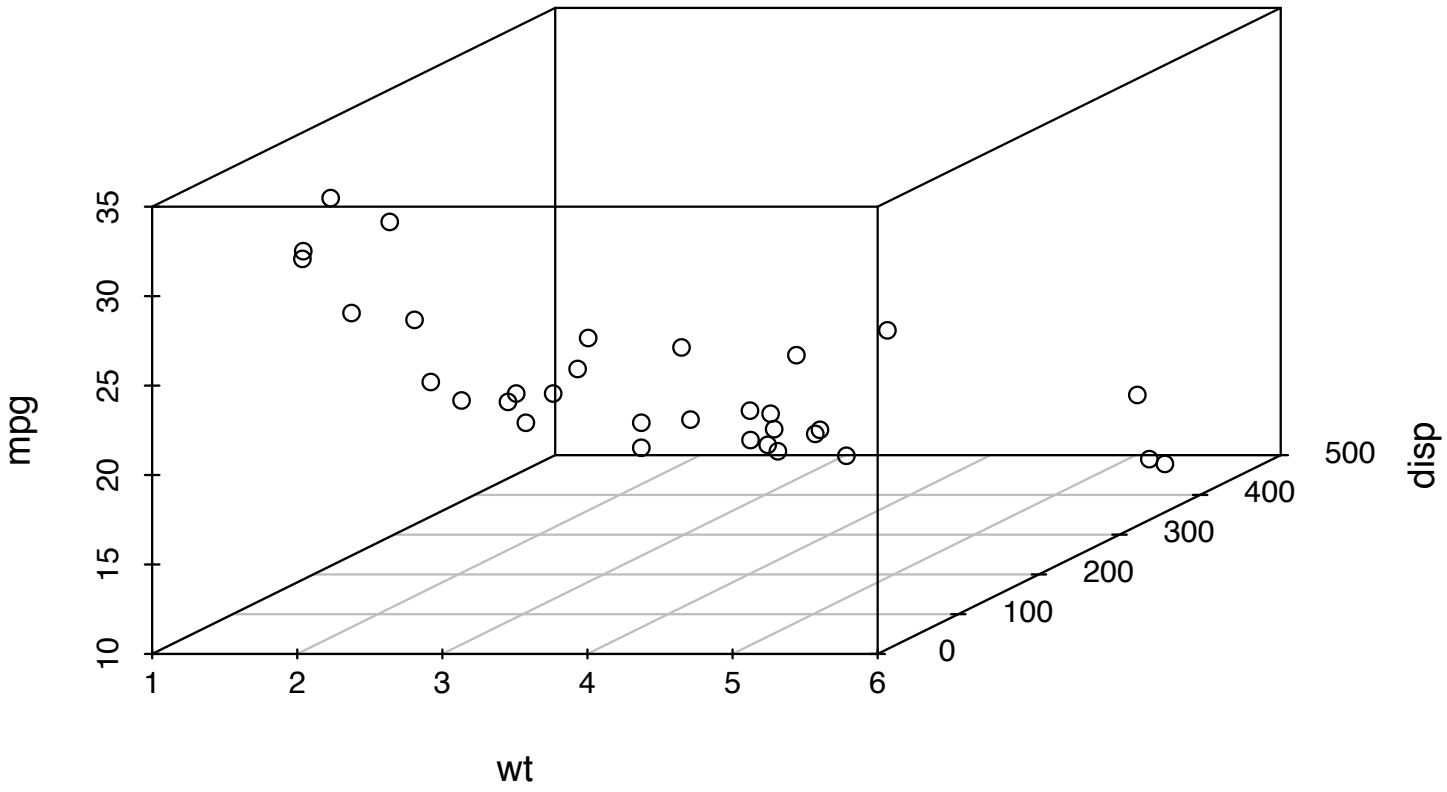


Example: Car road tests

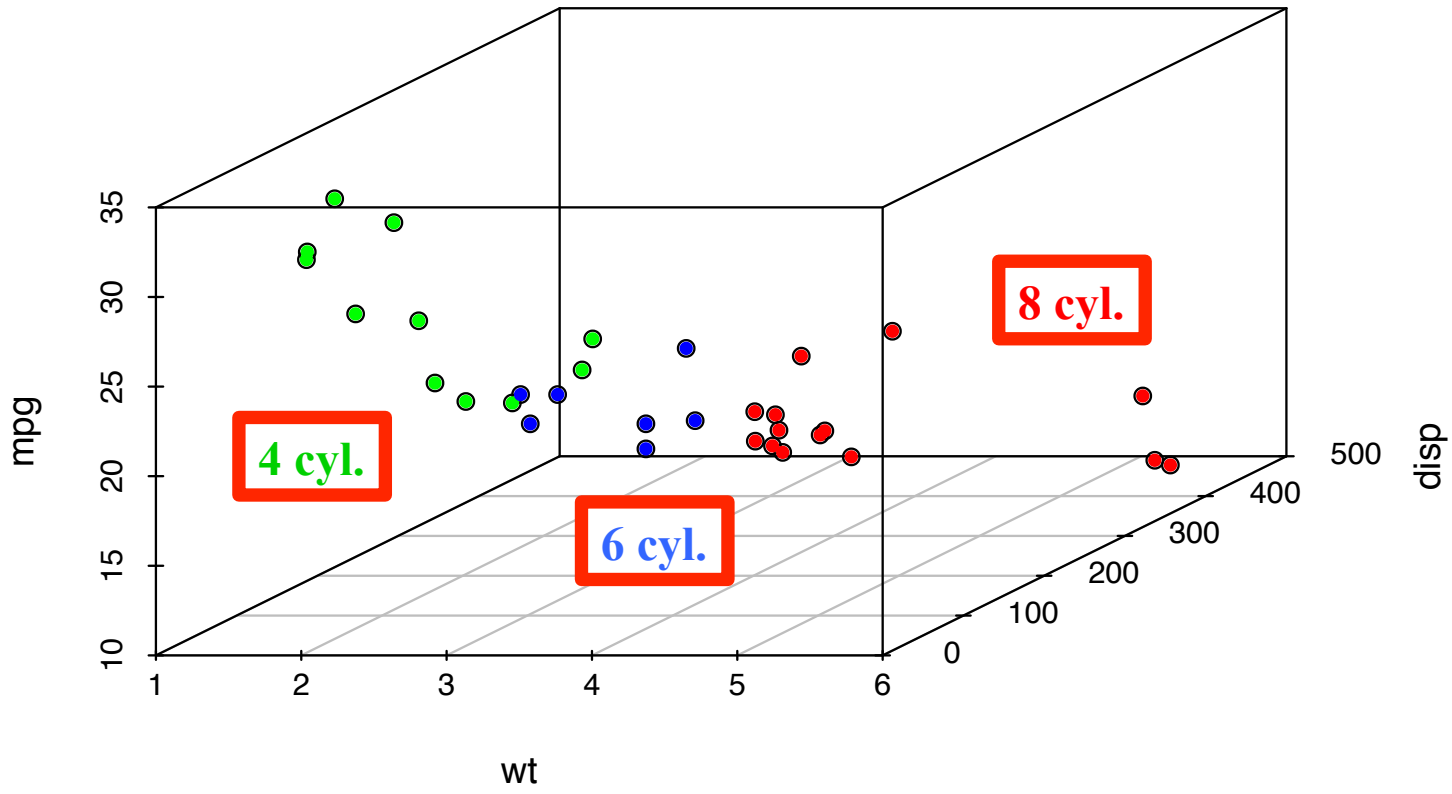
- Data from 1974 Motor Trend magazine. For each of 32 cars, data on 11 variables:
  - Miles per gallon.
  - Number of cylinders.
  - Displacement (cu. in.)
  - Gross horsepower.
  - Rear axle ratio.
  - Weight (lb / 1000).
  - 1 / 4 mile time.
  - V/S.
  - Transmission (0 = auto, 1 = manual).
  - Number of forward gears.
  - Number of carburetors.

The ‘mtcars’ data set from ‘scatterplot3d’ R package.

