CEU
Universidad
San Pablo

Multivariate Data Analysis

Session 0: Course outline

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid

Motivation for this course

http://lib.stat.cmu.edu/datasets/Plasma_Retinol - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://lib.stat.cmu.edu/datasets/Plasma_Retinol

Determinants of Plasma Retinol and Beta-Carotene Levels

Summary:
Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids migh
Plasma concentrations of the micronutrients varied widely from subject to subject. While plasma retinol levels varied by age and sex
We conclude that there is wide variability in plasma concentrations of these micronutrients in humans, and that much of this variabil

Authorization: Contact Authors

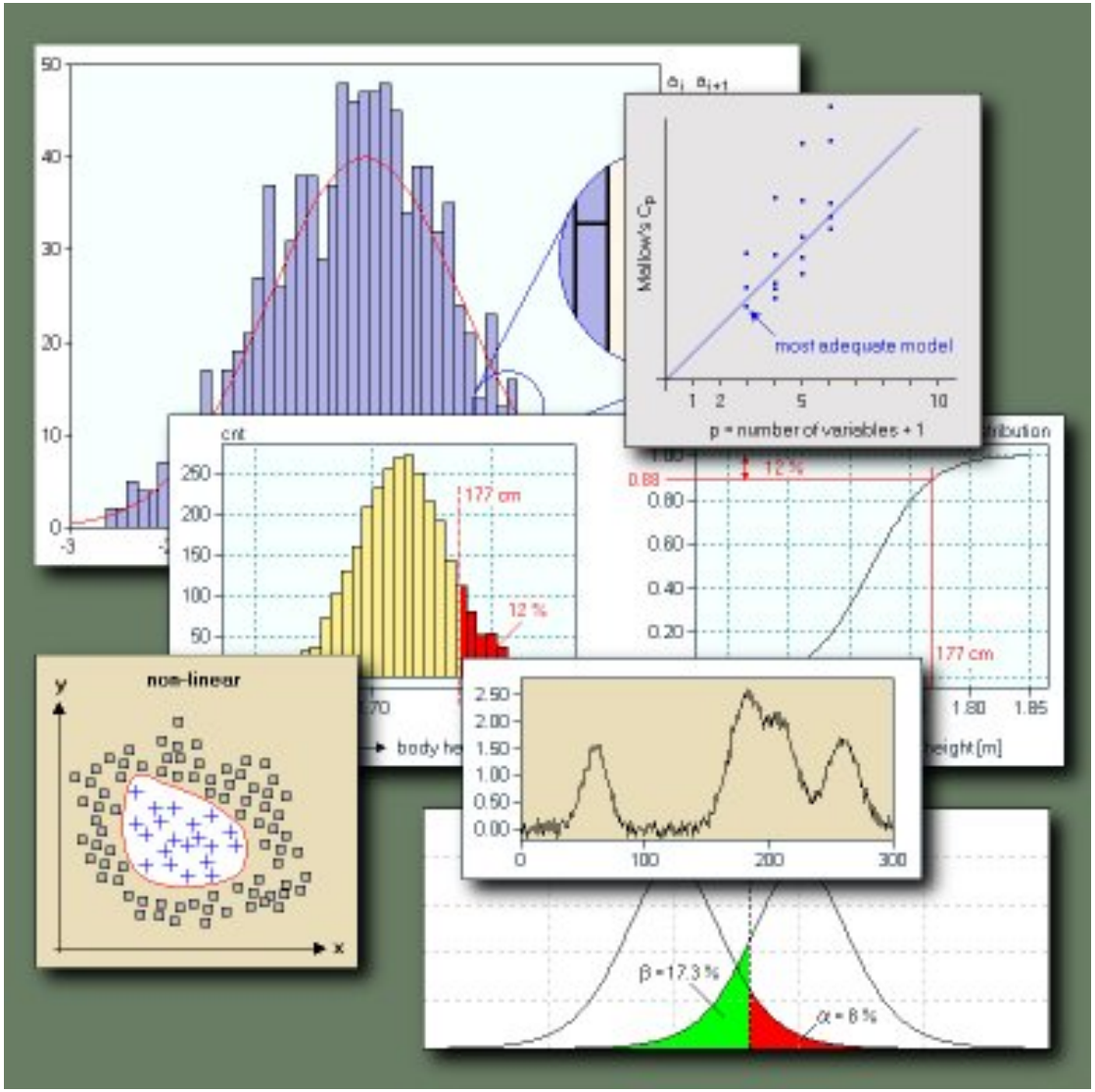
Reference: These data have not been published yet but a related reference is
Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. American Journal of E

Description: This datafile contains 315 observations on 14 variables. This data set can be used to demonstrate multiple regression, transfor

Variable Names in order from left to right:
AGE: Age (years)
SEX: Sex (1=Male, 2=Female).
SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
QUETELET: Quetelet (weight/(height^2))
VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
CALORIES: Number of calories consumed per day.
FAT: Grams of fat consumed per day.
FIBER: Grams of fiber consumed per day.
ALCOHOL: Number of alcoholic drinks consumed per week.
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

64	2	2	21.4838	1	1298.8	57	6.3	0	170.3	1945	890	200	915
76	2	1	23.87631	1	1032.5	50.1	15.8	0	75.8	2653	451	124	727
38	2	2	20.0108	2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562

Motivation for this course



Course outline

Course outline: Session 1

1. Introduction
    - 1.1. Types of variables
    - 1.2. Types of analysis and technique selection
    - 1.3. Descriptors (mean, covariance matrix)
    - 1.4. Variability and distance
    - 1.5. Linear dependence

2. Data Examination
    - 2.1. Graphical examination
    - 2.2. Missing Data
    - 2.3. Outliers
    - 2.4. Assumptions of multivariate analysis

Course outline: Session 2

3. Principal component analysis (PCA)
    - 3.1. Introduction
    - 3.2. Component computation
    - 3.3. Example
    - 3.4. Properties
    - 3.5. Extensions
    - 3.6. Relationship to SVD

4. Factor Analysis (FA)
    - 4.1. Introduction
    - 4.2. Factor computation
    - 4.3. Example
    - 4.4. Extensions
    - 4.5. Rules of thumb
    - 4.6. Comparison with PCA

Course outline: Session 3

5. Multidimensional Scaling (MDS)
    - 5.1. Introduction
    - 5.2. Metric scaling
    - 5.3. Example
    - 5.4. Nonmetric scaling
    - 5.5. Extensions
6. Correspondence analysis
    - 6.1. Introduction
    - 6.2. Projection search
    - 6.3. Example
    - 6.4. Extensions
7. Tensor analysis
    - 7.1 Introduction
    - 7.2 Parafac/Candecomp
    - 7.3 Example
    - 7.4 Extensions

Course outline: Session 4

8. Multivariate Analysis of Variance (MANOVA)
    - 8.1. Introduction
    - 8.2. Computations (1-way)
    - 8.3. Computations (2-way)
    - 8.4. Post-hoc tests
    - 8.5. Example
9. Canonical Correlation Analysis (CCA)
    - 9.1. Introduction
    - 9.2. Construction of the canonical variables
    - 9.3. Example
    - 9.4. Extensions
10. Latent Class Analysis (LCA)
    - 10.1. Introduction

Course Outline

Theory	Theory	Theory	Theory	Practice
Theory	Theory	Theory	Theory	Practice
Theory	Practice	Practice	Practice	Practice

Suggested readings: Overviews

It is suggested to read (before coming):

• H. Abdi. **Multivariate Analysis**. In: Lewis-Beck M., Bryman, A., Futing T. (Eds.) (2003). Encyclopedia of Social Sciences Research Methods. Thousand Oaks (CA): Sage.

• S. Sanogo and X.B. Yang. **Overview of Selected Multivariate Statistical Methods and Their Use in Phytopathological Research**. Phytopathology, 94: 1004-1006 (2004)

Resources

Data sets

http://www.models.kvl.dk/research/data

http://kdd.ics.uci.edu

Links to organizations, events, software, datasets

http://www.statsci.org/index.html

http://astro.u-strasbg.fr/~fmurtagh/mda-sw

http://lib.stat.cmu.edu

Lecture notes

http://www.nickfieller.staff.shef.ac.uk/sheff-only/pas6011-pas370.html

http://www.gseis.ucla.edu/courses/ed231a1/lect.html

Bibliography

- D. Peña. Análisis de datos multivariantes, Mc GrawHill, 2002
- B. Manly. Multivariate statistical methods: a primer. Chapman & Hall/CRC, 2004.
- J. Hair, W. Black, B. Babin, R. Anderson. Multivariate Data Analysis (6th ed), Prentice Hall, 2005.
- B. Everitt, G. Dunn. Applied multivariate data analysis. Hodder Arnold, 2001
- N. H. Timm. Applied multivariate analysis. Springer, 2004
- L. S. Meyers, G. C. Gamst, A. Guarino. Applied multivariate research: design and interpretation. Sage, 2005
- J. L. Schafer. Analysis of incomplete multivariate data. Chapman & Hall/CRC, 1997
- M. Bilodeau, D. Brenner. Theory of multivariate statistics. Springer, 2006

CEU
Universidad
San Pablo

Multivariate Data Analysis

Session 1: Introduction and data examination

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid

Course outline: Session 1

1. Introduction
   - 1.1. Types of variables
   - 1.2. Types of analysis and technique selection
   - 1.3. Descriptors (mean, covariance matrix)
   - 1.4. Variability and distance
   - 1.5. Linear dependence

2. Data Examination
   - 2.1. Graphical examination
   - 2.2. Missing Data
   - 2.3. Outliers
   - 2.4. Assumptions of multivariate analysis

1. Introduction

http://lib.stat.cmu.edu/datasets/Plasma_Retinol - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://lib.stat.cmu.edu/datasets/Plasma_Retinol

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. Ame

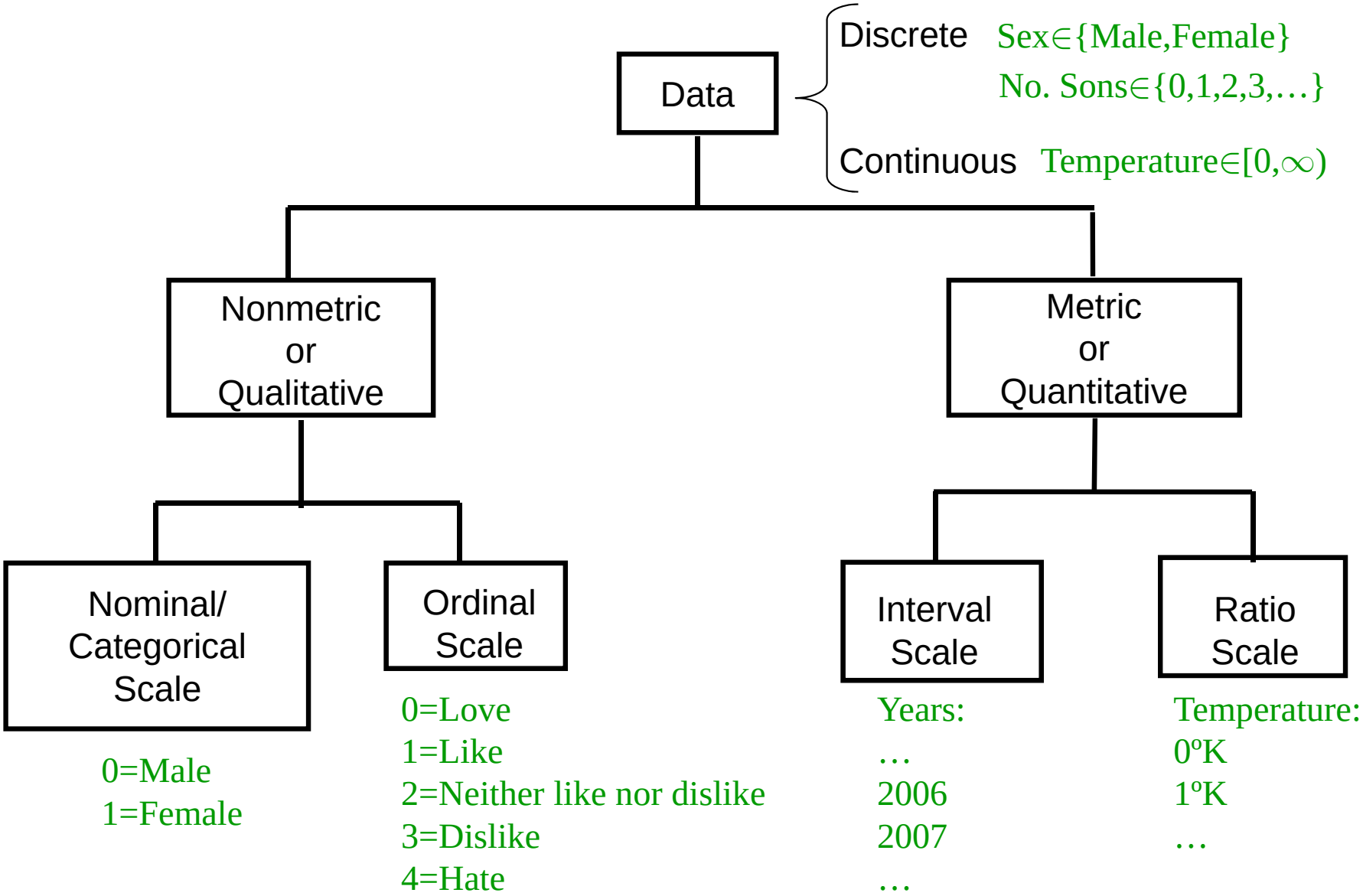
Description: This datafile contains 315 observations on 14 variables. This data set can be used to demonstrate multiple re

Variable Names in order from left to right:

AGE: Age (years)
SEX: Sex (1=Male, 2=Female).
SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
QUETELET: Quetelet (weight/(height^2))
VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
CALORIES: Number of calories consumed per day.
FAT: Grams of fat consumed per day.
FIBER: Grams of fiber consumed per day.
ALCOHOL: Number of alcoholic drinks consumed per week.
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

64	2	2	21.4838	1	1298.8	57	6.3	0	170.3	1945	890	200	915
76	2	1	23.87631	1	1032.5	50.1	15.8	0	75.8	2653	451	124	727
38	2	2	20.0108	2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	20.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741
57	1	1	31.73039	3	2072.9	106.7	9.6	0.9	420	1982	1105	120	679
66	2	1	21.78854	1	987.5	35.6	10.3	0	254.9	2120	1047	61	507
66	1	1	27.31916	3	1574.3	75	7.1	0	361.5	1388	980	108	852
64	1	2	31.44674	3	2868.5	128.8	15	20	379.5	3888	1545	211	1249

1.1 Introduction: Types of variables



1.1 Introduction: Types of variables

<u>Coding of categorical variables</u>

Hair Colour
{Brown, Blond, Black, Red}

$$\xrightarrow{\text{No order}} \left(x_{Brown}, x_{Blond}, x_{Black}, x_{Red}\right) \in \{0,1\}^4$$

Peter: Black
Molly: Blond
Charles: Brown

Peter: $\{0,0,1,0\}$
Molly: $\{0,1,0,0\}$
Charles: $\{1,0,0,0\}$

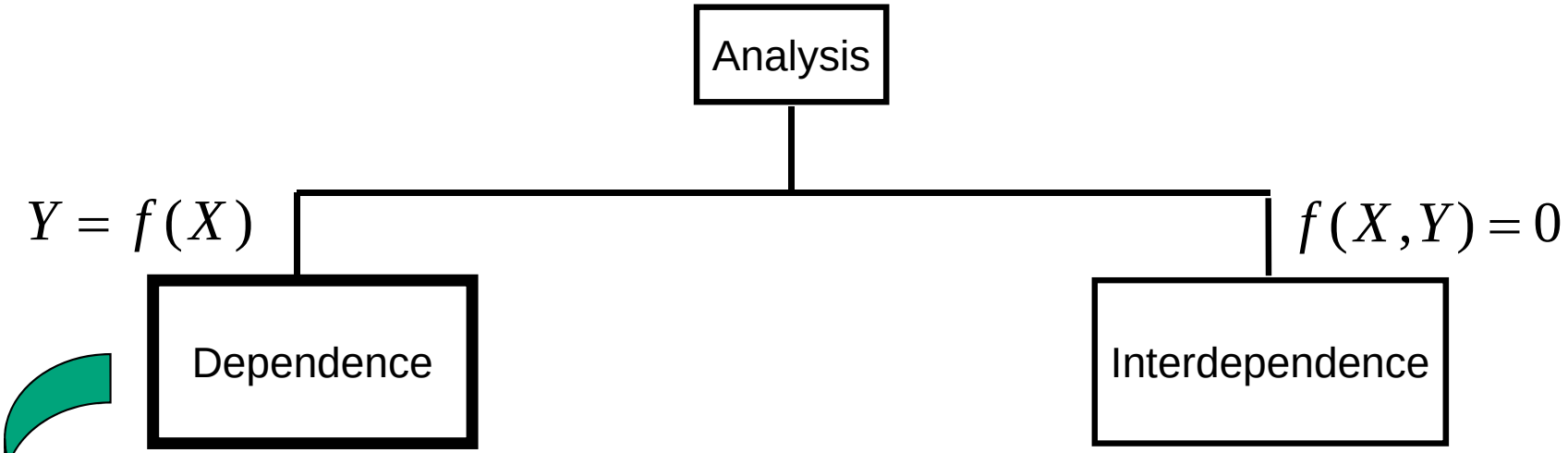
Company size
{Small, Medium, Big}

$$\xrightarrow{\text{Implicit order}} x_{size} \in \{0,1,2\}$$

Company A: Big
Company B: Small
Company C: Medium

Company A: 2
Company B: 0
Company C: 1

1.2 Introduction: Types of analysis



A variable or set of variables is identified as the dependent variable to be predicted or explained by other variables known as independent variables.

Example:
(No. Sons, House Type)=
f(Income, Social Status, Studies)

- Multiple Discriminant Analysis
- Logit/Logistic Regression
- Multivariate Analysis of Variance (MANOVA) and Covariance
- Conjoint Analysis
- Canonical Correlation
- Multiple Regression
- Structural Equations Modeling (SEM)

1.2 Introduction: Types of analysis

Analysis

$Y = f(X)$ — Dependence

$f(X,Y) = 0$ — Interdependence

Example: Who is similar to whom?
(No. Sons, House Type, Income, Social Status, Studies, …)

Involves the simultaneous analysis of all variables in the set, without distinction between dependent variables and independent variables.

- Principal Components and Common Factor Analysis
- Cluster Analysis
- Multidimensional Scaling (perceptual mapping)
- Correspondence Analysis
- Canonical Correlation

1.2 Introduction: Technique selection

- Multiple regression: a single metric variable is predicted by several metric variables.

Example:
No. Sons=f(Income, No. Years working)

- Structural Equation Modelling: several metric variables are predicted by several metric (known and latent) variables

Example:
(No. Sons, House m^2)=f(Income, No. Years working, (No. Years Married))

1.2 Introduction: Technique selection

- Multiple Analysis of Variance (MANOVA): Several metric variables are predicted by several categorical variables.

Example:
(Ability in Math, Ability in Physics)=f(Math textbook, Physics textbook, College)

- Discriminant analysis, Logistic regression: a single categorical (usually two-valued) variable is predicted by several metric independent variables

Example:
Purchaser (or non purchaser)=f(Income,No. Years working)

1.2 Introduction: Technique selection

- Canonical correlation: Several metric variables are predicted by several metric variables

Example:
(Grade Chemistry, Grade Physics)=f(Grade Math, Grade Latin)

- Conjoint Analysis: An ordinal variable (utility function) is predicted by several categorical/ordinal/metric variables

Example:
TV utility=f(Screen format, Screen size, Brand, Price)

- Classification Analysis: Predict categorical variable from several metric variables.

Example:
HouseType=f(Income,Studies)

1.2 Introduction: Technique selection

- Factor analysis/Principal Component Analysis: explain the variability of a set of observed metric variables as a function of unobserved variables (factors)

Example:
(Grade Math, Grade Latin, Grade Physics)=f(Intelligence, Maturity)

- Correspondence analysis: similar to factor analysis but with categorical data.

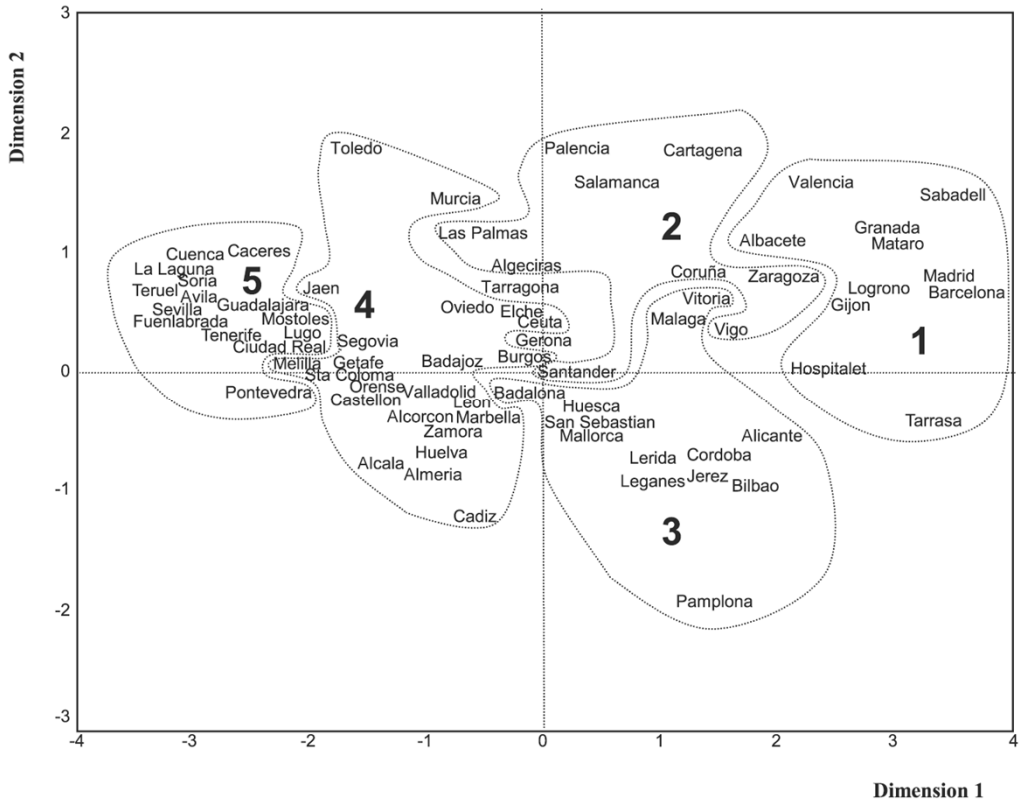
Example:
(Eye colour, Hair colour, Skin colour)=f(gene A, gene B)

- Cluster analysis: try to group individuals according to similar characteristics

Example:
(Grade Math, Grade Latin, Grade Physics, Grade Philosophy, Grade History)

1.2 Introduction: Technique selection

- Multidimensional scaling: Find representative factors so that the relative dissimilarities in the original space are as conserved as possible



Example:
(x,y)=f(City gross income, health indexes, population, political stability, …)

(Basic vector and matrix algebra)

http://lib.stat.cmu.edu/datasets/Plasma_Retinol - Microsoft Internet Explorer

Address http://lib.stat.cmu.edu/datasets/Plasma_Retinol

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. Ame

Description: This datafile contains 315 observations on 14 variables. This data set can be used to demonstrate multiple re

Variable Names in order from left to right:

- AGE: Age (years)
- SEX: Sex (1=Male, 2=Female).
- SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
- QUETELET: Quetelet (weight/(height^2))
- VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
- CALORIES: Number of calories consumed per day.
- FAT: Grams of fat consumed per day.
- FIBER: Grams of fiber consumed per day.
- ALCOHOL: Number of alcoholic drinks consumed per week.
- CHOLESTEROL: Cholesterol consumed (mg per day).
- BETADIET: Dietary beta-carotene consumed (mcg per day).
- RETDIET: Dietary retinol consumed (mcg per day)
- BETAPLASMA: Plasma beta-carotene (ng/ml)
- RETPLASMA: Plasma Retinol (ng/ml)

$\mathbf{x}^t$

64	2	2	21.4838 1	1298.8	57	6.3	0	170.3	1945	890	200	915	
76	2	1	23.87631	1	1832.5	58.1	15.8	8	75.8	2653	451	124	727
38	2	2	20.0108 2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721	
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	20.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741
57	1	1	31.73039	3	2072.9	106.7	9.6	0.9	420	1982	1105	120	679
66	2	1	21.78854	1	987.5	35.6	10.3	0	254.9	2120	1047	61	507
66	1	1	27.31916	3	1574.3	75	7.1	0	361.5	1388	980	108	852
64	1	2	31.44674	3	2868.5	128.8	15	20	379.5	3888	1545	211	1249

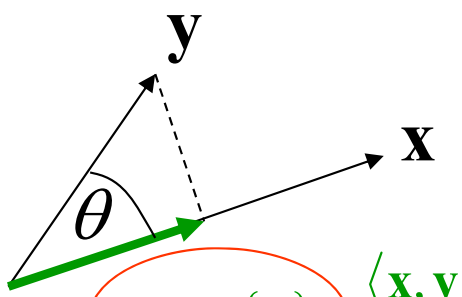
(Basic vector and matrix algebra)

Vector $\mathbf{x}$

Norm $\|x\| \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$

Internal product $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle = \mathbf{x} \cdot \mathbf{y} \triangleq \mathbf{x}^t \mathbf{y} = \sum_{i=1}^{N} x_i y_i = \|\mathbf{x}\| \|\mathbf{y}\| \cos\theta$

Dot product



$$\mathrm{Proj}_{\mathbf{x}}\{\mathbf{y}\} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|} \mathbf{u}_x = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|^2} \mathbf{x} = \mathbf{x}(\mathbf{x}^t \mathbf{x})^{-1} \mathbf{x}^t y$$

Orthogonality $\mathbf{x} \perp \mathbf{y} \Leftrightarrow \langle \mathbf{x}, \mathbf{y} \rangle = 0$

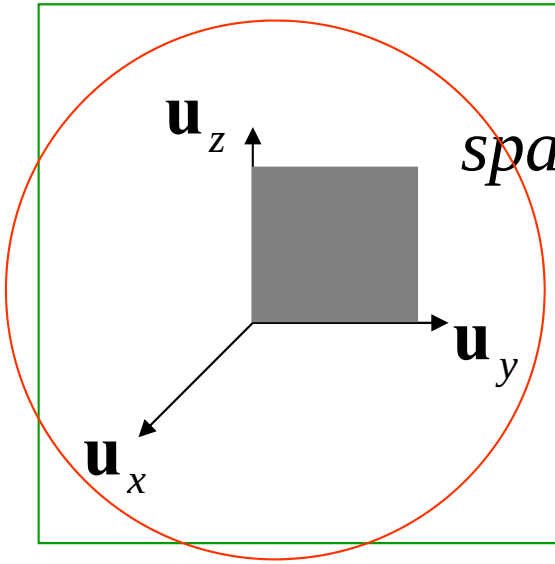
(Basic vector and matrix algebra)

Linear span

field

$$span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\} = \{\mathbf{x} = \underbrace{\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \ldots + \lambda_r\mathbf{x}_r}_{\text{Linearly dependent, i.e.,}} \mid \lambda_1, \lambda_2, \ldots, \lambda_r \in K\}$$

$$\mu_0\mathbf{x} + \mu_1\mathbf{x}_1 + \mu_2\mathbf{x}_2 + \ldots + \mu_r\mathbf{x}_r = \mathbf{0}$$



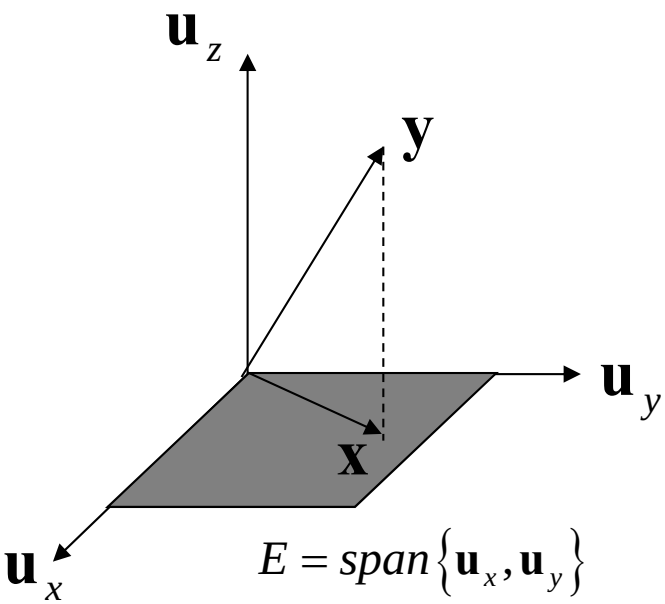
$$span\{\mathbf{u}_y, \mathbf{u}_z\}$$

$$span\{\mathbf{u}_x\} \perp span\{\mathbf{u}_y, \mathbf{u}_z\}$$

$$span\{\mathbf{u}_x\} = C^{\perp}\{span\{\mathbf{u}_y, \mathbf{u}_z\}\}$$

Complementary spaces

(Basic vector and matrix algebra)



Assuming that $\{\mathbf{u}_x, \mathbf{u}_y\}$ is a basis of the spanned space

$$\mathbf{x} = \mathrm{Proj}_E\{\mathbf{y}\} = \mathrm{Proj}_{\mathbf{u}_x}\{\mathbf{y}\} + \mathrm{Proj}_{\mathbf{u}_y}\{\mathbf{y}\}$$
$$= X(X^t X)^{-1} X^t \mathbf{y}$$

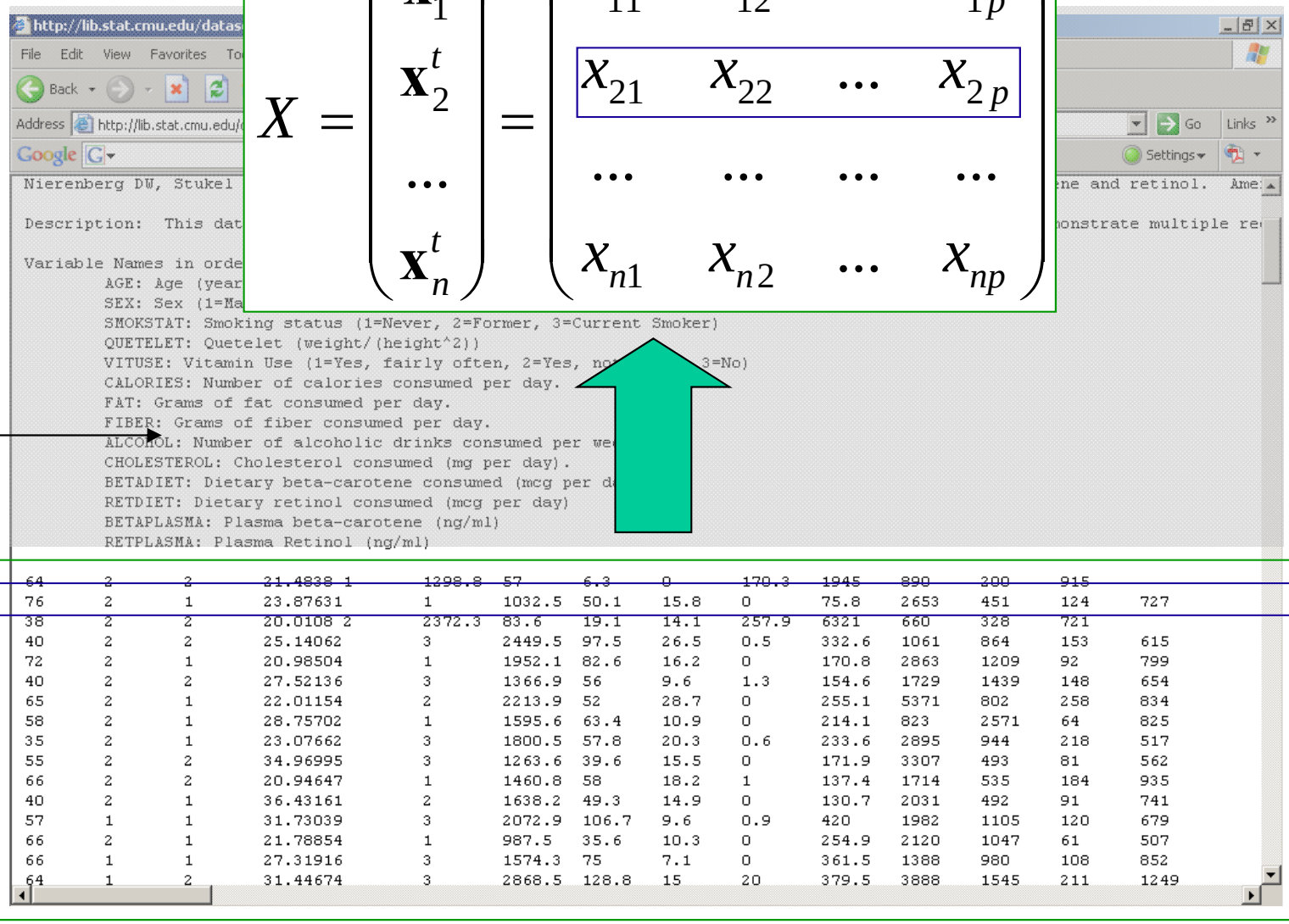
Basis vectors of E as columns

$$\boxed{(\mathbf{y} - \mathbf{x}) \perp E \Rightarrow (\mathbf{y} - \mathbf{x}) \in C^{\perp}\{E\}}$$

$$\boxed{\|\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y} - \mathbf{x}\|^2}$$