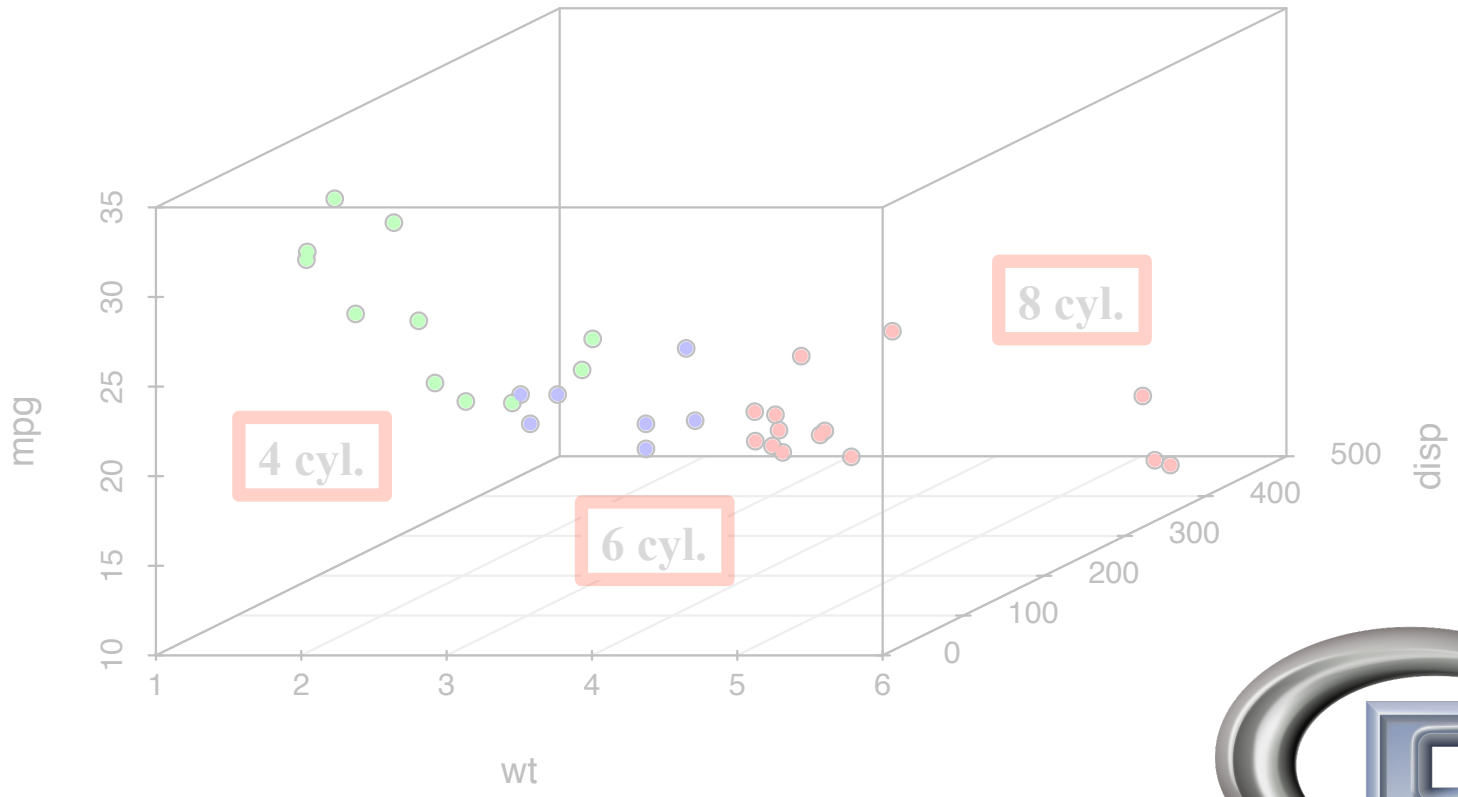
3D scatterplot



3D scatterplot



3D scatterplot

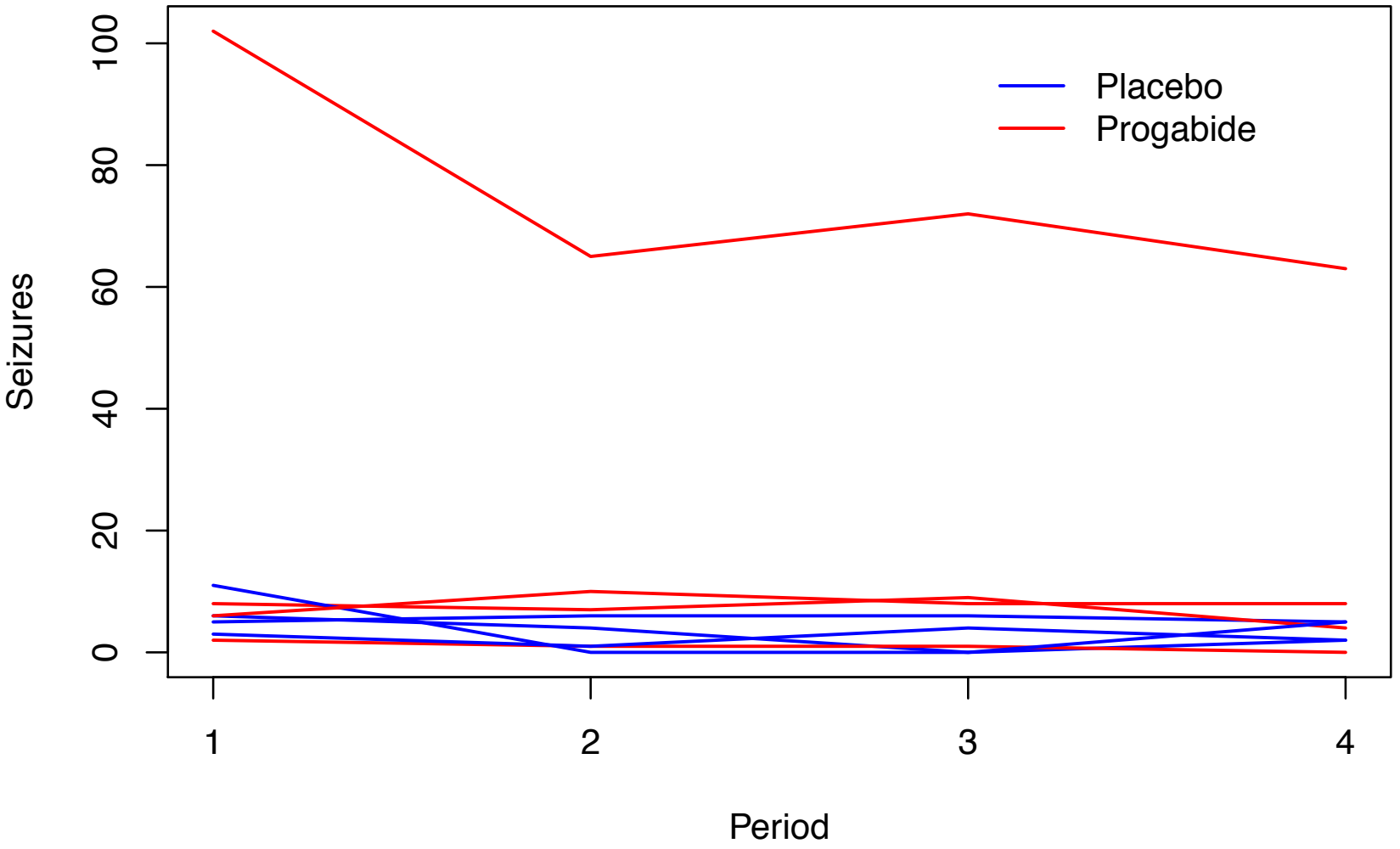


Example: Seizure rates

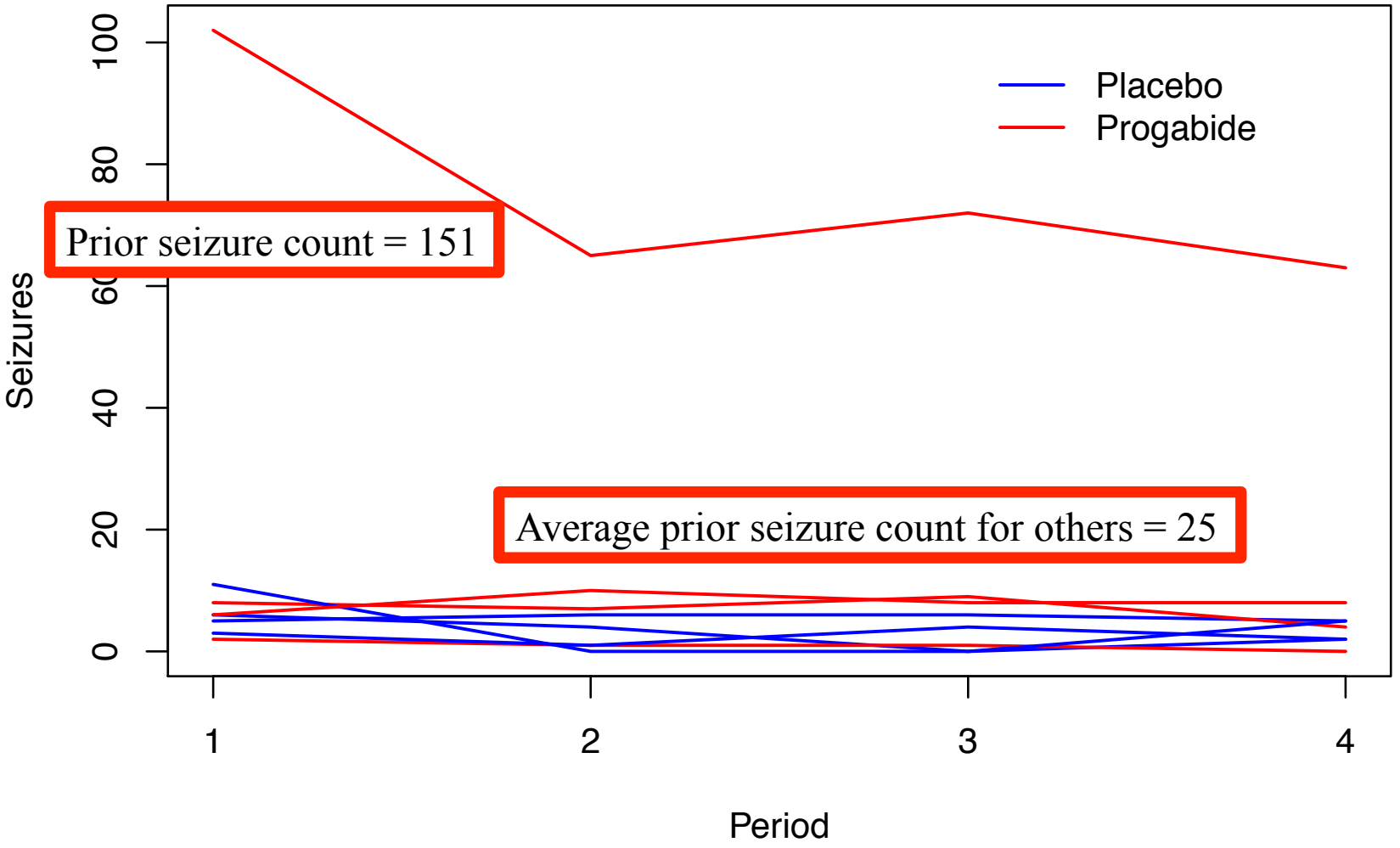
- Randomized clinical trial of anti-epileptic drug.
- 59 epileptic patients randomized to treatment (Progabide) or control (placebo).
  - Seizure counts in each of four two-week periods were recorded.
  - Also know subject's age and baseline seizure rate (the number of seizures before study began).
- There are *longitudinal / repeated measures* data.