1.3 Descriptors: Data representation

$$X = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \dots \\ \mathbf{x}_n^t \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$



Features

Individuals

X

$\mathbf{x}_2^t$

64	2	2	21.4838 1	1298.8	57	6.3	0	170.3	1945	890	200	915	
76	2	1	23.87631	1	1032.5	50.1	15.8	0	75.8	2653	451	124	727
38	2	2	20.0108 2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721	
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	20.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741
57	1	1	31.73039	3	2072.9	106.7	9.6	0.9	420	1982	1105	120	679
66	2	1	21.78854	1	987.5	35.6	10.3	0	254.9	2120	1047	61	507
66	1	1	27.31916	3	1574.3	75	7.1	0	361.5	1388	980	108	852
64	1	2	31.44674	3	2868.5	128.8	15	20	379.5	3888	1545	211	1249

1.3 Descriptors: Univariate analysis

$$X = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \dots \\ \mathbf{x}_n^t \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Sample mean $\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2}$

Sample standard deviation $s_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i2} - \bar{x}_2\right)^2}$

Sample variation coefficient $VC_2 = \sqrt{\frac{\bar{x}_2^2}{s_2^2}}$

If outliers

Robust statistics

m_2 Sample median

$$\Pr\{X_2 \le m_2\} \ge \frac{1}{2} \le \Pr\{X_2 \ge m_2\}$$

MAD_2 Sample Median Absolute Deviation

$$Median\{|x_2 - m_2|\}$$

1.3 Descriptors: Mean and covariance

$$X = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \dots \\ \mathbf{x}_n^t \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \longrightarrow \tilde{X} = \begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^t \\ (\mathbf{x}_2 - \bar{\mathbf{x}})^t \\ \dots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^t \end{pmatrix} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \dots & \dots & \dots & \dots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{pmatrix}$$

Matrix of centered data

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{pmatrix}^t \qquad \tilde{X} = X - \mathbf{1}\bar{\mathbf{x}}^t$$

$$\bar{\mathbf{x}} = \frac{1}{n} X^t \mathbf{1}$$

<u>Sample mean</u>

Vector of 1s

<u>Sample covariance</u>

$$s_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right) \left(x_{ik} - \bar{x}_k \right)$$

Measures how variables j and k are related

Symmetric, positive semidefinite

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t = \frac{1}{n} \tilde{X}^t \tilde{X}$$

$$\Sigma = E\left\{ (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^t \right\}$$

1.3 Descriptors: Covariance

$$X = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}$$ 3 variables
200 samples

$$\sigma_{13} = E\left\{(X_1 - \mu_1)(X_3 - \mu_3)\right\} = E\left\{\tilde{X}_1 \tilde{X}_3\right\}$$

$$X_1 \sim N(0,1) \qquad X_2 \sim N(0,1) \qquad X_3 \sim N(0,1)$$

$$S = \begin{pmatrix} 0.9641 & 0.0678 & -0.0509 \\ 0.0678 & 0.8552 & 0.0398 \\ -0.0509 & 0.0398 & 0.9316 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Sample covariance

Covariance

$$X_1 \sim N(0,1) \qquad X_2 = X_1 \qquad X_3 = -X_1$$

$$S = \begin{pmatrix} 0.9641 & 0.9641 & -0.9641 \\ 0.9641 & 0.9641 & -0.9641 \\ -0.9641 & -0.9641 & 0.9641 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}$$

$$X_1 \sim N(0,9) \qquad X_2 = X_1 \qquad X_3 = -X_1$$

$$S = \begin{pmatrix} 10.4146 & 10.4146 & -10.4146 \\ 10.4146 & 10.4146 & -10.4146 \\ -10.4146 & -10.4146 & 10.4146 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 9 & 9 & -9 \\ 9 & 9 & -9 \\ -9 & -9 & 9 \end{pmatrix}$$

1.3 Descriptors: Covariance

$$X_1 \sim N(1,2)$$
$$X_2 \sim N(2,3)$$
$$X_3 = X_1 - X_2$$

$$S = \begin{pmatrix} 1.6338 & -0.0970 & 1.5368 \\ -0.0970 & 2.8298 & -2.7329 \\ 1.5368 & -2.7329 & 4.2696 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 3 & -3 \\ 2 & -3 & 5 \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 2 \\ 0 & 3 & -3 \\ 2 & -3 & 5 \end{pmatrix} \right)$$

$$X_1 \sim N(\mu_1, \sigma_1^2)$$
$$X_2 \sim N(\mu_2, \sigma_2^2)$$
$$X_3 = a_1 X_1 + a_2 X_2$$

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ a_1\mu_1 + a_2\mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 & a_1\sigma_1^2 \\ 0 & \sigma_2^2 & a_2\sigma_2^2 \\ a_1\sigma_1^2 & a_2\sigma_2^2 & a_1^2\sigma_1^2 + a_2^2\sigma_2^2 \end{pmatrix} \right)$$

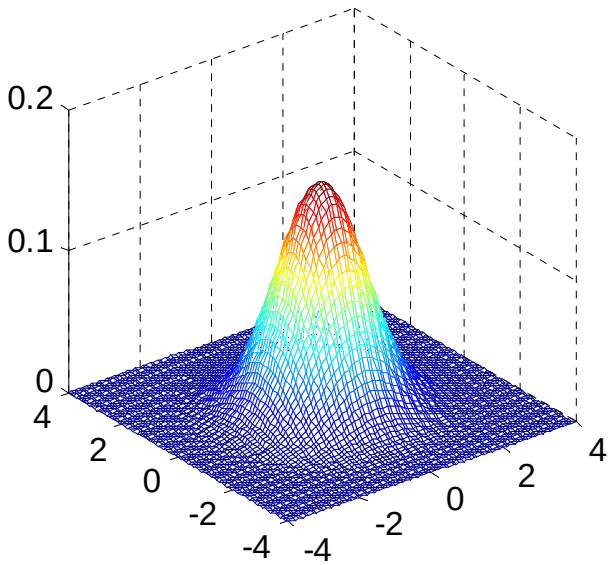
$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\left(2\pi |\Sigma|\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

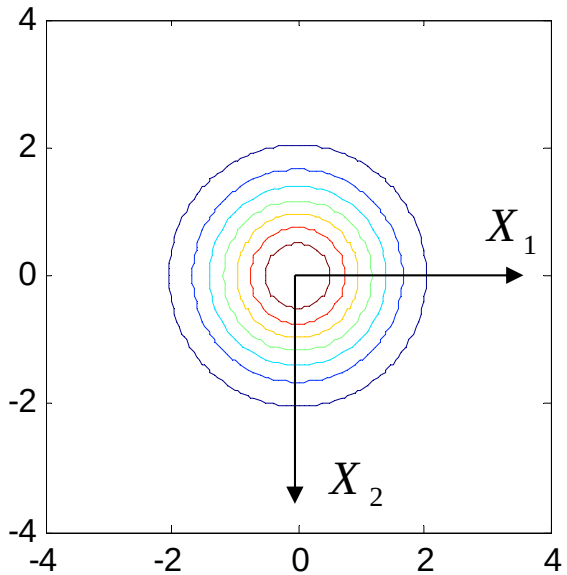
1.3 Descriptors: Covariance

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\left(2\pi|\Sigma|\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$





X_1 and X_2 are independent

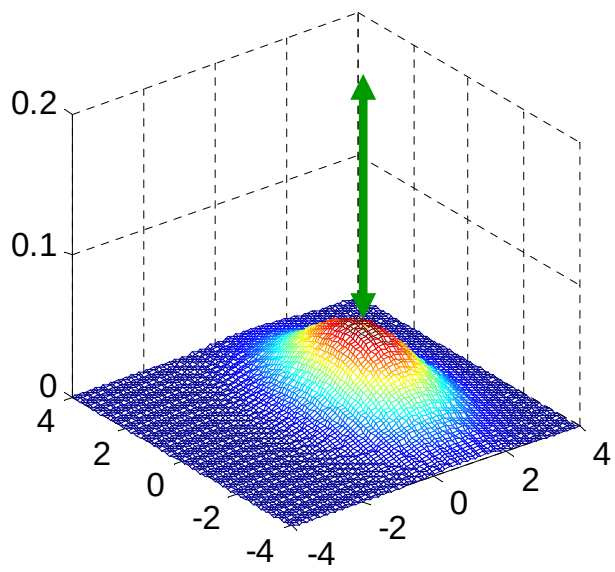
For multivariate Gaussians, covariance=0 implies independency

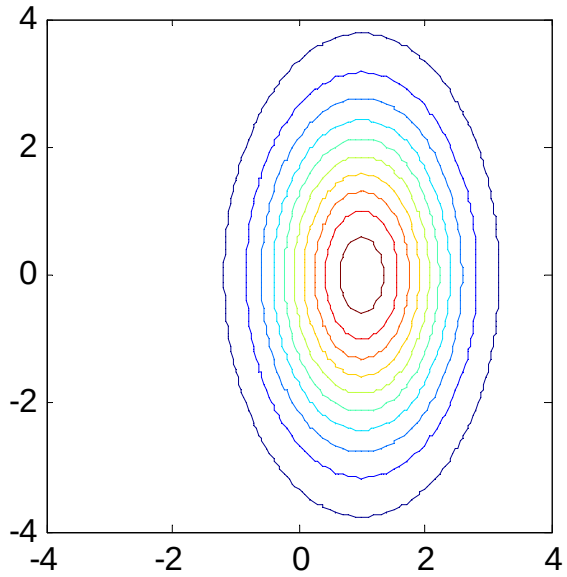
1.3 Descriptors: Covariance

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\left(2\pi|\Sigma|\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$





X_1 and X_2 are independent

1.3 Descriptors: Covariance

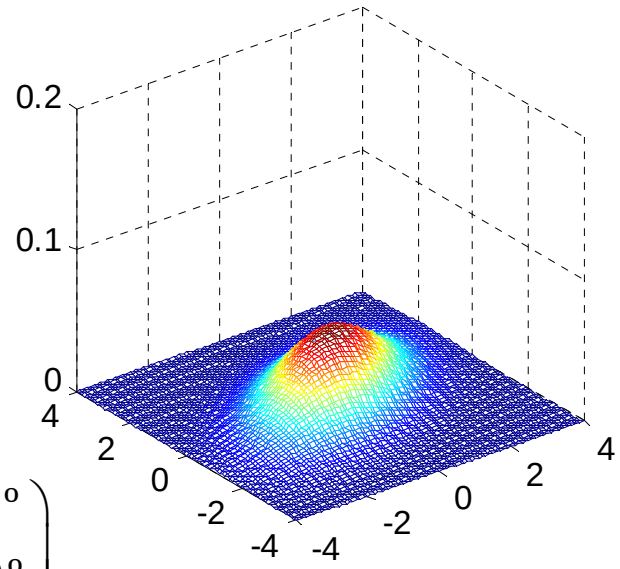
$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\left(2\pi|\Sigma|\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

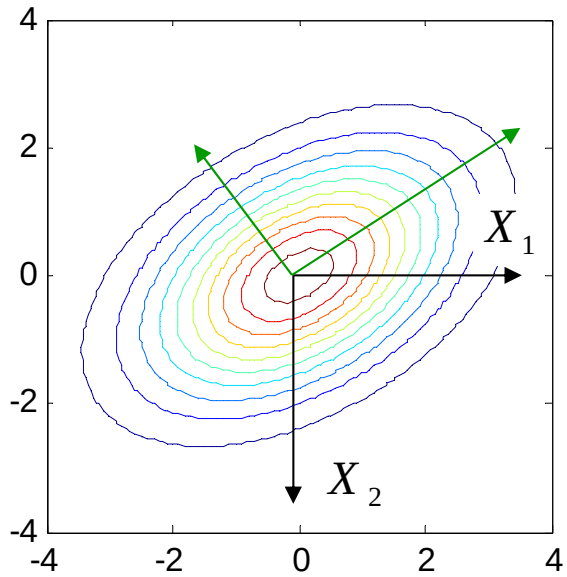
$$\boldsymbol{\mu} = \mathbf{0}$$

$$\Sigma = R \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} R^t$$

$$R = \begin{pmatrix} \cos 60^o & \sin 60^o \\ -\sin 60^o & \cos 60^o \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 2.5 & 0.866 \\ 0.866 & 1.5 \end{pmatrix}$$





X_1 and X_2 are NOT independent

BUT there exist two independent variables

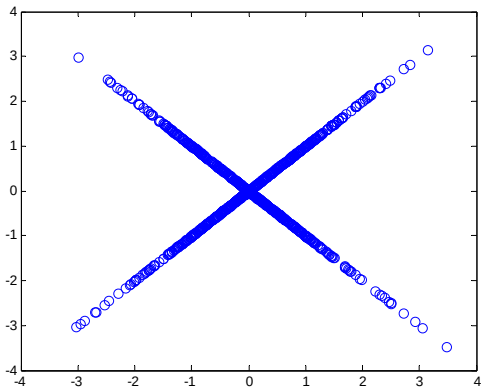
1.3 Descriptors: Covariance

Pitfalls of the covariance matrix

$$X_1 \sim N(0,1)$$

$$X_2 = \begin{cases} X_1 & p = 0.5 \\ -X_1 & 1-p = 0.5 \end{cases}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$X_1 \sim N(0,1)$$

$$X_2 = X_1^2$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\boxed{\begin{array}{l} Cov(X_1, X_2) = 0 \Rightarrow Uncorrelated \not\Rightarrow Independent \\ Cov(X_1, X_2) = 0 \wedge Gaussian \Rightarrow Independent \end{array}}$$

1.3 Descriptors: Covariance

<u>Redundant variables</u>

$$X_1 \sim N(0,1) \qquad X_2 \sim N(0,1) \qquad X_3 \sim N(0,1)$$

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad eig(\Sigma) = (1,1,1)$$

$$X_1 \sim N(0,1) \qquad X_2 = X_1 \qquad X_3 = -X_1$$

$$\Sigma = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \qquad eig(\Sigma) = (1,0,0)$$

$$\Sigma = R \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} R^t \qquad eig(\Sigma) = (3,1)$$

$$X_1 \sim N(1,2) \qquad X_2 \sim N(2,3) \qquad X_3 = X_1 - X_2$$

$$\Sigma = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 3 & -3 \\ 2 & -3 & 5 \end{pmatrix} \qquad eig(\Sigma) = (7.64, 2.35, 0)$$

1.4 Variability and distance

$\mathbf{x}_2^t$

$\mathbf{x}_{10}^t$

CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

64	2	2	21.4838 1	1298.8	57	6.3	0	170.3	1945	890	200	915	
76	2	1	23.87631	1	1032.5	50.1	15.8	0	75.8	2653	451	124	727
38	2	2	20.0108 2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721	
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	30.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741
57	1	1	31.73039	3	2072.9	106.7	9.6	0.9	420	1982	1105	120	679
66	2	1	21.78854	1	987.5	35.6	10.3	0	254.9	2120	1047	61	507
66	1	1	27.31916	3	1574.3	75	7.1	0	361.5	1388	980	108	852
64	1	2	31.44674	3	2868.5	128.8	15	20	379.5	3888	1545	211	1249

How far are they?

How far are they from the mean? $d(\mathbf{x}_i, \mathbf{x}_j)$

1-norm (Manhattan) $$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{n} \left| x_{is} - x_{js} \right|$$

Most used $\longrightarrow$ p-norm (Euclidean p=2) Minkowski $$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{s=1}^{n} \left(x_{is} - x_{js} \right)^p \right)^{\frac{1}{p}}$$

Infinity norm $$d(\mathbf{x}_i, \mathbf{x}_j) = \max_{s} \left| x_{is} - x_{js} \right|$$

1.4 Variability and distance

	Height (m)	Weight (kg)
Juan	1.80	80
John	1.70	72
Jean	1.65	81

$d(\mathbf{x}_i, \mathbf{x}_j)$	Juan	John	Jean
Juan	-----	8.0004	1.0112

	Height (cm)	Weight (kg)
Juan	180	80
John	170	72
Jean	165	81

$d(\mathbf{x}_i, \mathbf{x}_j)$	Juan	John	Jean
Juan	-----	11.3137	15.0333

Matrix-based distance $d^2(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t M^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)$

Euclidean distance $d^2(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t I^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t \left(\mathbf{x}_i - \mathbf{x}_j\right)$

Mahalanobis distance $d^2(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t \Sigma^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)$

Correntropy distance $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{N} k_\sigma \left(x_{ik} - x_{jk}\right)$

1.4 Variability and distance

Mahalanobis distance $d^2(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t \Sigma^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)$

$$\Sigma = \begin{pmatrix} \sigma^2_{height} & r\sigma_{height}\sigma_{weight} \\ r\sigma_{height}\sigma_{weight} & \sigma^2_{weight} \end{pmatrix} = \begin{pmatrix} 100 & 70 \\ 70 & 100 \end{pmatrix}$$

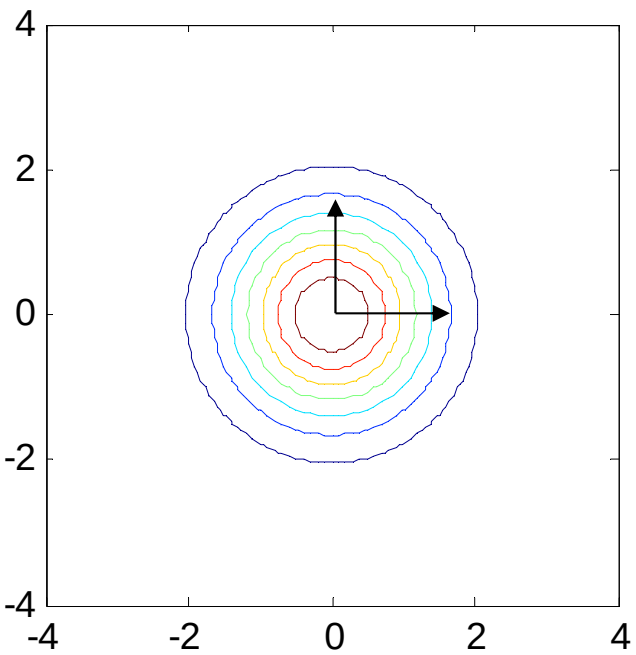
$\sigma_{height} = 10cm$

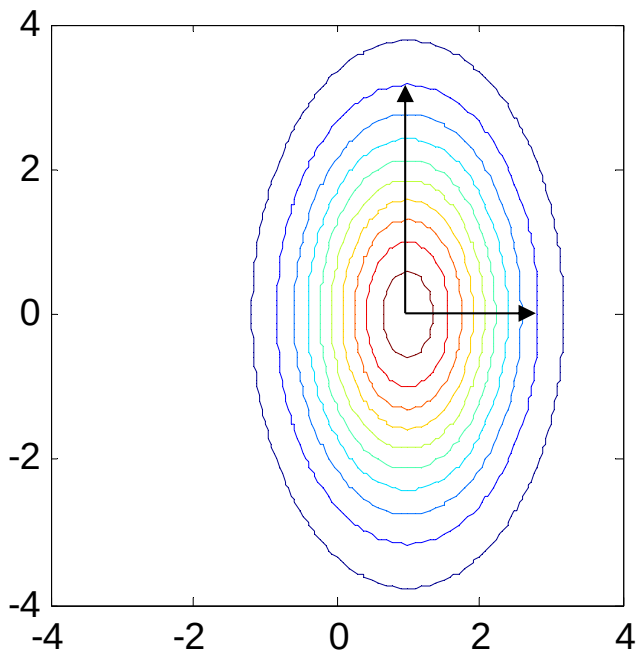
$\sigma_{weight} = 10kg$

$r = 0.7$

$d(\mathbf{x}_i, \mathbf{x}_j)$	Juan	John	Jean
Juan	-----	0.7529	4.8431

Independently of units!!





1.5 Linear dependence: Pair dependence

$$X_1 \sim N(0,1) \qquad X_2 = X_1 \qquad X_3 = -X_1 \qquad S = \begin{pmatrix} 0.9641 & 0.9641 & -0.9641 \\ 0.9641 & 0.9641 & -0.9641 \\ -0.9641 & -0.9641 & 0.9641 \end{pmatrix}$$

$$X_1 \sim N(0,9) \qquad X_2 = X_1 \qquad X_3 = -X_1 \qquad S = \begin{pmatrix} 10.4146 & 10.4146 & -10.4146 \\ 10.4146 & 10.4146 & -10.4146 \\ -10.4146 & -10.4146 & 10.4146 \end{pmatrix}$$

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{pmatrix} \longrightarrow R = \begin{pmatrix} 1 & \frac{s_{12}}{s_1 s_2} & \dots & \frac{s_{1p}}{s_1 s_p} \\ \frac{s_{21}}{s_2 s_1} & 1 & \dots & \frac{s_{2p}}{s_2 s_p} \\ \dots & \dots & \dots & \dots \\ \frac{s_{p1}}{s_p s_1} & \frac{s_{p2}}{s_p s_2} & \dots & 1 \end{pmatrix} = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

$$\boxed{r_{jk} = \frac{s_{jk}}{s_j s_k}}$$

$$D = \begin{pmatrix} s_1^2 & 0 & \dots & 0 \\ 0 & s_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_p^2 \end{pmatrix}$$

$$-1 \le r_{jk} \le 1$$

$$|r_{jk}| = 1 \Rightarrow x_j = a + b x_k$$

r_{jk} Is invariant to linear transformations of the variables

1.5 Linear dependence: Pair dependence

Example:

count =

11	11	9
7	13	11
14	17	20
11	13	9
43	51	69

…

Traffic count in three different places (thousands/day)

covariance=

643	980	1656
980	1714	2690
1656	2690	4627

More traffic?

correlation=

1.0000	0.9331	0.9599
0.9331	1.0000	0.9553
0.9599	0.9553	1.0000

1.5 Linear dependence: Multiple dependence

```
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

64  2  2  21.4838 1  1298.8  57      6.3   0     170.3  1945  890  200  915
76  2  1  23.87631   1       1032.5  50.1  15.8  0      75.8  2653  451  124  727
38  2  2  20.0108 2 2372.3   83.6    19.1  14.1  257.9  6321  660  328  721
40  2  2  25.14062   3       2449.5  97.5  26.5  0.5    332.6 1061  864  153  615
72  2  1  20.98504   1       1952.1  82.6  16.2  0      170.8 2863  1209 92   799
40  2  2  27.52136   3       1366.9  56    9.6   1.3    154.6 1729  1439 148  654
65  2  1  22.01154   2       2213.9  52    28.7  0      255.1 5371  802  258  834
58  2  1  28.75702   1       1595.6  63.4  10.9  0      214.1 823   2571 64   825
```

$$\hat{\tilde{X}}_1 = \beta_2 \tilde{X}_2 + \beta_3 \tilde{X}_3 + \ldots + \beta_p \tilde{X}_p = \boldsymbol{\beta}\tilde{\mathbf{X}}_{2,3,\ldots,p} \longrightarrow \boxed{\boldsymbol{\beta} = S_{2,3,\ldots,p}^{-1} S_1}$$

$$\hat{X}_1 = \bar{X}_1 + \hat{\tilde{X}}_1$$

Multiple correlation coefficient

$$\boxed{R^2_{1.2,3,\ldots p} = \frac{\sum_{i=1}^{n} (\hat{x}_{1i} - \bar{x}_1)^2}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}}$$

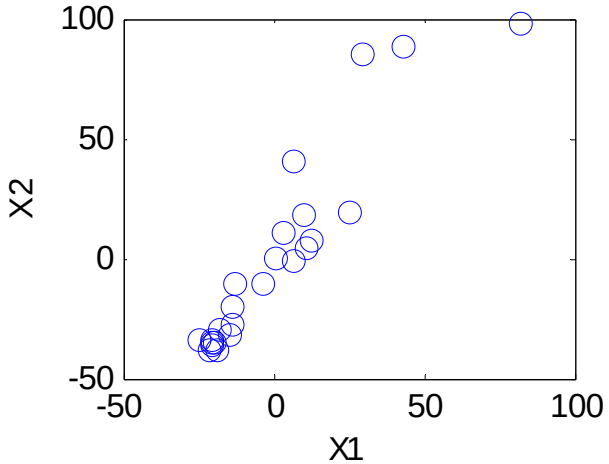
Variance of X_1 explained by a linear prediction

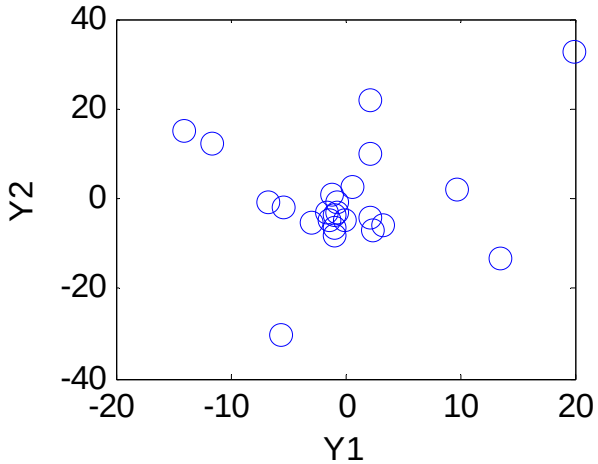
Total variance of X_1

$$S = \begin{pmatrix} s_1^2 & s_{12} & \ldots & s_{1p} \\ s_{21} & s_2^2 & \ldots & s_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ s_{p1} & s_{p2} & \ldots & s_p^2 \end{pmatrix}$$

1.5 Linear dependence: Pair dependence

Example:





$$R_{1.3} = 0.9599 \longrightarrow \text{As seen before}$$

$$R_{1.2,3} = 0.9615$$

Does X2 provide useful information on X1 once the influence of X3 is removed?

$$\hat{\tilde{X}}_1 = \beta_{13}\tilde{X}_3 \rightarrow Y_1 = \tilde{X}_1 - \hat{\tilde{X}}_1$$

$$\hat{\tilde{X}}_2 = \beta_{23}\tilde{X}_3 \rightarrow Y_2 = \tilde{X}_2 - \hat{\tilde{X}}_2$$

$$R_{Y_1.Y_2} = 0.1943 \qquad p_{value} = 0.363$$

No!

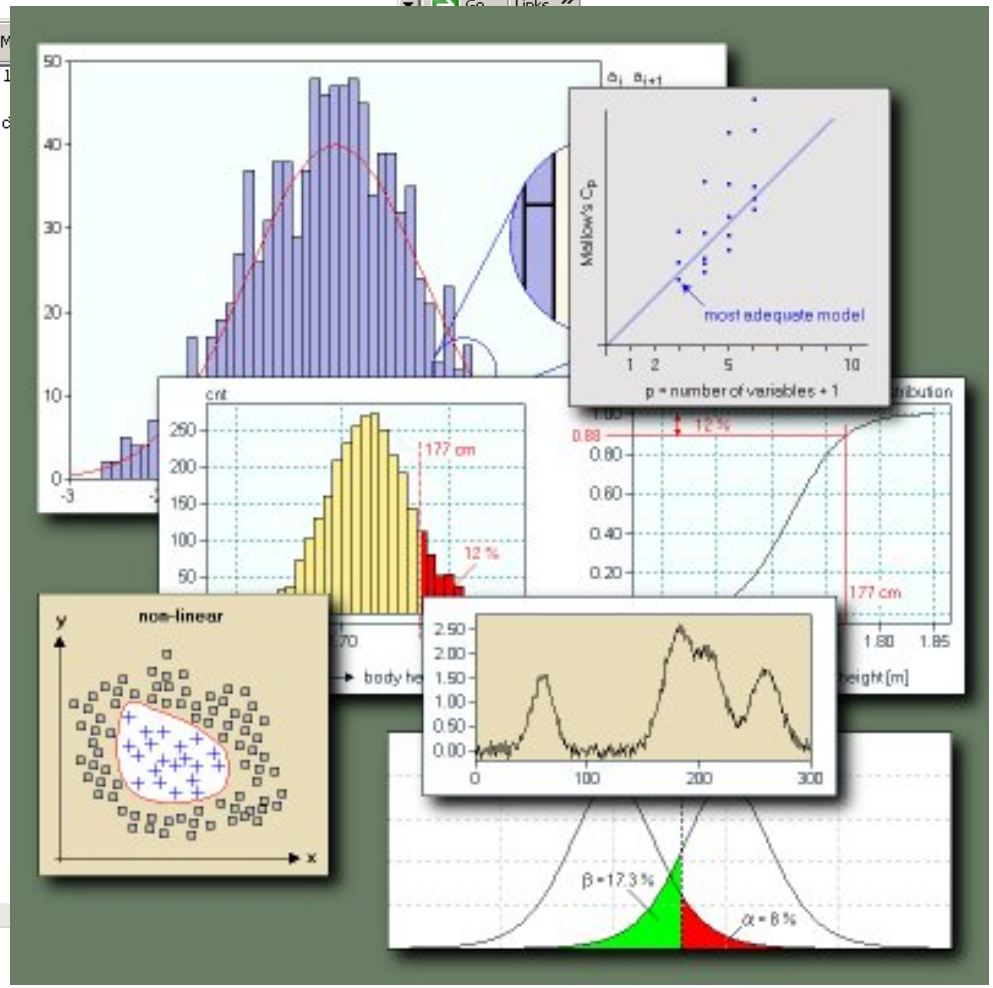
2. Data examination: Get acquainted with your data

http://lib.stat.cmu.edu/datasets/Plasma_Retinol - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://lib.stat.cmu.edu/datasets/Plasma_Retinol

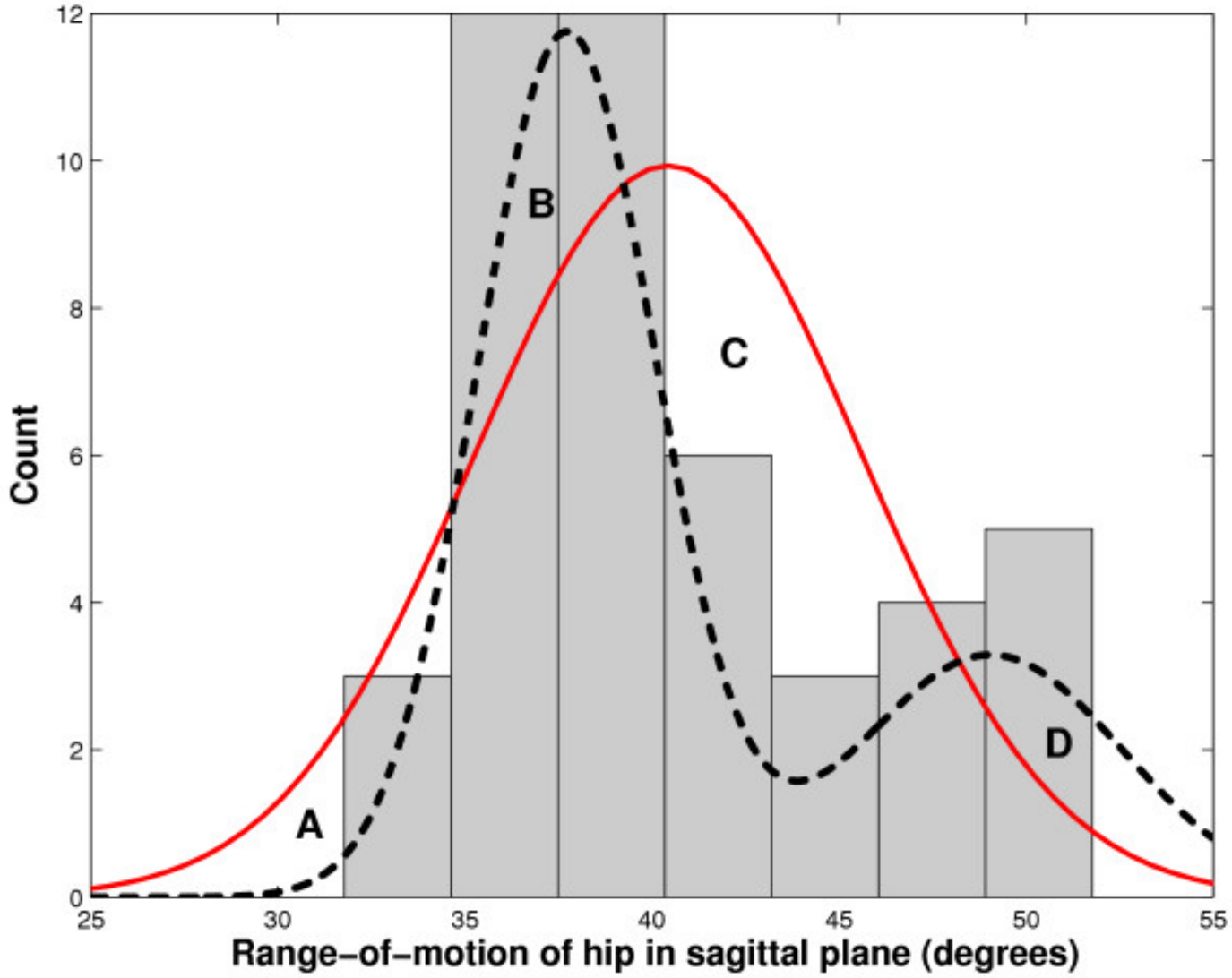
```
Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER.  Determinants of pl

Description:  This datafile contains 315 observations on 14 variables.  This d

Variable Names in order from left to right:
        AGE: Age (years)
        SEX: Sex (1=Male, 2=Female).
        SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
        QUETELET: Quetelet (weight/(height^2))
        VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
        CALORIES: Number of calories consumed per day.
        FAT: Grams of fat consumed per day.
        FIBER: Grams of fiber consumed per day.
        ALCOHOL: Number of alcoholic drinks consumed per week.
        CHOLESTEROL: Cholesterol consumed (mg per day).
        BETADIET: Dietary beta-carotene consumed (mcg per day).
        RETDIET: Dietary retinol consumed (mcg per day)
        BETAPLASMA: Plasma beta-carotene (ng/ml)
        RETPLASMA: Plasma Retinol (ng/ml)

64      2       2       21.4838 1       1298.8  57      6.3     0       170.3
76      2       1       23.87631        1       1032.5  50.1    15.8    0
38      2       2       20.0108 2       2372.3  83.6    19.1    14.1    257.9
40      2       2       25.14062        3       2449.5  97.5    26.5    0.5
72      2       1       20.98504        1       1952.1  82.6    16.2    0
40      2       2       27.52136        3       1366.9  56      9.6     1.3
65      2       1       22.01154        2       2213.9  52      28.7    0
58      2       1       28.75702        1       1595.6  63.4    10.9    0
35      2       1       23.07662        3       1800.5  57.8    20.3    0.6
55      2       2       34.96995        3       1263.6  39.6    15.5    0
66      2       2       20.94647        1       1460.8  58      18.2    1
40      2       1       36.43161        2       1638.2  49.3    14.9    0
57      1       1       31.73039        3       2072.9  106.7   9.6     0.9
66      2       1       21.78854        1       987.5   35.6    10.3    0
66      1       1       27.31916        3       1574.3  75      7.1     0
64      1       2       31.44674        3       2868.5  128.8   15      20
```



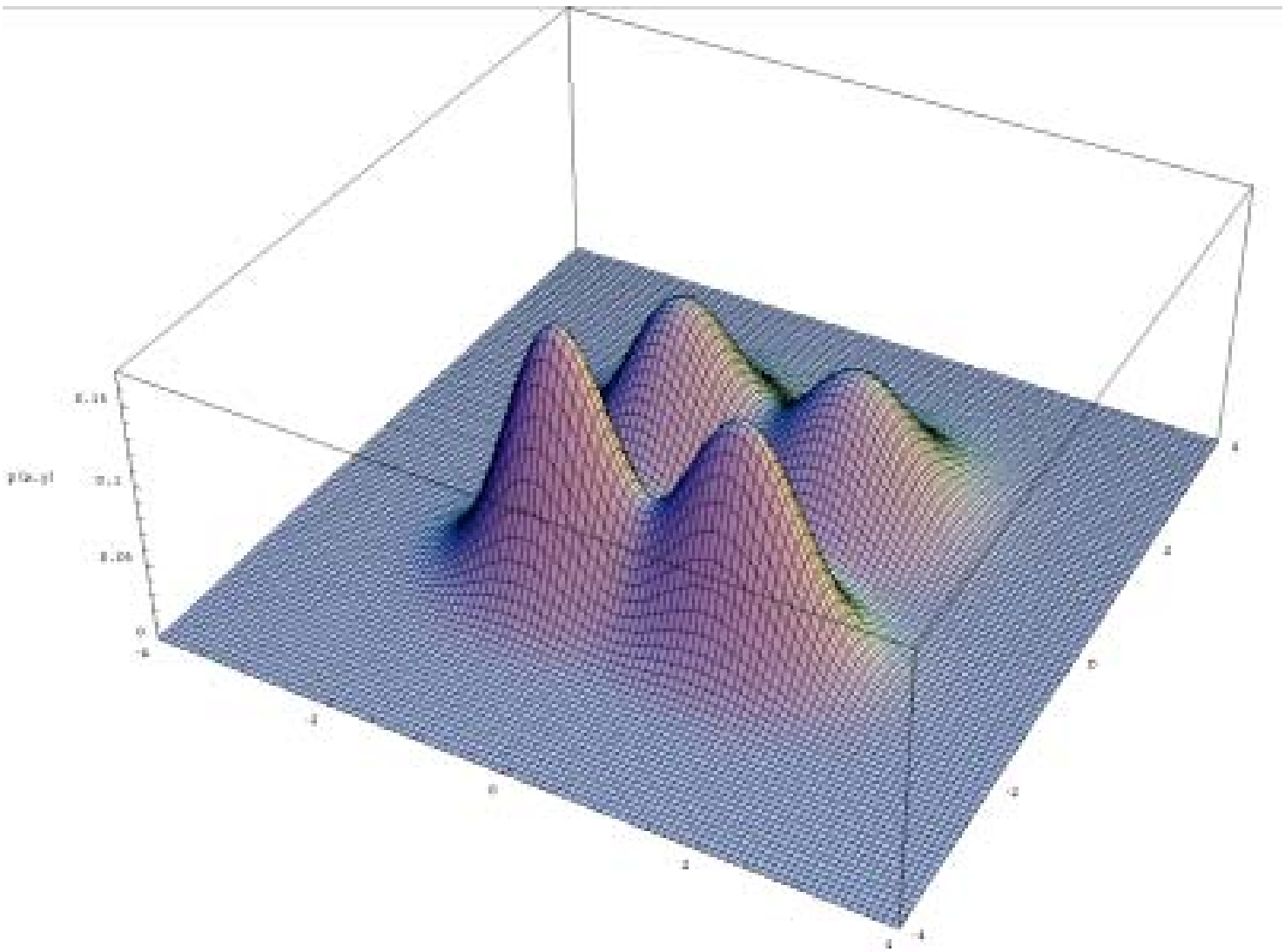
2.1 Graphical examination

- Univariate distribution plots
- Bivariate distribution plots
- Pairwise plots
  - Scatter plots
  - Boxplots
- Multivariate plots
  - Chernoff faces
  - Star plots

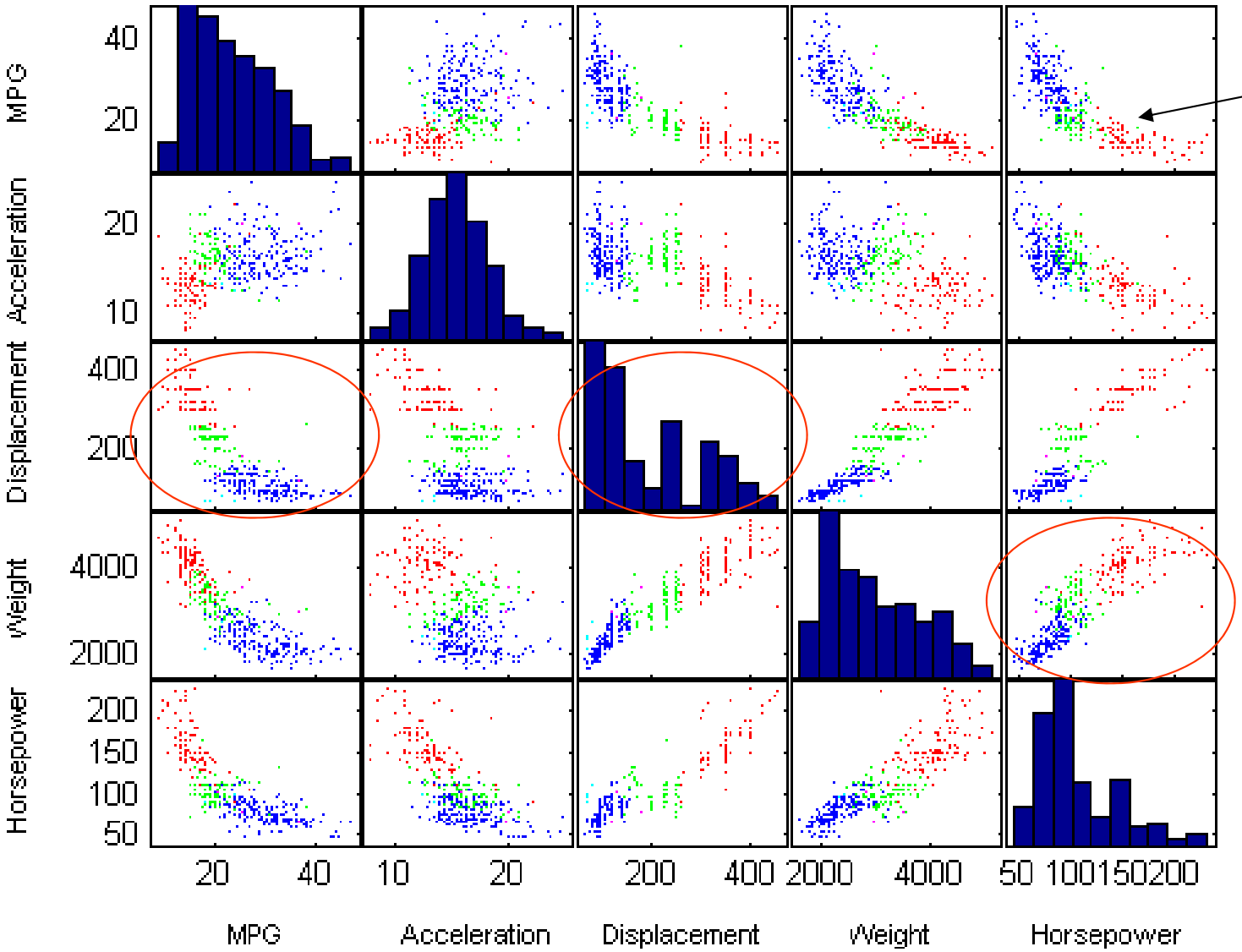
2.1 Graphical examination: Univariate distribution



2.1 Graphical examination: Bivariate distributions

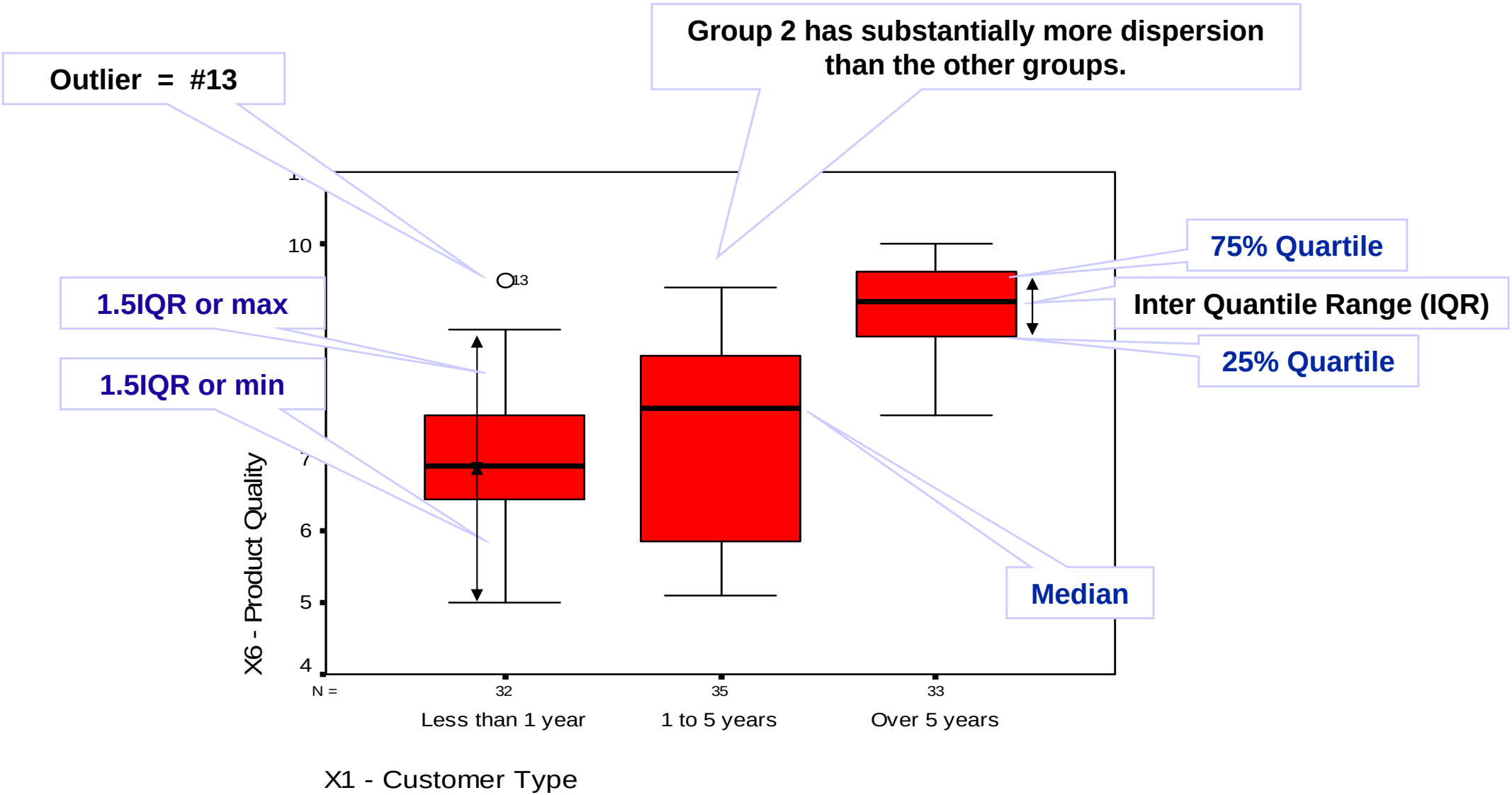


2.1 Graphical examination: Scatter plots

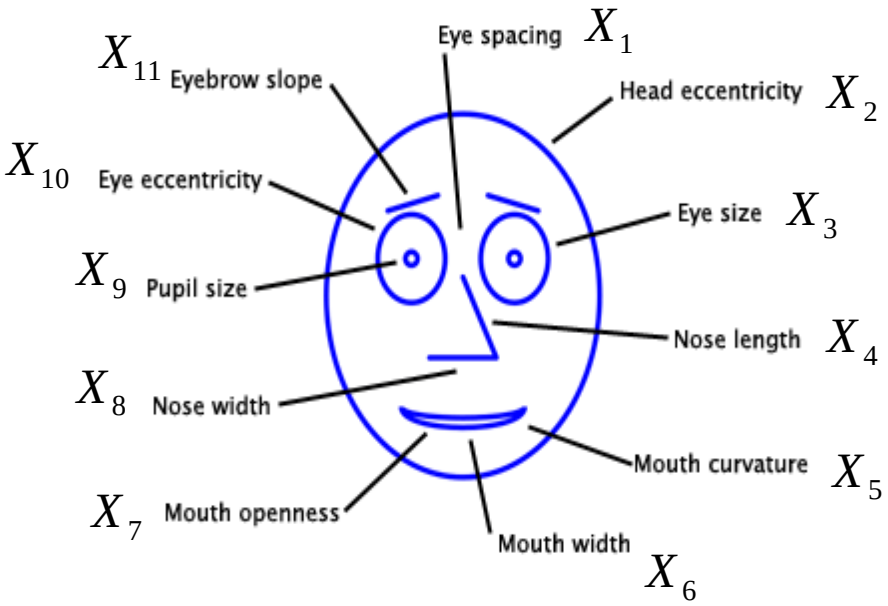


Coloured by class

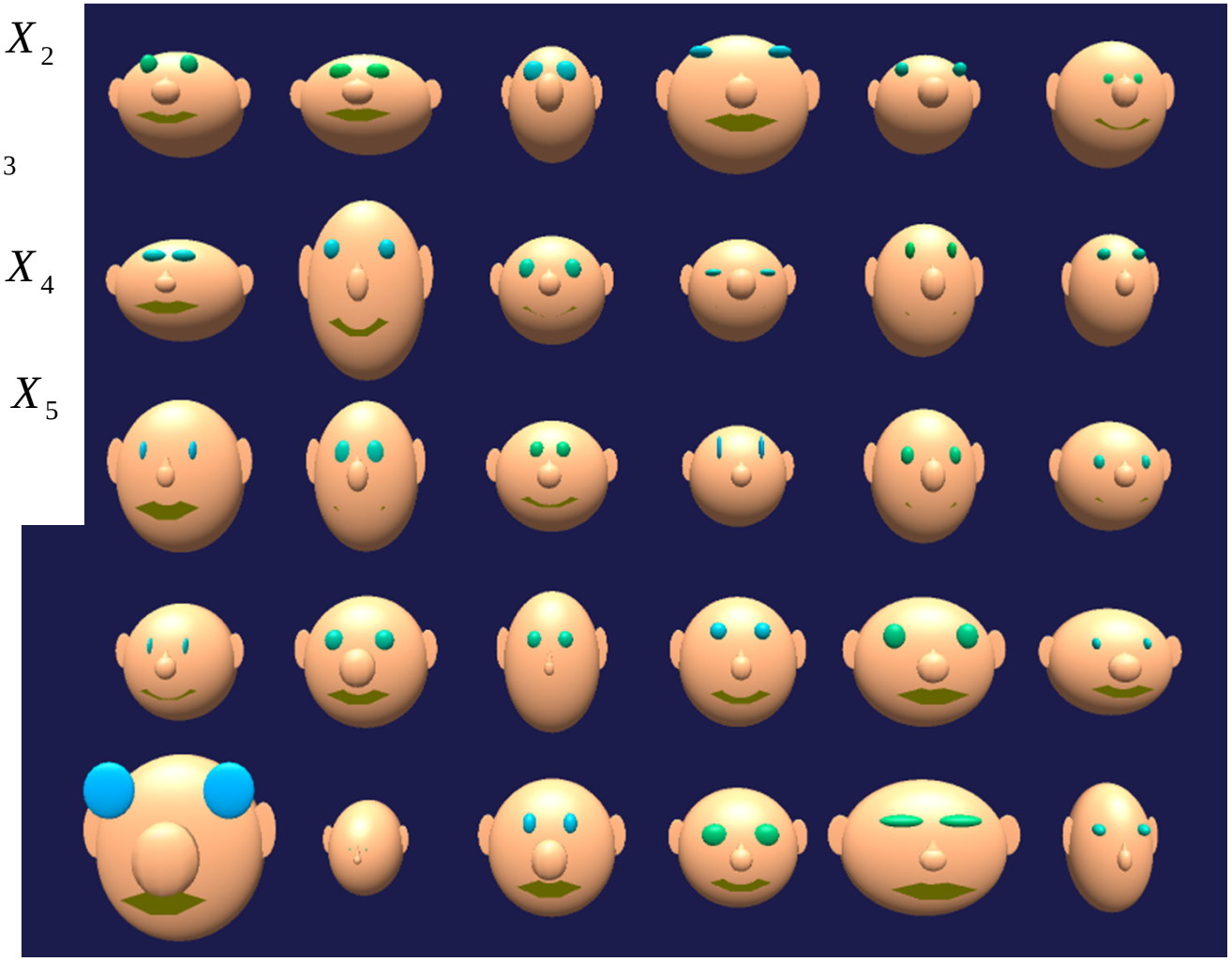
2.1 Graphical examination: Boxplots (Box-Whiskers)



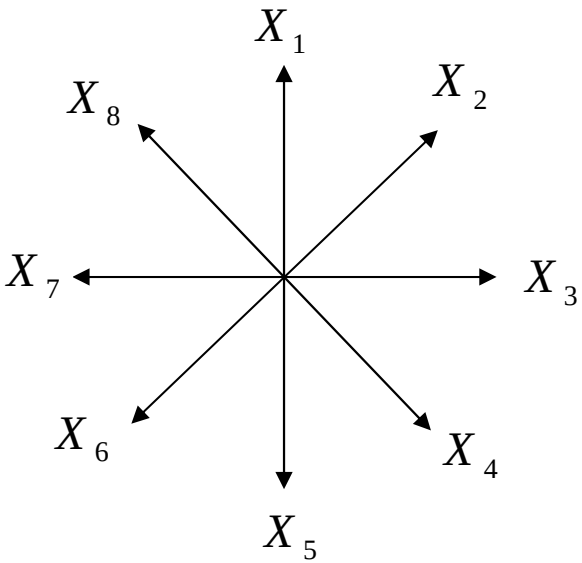
2.1 Graphical examination: Multivariate plots



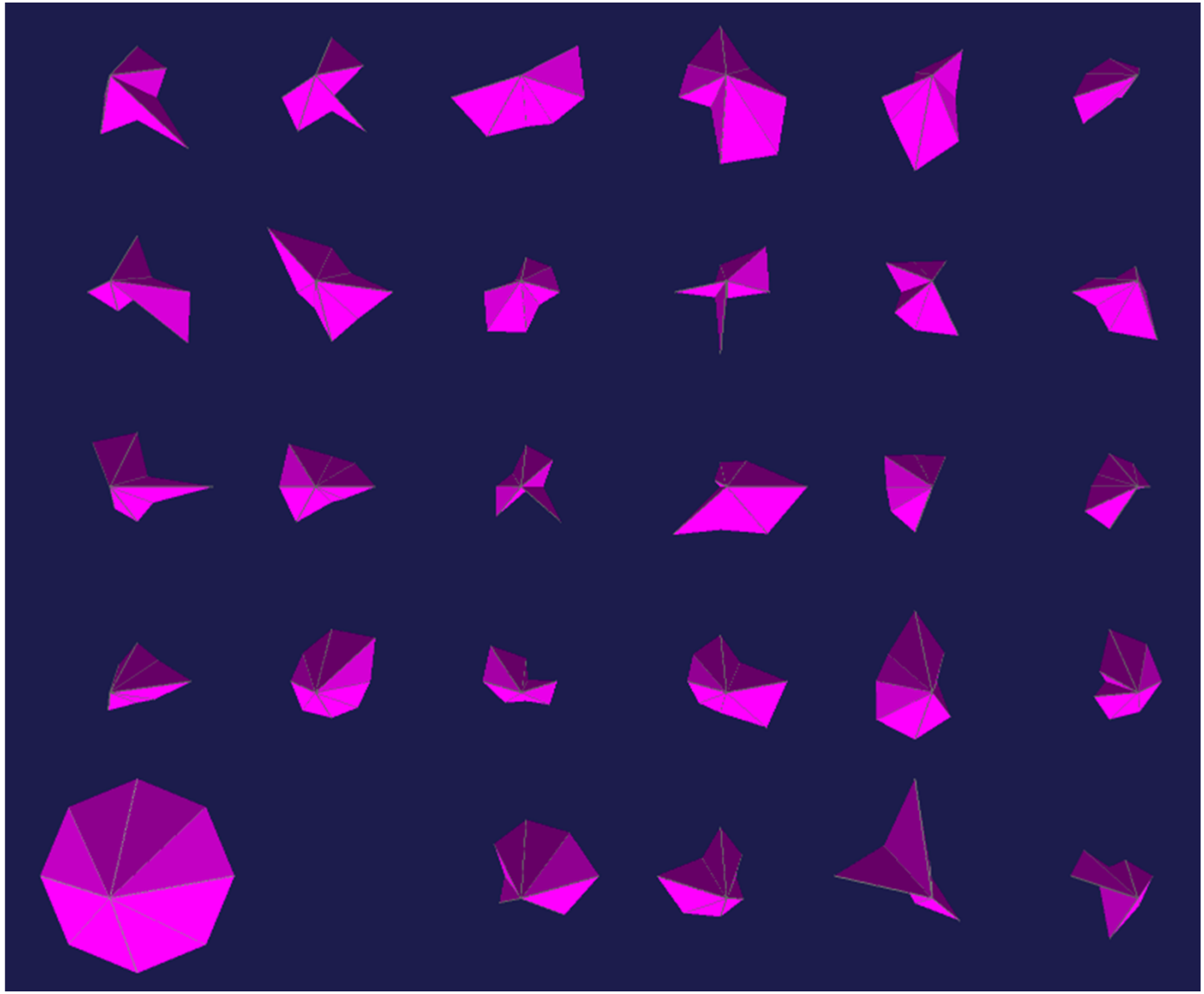
Chernoff Faces



2.1 Graphical examination: Multivariate plots



Star plots



2.2 Missing data

Types of missing data:

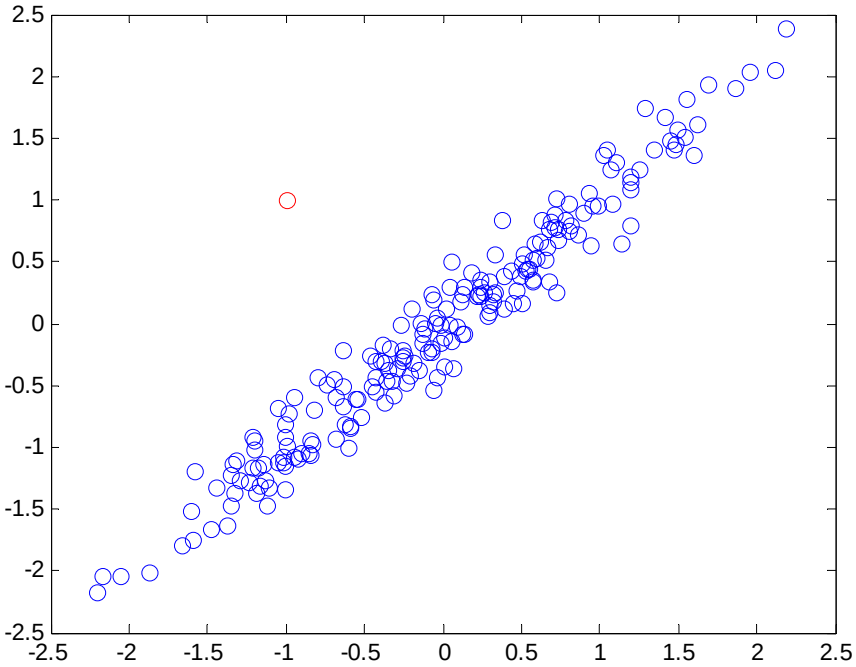
- Missing Completely At Random (MCAR)
- Missing at Random (MAR)

Strategies for handling missing data:

- use observations with complete data only
- delete case(s) and/or variable(s)
- estimate missing values (imputation):
  - + All-available
  - + Mean substitution
  - + Cold/Hot deck
  - + Regression (preferred for MCAR): Linear, Tree
  - + Expectation-Maximization (preferred for MAR)
  - + Multiple imputation (Markov Chain Monte Carlo, Bayesian)

2.3 Multivariate outliers

Model: Centroid+noise



Univariate detection

$$\frac{|x_i - median(x)|}{MAD(x)} > 4.5$$

Grubb's statistic (assumes normality)

$$\max \frac{|x_i - \bar{x}|}{s_x} > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{N},N-2}}{N-2+t^2_{\frac{\alpha}{N},N-2}}}$$

Critical value of Student's t distribution with N-2 degrees of freedom and a significance level

Multivariate detection

Number of variables

$$d^2(\mathbf{x}_i, \bar{\mathbf{x}}) = (\mathbf{x}_i - \bar{\mathbf{x}})^t S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) > p + 3\sqrt{2p}$$

2.3 Multivariate outliers

Model: Function+noise

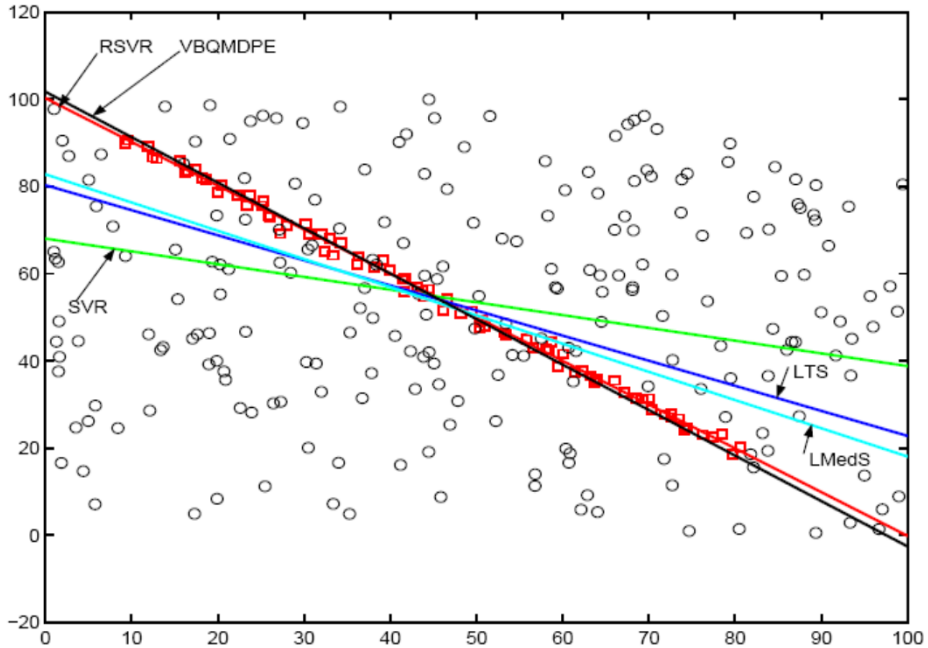
1. For each dimension
   a. Fit a model to data by regression

$$\hat{x}_1 = f(x_2, x_3, ..., x_p)$$

   b. Label all those points with large residuals as outliers

$$|\hat{x}_{1i} - x_{1i}| > \theta(p, N, \alpha)$$

   c. Go to step a until convergence



2. For each pair of dimensions
   a. Fit a model to data by regression

$$(\hat{x}_1, \hat{x}_2) = f(x_3, ..., x_p)$$

   b. Label all those points with large residuals as outliers

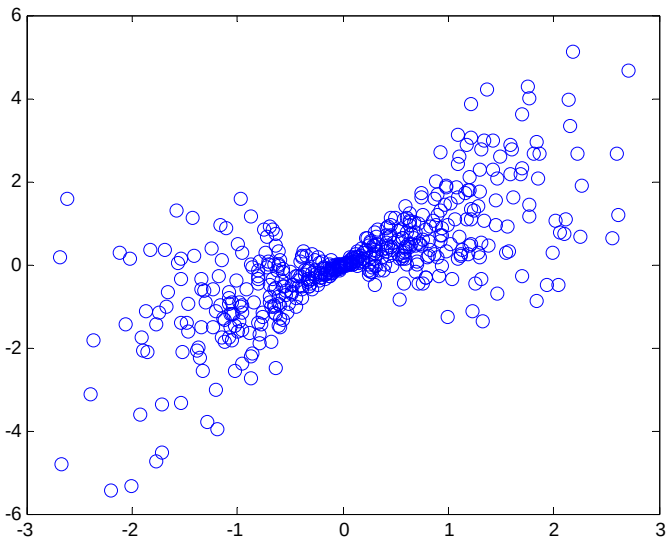
$$dist((\hat{x}_{1i}, \hat{x}_{2i}), (x_{1i}, x_{2i})) > \theta(p, N, \alpha)$$

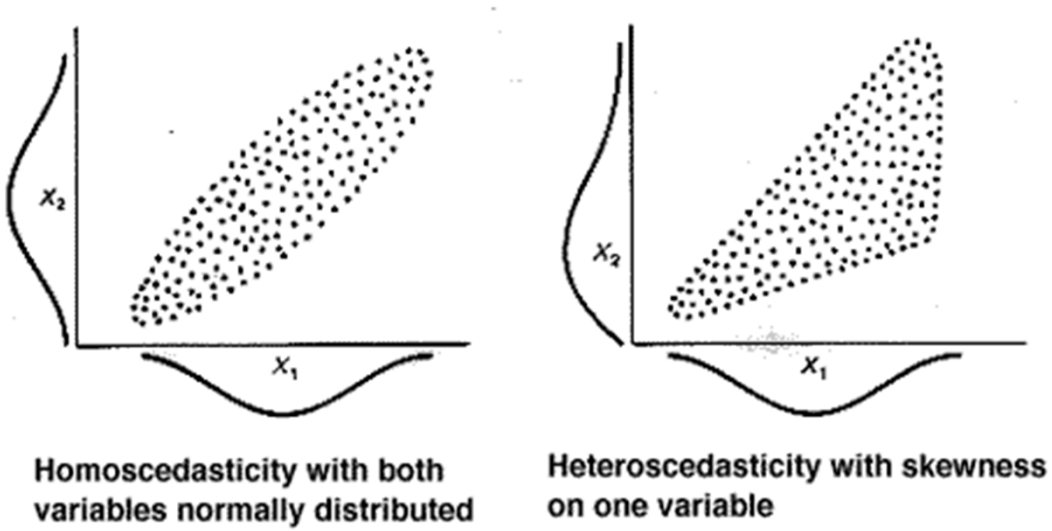
The threshold is a function of the number of variables, the number of samples, and the significance level

   c. Go to step a until convergence

2.4 Assumptions of multivariate analysis

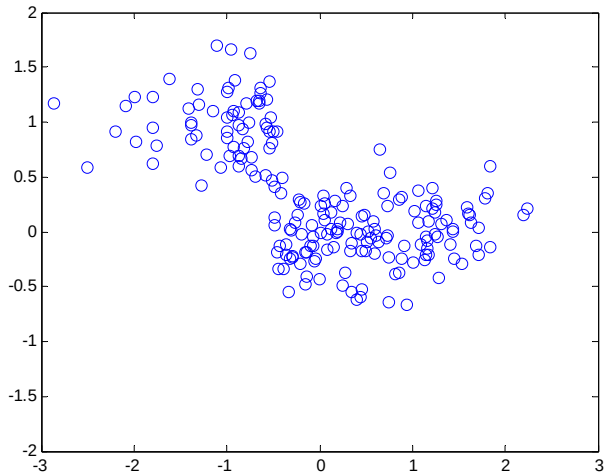
- Normality: the multivariate variable follows a multivariate Gaussian
  - Univariate variables, too
  - Tests: Shapiro-Wilks (1D), Kolmogorov-Smirnov(1D), Smith-Jain (nD)
- Homoscedasticity (Homoskedasticity): the variance of the dependent variables is the same across the range of predictor variables
  - Tests: Levene, Breusch-Pagan, White





2.4 Assumptions of multivariate analysis

- Linearity: All techniques based on correlation (multiple regression, logistic regression, factor analysis, structure equation modelling, principal component analysis, etc.) assume that the dependent variables depend linearly on the independent ones.
  - Test: Scatterplots
- Non-correlated errors: All prediction techniques assume that the prediction residual is independent of the predictors. This may not be true all over the predictor interval.