The 'epilepsy' data set from 'HSAUR2' R package.

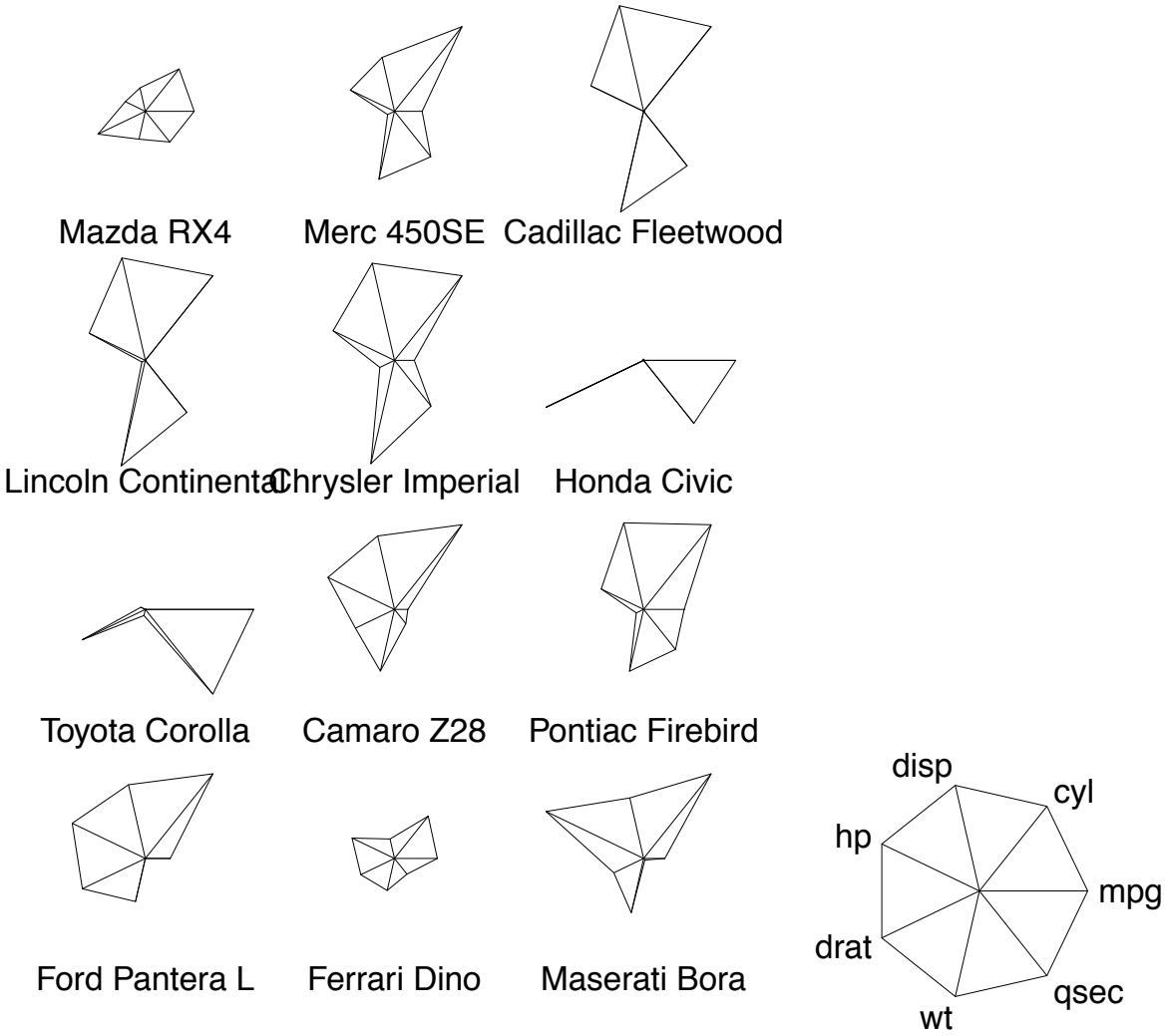
Growth curves



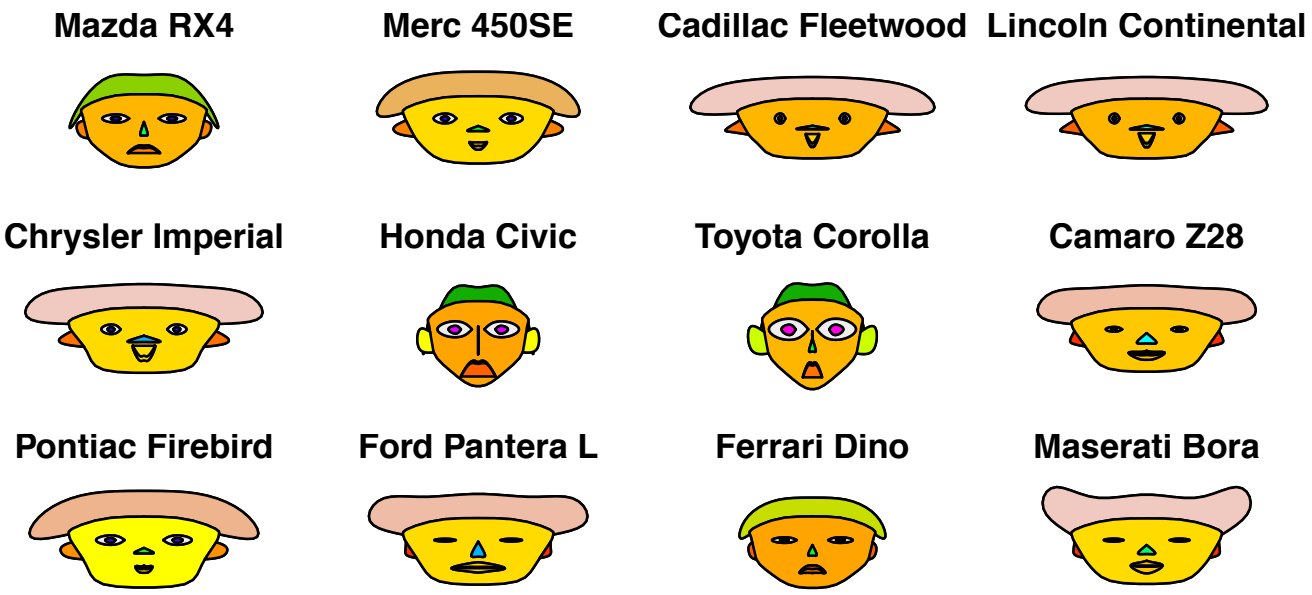
Growth curves



Stars plot (car road test data)



Faces plot (same data)



Height of face = mpg, width of face = cyl, structure of fact = disp, height of mouth = hp, width of mouth = drat, smiling = wt, height of eyes = qsec, width of eyes = mpg, height of hair = cyl, width of hair = disp, style of hair = hp, height of nose = drat, width of nose = wt, width of ear = qsec, height of ear = mpg.

Faces plot (same data)

Height of face = mpg, width of face = cyl, structure of fact = disp, height of mouth = hp, width of mouth = drat, smiling = wt, height of ey of eyes = mpg, height of hair = cyl, width of hair = disp, style height of nose = drat, width of nose = wt, width of ear = qsec mpg.

Distance

- *Distance* plays a major role in all of the methods we will consider this semester.
- Standard Euclidean distance is appropriate when variables vary equally and are not correlated.
- In general, statistical data will exhibit variable-specific variability and (often) correlation between variables.
- Need a way to quantify *statistical distance*.

Let $P = (x_1, x_2, \ldots, x_p)$ be a point with p coordinates. The straight-line distance from P to the origin $O = (0, 0, \ldots, 0)$ is

$$d(O, P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$$

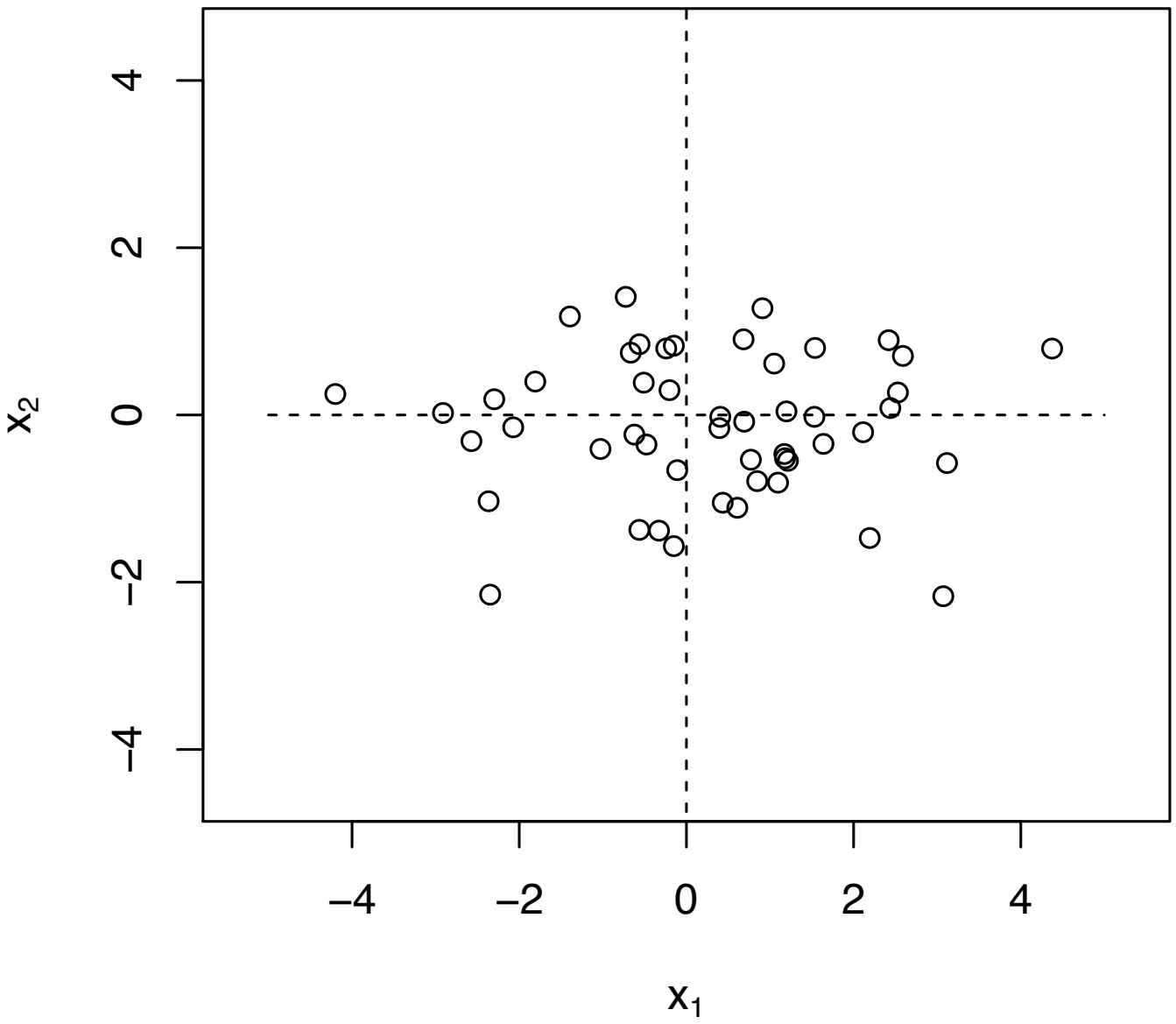
All points $(x_1, x_2, \ldots, x_p)$ that lie a constant distance c^2 from the origin satisfy

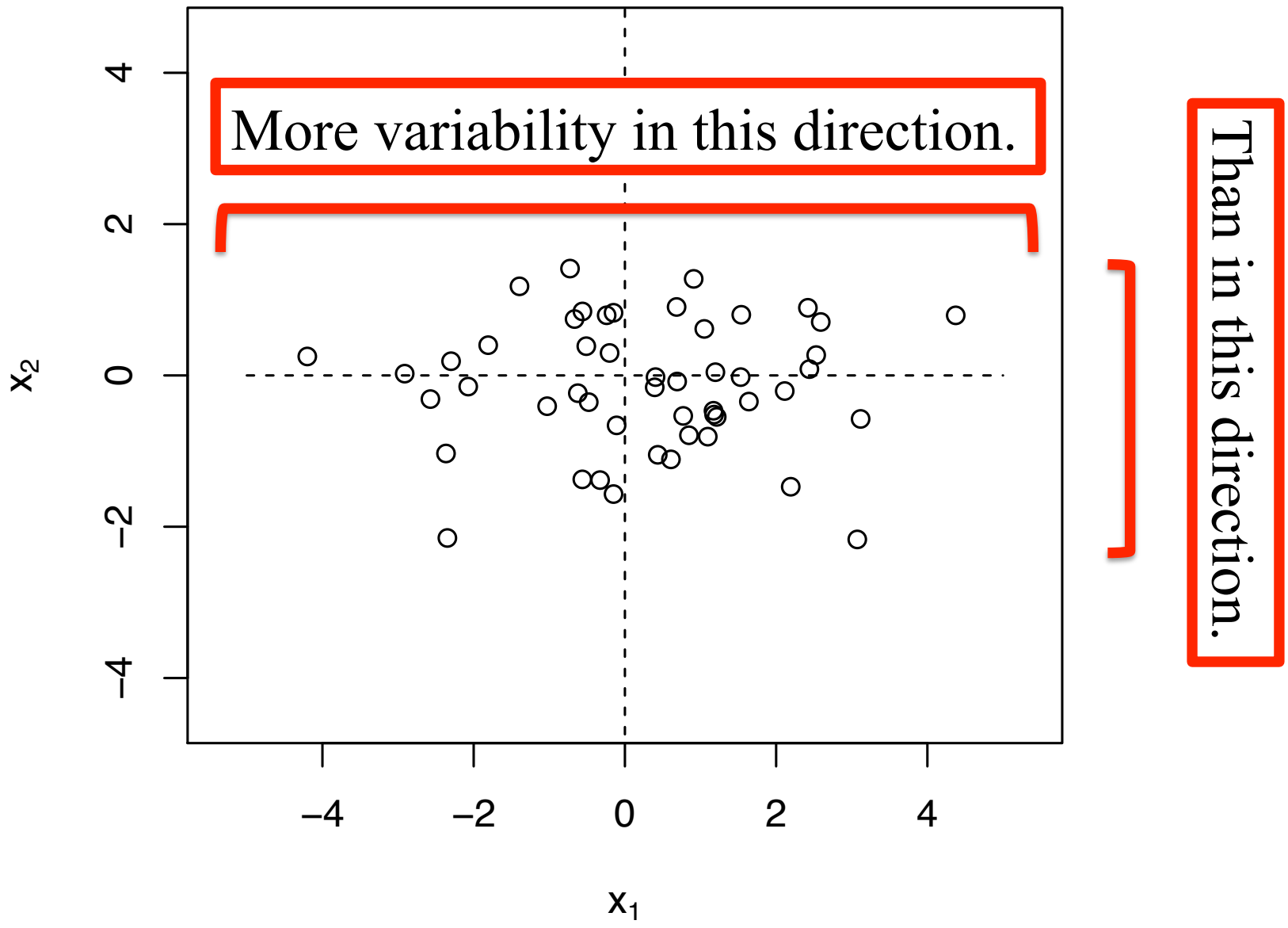
$$d^2(O, P) = x_1^2 + x_2^2 + \cdots x_p^2 = c^2$$