2.4 Assumptions of multivariate analysis

- The solution to most assumption violations is provided by data transformations.

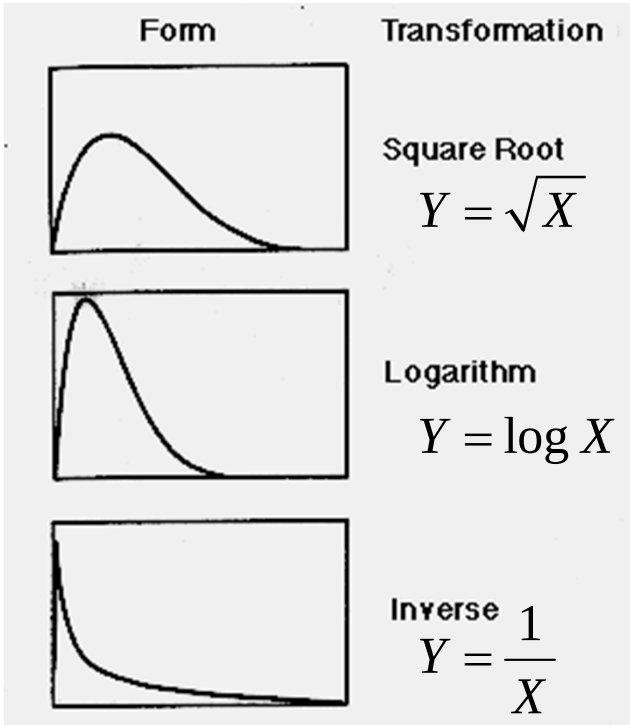


Table of sample pdfs and suggested transformation

Multivariate standardization

$$\mathbf{Y} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}^t) S_X^{-\frac{1}{2}}$$

Course outline: Session 1

1. Introduction
   - 1.1. Types of variables
   - 1.2. Types of analysis and technique selection
   - 1.3. Descriptors (mean, covariance matrix)
   - 1.4. Variability and distance
   - 1.5. Linear dependence

2. Data Examination
   - 2.1. Graphical examination
   - 2.2. Missing Data
   - 2.3. Outliers
   - 2.4. Assumptions of multivariate analysis

CEU
Universidad
San Pablo

Multivariate Data Analysis

Session 2: Principal Component Analysis and Factor Analysis

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid

Course outline: Session 2

3. Principal component analysis (PCA)
    - 3.1. Introduction
    - 3.2. Component computation
    - 3.3. Example
    - 3.4. Properties
    - 3.5. Extensions
    - 3.6. Relationship to SVD

4. Factor Analysis (FA)
    - 4.1. Introduction
    - 4.2. Factor computation
    - 4.3. Example
    - 4.4. Extensions
    - 4.5. Rules of thumb
    - 4.6. Comparison with PCA

3.1 PCA: Introduction

http://lib.stat.cmu.edu/datasets/Plasma_Retinol - Microsoft Internet Explorer

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. Ame

Description: This datafile contains 315 observations on 14 variables. This data set can be used to demonstrate multiple re

Variable Names in order from left to right:

AGE: Age (years)
SEX: Sex (1=Male, 2=Female).
SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
QUETELET: Quetelet (weight/(height^2))

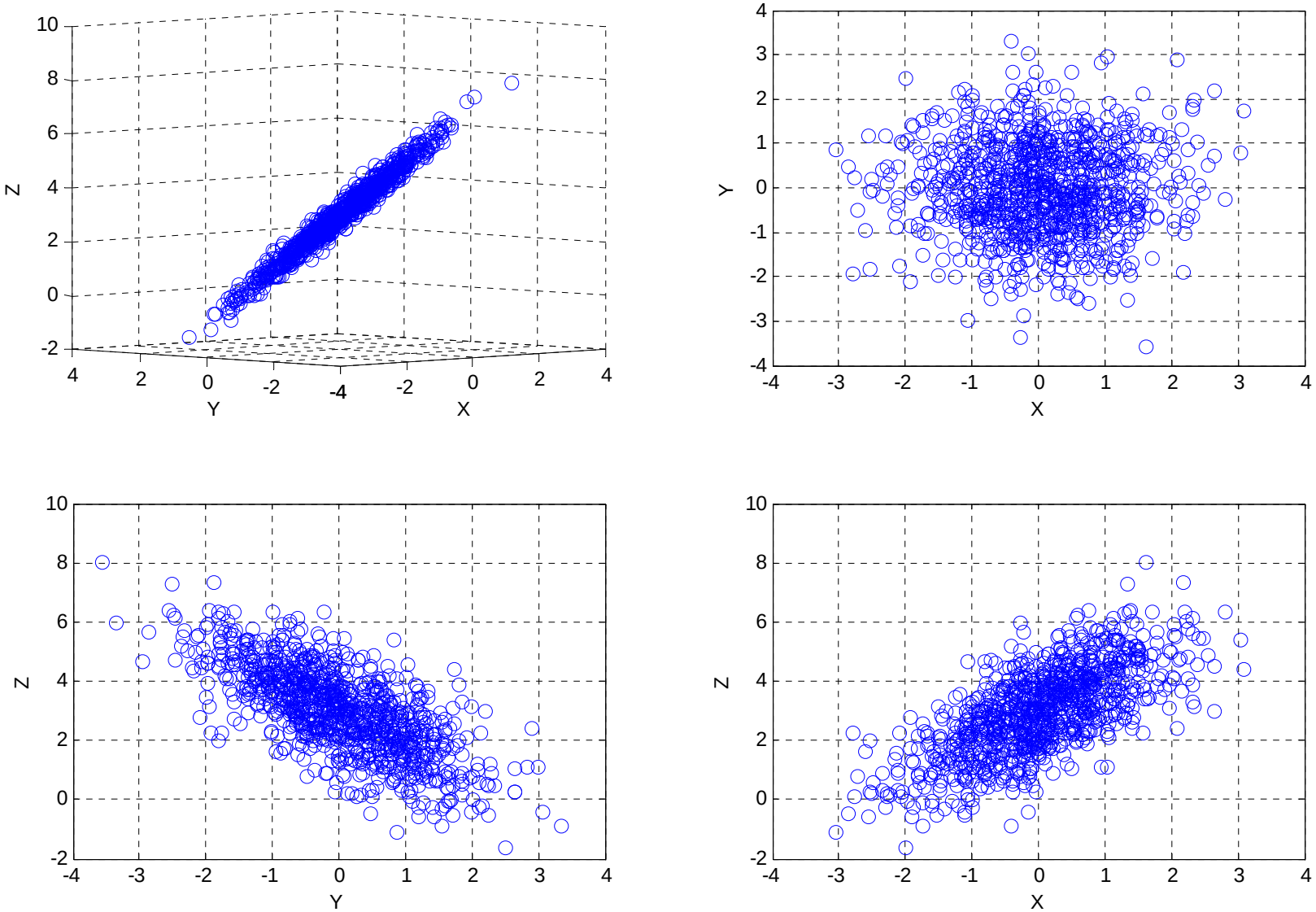
Can we capture the information provided by the original p variables with a fewer number of variables?

BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

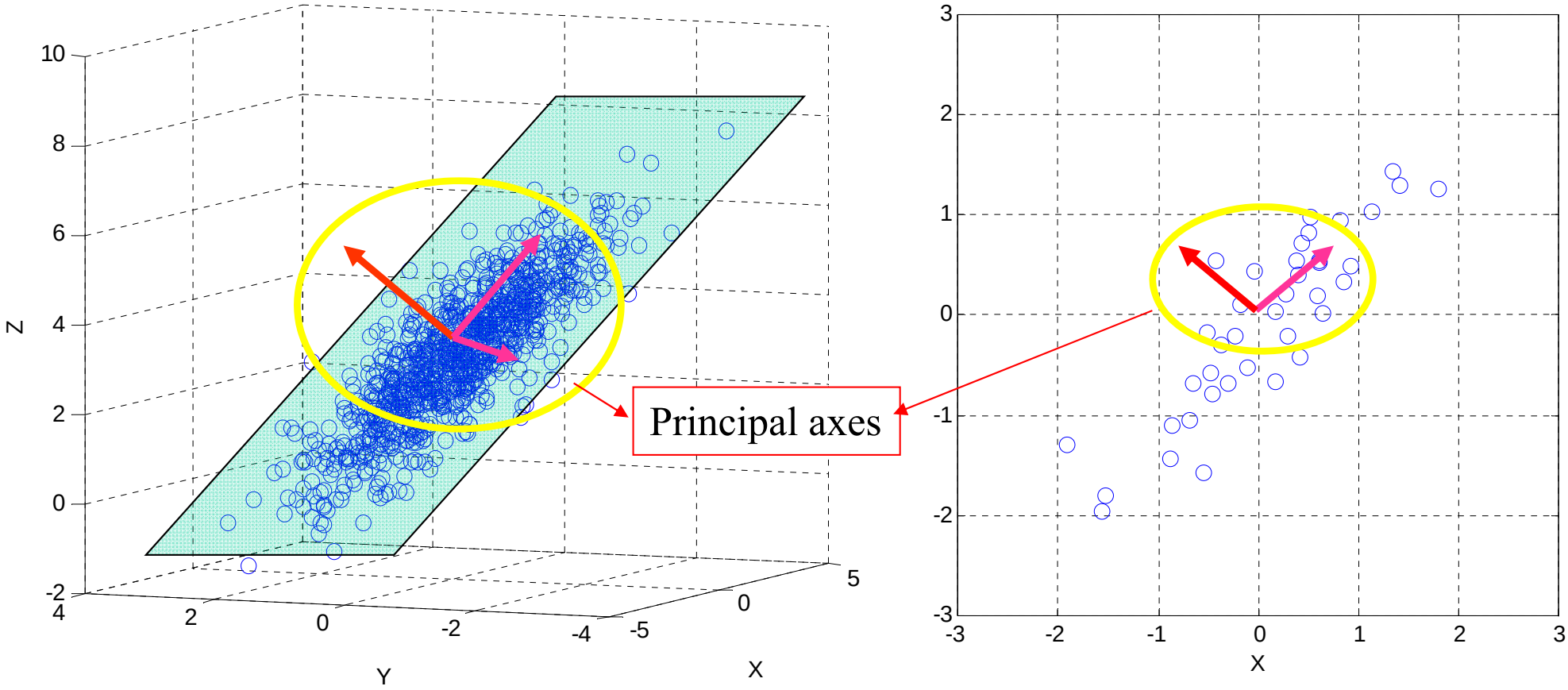
$\mathbf{x}^t$

64	2	2	21.4838	1	1298.8	57	6.3	0	170.3	1945	890	200	915
76	2	1	23.87631	1	1832.5	58.1	15.8	8	75.8	2653	451	124	727
38	2	2	20.0108	2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	20.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741
57	1	1	31.73039	3	2072.9	106.7	9.6	0.9	420	1982	1105	120	679
66	2	1	21.78854	1	987.5	35.6	10.3	0	254.9	2120	1047	61	507
66	1	1	27.31916	3	1574.3	75	7.1	0	361.5	1388	980	108	852
64	1	2	31.44674	3	2868.5	128.8	15	20	379.5	3888	1545	211	1249

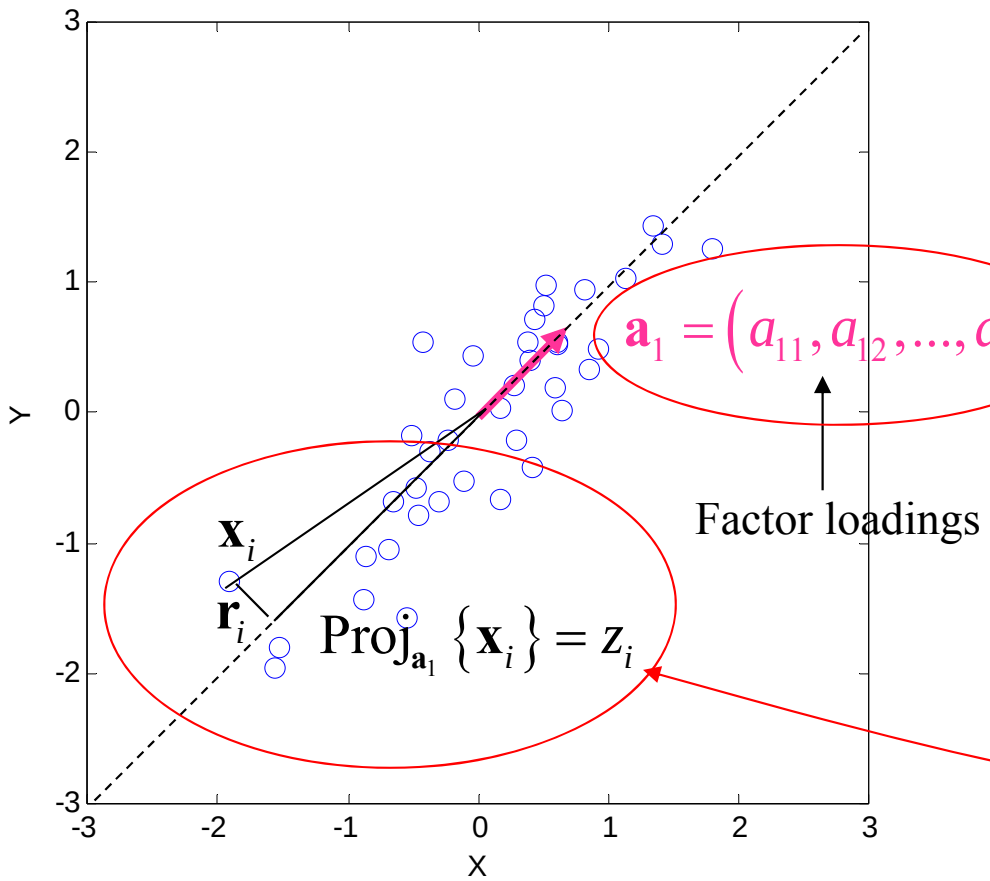
3.1 PCA: Introduction



3.1 PCA: Introduction



3.1 PCA: Introduction



$$\mathrm{Proj}_{\mathbf{a}_1}\{\mathbf{x}_i\} = \overbrace{\langle \mathbf{x}_i, \mathbf{a}_1 \rangle}^{z_i} \mathbf{a}_1 = z_i \mathbf{a}_1$$

$$\mathbf{x}_i = z_i \mathbf{a}_1 + \mathbf{r}_i$$

$$\|\mathbf{x}_i\|^2 = \|z_i \mathbf{a}_1\|^2 + \|\mathbf{r}_i\|^2 = z_i^2 + \|\mathbf{r}_i\|^2$$

$$\boxed{\mathbf{a}_1^* = \arg\min_{\mathbf{a}_1} \sum_{i=1}^{n} \|\mathbf{r}_i\|^2} = \arg\min_{\mathbf{a}_1} \sum_{i=1}^{n} \left(\|\mathbf{x}_i\|^2 - z_i^2 \right)$$

$$= \arg\max_{\mathbf{a}_1} \sum_{i=1}^{n} z_i^2 = \boxed{\arg\max_{\mathbf{a}_1} Var\{Z\}}$$

3.2 PCA: Component computation

Computation of the first component

$$z_1 = \langle \mathbf{x}_1, \mathbf{a}_1 \rangle = \mathbf{x}_1^t \mathbf{a}_1$$
$$z_2 = \langle \mathbf{x}_2, \mathbf{a}_1 \rangle = \mathbf{x}_2^t \mathbf{a}_1$$
$$\dots$$
$$z_n = \langle \mathbf{x}_n, \mathbf{a}_1 \rangle = \mathbf{x}_n^t \mathbf{a}_1$$

$$\begin{pmatrix} z_1 \\ z_2 \\ \dots \\ z_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \dots \\ \mathbf{x}_n^t \end{pmatrix} \mathbf{a}_1$$

Data matrix

$$\mathbf{z} = X\mathbf{a}_1$$

Z Sample mean: $\bar{\mathbf{x}} = \mathbf{0} \Rightarrow \bar{z} = \bar{\mathbf{x}}\mathbf{a}_1 = 0$

Sample covariance matrix

Z Sample variance: $s_z^2 = \frac{1}{n}\mathbf{z}^t\mathbf{z} = \frac{1}{n}\mathbf{a}_1^t X^t X \mathbf{a}_1 = \mathbf{a}_1^t S_x \mathbf{a}_1 = \lambda$

Largest eigenvalue

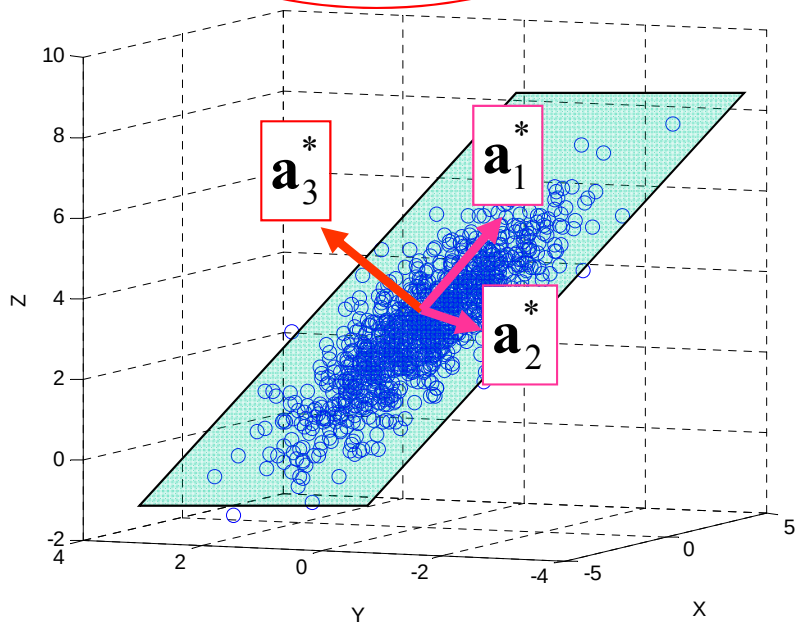
$$\mathbf{a}_1^* = \arg\max_{\mathbf{a}_1} s_z^2 \quad s.t. \quad \|\mathbf{a}_1\|^2 = 1$$

$$= \arg\max_{\mathbf{a}_1} \underbrace{\left(\mathbf{a}_1^t S_x \mathbf{a}_1 - \lambda\left(\mathbf{a}_1^t \mathbf{a}_1 - 1\right)\right)}_{F} \longrightarrow \frac{\partial F}{\partial \mathbf{a}_1} = \mathbf{0} \Rightarrow \boxed{\begin{aligned} S_x \mathbf{a}_1^* &= \lambda \mathbf{a}_1^* \\ \lambda &= \mathbf{a}_1^{*t} S_x \mathbf{a}_1^* \end{aligned}}$$

3.2 PCA: Component computation

$$\mathbf{a}_1^*, \mathbf{a}_2^* = \arg\max_{\mathbf{a}_1, \mathbf{a}_2} s_{z_1}^2 + s_{z_2}^2 \quad s.t. \quad \|\mathbf{a}_1\|^2 = 1, \quad \|\mathbf{a}_2\|^2 = 1$$

$$= \arg\max_{\mathbf{a}_1, \mathbf{a}_2} \underbrace{\left(\mathbf{a}_1^t S_x \mathbf{a}_1 + \mathbf{a}_2^t S_x \mathbf{a}_2 - \lambda_1\left(\mathbf{a}_1^t \mathbf{a}_1 - 1\right) - \lambda_2\left(\mathbf{a}_2^t \mathbf{a}_2 - 1\right)\right)}_{F}$$



$$S_x \mathbf{a}_1^* = \lambda_1 \mathbf{a}_1^* \qquad \lambda_1 = \mathbf{a}_1^{*t} S_x \mathbf{a}_1^*$$

$$S_x \mathbf{a}_2^* = \lambda_2 \mathbf{a}_2^* \qquad \lambda_2 = \mathbf{a}_2^{*t} S_x \mathbf{a}_2^*$$

Largest two eigenvalues

In general, $\mathbf{a}_i^*$ are the eigenvectors sorted by eigenvalue descending order

3.2 PCA: Component computation

$\mathbf{x}^t$

RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

64	2	2	21.4838 1	1298.8	57	6.3	0	170.3	1945	890	200	915	
76	2	1	23.87631	1	1032.5	58.1	15.8	8	75.8	2653	451	124	727
38	2	2	20.0108 2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721	
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	20.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741
57	1	1	31.73039	3	2072.9	106.7	9.6	0.9	420	1982	1105	120	679
66	2	1	21.78854	1	987.5	35.6	10.3	0	254.9	2120	1047	61	507
66	1	1	27.31916	3	1574.3	75	7.1	0	361.5	1388	980	108	852
64	1	2	31.44674	3	2868.5	128.8	15	20	379.5	3888	1545	211	1249

$$\mathbf{z}_1 = X\mathbf{a}_1^* \quad \mathbf{z}_2 = X\mathbf{a}_2^* \quad \ldots \quad \mathbf{z}_p = X\mathbf{a}_p^*$$

Largest variance of all z variables

Smallest variance of all z variables

$$\boxed{Z = XA^*}$$

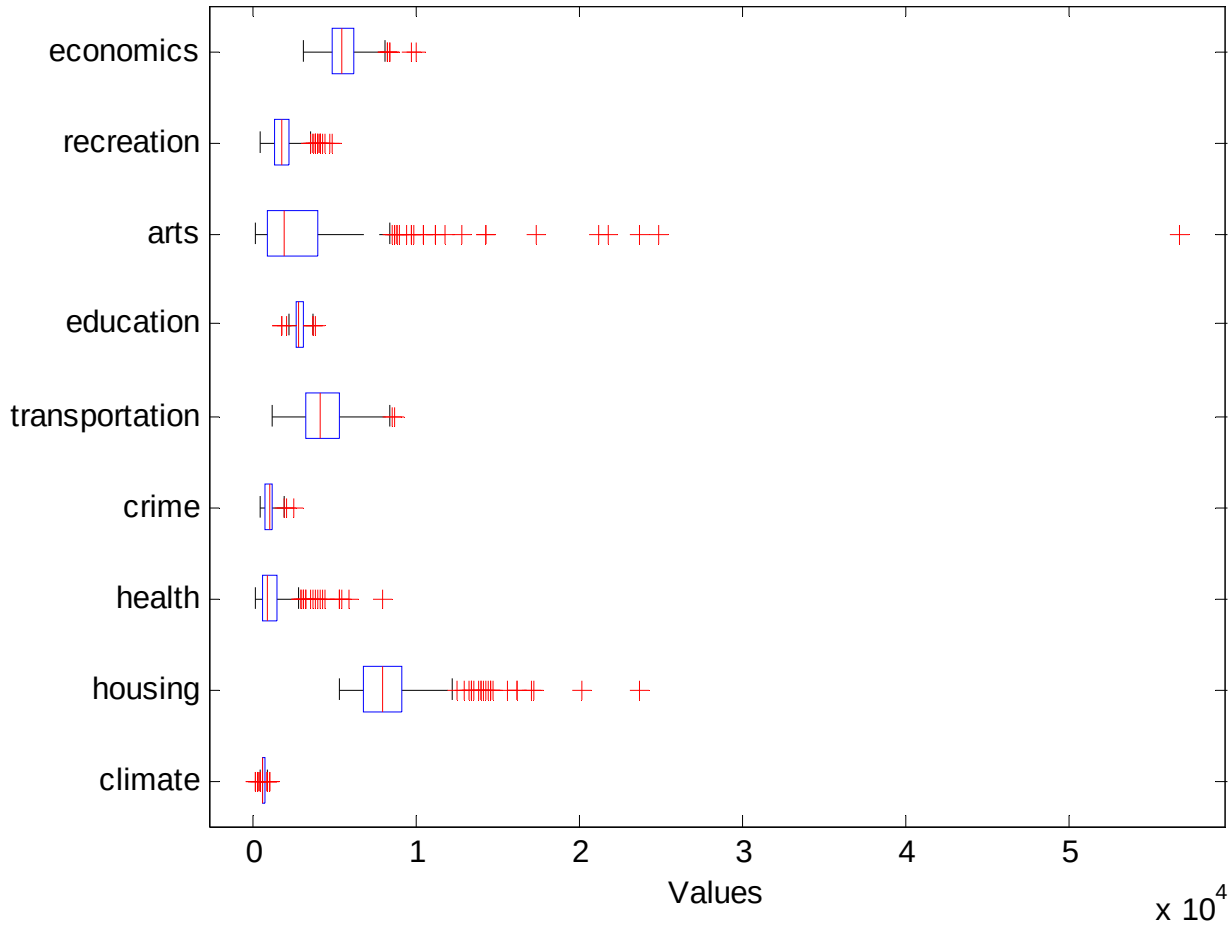
Matrix with the projection directions as columns

New data matrix (size=nxp)

Data matrix (size=nxp)

3.3 PCA: Example

Nine different indices of the quality of life in 329 U.S. cities. These are climate, housing, health, crime, transportation, education, arts, recreation, and economics. For each index, higher is better



3.3 PCA: Example

Factor loadings

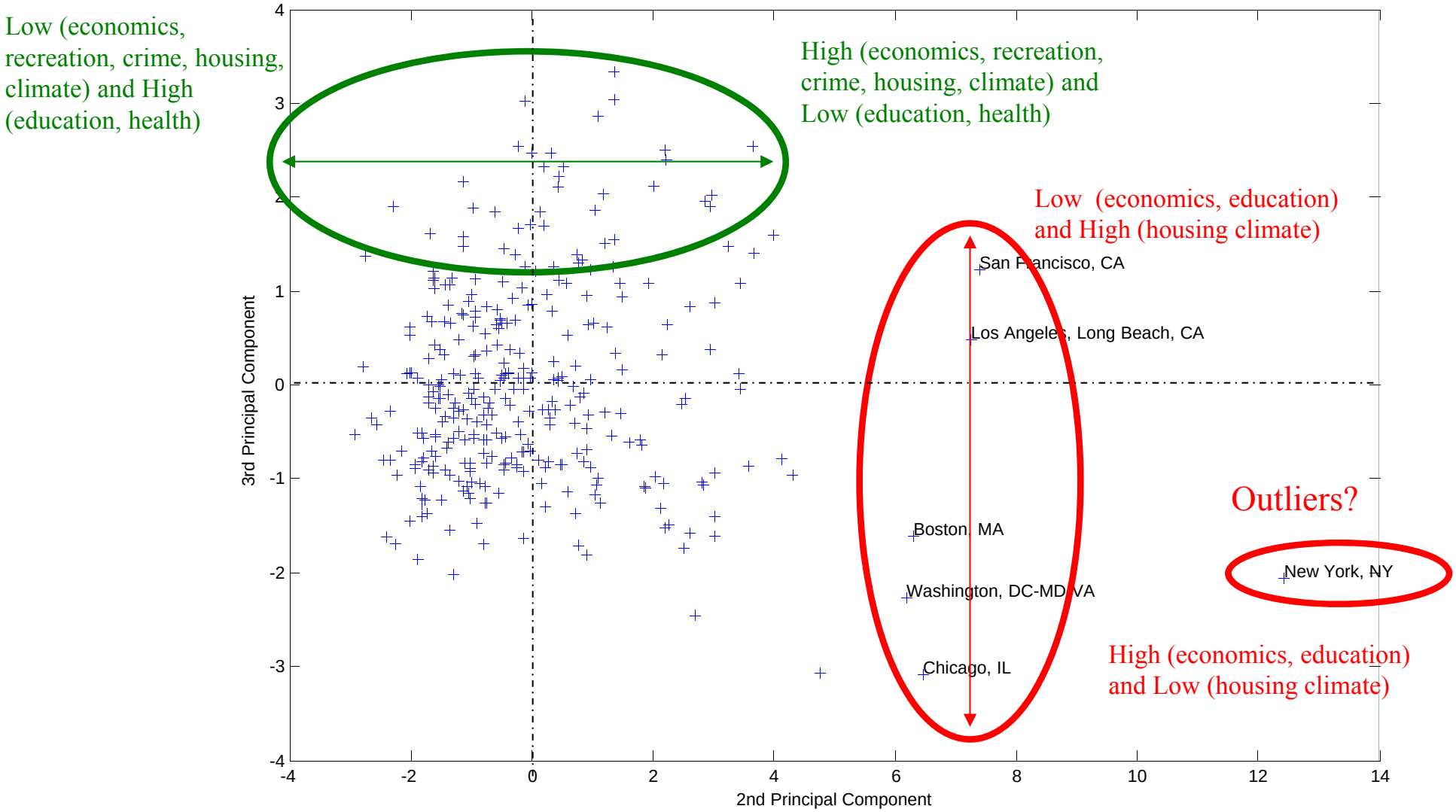
	$\mathbf{a}_1^*$	$\mathbf{a}_2^*$	$\mathbf{a}_3^*$
climate	0.2064	0.2178	-0.6900
housing	0.3565	0.2506	-0.2082
health	0.4602	-0.2995	-0.0073
crime	0.2813	0.3553	0.1851
transportation	0.3512	-0.1796	0.1464
education	0.2753	-0.4834	0.2297
arts	0.4631	-0.1948	-0.0265
recreation	0.3279	0.3845	-0.0509
economics	0.1354	0.4713	0.6073