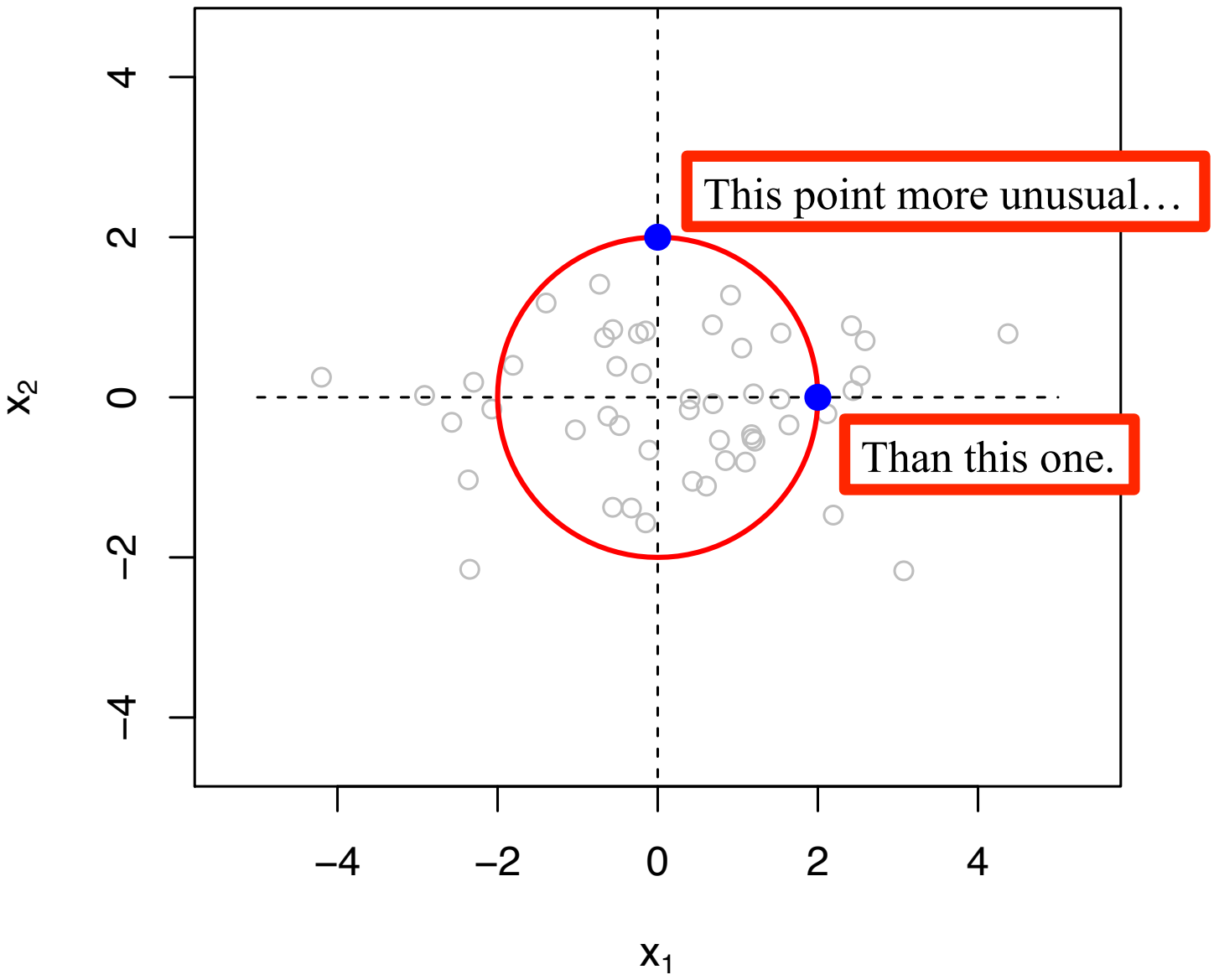
This is a hypersphere (circle if $p = 2$). The straight-line distance between two arbitrary points $P = (x_1, x_2, \ldots, x_p)$ and $Q = (y_1, y_2, \ldots, y_p)$ is

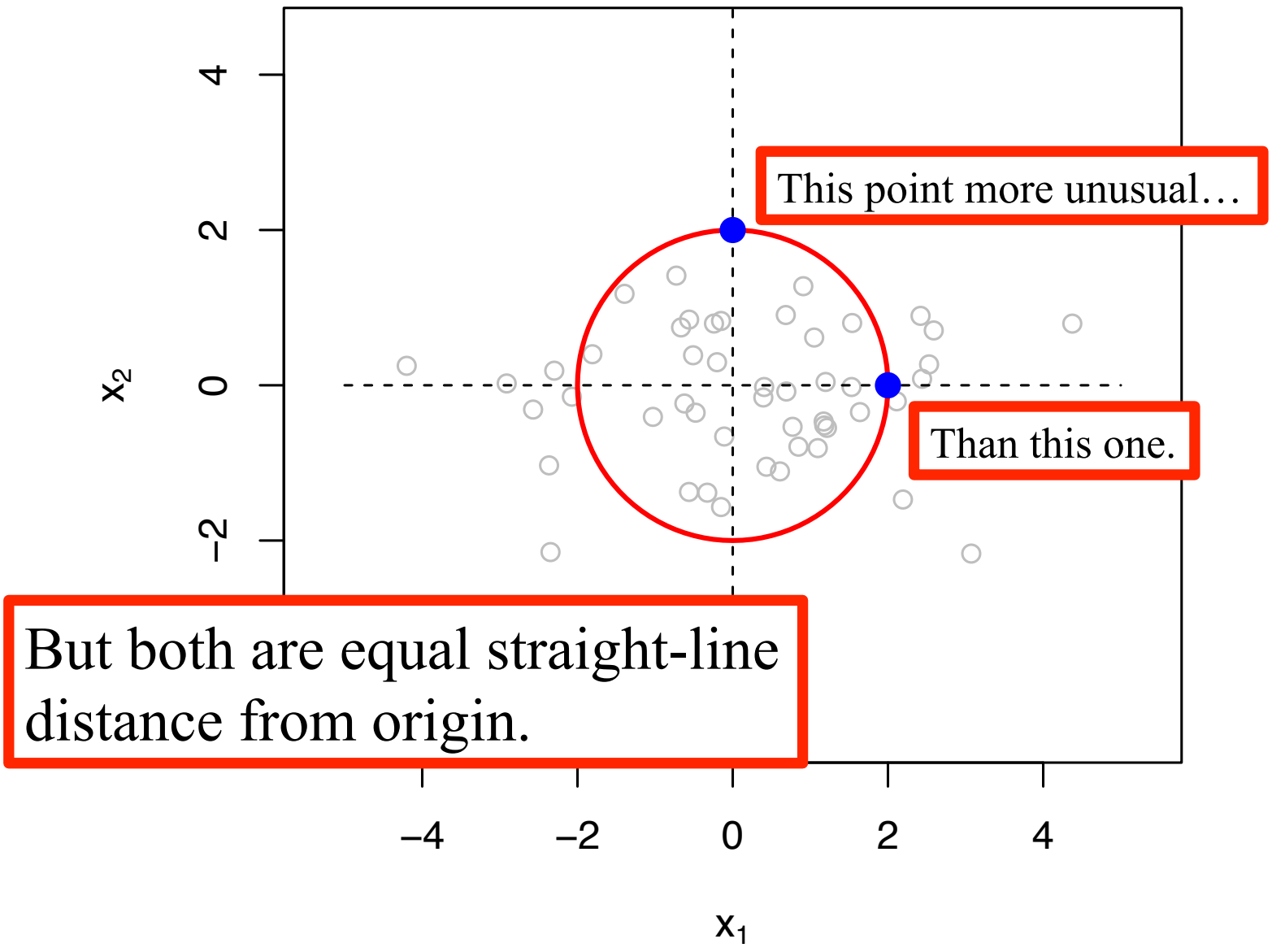
$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

Note that straight-line (Euclidean) distance weights each variable equally and does not accommodate correlation between the variables.


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1


More variability in this direction.
Than in this direction.
x_2
x_1
4
2
0
−2
−4
−4
−2
0
2
4


This point more unusual…
Than this one.
x_2
x_1
4
2
0
−2
−4
−4
−2
0
2
4


This point more unusual…
Than this one.
But both are equal straight-line distance from origin.
x2
x1
4
2
0
−2
−4
−2
0
2
4

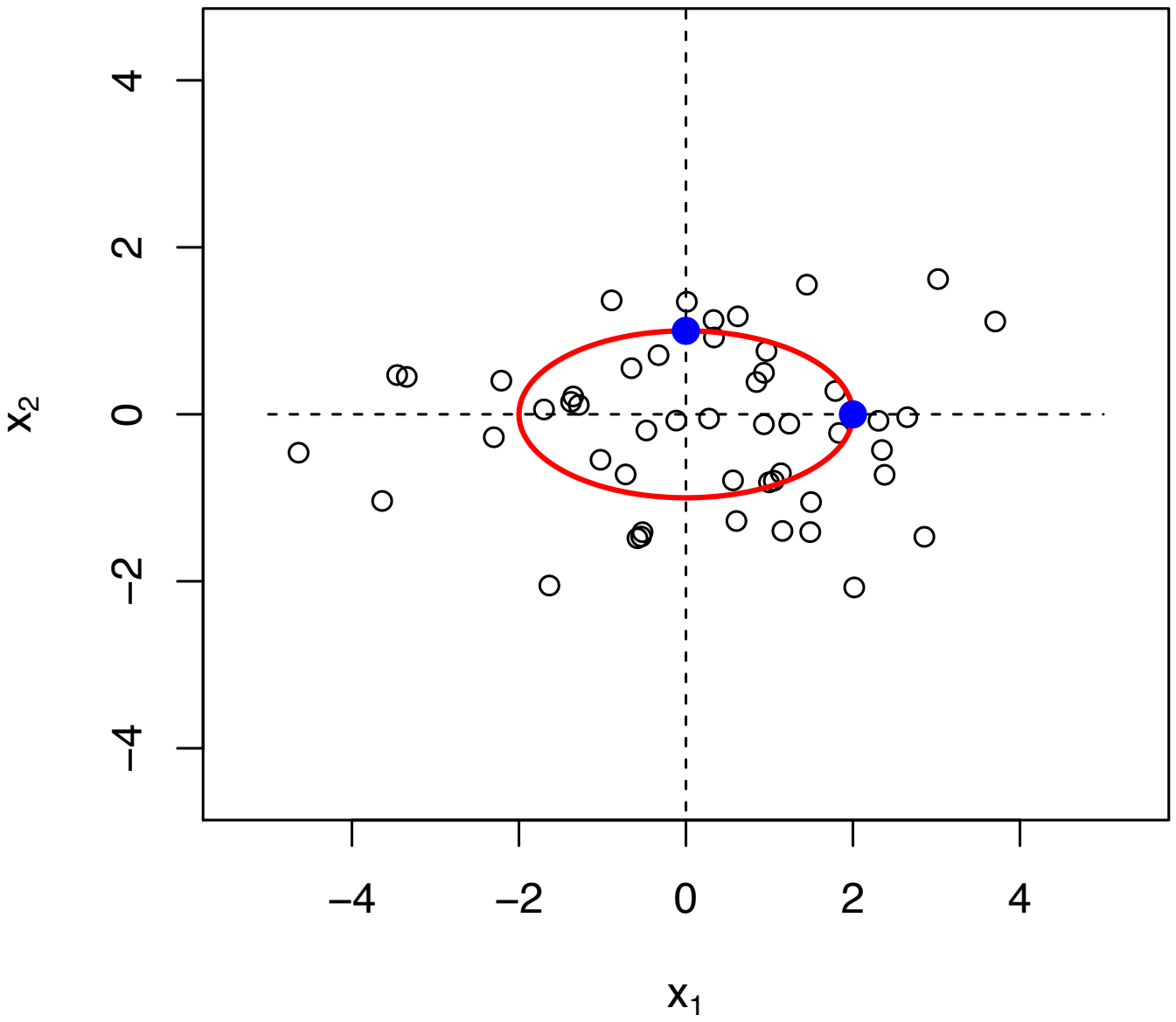
In previous example, x_1 was more variable than x_2. Seems reasonable to weight x_2 more heavily than x_1 when computing a distance to origin. One possibility:

$$d(O,P) = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$

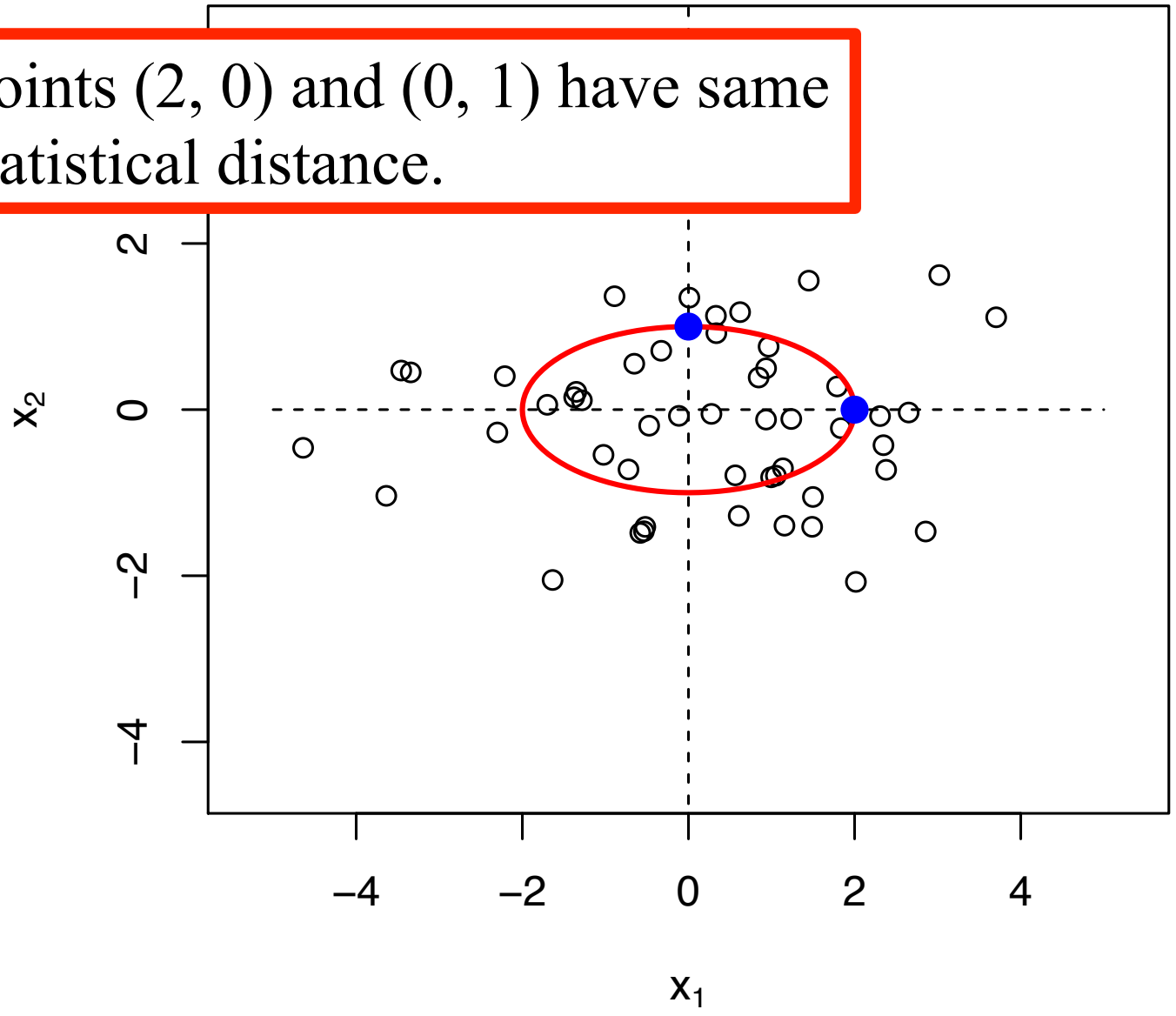
Now all points (x_1, x_2) that lie a constant *statistical* distance c^2 from the origin satisfy

$$\frac{x_{11}^2}{s_{11}} + \frac{x_{22}^2}{s_{22}} = c^2$$

This is now an ellipse, not a circle.


4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1

Points (2, 0) and (0, 1) have same statistical distance.