All positive, i.e., weighted average

Difference between (economics, recreation, crime, housing, climate) vs (education, health)

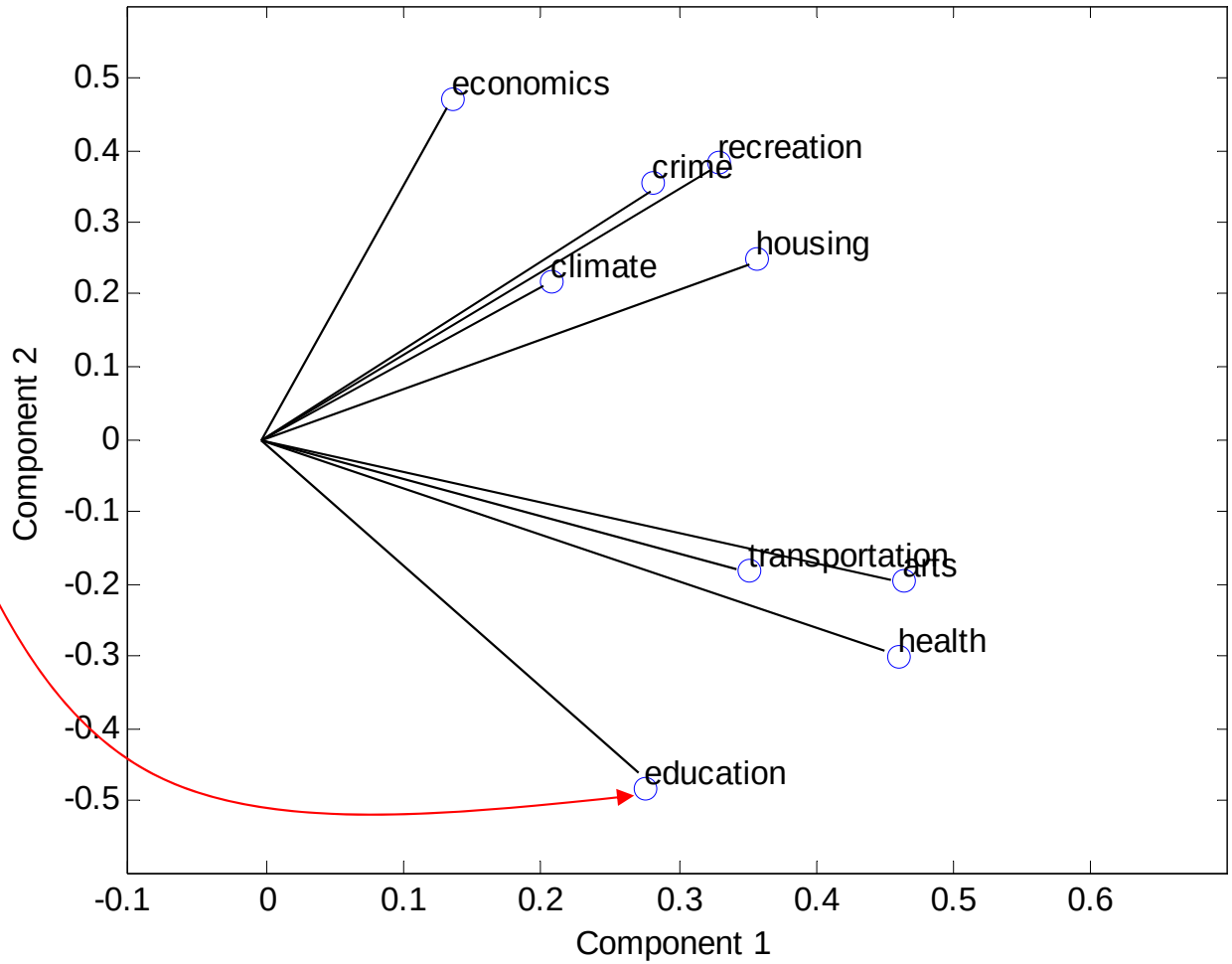
Difference between (economics, education) vs (housing, climate)

3.3 PCA: Example



3.3 PCA: Example

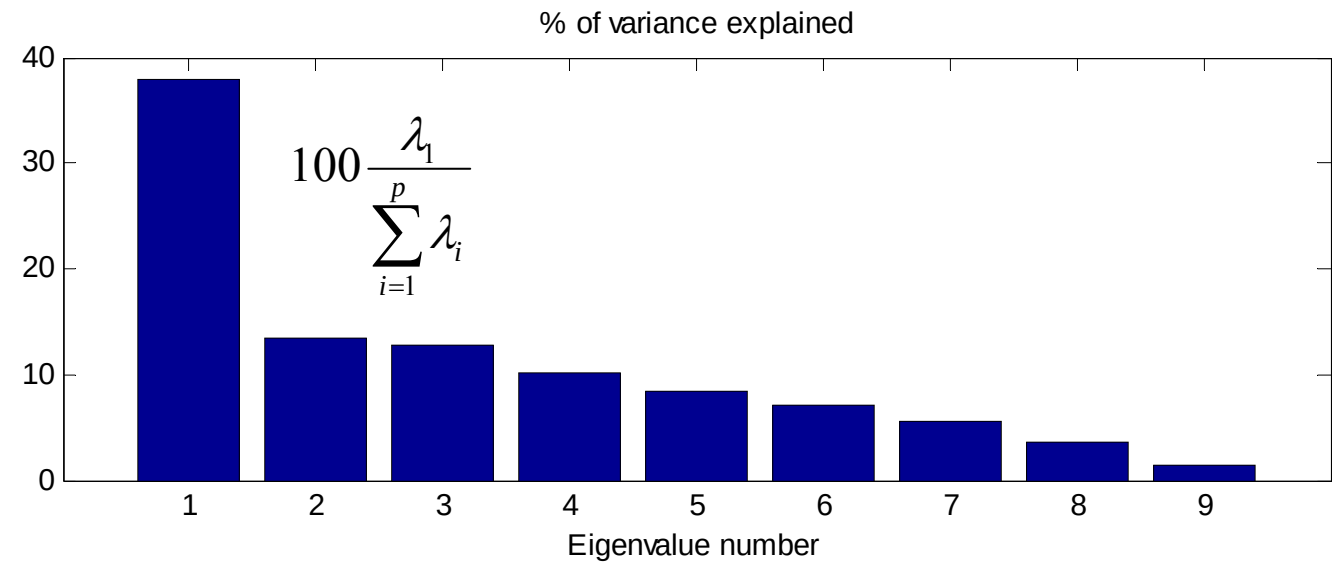
	$\mathbf{a}_1^*$	$\mathbf{a}_2^*$
climate	0.2064	0.2178
housing	0.3565	0.2506
health	0.4602	-0.2995
crime	0.2813	0.3553
transportation	0.3512	-0.1796
education	0.2753	-0.4834
arts	0.4631	-0.1948
recreation	0.3279	0.3845
economics	0.1354	0.4713

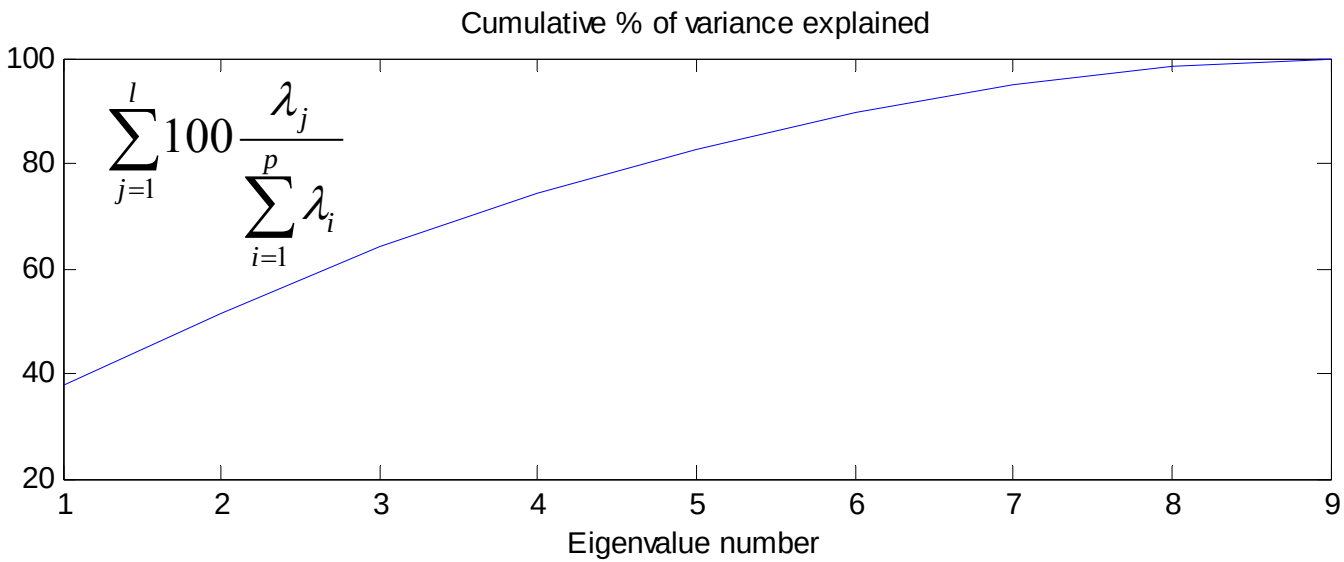


3.4 PCA: Properties

$$tr\{S_x\} = s_{x_1}^2 + s_{x_2}^2 + \ldots + s_{x_p}^2$$
$$= s_{z_1}^2 + s_{z_2}^2 + \ldots + s_{z_p}^2$$
$$= \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

Total variance



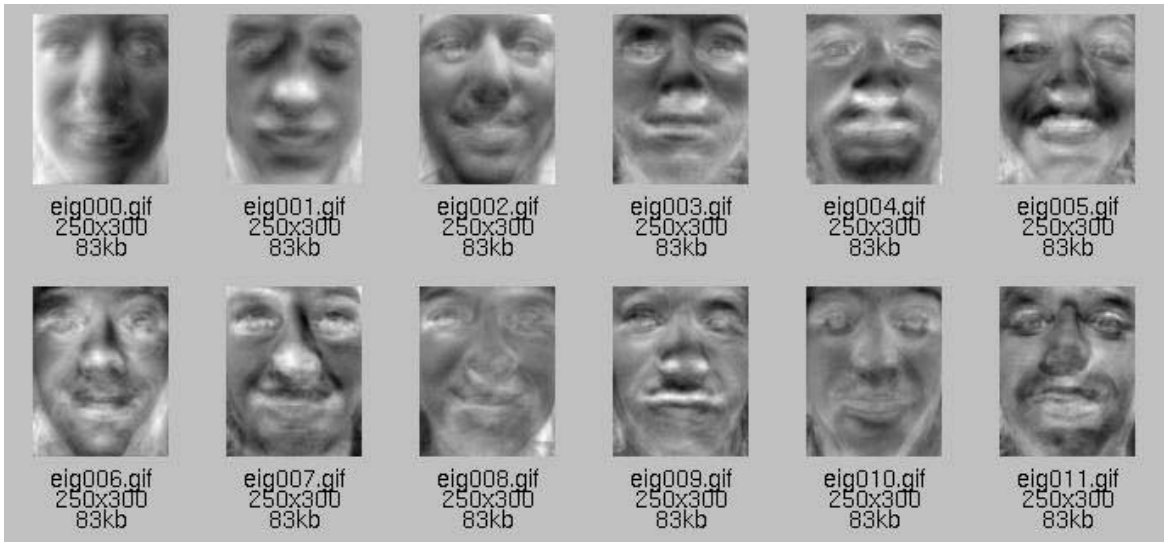


3.4 PCA: Properties

Data restoration

$$\mathbf{x} = \sum_{i=1}^{p} z_i \mathbf{a}_i^*$$

PCA



$= z_1$

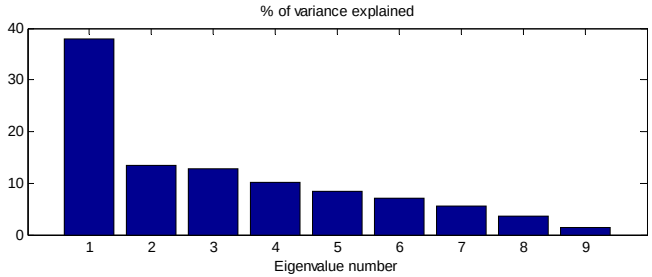
$+ z_2$

$+ \ldots$

3.4 PCA: Properties

Data compression and denoising

$$\mathbf{x} = \sum_{i=1}^{p'} z_i \mathbf{a}_i^*$$

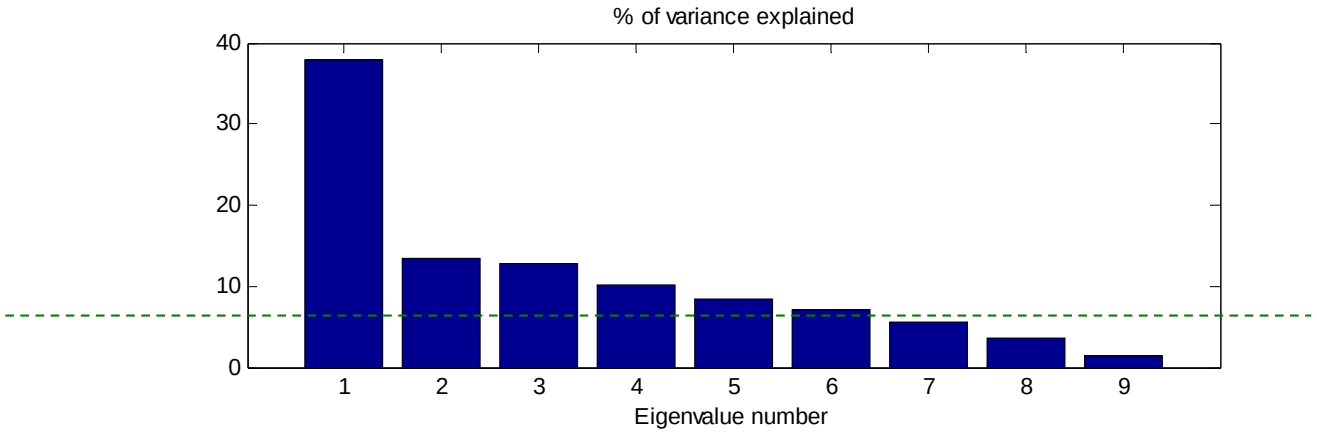


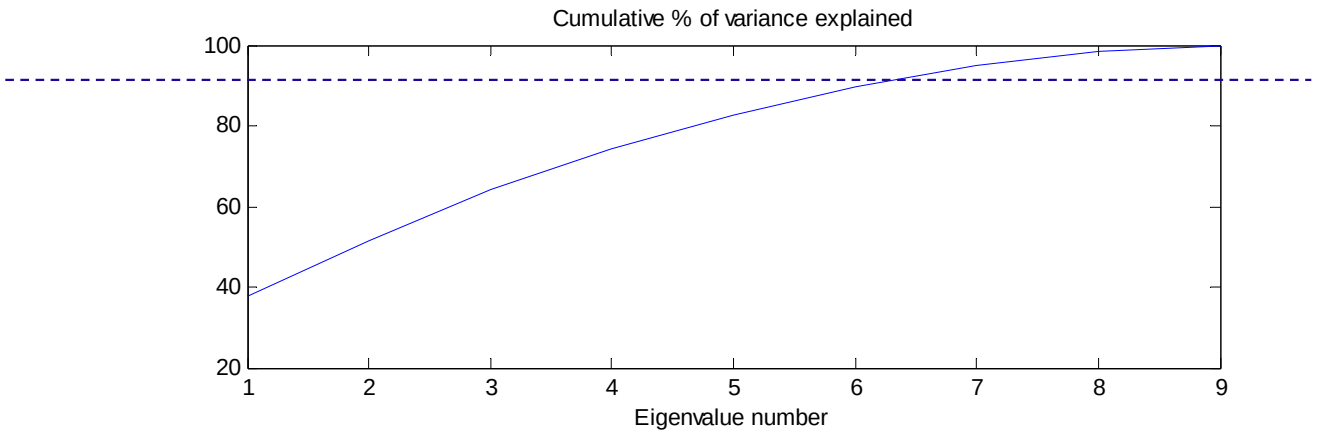
PCA

3.4 PCA: Properties

Criteria:

- Explained variance of a single eigenvalue below a threshold
- Total accounted variance above a threshold

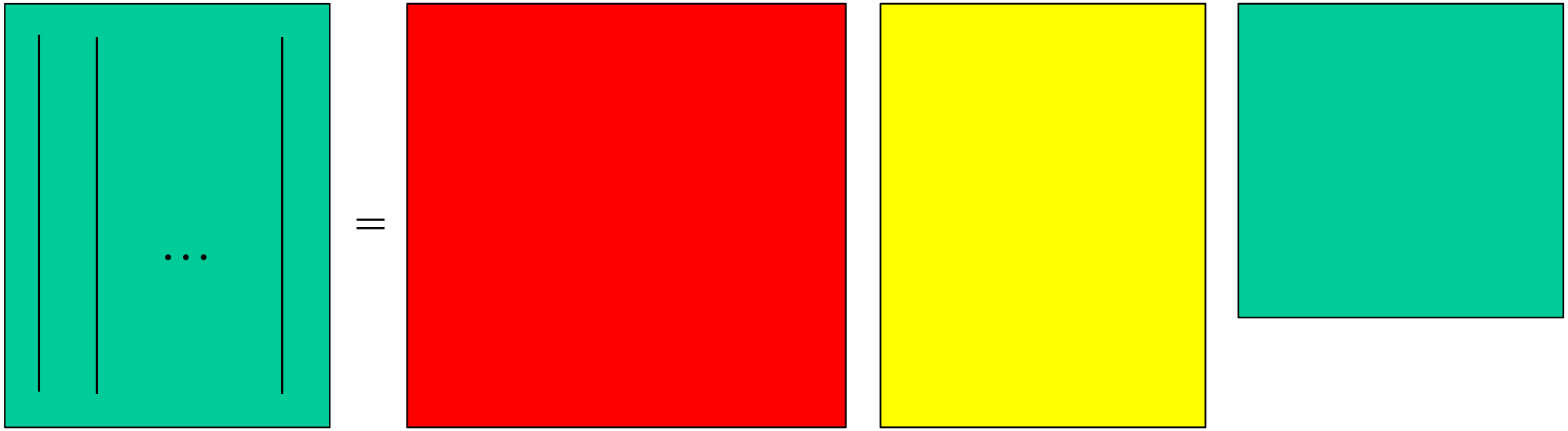




3.5 PCA: Extensions

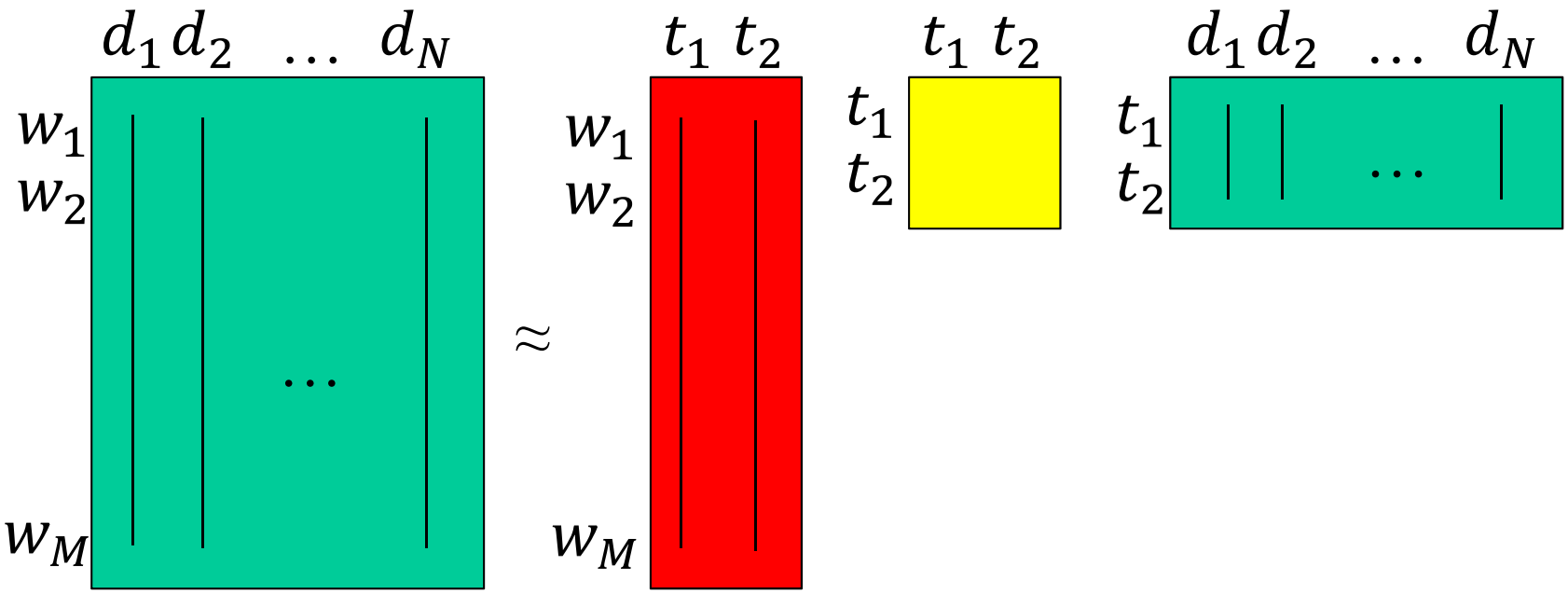
- Use the correlation matrix instead of the covariance matrix (in this way the influence of a variable with a extreme variance is avoided).
- PCA of binary data: PCA is not well suited to non real data
- Sparse PCA: Produce sparse factor loadings
- Noisy PCA or Robust PCA: Input vectors contaminated by additive noise
- Incremental PCA: Recompute easily the PCA as new data comes in
- Probabilistic PCA: Consider the probability distribution of the input data vectors
- Assymetric PCA: Consider the unbalanced distribution of the samples
- Generalized PCA: Use nonlinear projections
- Kernel PCA: Use any kernel instead of the covariance matrix
- Principal curves analysis: Use projections onto any curve instead of a line
- Projection pursuit: Find interesting (non-Gaussian) projection directions
- Correlational PCA: based on Correlational Embedding Analysis
- Localized PCA: PCA of local neighbourhoods

3.6 Relationship to SVD



X	=	W	D	Ξ^t
$\mathrm{M} \times N$		$\mathrm{M} \times M$	$\mathrm{M} \times N$	$N \times N$
	=	$Eig\{\frac{1}{N}XX^t\}$	D	$Eig\{\frac{1}{N}X^tX\}^t$

3.6 Relationship to SVD



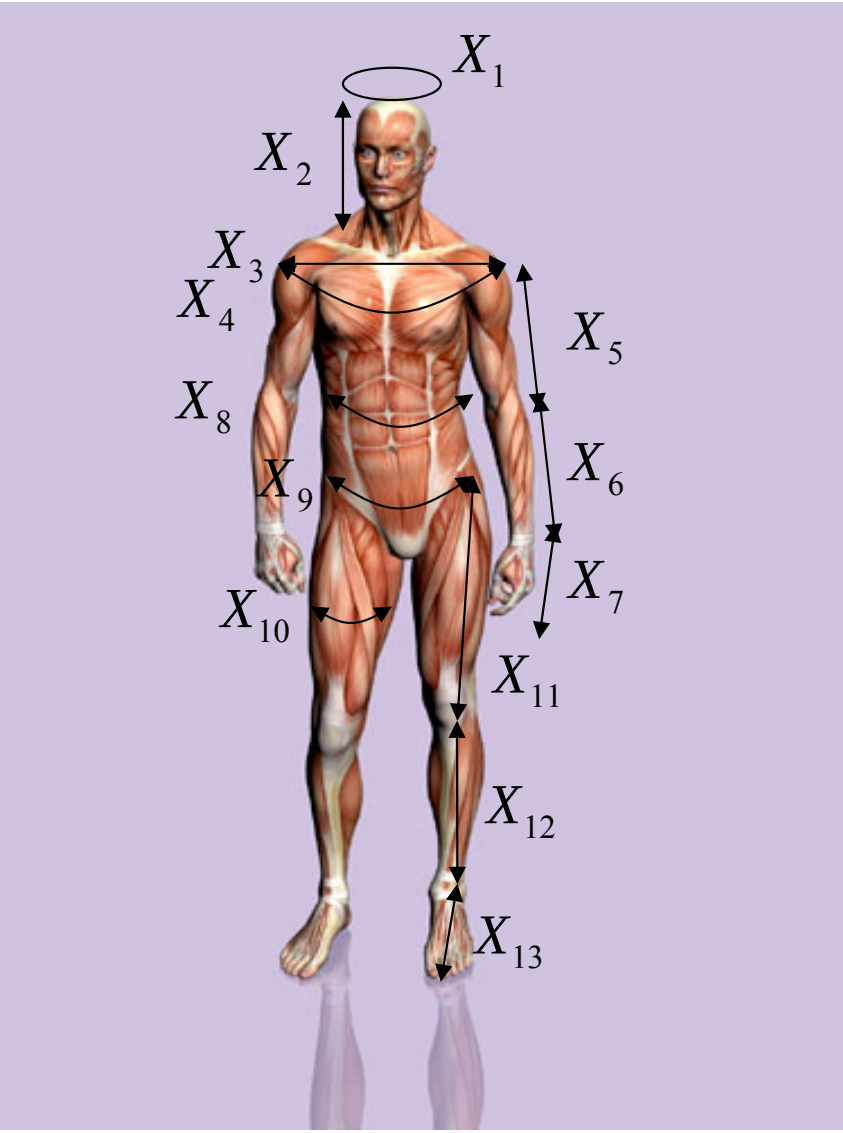
X	≈	W’	D’	Ξ'^t	(SVD)
$M \times N$		$M \times m$	$m \times m$	$m \times N$	
X	≈	X	A^*	B^t	(PCA)
$M \times N$		$M \times N$	$N \times m$	$m \times N$	

Course outline: Session 2

3. Principal component analysis (PCA)
    - 3.1. Introduction
    - 3.2. Component computation
    - 3.3. Example
    - 3.4. Properties
    - 3.5. Extensions
    - 3.6. Relationship to SVD

4. Factor Analysis (FA)
    - 4.1. Introduction
    - 4.2. Factor computation
    - 4.3. Example
    - 4.4. Extensions
    - 4.5. Rules of thumb
    - 4.6. Comparison with PCA

4.1 FA: Introduction



$$\underbrace{(X_1,\ldots,X_{13})}_{\text{Observable variables}} = f\,\underbrace{(gene_1, gene_2, food)}_{\text{Non-observable (latent) variables}}$$

$$\boxed{\mathbf{X} = \boldsymbol{\mu}_X + \Lambda\mathbf{f} + \boldsymbol{\varepsilon}}$$

Average — Load matrix — Noise

4.1 FA: Introduction

$$\underset{N_p(\boldsymbol{\mu}_X,\Sigma_X)}{\mathbf{X}} = \boldsymbol{\mu}_X + \Lambda \underset{N_m(\mathbf{0},I)}{\mathbf{f}} + \underset{N_p(\mathbf{0},\Sigma_\varepsilon)}{\boldsymbol{\varepsilon}} \longrightarrow \overset{m}{\begin{array}{c} \\ \\ \\ \\ \end{array}}$$

$$p\begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_{13} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_{13} \end{pmatrix} + \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \dots & \dots & \dots \\ \lambda_{13,1} & \lambda_{13,2} & \lambda_{13,3} \end{pmatrix} p \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} m + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{13} \end{pmatrix} p$$

<u>Properties</u>:

- Factors are uncorrelated/independent $\quad E\left\{\mathbf{F}\mathbf{F}^t\right\} = I$
- Factors and noise are uncorrelated $\quad E\left\{\boldsymbol{\varepsilon}\mathbf{F}^t\right\} = 0$
- The load matrix is the covariance between the observed variables and the factors $\quad \Lambda = E\left\{(\mathbf{X}-\boldsymbol{\mu}_X)\mathbf{F}^t\right\}$
- The variance of the observed variables can be explained by the loading matrix and the variance of the noise

$$\Sigma_X = \Lambda\Lambda^t + \Sigma_\varepsilon$$

Commonality

$$\sigma^2_{X_i} = \sum_{j=1}^{m} \lambda^2_{ij} + \sigma^2_{\varepsilon_i} = h^2_i + \sigma^2_{\varepsilon_i}$$

4.1 FA: Introduction

$$\mathbf{X} = \boldsymbol{\mu}_X + \Lambda \mathbf{f} + \boldsymbol{\varepsilon}$$

Properties:

• The load matrix and factors are not uniquely specified: any rotation of the factors can be compensated by the load matrix

$$\mathbf{X} = \boldsymbol{\mu}_X + \Lambda \mathbf{f} + \boldsymbol{\varepsilon} = \boldsymbol{\mu}_X + (\Lambda H^t)(H\mathbf{f}) + \boldsymbol{\varepsilon}$$

Any orthogonal matrix

Solution:

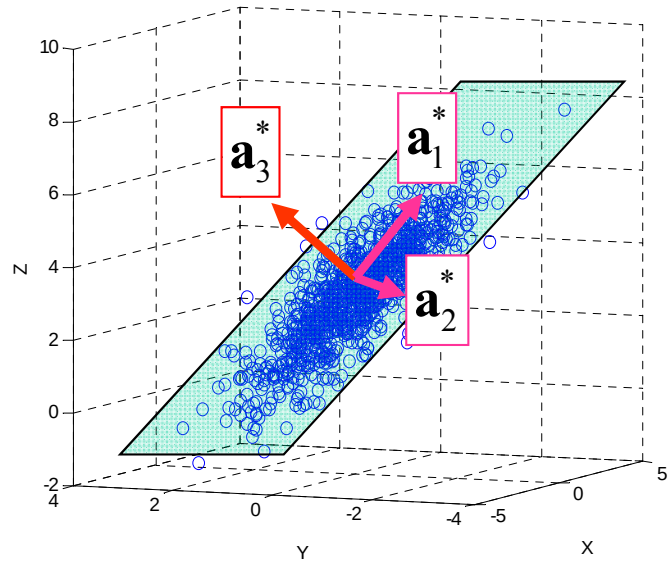
1. Impose that $\Lambda^t \Lambda$ is a diagonal matrix: Principal factor method
2. Impose that $\Lambda^t \Sigma_\varepsilon^{-1} \Lambda$ is a diagonal matrix: Maximum-Likelihood method

This matrix provides the possibility of rotating the components so that we achieve a certain property (having the maximum number of zeros, …) that helps us to understand the factors. There are several criteria offered by the programs: varimax, equamax, parsimax, quartimax, orthomax, …

4.1 FA: Introduction

$$\min_{b_{ij}} \sum_{f=1}^{m} \left(\sum_{i=1}^{p} b_{ij}^4 - \frac{\gamma}{p} \left(\sum_{i=1}^{p} b_{ij}^2 \right)^2 \right)$$

$$b_{ij}^2 = \frac{\lambda_{ij}^2}{\sum_{k=1}^{m} \lambda_{ik}^2}$$



Orthomax	Orthogonal rotation that maximizes a criterion based on the variance of the loadings.	γ
Parsimax	Special case of the orthomax rotation	$\gamma = \frac{p(m-1)}{p+m-2}$
Quartimax	Minimizes the number of factors needed to explain a variable. Special case of orthomax.	$\gamma = 0$
Varimax	Maximizes the variance of the squared loadings of each factor. Special case of orthomax.	$\gamma = 1$
Equimax	Compromise between quatimax and varimax. Special case of orthomax.	$\gamma = \frac{m}{2}$
Promax	Allows for oblique factors (they are less interpretable)	

4.2 FA: Factor computation

<u>Principal factor method</u>

$$\Sigma_X = \Lambda\Lambda^t + \Sigma_\varepsilon \longrightarrow \boxed{\text{Solve for } \Lambda \text{ in } S_X - \hat{\Sigma}_\varepsilon = \Lambda\Lambda^t} \text{ s.t. } \Lambda^t\Lambda \text{ is diagonal}$$

Option a: $\hat{\sigma}^2_{\varepsilon_i} = 0$
Option b: regression residual

This step is also known as commonality estimation since once the variance of the noise is known, the commonalities are also known

$$\sigma^2_{X_i} = h_i^2 + \sigma^2_{\varepsilon_i} \longrightarrow s^2_{X_i} = h_i^2 + \hat{\sigma}^2_{\varepsilon_i}$$

$$\longrightarrow h_i^2 = s^2_{X_i} - \hat{\sigma}^2_{\varepsilon_i}$$

4.2 FA: Factor computation

<u>Maximum Likelihood method</u>

$\mathbf{X} \sim N\left(\boldsymbol{\mu}_X, \Sigma_X\right) \longrightarrow$ Likelihood of observing a single individual

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\left(2\pi\left|\Sigma_X\right|\right)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_X)^t \Sigma_X^{-1}(\mathbf{x}-\boldsymbol{\mu}_X)\right)$$

Likelihood of observing all individuals

$$f_{\mathbf{X}}(X) = \prod_{i=1}^{n} f_X(\mathbf{x}_i)$$

Log-likelihood of observing all individuals

$$L_{\mathbf{X}}(X) = -\frac{pn}{2}\log 2\pi - \frac{n}{2}\log\left|\Sigma_X\right| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu}_X)^t \Sigma_X^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_X)$$

$$\Sigma_X = \Lambda\Lambda^t + \Sigma_\varepsilon$$

4.2 FA: Factor computation

<u>Maximum Likelihood method</u>

$$\left.\begin{aligned}\frac{\partial L_{\mathbf{X}}(X)}{\partial \Lambda} &= 0\\ \frac{\partial L_{\mathbf{X}}(X)}{\partial \Sigma_{\varepsilon}} &= 0\end{aligned}\right\} \longrightarrow \left\{\begin{array}{ll}1. & \text{Estimate an initial guess for } \hat{\Lambda}\\ 2. & \text{Estimate } \hat{\Sigma}_{\varepsilon} \text{ with the current guess of } \hat{\Lambda}\\ & \qquad \hat{\Sigma}_{\varepsilon} = S_X - \hat{\Lambda}\hat{\Lambda}^t\\ 3. & \text{Estimate } \hat{\Lambda} \text{ with the current guess of } \hat{\Sigma}_{\varepsilon}\,.\\ & \text{Solve for } \hat{\Lambda} \text{ in}\\ & \qquad \left(\hat{\Sigma}_{\varepsilon}^{-\frac{1}{2}}(S-I)\hat{\Sigma}_{\varepsilon}^{-\frac{1}{2}}\right)\left(\hat{\Sigma}_{\varepsilon}^{-\frac{1}{2}}\hat{\Lambda}\right) = \left(\hat{\Sigma}_{\varepsilon}^{-\frac{1}{2}}\hat{\Lambda}\right)\left(\hat{\Lambda}^t\hat{\Sigma}_{\varepsilon}^{-1}\hat{\Lambda}\right)\\ 4. & \text{Return to Step 2 till convergence}\end{array}\right.$$

4.3 FA: Example

120 students have each taken **five exams**, the first two covering **mathematics**, the next two on **literature**, and a **comprehensive** fifth exam. It seems reasonable that the five grades for a given student ought to be related. Some students are good at both subjects, some are good at only one, etc. The goal of this analysis is to determine if there is quantitative evidence that **the students' grades on the five different exams are largely determined by only two types of ability**.

```
Loadings =

    0.6289    0.3485
    0.6992    0.3287
    0.7785   -0.2069
    0.7246   -0.2070
    0.8963   -0.0473
```

A factor distinguishing those good in mathematics and bad in literature or viceversa

An overall factor affecting to all exams (mainly the comprehensive, then literature and finally mathematics). This can be interpreted as a general intelligence factor.

```
NoiseVar =

    0.4829
    0.4031
    0.3512
    0.4321
    0.1944
```

The comprehensive exam is the easiest to predict with these two factors

Normalized to 1 (1=no variance reduction obtained by commonalities; 0=all variance is explained by commonalities)

4.3 FA: Example

Example:

count =

11	11	9
7	13	11
14	17	20
11	13	9
43	51	69

…

Traffic count in three different places (thousands/day)

```
Loadings =      NoiseVar =

    0.9683          0.0624
    0.9636          0.0714
    0.9913          0.0172
```

As expected X3 is highly related to the main factor, and it mostly explains the traffic in X1 and X2. The noise is coming from the red arrows.