In previous example, x_1 was more variable than x_2. Seems reasonable to weight x_2 more heavily than x_1 when computing a distance to origin. One possibility:

$$d(O,P) = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$

Now all points (x_1, x_2) that lie a constant *statistical* distance c^2 from the origin satisfy

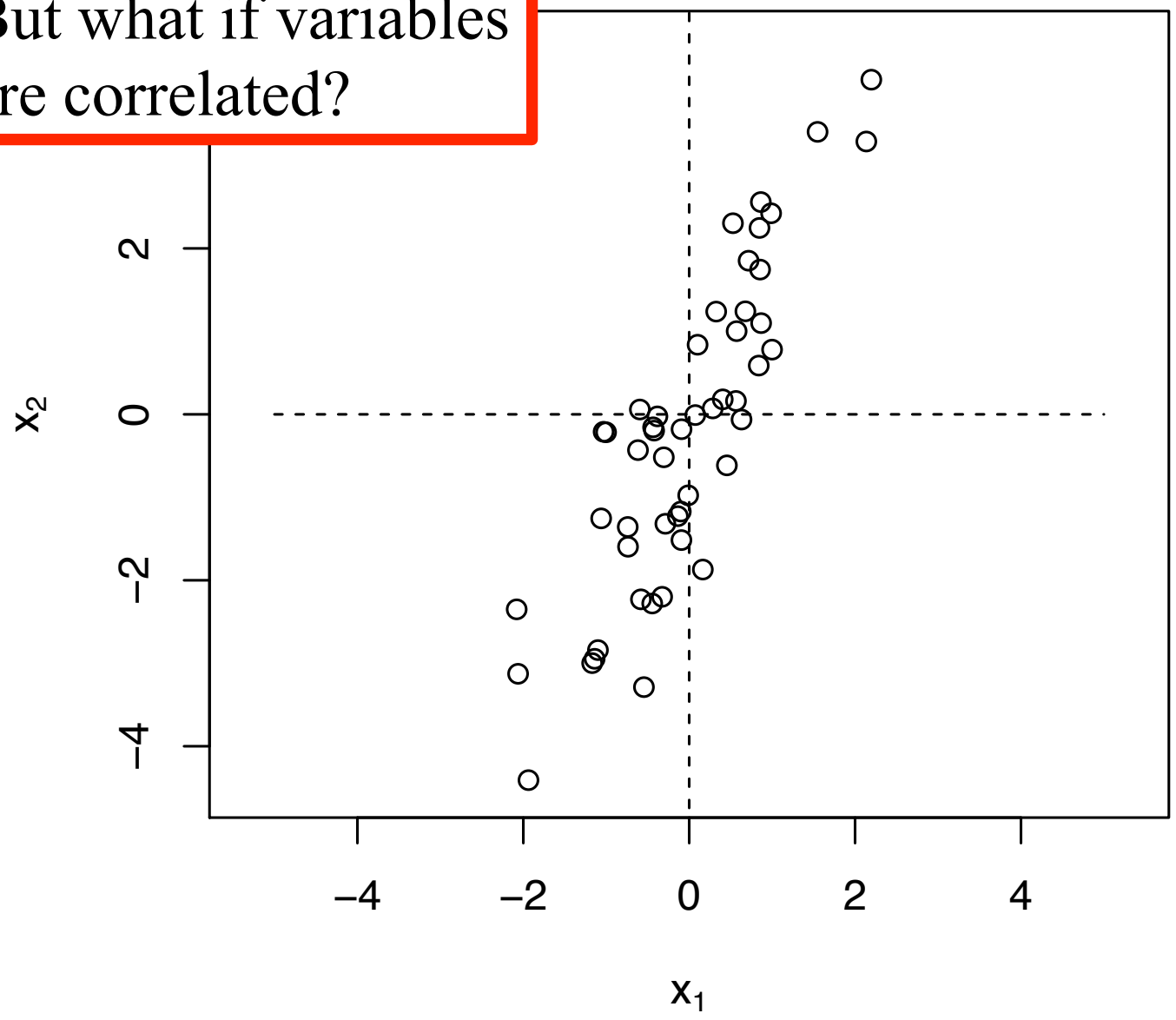
$$\frac{x_{11}^2}{s_{11}} + \frac{x_{22}^2}{s_{22}} = c^2$$

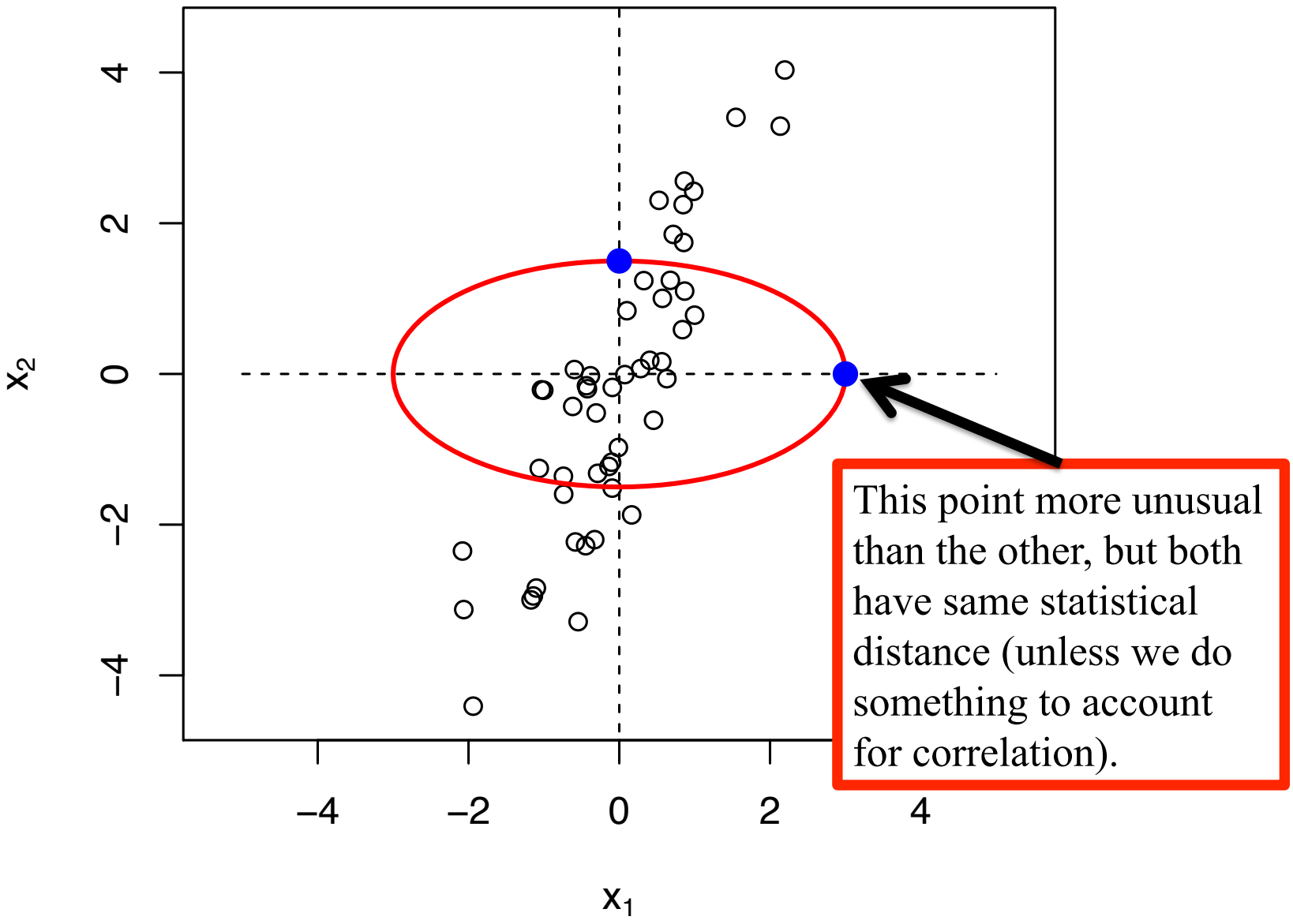
This is now an ellipse, not a circle.

More generally, let $P = (x_1, x_2, \ldots, x_p)$, and let $Q = (y_1, y_2, \ldots, y_p)$ be a *fixed* point (could be origin). Assume the x variables are uncorrelated.