4.4 FA: Extensions

- Multiple factor analysis (MFA): Study common factors along several datasets about the same individuals
- Hierarchical FA and MFA (HFA and HMFA), Higher order FA: Apply FA to the z scores obtained after a first FA
- Nonlinear FA: Nonlinear relationship between factors and observed variables.
- Mixture of Factor Analyzers: formulation of the problem as a Gaussian mixture.

4.5 Rules of thumb

- At least 10 times as many subjects as you have variables.
- At least 6 variables per expected factor (if loadings are low, you need more)
- At least 3 variables should correlate well with each factor
- Each factor should have at least 3 variables that load well.
- If a variable correlates well with several factors, you need more subjects to provide significance.
- The size of commonalities is proportional to the number of subjects.

4.6 Comparison with PCA

Principal Component Analysis:

- Selection of number of factors a posteriori
- Linear model
- Rotation is possible at the end
- Output variables are orthogonal
- Many extensions

Factor analysis:

- Selection of the number of factors a priori
- Linear model
- Rotation is possible at the end
- Output variables may not be orthogonal
- Few extensions
- Assumption of normality

Course outline: Session 2

3. Principal component analysis (PCA)
    - 3.1. Introduction
    - 3.2. Component computation
    - 3.3. Example
    - 3.4. Properties
    - 3.5. Extensions
    - 3.6. Relationship to SVD

4. Factor Analysis (FA)
    - 4.1. Introduction
    - 4.2. Factor computation
    - 4.3. Example
    - 4.4. Extensions
    - 4.5. Rules of thumb
    - 4.6. Comparison with PCA

CEU
Universidad
San Pablo

Multivariate Data Analysis

Session 3: Multidimensional scaling, correspondence analysis, tensor analysis

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid

Course outline: Session 3

5. Multidimensional Scaling (MDS)
    - 5.1. Introduction
    - 5.2. Metric scaling
    - 5.3. Example
    - 5.4. Nonmetric scaling
    - 5.5. Extensions
6. Correspondence analysis
    - 6.1. Introduction
    - 6.2. Projection search
    - 6.3. Example
    - 6.4. Extensions
7. Tensor analysis
    - 7.1 Introduction
    - 7.2 Parafac/Candecomp
    - 7.3 Example
    - 7.4 Extensions

5.1 MDS: Introduction

Factor Analysis
- Requires input Gaussianity
- Distance between individuals through the covariance matrix
- Individuals are directly observed

Multidimensional Scaling
- Does not require Gaussianity
- Any distance among individuals

- Individuals are not observed, but their relative distances are

$$\mathbf{x}_1 = (\ldots)$$
$$\mathbf{x}_2 = (\ldots)$$
$$\mathbf{x}_3 = (\ldots)$$

$d(\mathbf{x}_i, \mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	0.3	2
$\mathbf{x}_2$	0.3	0	1.5
$\mathbf{x}_3$	2	1.5	0

It may not be a true (in the mathematical sense) distance measure but a dissimilarity measure.

5.2 MDS: Metric scaling

Zero-mean Data Matrix

Similarity Matrix

Dissimilarity (Euclidean distance) Matrix $\longrightarrow$ $\boxed{\text{MDS=PCA}}$

$$X \longrightarrow \tilde{X} \longrightarrow Q = \tilde{X}\tilde{X}^t \longrightarrow D$$

Data matrix

$$q_{ij} = \mathbf{x}_i^t \mathbf{x}_j$$

$$\mathbf{x}_i \approx \mathbf{x}_j \Rightarrow q_{ij} \uparrow$$

$$\mathbf{x}_i \neq \mathbf{x}_j \Rightarrow q_{ij} \downarrow$$

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

$$Q = -\frac{1}{2}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right) D \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^t\right)^t$$

Q is positive semidefinite iff D has been built with the Euclidean distance

Assuming Q is positive semidefinite

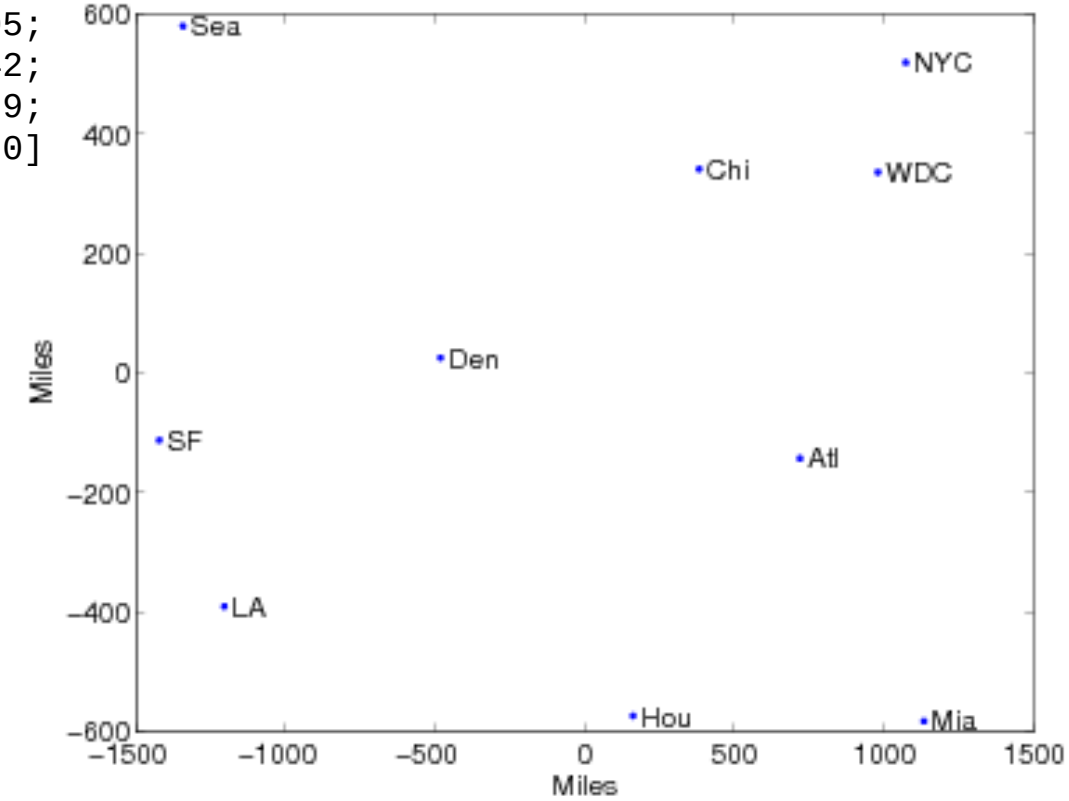
$$Q = PDP^t = \left(PD^{\frac{1}{2}}\right)\left(PD^{\frac{1}{2}}\right)^t$$

$$\tilde{Y} = PD^{\frac{1}{2}}$$

Equal up to a rotation

5.3 MDS: Example

```
cities = {'Atl','Chi','Den','Hou','LA','Mia','NYC','SF','Sea','WDC'};
D = [   0  587 1212  701 1936  604  748 2139 2182  543;
      587    0  920  940 1745 1188  713 1858 1737  597;
     1212  920    0  879  831 1726 1631  949 1021 1494;
      701  940  879    0 1374  968 1420 1645 1891 1220;
     1936 1745  831 1374    0 2339 2451  347  959 2300;
      604 1188 1726  968 2339    0 1092 2594 2734  923;
      748  713 1631 1420 2451 1092    0 2571 2408  205;
     2139 1858  949 1645  347 2594 2571    0  678 2442;
     2182 1737 1021 1891  959 2734 2408  678    0 2329;
      543  597 1494 1220 2300  923  205 2442 2329    0]
```



5.3 MDS: Example

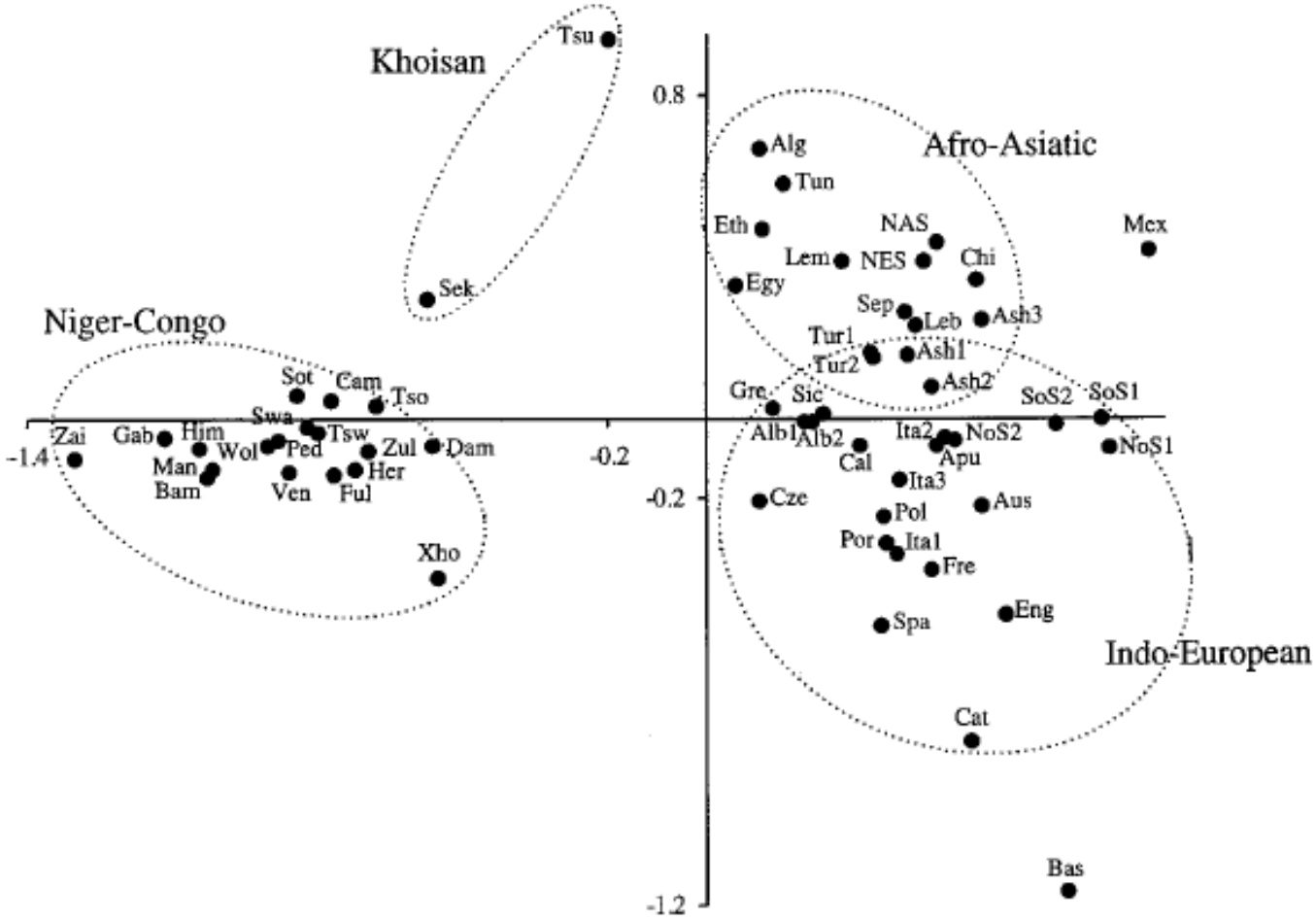


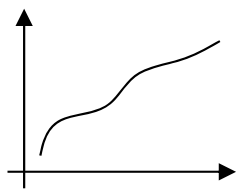
Figure 2 Multidimensional scaling analysis of 58 samples tested for Y chromosome–specific p49a,f/*Taq*I polymorphism. Genetic and linguistic distances are significantly correlated ($r = .567$, $P < .001$).

5.4 MDS: Nonmetric scaling

$d(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	0.3	2
$\mathbf{x}_2$	0.3	0	1.5
$\mathbf{x}_3$	2	1.5	0

$$\delta(\mathbf{x}_i,\mathbf{x}_j) = f\left(d(\mathbf{x}_i,\mathbf{x}_j)\right)$$

Nonlinear, monotonically increasing and unknown



$\delta(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$	0	0.4	1.8
$\mathbf{x}_2$	0.4	0	1.2
$\mathbf{x}_3$	1.8	1.2	0

MDS

$\hat{d}(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\mathbf{x}_1$			
$\mathbf{x}_2$			
$\mathbf{x}_3$			

STRESS Objective function

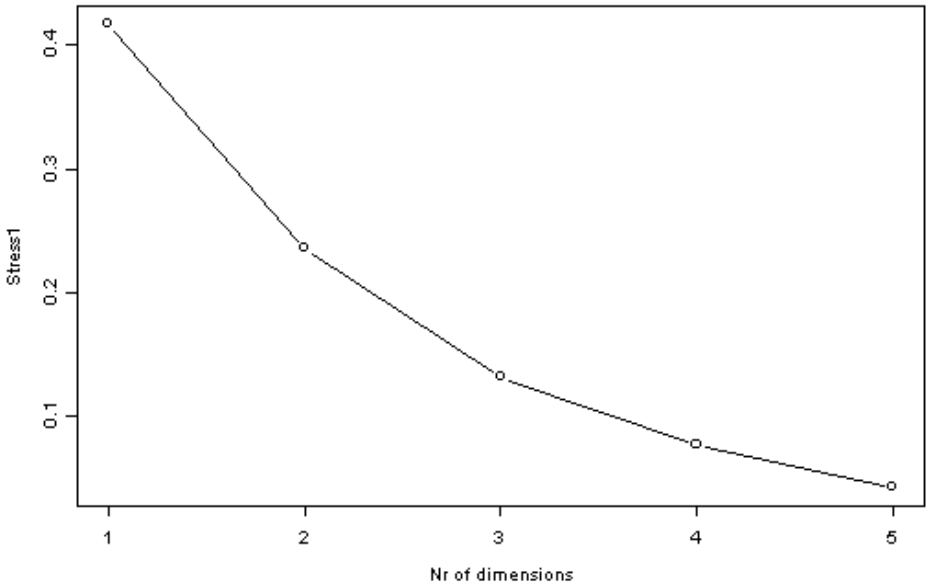
$$\hat{d}_{ij}^{*} = \arg\min_{\hat{d}_{ij}} \frac{\overbrace{\sum_{i=1}^{n}\sum_{j=i+1}^{n}(\delta_{ij}-\hat{d}_{ij})^2}^{\text{STRESS Objective function}}}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{d}_{ij}^2}$$

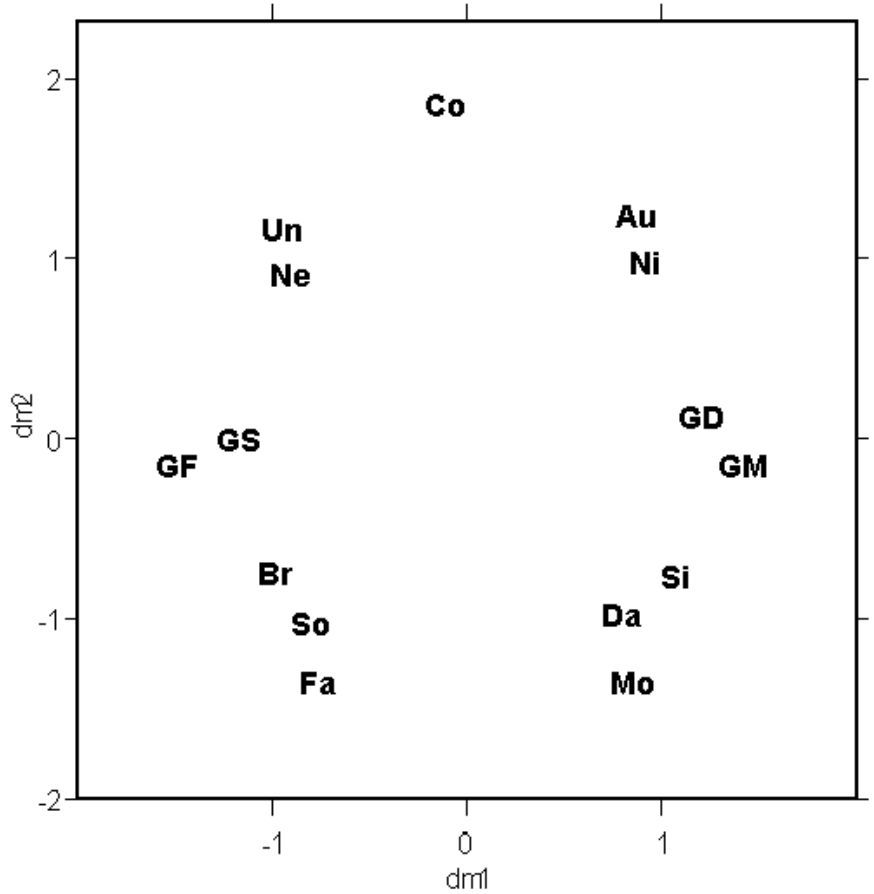
Solution:

$$\hat{\mathbf{x}}_i^{(k+1)} = \hat{\mathbf{x}}_i^{(k)} + \frac{\alpha}{n-1}\sum_{j=1,\, j\neq i}^{n}\left(1-\frac{\hat{d}_{ij}}{\delta_{ij}}\right)(\hat{\mathbf{x}}_j^{(k)}-\hat{\mathbf{x}}_i^{(k)})$$

5.4 MDS: Nonmetric scaling

Example: Sort (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son and uncle) according to their "meaning"





5.5 MDS: Extensions

- Classical MDS: D is computed using Euclidean distance.
- Metric MDS: D is computed using a distance but may be not Euclidean
- Nonmetric MDS: D is not a distance but provides the right ranks
- Replicated MDS: Several experts are consulted to build D
- Weighted MDS: It allows experts to apply different nonlinear transformations.
- ISOMAP: Uses geodesic distance
- MDS by local patches: approximate data locally by small patches and embed these patches in a low dimensional space.

Course outline: Session 3

5. Multidimensional Scaling (MDS)
    - 5.1. Introduction
    - 5.2. Metric scaling
    - 5.3. Example
    - 5.4. Nonmetric scaling
    - 5.5. Extensions
6. Correspondence analysis
    - 6.1. Introduction
    - 6.2. Projection search
    - 6.3. Example
    - 6.4. Extensions
7. Tensor analysis
    - 7.1 Introduction
    - 7.2 Parafac/Candecomp
    - 7.3 Example
    - 7.4 Extensions

6.1 CA: Introduction

Table 1: Science Doctorates in the USA, 1960-1975

Discipline/Year	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	134	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Agriculture	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

8-dimensional vector: which topics are closer?

12-dimensional vector: which years are more similar?

F Frequency matrix summing up to 1

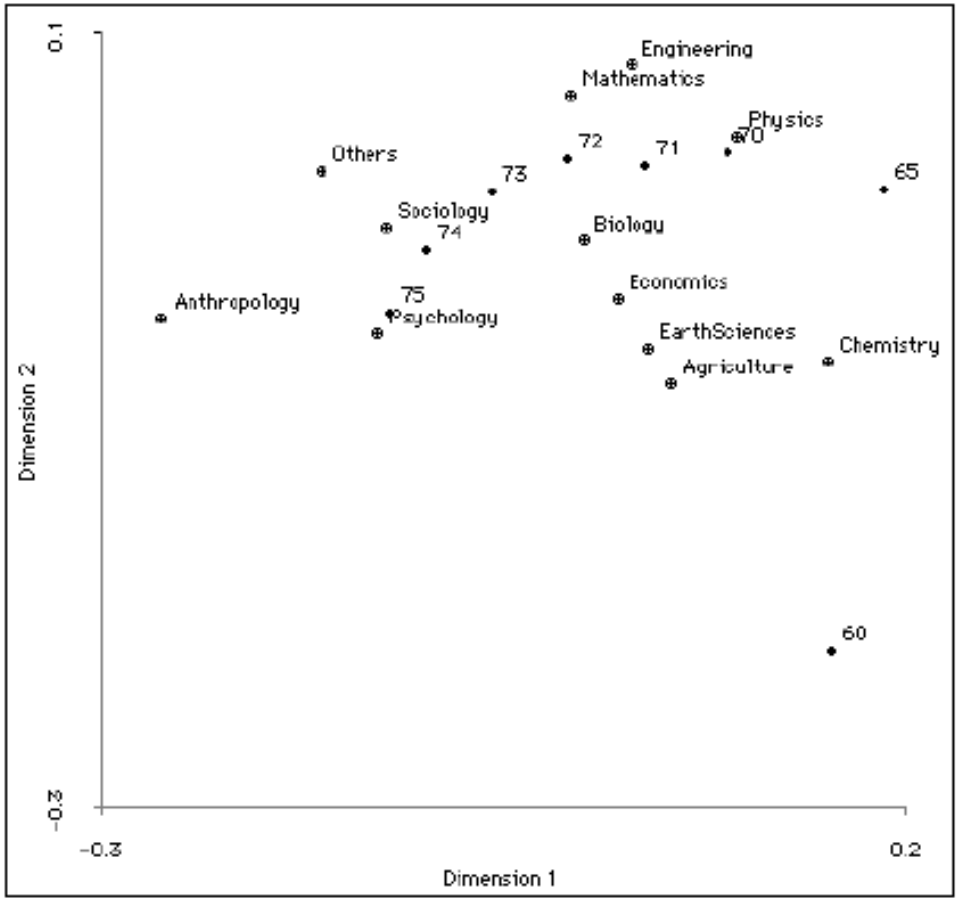


Figure 1: Correspondence Analysis of Doctorate Data

6.2 CA: Projection search

Table 1: Science Doctorates in the USA, 1960-1975

Discipline/Year	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	134	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Agriculture	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

I

J

1. We should consider <u>the structure of the rows</u>. For instance, the following two rows should be equivalent.

```
--0.05-0.05--> 0.10 = f_i.
--0.45-0.45--> 0.90 Row mass
```

For achieving this we devide by the row sum

```
0.50 0.50
0.50 0.50
```

In matrix form, we define a new frequency matrix

$$R = D_{row}^{-1} F$$

Where D_{row} is a diagonal matrix with the row sums.

$$D_{row} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.9 \end{pmatrix}$$

6.2 CA: Projection search

Table 1: Science Doctorates in the USA, 1960-1975

Discipline/Year	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	134	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Agriculture	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

I

J

2. A change in probability from 0.6001 to 0.6101 is not that much as a change from 0.0001 to 0.0101, i.e., <u>we have to weight (divide) attributes by the relative frequency</u> of the attribute.

```
0.6001 0.0001 0.3998
0.6101 0.0101 0.3798

1.2101 0.0102 0.7796 = f.j
```

Column mass

After division we have

```
0.4959 0.0098 0.5128
0.5041 0.9902 0.4872
```

$$D_{col} = \begin{pmatrix} 1.2101 & 0 & 0 \\ 0 & 0.0102 & 0 \\ 0 & 0 & 0.7796 \end{pmatrix}$$

We will later use the diagonal matrix D_{col} whose values are the column sums.

6.2 CA: Projection search

Distance between two rows

$$d^2_{\chi^2}(\mathbf{r}_a,\mathbf{r}_b) = \left(\mathbf{r}_a - \mathbf{r}_b\right)^t D_{col}^{-1}\left(\mathbf{r}_a - \mathbf{r}_b\right) = d^2_{Euclidean}(D_{col}^{-\frac{1}{2}}\mathbf{r}_a, D_{col}^{-\frac{1}{2}}\mathbf{r}_b)$$

$$= d^2_{Euclidean}\left(D_{col}^{-\frac{1}{2}}\frac{\mathbf{f}_a}{f_{a.}}, D_{col}^{-\frac{1}{2}}\frac{\mathbf{f}_b}{f_{b.}}\right) = \sum_{j=1}^{J}\left(\frac{f_{aj}}{f_{a.}} - \frac{f_{bj}}{f_{b.}}\right)^2 \frac{1}{f_{.j}}$$

We may transform directly the matrix F into some other matrix whose rows are the ones needed for the Eucliden distance

$$F \longrightarrow \boxed{\begin{array}{l} Y = D_{row}^{-1} F D_{col}^{-\frac{1}{2}} \\ d^2_{\chi^2}(\mathbf{r}_a,\mathbf{r}_b) = d^2_{Euclidean}(\mathbf{y}_a,\mathbf{y}_b) \end{array}}$$

$$y_{ij} = \frac{f_{ij}}{f_{i.} f_{.j}^{-\frac{1}{2}}}$$

6.2 CA: Projection search

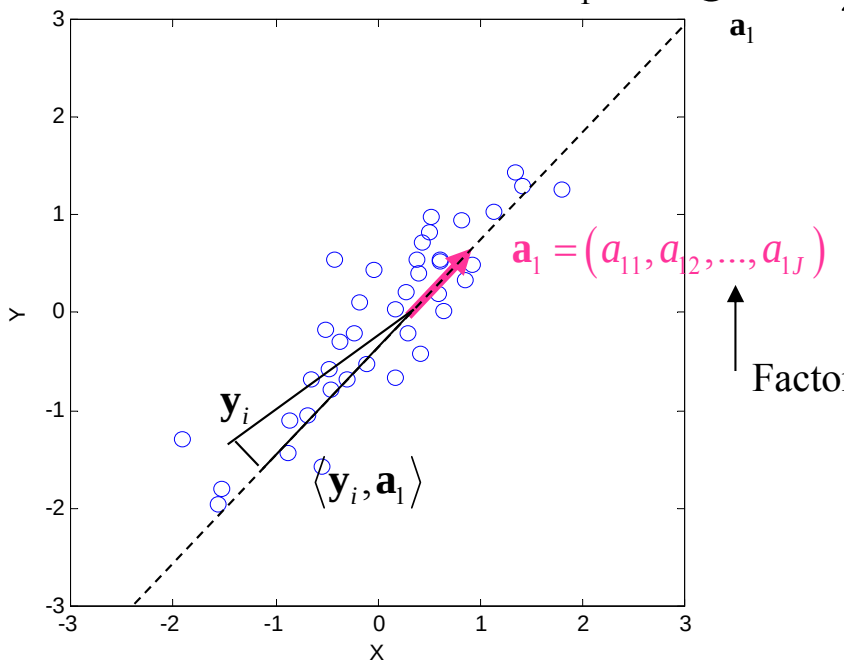
<u>Search of a projection direction</u>

Y $\xrightarrow{\text{PCA}}$ $\mathbf{a}_1^* = \arg\max_{\mathbf{a}_1} \sum_{i=1}^{I} \langle \mathbf{y}_i, \mathbf{a}_1 \rangle^2 = \arg\max_{\mathbf{a}_1} \mathbf{a}_1^t Y^t Y \mathbf{a}_1$

Data matrix

Equal weight to all rows. These are directions of maximum inertia (instead of maximum variance)

Y $\xrightarrow{\text{CA}}$ $\mathbf{a}_1^* = \arg\max_{\mathbf{a}_1} \sum_{i=1}^{I} f_{i.} \langle \mathbf{y}_i, \mathbf{a}_1 \rangle^2 = \arg\max_{\mathbf{a}_1} \mathbf{a}_1^t Y^t D_{row} Y \mathbf{a}_1$



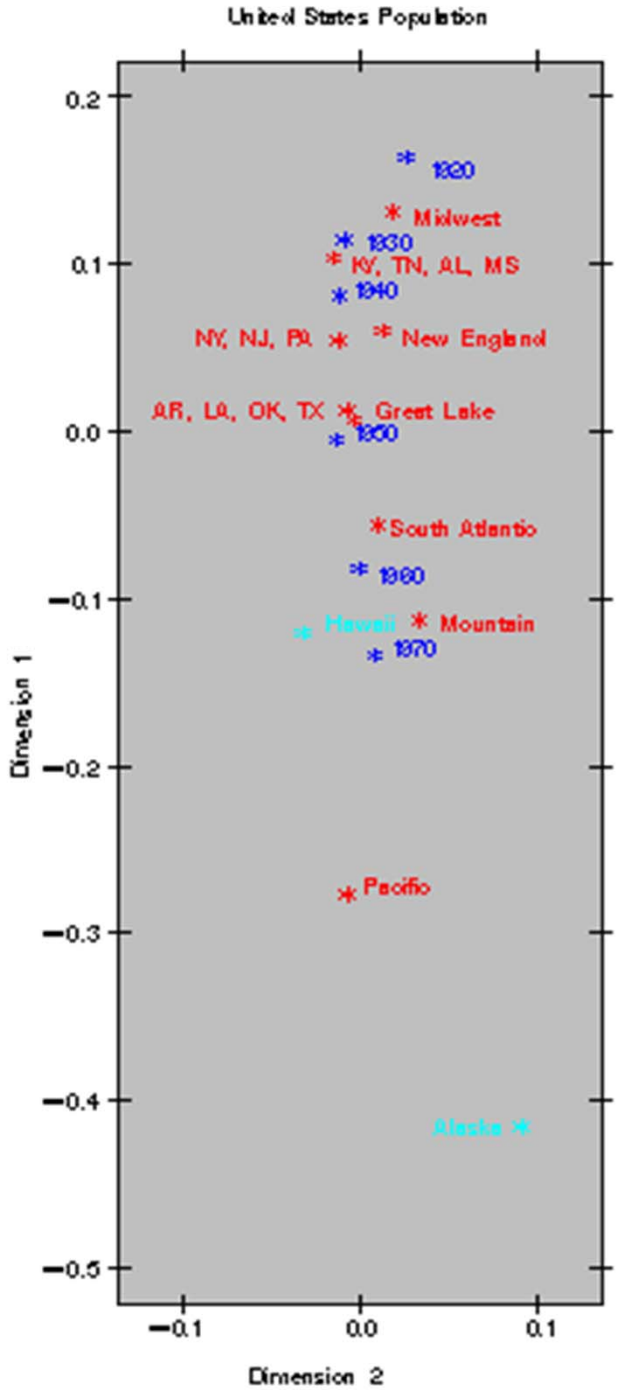
Factor loadings

This analysis can also be done by columns, instead of rows

6.3 CA: Example

Contingency Table							
Percents	1920	1930	1940	1950	1960	1970	Sum
New England	0.830	0.916	0.946	1.045	1.179	1.328	6.245
NY, NJ, PA	2.497	2.946	3.089	3.382	3.833	4.173	19.921
Great Lake	2.409	2.838	2.987	3.410	4.064	4.516	20.224
Midwest	1.407	1.492	1.516	1.577	1.727	1.831	9.550
South Atlantic	1.569	1.772	1.999	2.376	2.914	3.441	14.071
KY, TN, AL, MS	0.998	1.109	1.209	1.284	1.352	1.436	7.388
AR, LA, OK, TX	1.149	1.366	1.466	1.631	1.902	2.167	9.681
Mountain	0.374	0.415	0.466	0.569	0.769	0.929	3.523
Pacific	0.625	0.919	1.092	1.625	2.282	2.855	9.398
Sum	11.859	13.773	14.771	16.900	20.020	22.677	100.000

United States population



6.4 CA: Extensions

- Multiple CA: Extension to more than two variables.
- Joint CA: Another extension to more than two variables.
- Detrended CA: Postprocessed CA to remove the “arch” effect

Course outline: Session 3

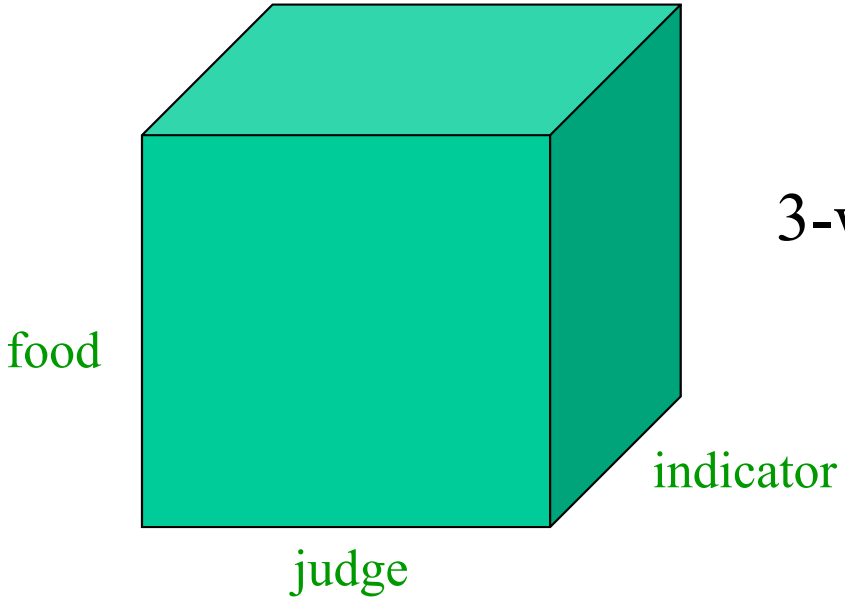
5. Multidimensional Scaling (MDS)
    - 5.1. Introduction
    - 5.2. Metric scaling
    - 5.3. Example
    - 5.4. Nonmetric scaling
    - 5.5. Extensions
6. Correspondence analysis
    - 6.1. Introduction
    - 6.2. Projection search
    - 6.3. Example
    - 6.4. Extensions
7. Tensor analysis
    - 7.1 Introduction
    - 7.2 Parafac/Candecomp
    - 7.3 Example
    - 7.4 Extensions

7.1 Tensor analysis: Introduction

Examples:

- Scores of n subjects, at m tests, at p time points.
- Scores of n air quality indicators on m time points at p locations.
- Scores of n judges on m quality indicators for p food products.

Score=f(food, judge, indicator)

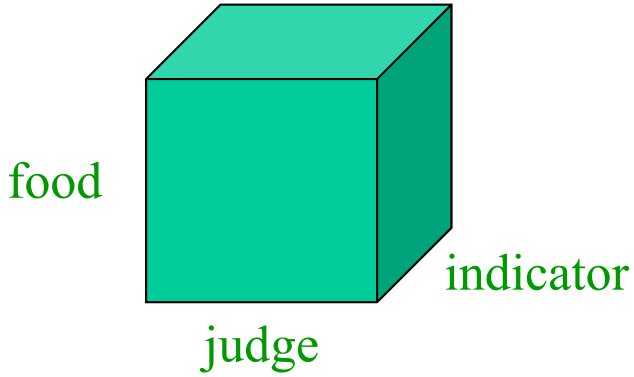


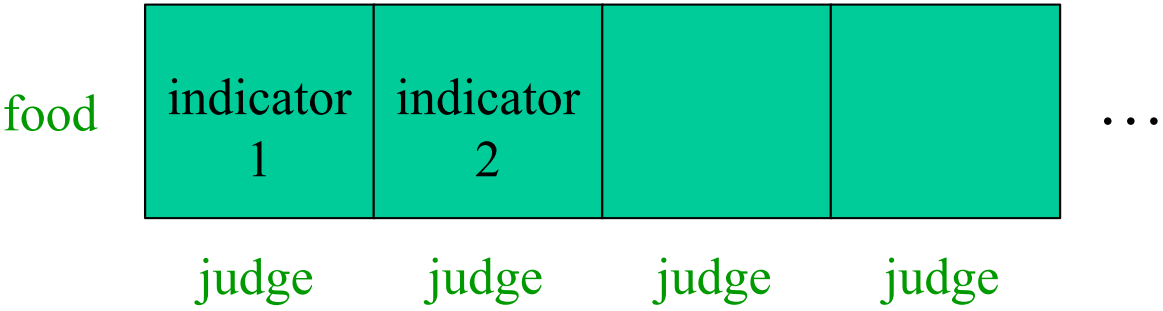
3-way tensor

7.1 Tensor analysis: Introduction

<u>Traditional analysis</u>

- Unfolding





- PCA Analysis

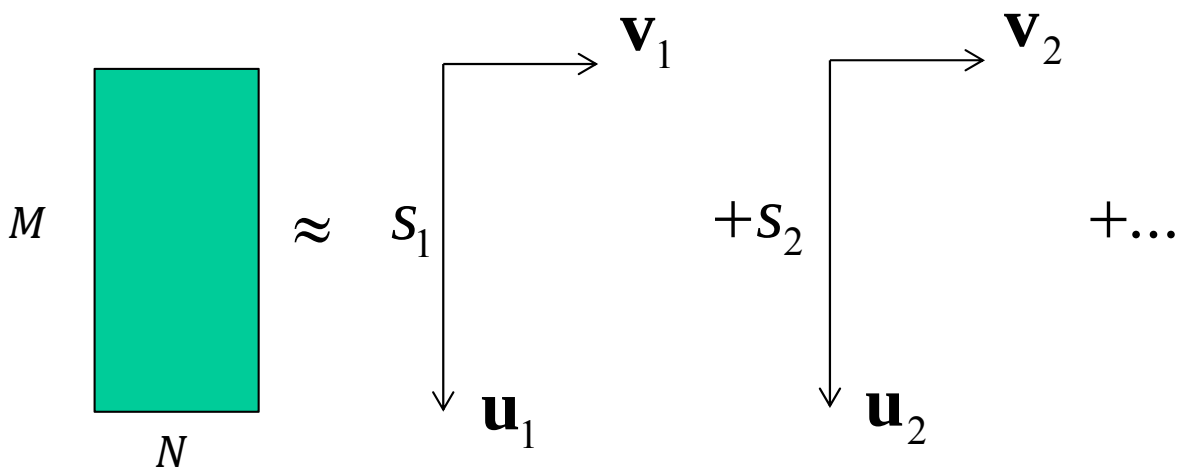
7.2 Tensor analysis: PARAFAC

<u>SVD (PCA) revisited</u>

$$X \approx USV^t = s_1\mathbf{u}_1\mathbf{v}_1^t + s_2\mathbf{u}_2\mathbf{v}_2^t + \ldots + s_m\mathbf{u}_m\mathbf{v}_m^t = \sum_{k=1}^{m} s_k\mathbf{u}_k \otimes \mathbf{v}_k$$

$\uparrow$ $\mathrm{M} \times N$ $\qquad$ $\uparrow$ $(\mathrm{M} \times 1) \times (1 \times N)$ $\qquad$ $\uparrow$

Outer product $\quad (\mathbf{u} \otimes \mathbf{v})_{ij} = u_i v_j$



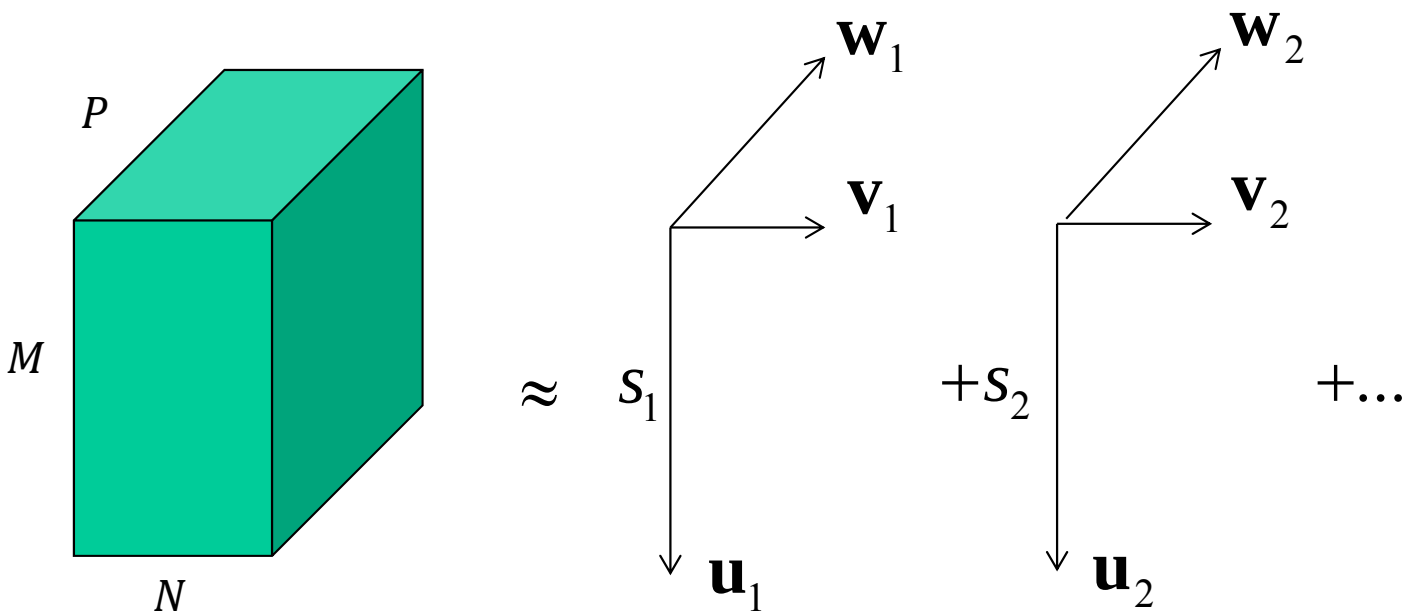
7.2 Tensor analysis: PARAFAC

PARAFAC

$$X \approx \sum_{k=1}^{m} s_k \mathbf{u}_k \otimes \mathbf{v}_k \otimes \mathbf{w}_k \qquad \mathbf{u}_k \in \mathbb{R}^M, \mathbf{v}_k \in \mathbb{R}^N, \mathbf{w}_k \in \mathbb{R}^P$$

$\uparrow$ $\mathrm{M} \times N \times P$

$$\left(\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w}\right)_{ijk} = u_i v_j w_k \quad \text{Outer product}$$



7.2 Tensor analysis: PARAFAC

<u>SVD(PCA) vs PARAFAC</u>

- PARAFAC cannot be found analytically.
- PARAFAC has to be found numerically (Alternating Least Squares).
- The best rank-m approximation may not exist.
- The m components are not ordered.
- The best rank-(m-1) approximation is not a subset of the rank-m solution.
- PARAFAC does not maximize the explained variance.
- PARAFAC minimizes the approximation error.
- PARAFAC solution is unique (PCA is not, rotations).
- PARAFAC factorization may be real or complex-valued.

7.2 Tensor analysis: PARAFAC

Preprocessing for SVD (PCA)

$$X \approx USV^t$$

$\uparrow$

$\mathrm{M} \times N$ M individuals, N variables

Center and standardize columns of X

$$x'_{ij} = \frac{x_{ij} - x_{\bullet j}}{\sigma_{\bullet j}}$$

Why center? Scores are normally relative (origin is arbitrary).

Why normalize? Assures equal influence of each variable.

Preprocessing for PARAFAC

More possibilities, depending on interest

$$x'_{ijk} = \frac{x_{ijk} - x_{\bullet jk}}{\sigma_{\bullet jk}} \qquad x'_{ijk} = \frac{x_{ijk} - x_{\bullet\bullet k}}{\sigma_{\bullet\bullet k}} \qquad \ldots$$

7.2 Tensor analysis: PARAFAC

How to choose the number of components

$$X \approx \sum_{k=1}^{m} s_k \mathbf{u}_k \otimes \mathbf{v}_k \otimes \mathbf{w}_k$$

$$X = \hat{X} + E$$

$$R^2 = \frac{\|X\|_F^2 - \|E\|_F^2}{\|X\|_F^2}$$

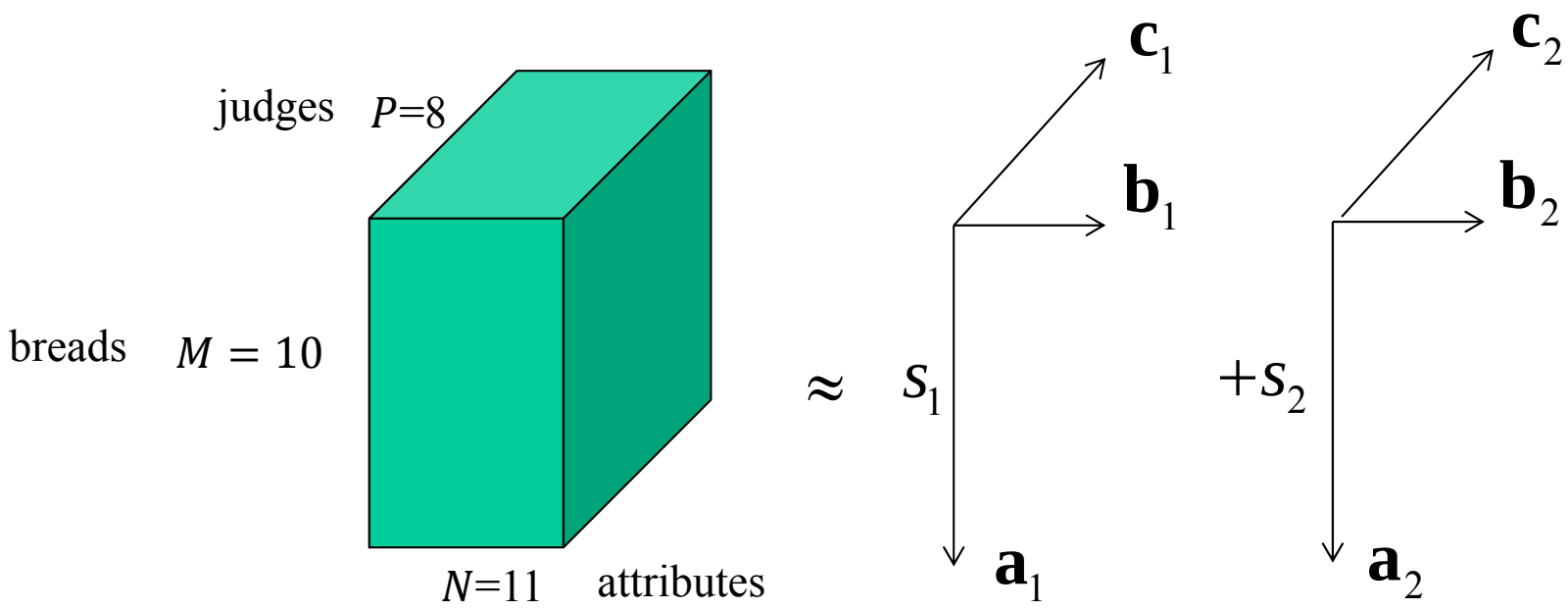
m	1	2	3	4	5
R^2	0.102	0.164	0.187	0.189	0.191

Chosen m

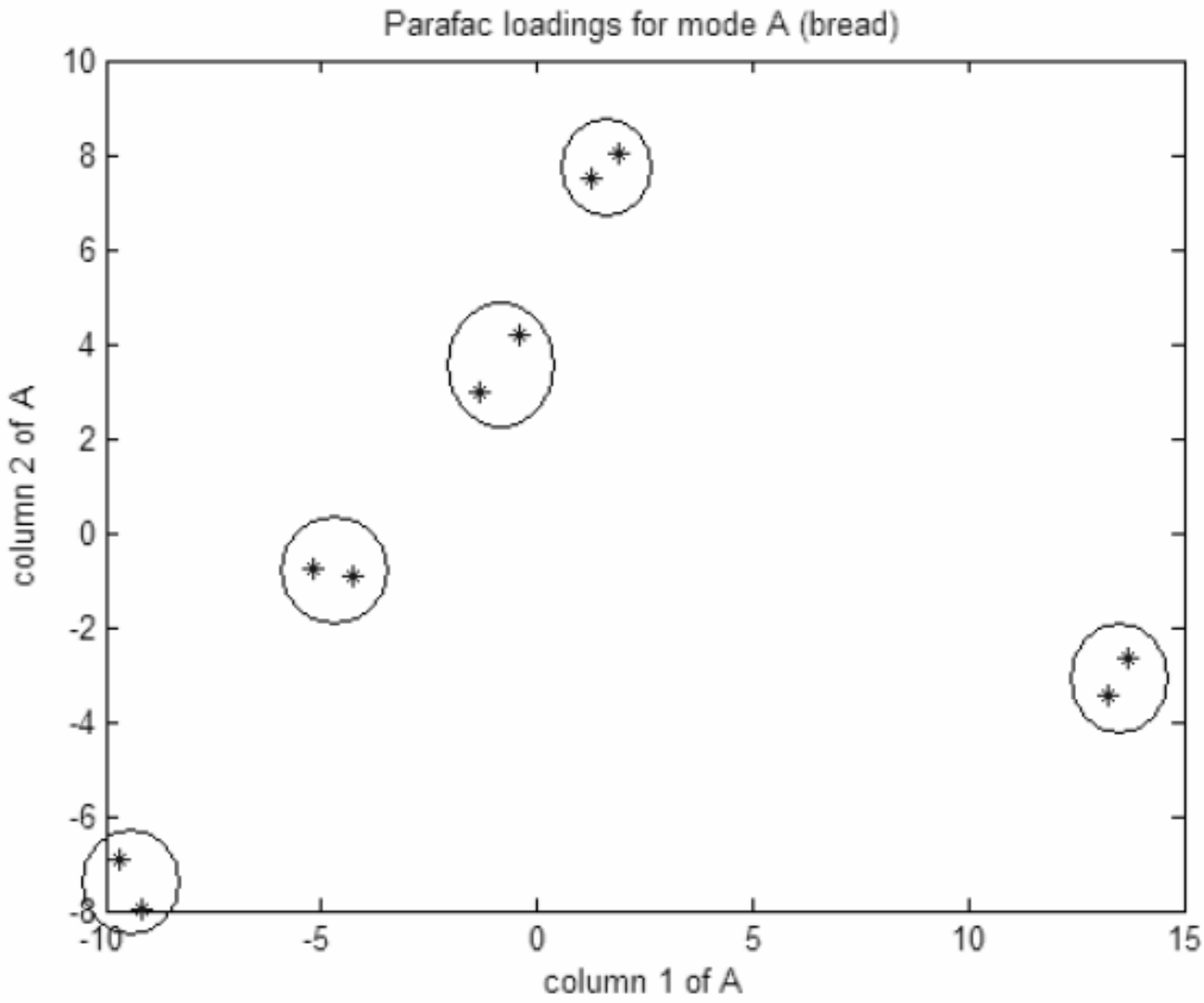
7.3 Tensor analysis: Example

5 different kinds of breads are judged by 8 judges on 11 attributes.
There are 2 replicates (→10 breads).

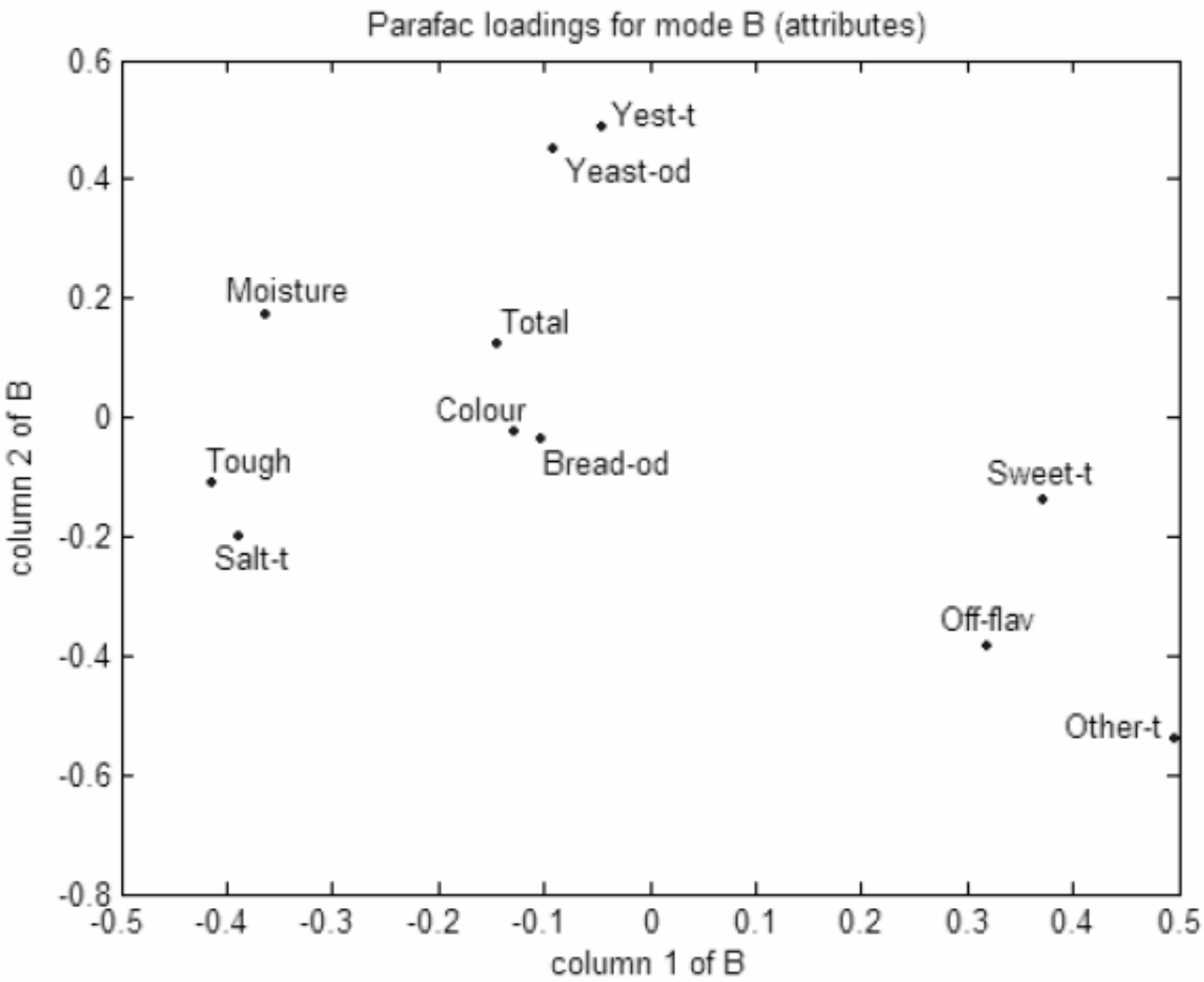
$$\underset{\substack{\uparrow \\ \mathrm{M} \times N \times P}}{X} \approx \sum_{k=1}^{2} s_k \mathbf{a}_k \otimes \mathbf{b}_k \otimes \mathbf{c}_k$$



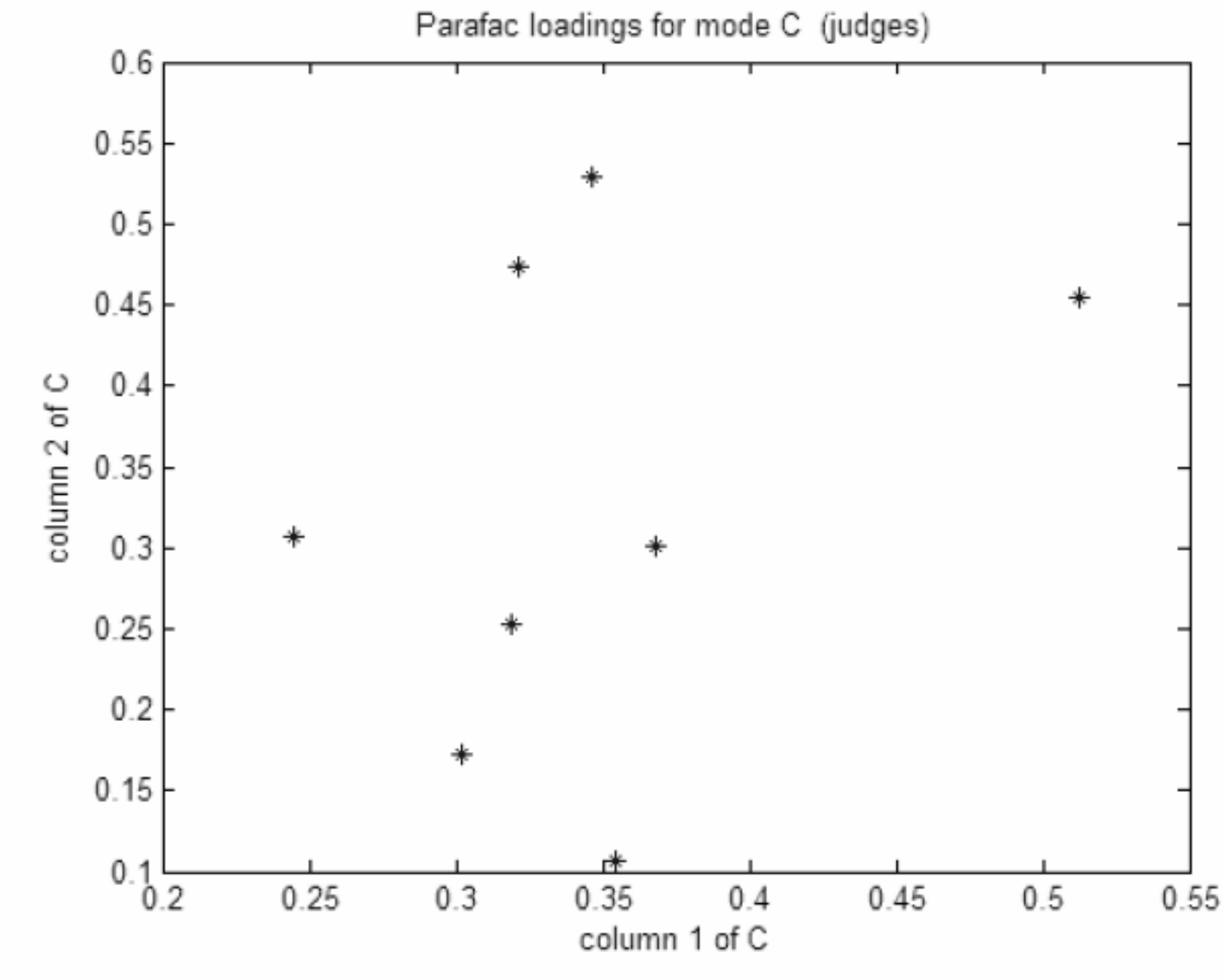
7.3 Tensor analysis: Example



7.3 Tensor analysis: Example



7.3 Tensor analysis: Example



7.4 Tensor analysis: Extensions

PARAFAC

$$X \approx \sum_{k=1}^{m} s_k \mathbf{u}_k \otimes \mathbf{v}_k \otimes \mathbf{w}_k \qquad \mathbf{u}_k \in \mathbb{R}^M, \mathbf{v}_k \in \mathbb{R}^N, \mathbf{w}_k \in \mathbb{R}^P$$

Tucker model

$$X \approx \sum_{k_1=1}^{m_1} \sum_{k_2=1}^{m_2} \sum_{k_3=1}^{m_3} s_{k_1 k_2 k_3} \mathbf{u}_{k_1} \otimes \mathbf{v}_{k_2} \otimes \mathbf{w}_{k_3}$$

Principal Tensor Analysis (Multiway, multivariate analysis)

General name for this kind of algorithms to analyze tensors.

7.4 Tensor analysis: Extensions

<u>Difference with ANOVA:</u>

In ANOVA, interactions are modeled as δ_{ijk} while in PARAFAC, they are modeled as $a_i b_j c_k$. ANOVA allows more general interaction models, but if the multiplicative model is true, then PARAFAC can be better interpreted.

Course outline: Session 3

5. Multidimensional Scaling (MDS)
    - 5.1. Introduction
    - 5.2. Metric scaling
    - 5.3. Example
    - 5.4. Nonmetric scaling
    - 5.5. Extensions
6. Correspondence analysis
    - 6.1. Introduction
    - 6.2. Projection search
    - 6.3. Example
    - 6.4. Extensions
7. Tensor analysis
    - 7.1 Introduction
    - 7.2 Parafac/Candecomp
    - 7.3 Example
    - 7.4 Extensions

CEU
Universidad
San Pablo

Multivariate Data Analysis

Session 4: MANOVA, Canonical Correlation, Latent Class Analysis

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid

Course outline: Session 4

8. Multivariate Analysis of Variance (MANOVA)
    - 8.1. Introduction
    - 8.2. Computations (1-way)
    - 8.3. Computations (2-way)
    - 8.4. Post-hoc tests
    - 8.5. Example
9. Canonical Correlation Analysis (CCA)
    - 9.1. Introduction
    - 9.2. Construction of the canonical variables
    - 9.3. Example
    - 9.4. Extensions
10. Latent Class Analysis (LCA)
    - 10.1. Introduction

8.1 MANOVA: Introduction

Several metric variables are predicted by several categorical variables.

Example:
(Ability in Math, Ability in Physics)=f(Math textbook, Physics textbook, College)

1. What are the main effects of the independent variables?
2. What are the interactions among the independent variables?
3. What is the importance of the dependent variables?
4. What is the strength of association between dependent variables?