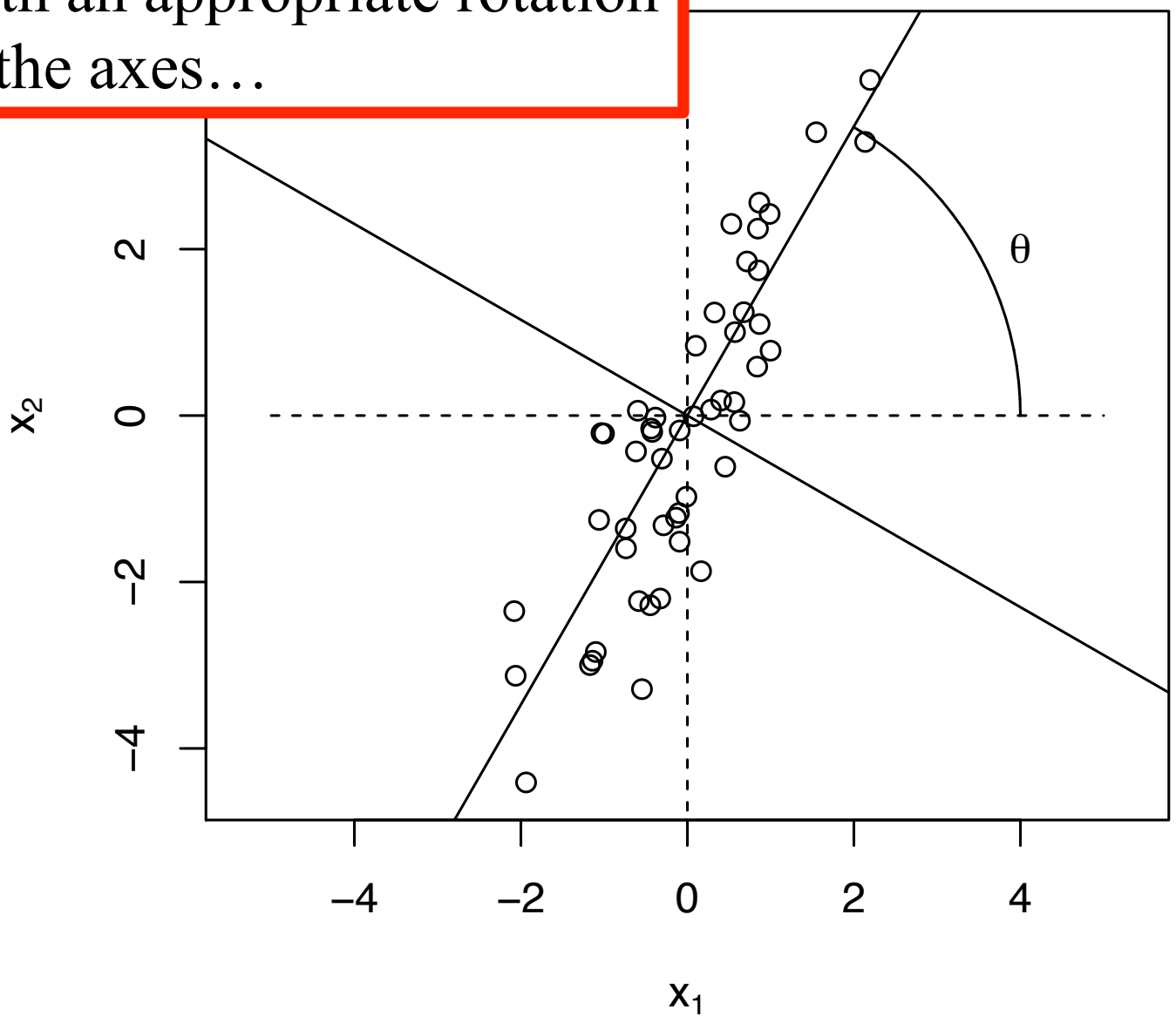
$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}$$

But what if variables are correlated?

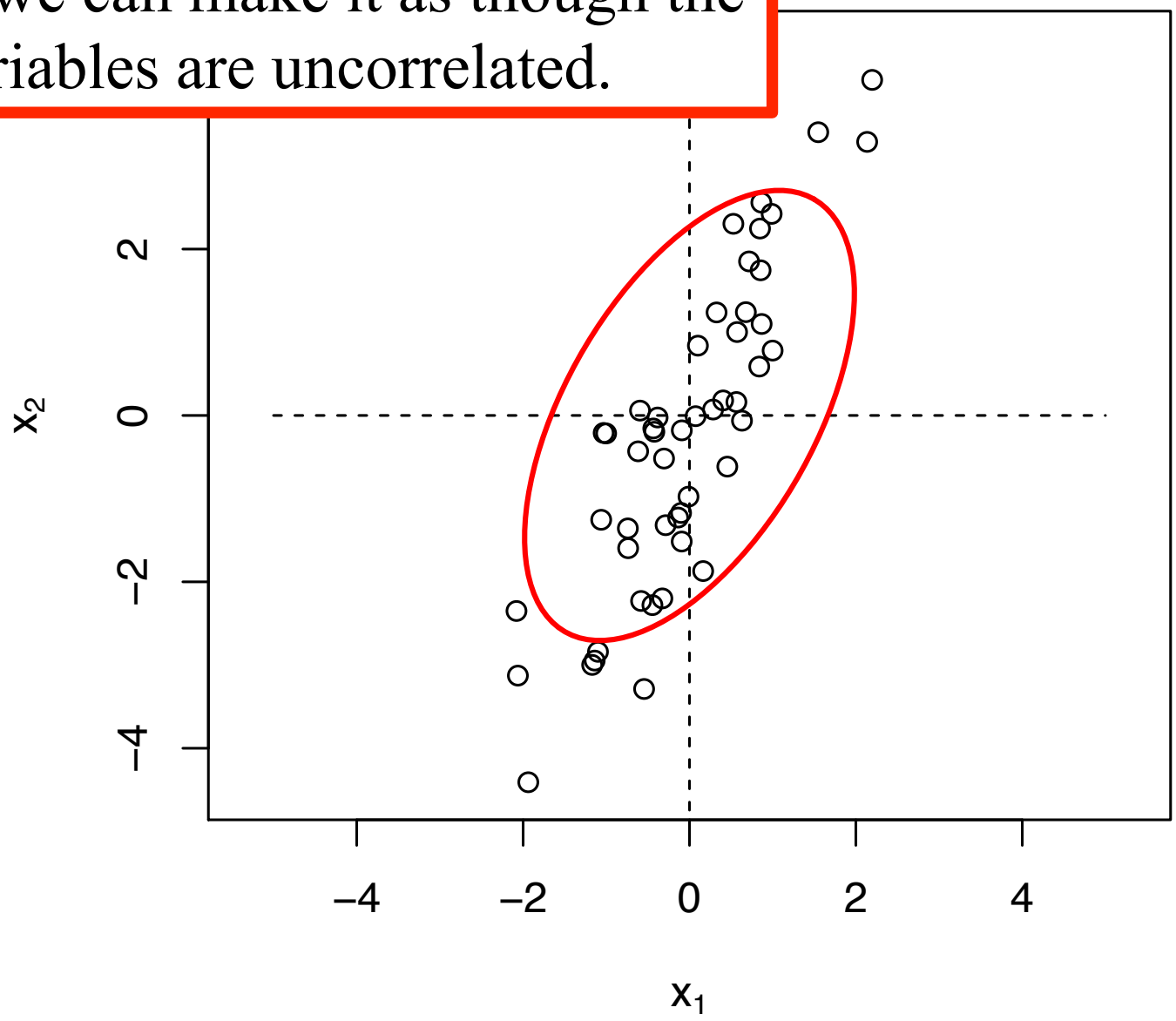



4
2
0
−2
−4
x_2
−4
−2
0
2
4
x_1
This point more unusual than the other, but both have same statistical distance (unless we do something to account for correlation).

With an appropriate rotation of the axes…



…we can make it as though the variables are uncorrelated.



Let θ be the angle of rotation from the previous figure. We can obtain rotated versions of our variable coordinates by

$$\widetilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$
$$\widetilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$$

Let $\widetilde{s}_{11}$ and $\widetilde{s}_{22}$ be the sample variances of the $\widetilde{x}_1$ and $\widetilde{x}_2$ measurements. Then, the distance to origin becomes

$$d(O, P) = \sqrt{\frac{\widetilde{x}_1^2}{\widetilde{s}_{11}} + \frac{\widetilde{x}_2^2}{\widetilde{s}_{22}}}$$

And there are constants a_{11}, a_{12}, a_{22} such that we can write

$$d(O, P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

Let θ be the angle of rotation from the previous figure. We can obtain rotated versions of our variable coordinates by

$$\widetilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$
$$\widetilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$$

Let $\widetilde{s}_{11}$ and $\widetilde{s}_{22}$ be the sample variances of the $\widetilde{x}_1$ and $\widetilde{x}_2$ measurements. Then, the distance to origin becomes

$$d(O, P) = \sqrt{\frac{\widetilde{x}_1^2}{\widetilde{s}_{11}} + \frac{\widetilde{x}_2^2}{\widetilde{s}_{22}}}$$

And there are constants a_{11}, a_{12}, a_{22} such that we can write

$$d(O, P) = \sqrt{a_{11}x_1^2 + \boxed{2a_{12}x_1x_2} + a_{22}x_2^2}$$

A new cross product term

Let θ be the angle of rotation from the previous figure. We can obtain rotated versions of our variable coordinates by

$$\widetilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$
$$\widetilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$$

Let $\widetilde{s}_{11}$ and $\widetilde{s}_{22}$ be the sample variances of the $\widetilde{x}_1$ and $\widetilde{x}_2$ measurements. Then, the distance to origin becomes

$$d(O, P) = \sqrt{\frac{\widetilde{x}_1^2}{\widetilde{s}_{11}} + \frac{\widetilde{x}_2^2}{\widetilde{s}_{22}}}$$

And there are constants a_{11}, a_{12}, a_{22} such that we can write

$$d(O, P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

Similarly, for a fixed point $Q = (y_1, y_2)$

$$d(P, Q) = \sqrt{a_{11}(x_1 - y_y)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$