	Math Text A	Math Text B
Physics Text A College A	(9,9) (7,9) (10,6) (6,7)	(7,7) (4,5) (10,10) (9,9)
Physics Text A College B	(3,1) (5,5) (5,5) (5,5)	(6,7) (8,7) (8,8) (9,8)
Physics Text B College A	(2,8) (9,10) (10,10) (6,9)	(9,6) (5,4) (1,3) (8,8)
Physics Text B College B	(10,8) (7,5) (5,5) (6,5)	(6,6) (7,7) (8,3) (9,7)

8.1 MANOVA: Introduction

1-way

Math Text A	Math Text B
(9,9) (7,9) (10,6) (6,7)	(7,7) (4,5) (10,10) (9,9)

2-way

	Math Text A	Math Text B
Physics Text A	(9,9) (7,9) (10,6) (6,7)	(7,7) (4,5) (10,10) (9,9)
Physics Text A	(3,1) (5,5) (5,5) (5,5)	(6,7) (8,7) (8,8) (9,8)

3-way

	Math Text A	Math Text B
Physics Text A College A	(9,9) (7,9) (10,6) (6,7)	(7,7) (4,5) (10,10) (9,9)
Physics Text A College B	(3,1) (5,5) (5,5) (5,5)	(6,7) (8,7) (8,8) (9,8)
Physics Text B College A	(2,8) (9,10) (10,10) (6,9)	(9,6) (5,4) (1,3) (8,8)
Physics Text B College B	(10,8) (7,5) (5,5) (6,5)	(6,6) (7,7) (8,3) (9,7)

8.1 MANOVA: Introduction

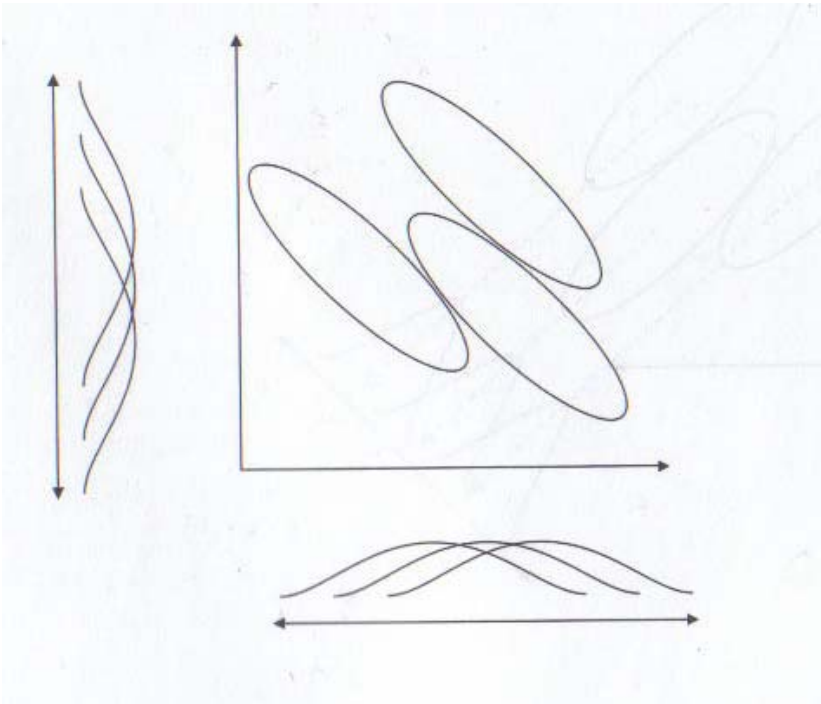
Analysis technique selection guide

	Number of Dependent Variables	
Number of Groups in Independent Variable	**One (Univariate)**	**Two or More (Multivariate)**
Two Groups (Specialized Case)	Student's t-test	Hotelling's T^2 test
Two or More Groups (Generalized Case)	Analysis of Variance (ANOVA)	Multivariate Analysis of Variance (MANOVA)

8.1 MANOVA: Introduction

<u>Why not multiple ANOVAs?</u>

1. Independent ANOVAs cannot “see” covariation patterns among dependent variables.
2. MANOVA may identify small differences while independent ANOVAs may not
3. MANOVA is sensitive to mean differences, the direction and size of correlations among dependents.
4. Running multiple ANOVAs results in increasing Type I errors (multiple testing)



8.1 MANOVA: Introduction

Assumptions:

- Normality: Data is assumed to be normally distributed. However, MANOVA is rather robust to non-Gaussian distributions (particularly so if cell size>20 or 30). Outliers should be removed before applying MANOVA.
- Homogeneity of covariances: Dependent variables have the same covariance matrix within each cell. MANOVA is relatively robust if the number of samples per cell is relatively the same.
- Sample size per group is a critical issue (design of experiments): must be larger than the number of dependent variables, approximately equal number of samples per cell. Large samples make MANOVA more robust to violations.

8.1 MANOVA: Introduction

Assumptions:

- Linearity: MANOVA assumes linear relationships among all pairs of dependent variables, all pairs of covariates, and all pairs of dependent variable-covariate within a cell.
  - MANOVA works best if depedent variables are moderately correlated.
  - If the variables are not correlated, use independent ANOVAs-
  - If the variables are too correlated (>0.8), the covariance matrix becomes nearly singular and calculations are ill-conditioned. (Remove collinearity by PCA or similar; then run independent ANOVAs)

8.2 MANOVA: Computations (1-way)

	Cell 1 from	Cell 2 from	Cell k from
	$N_p(\boldsymbol{\mu}_1, \Sigma)$	$N_p(\boldsymbol{\mu}_2, \Sigma)$	$N_p(\boldsymbol{\mu}_k, \Sigma)$
Measurements	$\mathbf{x}_{11}$	$\mathbf{x}_{21}$	$\mathbf{x}_{k1}$
	$\mathbf{x}_{12}$	$\mathbf{x}_{22}$	$\mathbf{x}_{k2}$
	$\cdots$	$\cdots$	$\cdots$
	$\mathbf{x}_{1n}$	$\mathbf{x}_{2n}$	$\mathbf{x}_{kn}$

Measurement model

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}$$

$$\sum_{i=1}^{k} \boldsymbol{\alpha}_i = 0$$

Cell mean $\bar{\mathbf{x}}_{i.} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_{ij}$

Total mean $\bar{\mathbf{x}}_{..} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} \mathbf{x}_{ij}$

What are the main effects of the independent variables?

$$\hat{\boldsymbol{\alpha}}_i = \bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{..}$$

<u>MANOVA Hypothesis Testing</u>

$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$

H_1 : At least two means are different

8.2 MANOVA: Computations (1-way)

<u>MANOVA Hypothesis Testing</u>

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$$

H_1 : At least two means are different

$$S = H + E$$

Total covariance

Among cells covariance

Within cell covariance

We reject H_0 if $\Lambda = \dfrac{|E|}{|S|} < \Lambda_{p,\alpha,\nu_H,\nu_E}$

Wilks' lambda

$\nu_H = k - 1$

$\nu_E = k(n-1)$

There are three other ways of testing the MANOVA hypothesis:

- Pillai's test
- Lawley and Hotelling's test
- Roy's greatest root

$$S = \frac{1}{kn}\sum_{i=}^{k}\sum_{j=1}^{n}\left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..}\right)\left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..}\right)^t$$

$$H = \frac{1}{k}\sum_{i=}^{k}\left(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..}\right)\left(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..}\right)^t$$

$$E = \frac{1}{kn}\sum_{i=}^{k}\sum_{j=1}^{n}\left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.}\right)\left(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.}\right)^t$$

$$E\left\{\frac{E}{nk-k}\right\} = \Sigma$$

8.2 MANOVA: Computations (1-way)

Table A.9. Lower Critical Values of Wilks Λ, $\alpha = .05$

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_s$ are eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. Reject H_0 if $\Lambda \leq$ table value. [a] Multiply entry by 10^{-3}.

	ν_H											
ν_E	1	2	3	4	5	6	7	8	9	10	11	12
						$p = 1$						
1	6.16[a]	2.50[a]	1.54[a]	1.11[a]	.868[a]	.712[a]	.603[a]	.523[a]	.462[a]	.413[a]	.374[a]	.341[a]
2	.098	.050	.034	.025	.020	.017	.015	.013	.011	.010	9.28[a]	8.51[a]
3	.229	.136	.097	.076	.062	.053	.046	.041	.036	.033	.030	.028
4	.342	.224	.168	.135	.113	.098	.086	.076	.069	.063	.058	.053
5	.431	.302	.236	.194	.165	.144	.128	.115	.104	.096	.088	.082
6	.501	.368	.296	.249	.215	.189	.169	.153	.140	.129	.119	.111
7	.556	.425	.349	.298	.261	.232	.209	.190	.175	.161	.150	.140
8	.601	.473	.396	.343	.303	.271	.246	.225	.208	.193	.180	.169
9	.638	.514	.437	.382	.341	.308	.281	.258	.239	.223	.209	.196
10	.668	.549	.473	.418	.376	.341	.313	.289	.269	.251	.236	.222
11	.694	.580	.505	.450	.407	.372	.343	.318	.297	.278	.262	.247
12	.717	.607	.534	.479	.436	.400	.370	.345	.323	.304	.286	.271
13	.736	.631	.560	.506	.462	.426	.396	.370	.347	.327	.310	.294

8.2 MANOVA: Computations (1-way)

Are there dependent variables that are not affected by the independent variables?

What is the strength of association between dependent variables and independent variables?

What is the relationship among the cell averages?

Example: $\boldsymbol{\mu}_2 = \dfrac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_3}{2}$

$$H_0 : \boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 = \mathbf{0}$$

$$H_1 : \boldsymbol{\mu}_1 - 2\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3 \neq \mathbf{0}$$

If $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$ is rejected, ANOVA tests can be performed on each component individually

$$R^2 = 1 - \Lambda$$

$$H_0 : c_1\boldsymbol{\mu}_1 + c_2\boldsymbol{\mu}_2 + \ldots + c_k\boldsymbol{\mu}_k = \mathbf{0}$$

$$H' = \frac{n}{\sum_{i=1}^{k} c_i^2}\left(\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_{i.}\right)\left(\sum_{i=1}^{k} c_i \bar{\mathbf{y}}_{i.}\right)^t$$

We reject H_0 if $\Lambda = \dfrac{|E|}{|E + H'|} < \Lambda_{p,\alpha,1,\nu_E}$

8.3 MANOVA: Computations (2-way)

Measurements	B=1	B=2		B=b
A=1	$\mathbf{x}_{111}$ … $\mathbf{x}_{11n}$	$\mathbf{x}_{121}$ … $\mathbf{x}_{12n}$		$\mathbf{x}_{1b1}$ … $\mathbf{x}_{1bn}$
	…	…		…
A=a	$\mathbf{x}_{a11}$ … $\mathbf{x}_{a1n}$	$\mathbf{x}_{a21}$ … $\mathbf{x}_{a2n}$		$\mathbf{x}_{ab1}$ … $\mathbf{x}_{abn}$

Measurement model

$$\boxed{\mathbf{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_i + \boldsymbol{\gamma}_{ij} + \boldsymbol{\varepsilon}_{ijk}}$$

Independent effects

Interaction

$$\sum_{i=1}^{a} \boldsymbol{\alpha}_i = \sum_{j=1}^{b} \boldsymbol{\beta}_j = \sum_{i=1}^{a} \boldsymbol{\gamma}_{ij} = \sum_{j=1}^{b} \boldsymbol{\gamma}_{ij} = \mathbf{0}$$

Cell averages	B=1	B=2		B=b	
A=1	$\overline{\mathbf{x}}_{11.}$	$\overline{\mathbf{x}}_{12.}$		$\overline{\mathbf{x}}_{1b.}$	$\overline{\mathbf{x}}_{1..}$
	…	…		…	
A=a	$\overline{\mathbf{x}}_{a1.}$	$\overline{\mathbf{x}}_{a2.}$		$\overline{\mathbf{x}}_{ab.}$	$\overline{\mathbf{x}}_{a..}$
	$\overline{\mathbf{x}}_{.1.}$	$\overline{\mathbf{x}}_{.2.}$		$\overline{\mathbf{x}}_{.b.}$	$\overline{\mathbf{x}}_{...}$

8.3 MANOVA: Computations (2-way)

Cell averages	B=1	B=2		B=b	
A=1	$\bar{\mathbf{x}}_{11.}$	$\bar{\mathbf{x}}_{12.}$		$\bar{\mathbf{x}}_{1b.}$	$\bar{\mathbf{x}}_{1..}$
	$\cdots$	$\cdots$		$\cdots$	
A=a	$\bar{\mathbf{x}}_{a1.}$	$\bar{\mathbf{x}}_{a2.}$		$\bar{\mathbf{x}}_{ab.}$	$\bar{\mathbf{x}}_{a..}$
	$\bar{\mathbf{x}}_{.1.}$	$\bar{\mathbf{x}}_{.2.}$		$\bar{\mathbf{x}}_{.b.}$	$\bar{\mathbf{x}}_{...}$

What are the main effects of the independent variables?

$$\hat{\boldsymbol{\alpha}}_i = \bar{\mathbf{x}}_{i..} - \bar{\mathbf{x}}_{...} \qquad \hat{\boldsymbol{\beta}}_j = \bar{\mathbf{x}}_{.j.} - \bar{\mathbf{x}}_{...}$$

What are the interactions among independent variables?

$$\hat{\boldsymbol{\gamma}}_{ij} = \bar{\mathbf{x}}_{ij.} - (\bar{\mathbf{x}}_{...} + \hat{\boldsymbol{\alpha}}_i + \hat{\boldsymbol{\beta}}_j)$$

8.3 MANOVA: Computations (2-way)

$$T = H_A + H_B + H_{AB} + E$$

Total variation

Variation due to A

Variation due to B

Variation due to interactions

Noise variation

Source	Sum of Squares and Products Matrix	df
A	$\mathbf{H}_A = nb \sum_i (\bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{...})(\bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{...})'$	$a - 1$
B	$\mathbf{H}_B = na \sum_j (\bar{\mathbf{y}}_{.j.} - \bar{\mathbf{y}}_{...})(\bar{\mathbf{y}}_{.j.} - \bar{\mathbf{y}}_{...})'$	$b - 1$
AB	$\mathbf{H}_{AB} = n \sum_{ij} (\bar{\mathbf{y}}_{ij.} - \bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{.j.} + \bar{\mathbf{y}}_{...}) \times (\bar{\mathbf{y}}_{ij.} - \bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{.j.} + \bar{\mathbf{y}}_{...})'$	$(a - 1)(b - 1)$
Error	$\mathbf{E} = \sum_{ijk} (\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{ij.})(\bar{\mathbf{y}}_{ijk} - \bar{\mathbf{y}}_{ij.})'$	$ab(n - 1)$
Total	$\mathbf{T} = \sum_{ijk} (\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{...})(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{...})'$	$abn - 1$

8.3 MANOVA: Computations (2-way)

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \ldots = \boldsymbol{\alpha}_a \longrightarrow \Lambda_A = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_A|} \text{ is } \Lambda_{p,a-1,ab(n-1)},$$

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_b \longrightarrow \Lambda_B = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_B|} \text{ is } \Lambda_{p,b-1,ab(n-1)},$$

$$H_0 : \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \ldots = \boldsymbol{\gamma}_{ab} \longrightarrow \Lambda_{AB} = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_{AB}|} \text{ is } \Lambda_{p,(a-1)(b-1),ab(n-1)}.$$

$$R^2 = 1 - \Lambda$$

What is the strength of association between dependent variables and independent variables?

What is the relationship among the cell averages?

$$H_0 : c_{1..}\boldsymbol{\mu}_{1..} + c_{2..}\boldsymbol{\mu}_{2..} + \ldots + c_a\boldsymbol{\mu}_{a..} = \mathbf{0}$$

8.3 MANOVA: Other computations

Considers differences over all the characteristic roots:

- Wilks' lambda: most commonly used statistic for overall significance
- Hotelling's trace
- Pillai's criterion: more robust than Wilks'; should be used when sample size decreases, unequal cell sizes or homogeneity of covariances is violated

Tests for differences on only the first discriminant function

- Roy's greatest characteristic root: most appropriate when DVs are strongly interrelated on a single dimension (only one source of variation is expected). Highly sensitive to violation of assumptions - most powerful when all assumptions are met

8.4 MANOVA: Post-hoc tests

What to do once we know there are significant differences among cells?

- Resort to univariate tests.
  - Detect which component makes the difference with ANOVA $H_0 : \mu_{1l} = \mu_{2l} = \ldots = \mu_{kl}$
  - For the component making the difference, detect which pair of averages are different: Scheffé, Fisher LSD (Least Significant Difference), Dunnet, Tukey HSD (Honest Significant Difference), Games-Howell.
- Resort to two-cell tests.
  - Hotelling T^2. $H_0 : \boldsymbol{\mu}_i = \boldsymbol{\mu}_j$
- Use Linear Discriminant Analysis.

$$\begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \ldots \\ \mu_{1p} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \ldots \\ \mu_{2p} \end{pmatrix} = \ldots = \begin{pmatrix} \mu_{k1} \\ \mu_{k2} \\ \ldots \\ \mu_{kp} \end{pmatrix}$$

$$H_0 : \mu_{il} = \mu_{jl}$$

8.5 MANOVA: Example

1-way

High education	Scientific American	(205,9,34) (203,20,21),…
	Fortune	…
	The New Yorker	…
Medium education	Sports Illustrated	…
	Newsweek	…
	People Weekly	…
Low education	National Enquirer	…
	Grit	…
	True Confessions	…

(words/ad, words/sentence, #words with >=3 syllables)

$$H_0 : \boldsymbol{\mu}_{ScientificAmerican} = \boldsymbol{\mu}_{Fortune} = \ldots = \boldsymbol{\mu}_{TrueConfessions}$$

$$\Lambda = 0.371$$

$$p_{value} = 0.004$$ Not all magazines are equal in the three variables

ANOVA words/ad:

$$H_0 : \mu_{words/ad\ ScientificAmerican} = \mu_{words/ad\ Fortune} = \ldots$$

$$F = 4.127 \quad p_{value} = 0.001$$

ANOVA words/sentence:

$$H_0 : \mu_{words/sentence\ ScientificAmerican} = \ldots$$

$$F = 1.643 \quad p_{value} = 0.140$$

ANOVA #words with >=3 syllables:

$$H_0 : \mu_{words>=3syll\ ScientificAmerican} = \ldots$$

$$F = 3.694 \quad p_{value} = 0.002$$

There are differences in the number of words per ad and the number of words with 3 or more syllables

8.5 MANOVA: Example

ANOVA words/ad:

$$H_0 : \mu_{words/ad\ ScientificAmerican} = \mu_{words/ad\ Fortune} = \ldots$$

$$H_0 : \mu_{words/ad\ ScientificAmerican} = \mu_{words/ad\ NewYorker}$$
$$H_0 : \mu_{words/ad\ Fortune} = \mu_{words/ad\ NewYorker}$$
$$H_0 : \mu_{words/ad\ ScientificAmerican} = \mu_{words/ad\ Grit}$$

ANOVA #words with >=3 syllables:

$$H_0 : \mu_{words>=3syll\ ScientificAmerican} = \ldots$$

$$H_0 : \mu_{\#words>=3syll\ ScientificAmerican} = \mu_{\#words>=3syll\ NewYorker}$$

Very few significant differences were seen between magazines for two of the dependent variables (total words per ad and number of 3 syllable words per ad), and all magazines were statistically non-different for one (number of words per sentence). <u>Therefore, one cannot state that magazines placed into the high education group have significantly more words per advertisement than magazines in either medium or low education groups.</u>

8.5 MANOVA: Example

1-way

High education	(205,9,34) (203,20,21) …
Medium education	…
Low education	…

$$H_0 : \boldsymbol{\mu}_{ScientificAmerican} = \boldsymbol{\mu}_{Fortune} = \ldots = \boldsymbol{\mu}_{TrueConfessions}$$

Restart

Course outline: Session 4

8. Multivariate Analysis of Variance (MANOVA)
    - 8.1. Introduction
    - 8.2. Computations (1-way)
    - 8.3. Computations (2-way)
    - 8.4. Post-hoc tests
    - 8.5. Example
9. Canonical Correlation Analysis (CCA)
    - 9.1. Introduction
    - 9.2. Construction of the canonical variables
    - 9.3. Example
    - 9.4. Extensions
10. Latent Class Analysis (LCA)
    - 10.1. Introduction

9.1 CCA: Introduction

House Structure

- Number of people (X_1)
- Number of people above 14 years old (X_2)
- Education of the main family person (X_3)
- Number of working people (X_4)

House Spendings

- Food (Y_1)
- Clothes (Y_2)
- Houseware (Y_3)
- Transport and Communications (Y_4)
- Free time and Education (Y_5)

$$\xrightarrow{\text{Latent variables}} X_1^*, X_2^*, X_3^*, X_4^*$$

$$\updownarrow \text{Maximum correlation}$$

$$\xrightarrow{\text{Latent variables}} Y_1^*, Y_2^*, Y_3^*, Y_4^*$$

$$\boxed{\begin{array}{ccc} \mathbf{X} \sim N_{p_X}(0, \Sigma_X) & \longrightarrow & \mathbf{X}^* = A\mathbf{X} \\ \updownarrow \Sigma_{XY} & & \updownarrow \text{Maximum correlation} \\ \mathbf{Y} \sim N_{p_Y}(0, \Sigma_Y) & \longrightarrow & \mathbf{Y}^* = B\mathbf{Y} \end{array}}$$

9.2 CCA: Construction of Canonical Variables

First canonical variables

Two samples

$$\mathbf{x} \sim N_{p_X}(0, \Sigma_X) \longrightarrow x_1^* = \boldsymbol{\alpha}_1^t \mathbf{x}$$

$$\Sigma_{XY}$$

$$\mathbf{y} \sim N_{p_Y}(0, \Sigma_Y) \longrightarrow y_1^* = \boldsymbol{\beta}_1^t \mathbf{y}$$

$$Corr\left\{x_1^*, y_1^*\right\} = \frac{E\left\{x_1^* y_1^{*t}\right\}}{\sqrt{E\left\{x_1^* x_1^{*t}\right\} E\left\{y_1^* y_1^{*t}\right\}}}$$

$$= \frac{\boldsymbol{\alpha}_1^t \Sigma_{XY} \boldsymbol{\beta}_1}{\sqrt{\left(\boldsymbol{\alpha}_1^t \Sigma_X \boldsymbol{\alpha}_1\right)\left(\boldsymbol{\beta}_1^t \Sigma_Y \boldsymbol{\beta}_1\right)}}$$

$$\boldsymbol{\alpha}_1^*, \boldsymbol{\beta}_1^* = \underset{\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1}{\arg\max} \left(Corr\left\{x_1^*, y_1^*\right\}\right)^2 \quad s.t. \quad Var\left\{x_1^*\right\} = Var\left\{y_1^*\right\} = 1$$

$$= \underset{\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1}{\arg\max} \frac{\left(\boldsymbol{\alpha}_1^t \Sigma_{XY} \boldsymbol{\beta}_1\right)^2}{\left(\boldsymbol{\alpha}_1^t \Sigma_X \boldsymbol{\alpha}_1\right)\left(\boldsymbol{\beta}_1^t \Sigma_Y \boldsymbol{\beta}_1\right)} - \lambda\left(\boldsymbol{\alpha}_1^t \Sigma_X \boldsymbol{\alpha}_1 - 1\right) - \eta\left(\boldsymbol{\beta}_1^t \Sigma_Y \boldsymbol{\beta}_1 - 1\right)$$

Sol: $\left(Corr\left\{x_1^*, y_1^*\right\}\right)^2 = \lambda^2 = \eta^2$ $\quad \left(\Sigma_X^{-1} \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}\right) \boldsymbol{\alpha}_1^* = \lambda^2 \boldsymbol{\alpha}_1^*$ $\quad \left(\Sigma_Y^{-1} \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY}\right) \boldsymbol{\beta}_1^* = \lambda^2 \boldsymbol{\beta}_1^*$

Eigenvectors associated to the largest eigenvalues

9.2 CCA: Construction of Canonical Variables

Second canonical variables

$$\boldsymbol{\alpha}_2^*, \boldsymbol{\beta}_2^* = \arg\max_{\boldsymbol{\alpha}_2, \boldsymbol{\beta}_2} \left(Corr\left\{ x_2^*, y_2^* \right\} \right)^2$$
$$s.t. \quad Var\left\{ x_2^* \right\} = Var\left\{ y_2^* \right\} = 1$$
$$Corr\left\{ x_1^*, x_2^* \right\} = Corr\left\{ y_1^*, y_2^* \right\} = 0$$

Sol:
The following largest eigenvalue

$$\left(\Sigma_X^{-1} \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX} \right) \boldsymbol{\alpha}_2^* = \lambda_2^2 \boldsymbol{\alpha}_2^*$$
$$\left(\Sigma_Y^{-1} \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} \right) \boldsymbol{\beta}_2^* = \lambda_2^2 \boldsymbol{\beta}_2^*$$

Up to $r = \min\left\{ p_X, p_Y \right\}$ canonical variables

Sol: All eigenvalues in descending order

Hypothesis testing

$$H_0 : \lambda_i > 0 \quad i = 1, 2, \ldots, s; \quad \lambda_{s+1} = \ldots = \lambda_r = 0$$
$$H_1 : \lambda_i > 0 \quad i = 1, 2, \ldots, s; \quad \lambda_{s+1} + \ldots + \lambda_r \neq 0$$

$$l = -\left(n - \tfrac{1}{2}\left(p_X + p_Y + 3 \right) \right) \sum_{i=s+1}^{r} \log(1 - \lambda_i^2)$$
$$l \sim \chi^2_{(p_X - s)(p_Y - s)}$$

9.3 CCA: Example

House Structure

- Number of people (X_1)
- Number of people above 14 years old (X_2)
- Education of the main family person (X_3)
- Number of working people (X_4)

$$R_X = \begin{pmatrix} 1 & & & \\ 0.55 & 1 & & \\ 0.11 & 0.04 & 1 & \\ 0.53 & -0.11 & 0.00 & 1 \end{pmatrix}$$

House Spendings

- Food (Y_1)
- Clothes (Y_2)
- Houseware (Y_3)
- Transport and Communications (Y_4)
- Free time and Education (Y_5)

$$R_Y = \begin{pmatrix} 1 & & & & \\ 0.29 & 1 & & & \\ 0.13 & 0.25 & 1 & & \\ 0.23 & 0.23 & 0.35 & 1 & \\ 0.33 & 0.32 & 0.22 & 0.36 & 1 \end{pmatrix}$$

Number of samples $n = 75$

$$R_{XY} = \begin{pmatrix} 0.46 & 0.34 & 0.05 & 0.33 & 0.29 \\ 0.03 & 0.18 & -0.02 & 0.13 & 0.17 \\ 0.22 & 0.32 & 0.51 & 0.26 & 0.23 \\ 0.40 & 0.14 & -0.02 & 0.25 & 0.17 \end{pmatrix}$$

Individually, the two more correlated variables are the eduction of the main person of the family and the spending in houseware.

9.3 CCA: Example

	$\boldsymbol{\alpha}_1^*$	$\boldsymbol{\alpha}_2^*$	$\boldsymbol{\alpha}_3^*$	$\boldsymbol{\alpha}_4^*$
Number of people (X_1)	0.810	–0.501	0.348	1.413
Number of people above 14 years old (X_2)	–0.286	0.101	–1.175	–0.825
Education of the main family person (X_3)	0.586	0.788	–0.010	–0.245
Number of working people (X_4)	0.077	–0.303	–0.093	–1.411

Correlation between the two canonical variables →

$\lambda =$	0.663	0.457	0.232	0.109

	$\boldsymbol{\beta}_1^*$	$\boldsymbol{\beta}_2^*$	$\boldsymbol{\beta}_3^*$	$\boldsymbol{\beta}_4^*$
Food (Y_1)	0.592	–0.459	0.757	0.119
Clothes (Y_2)	0.332	0.057	–0.544	0.799
Houseware (Y_3)	0.241	0.971	0.407	–0.088
Transport and Communications (Y_4)	0.293	–0.300	–0.287	–0.726
Free time and Education (Y_5)	0.032	0.032	–0.581	–0.248

9.3 CCA: Example

$$\lambda = 0.663 \quad 0.457 \quad 0.232 \quad 0.109$$

$$TotalVariance = 0.663^2 + 0.457^2 + 0.232^2 + 0.109^2 = 0.714$$

Variance explained by the first canonical variable $\dfrac{0.663^2}{0.714} = 61.6\%$

Variance explained by the first two canonical variables $\dfrac{0.663^2 + 0.457^2}{0.714} = 90.8\%$

$$H_0 : \lambda_1 > 0; \lambda_2, \lambda_3, \lambda_4 = 0$$

$$l = -\left(75 - \tfrac{1}{2}(5+4+3)\right)\left(\log(1-0.457^2) + \log(1-0.232^2) + \log(1-0.109^2)\right) = 20.81$$

$$l \sim \chi^2_{(4-1)(5-1)} \Rightarrow \Pr\{l \geq 20.81\} = 0.0534 \longrightarrow \text{We reject } H_0$$

$$H_0 : \lambda_1, \lambda_2 > 0, \lambda_3, \lambda_4 = 0$$

$$l = -\left(75 - \tfrac{1}{2}(5+4+3)\right)\left(\log(1-0.232^2) + \log(1-0.109^2)\right) = 4.64$$

$$l \sim \chi^2_{(4-2)(5-2)} \Rightarrow \Pr\{l \geq 4.64\} = 0.5907 \longrightarrow \text{We cannot reject } H_0$$

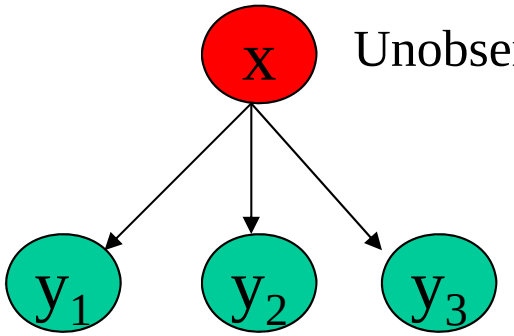
9.3 CCA: Extensions

- Kernel CCA: input variables are first mapped by a kernel onto a larger dimensional space.
- Generalized CCA: Study more than two groups of variables
- Rotated CCA: Rotate the basis vectors of CCA
- Nonlinear CCA: Use nonlinear projections of the input variables
- Sparse CCA and CCA for Multilabel classification: Sparse loadings
- Canonical Variate Analysis: find linear transformations of the input variables so that the ratio between the inter-group and intra-group variations is maximized.
- Multivariate linear regression (MLR): not only understand underlying variables but be able to predict.
- LInear Structural RELation model (LISREL): MLR with latent variables from both variable sets.

Course outline: Session 4

8. Multivariate Analysis of Variance (MANOVA)
    - 8.1. Introduction
    - 8.2. Computations (1-way)
    - 8.3. Computations (2-way)
    - 8.4. Post-hoc tests
    - 8.5. Example
9. Canonical Correlation Analysis (CCA)
    - 9.1. Introduction
    - 9.2. Construction of the canonical variables
    - 9.3. Example
    - 9.4. Extensions
10. Latent Class Analysis (LCA)
    - 10.1. Introduction

10.1 Latent Class Analysis: Introduction



Unobserved: discrete

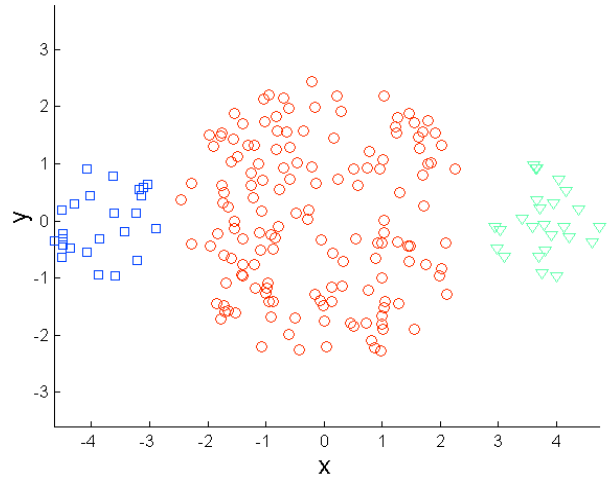
Observed: discrete $\longrightarrow$ Naive Bayesian network

$$P(\mathbf{Y}=\mathbf{y}) = \sum_{x=1}^{C} P(X=x)P(\mathbf{Y}=\mathbf{y} \mid X=x)$$

Observed: continuous $\longrightarrow$ Finite Mixture Model

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{x=1}^{C} P(X=x)\, f_{\mathbf{Y}|x}(\mathbf{y} \mid x)$$

1. C=1
2. Estimate model parameters
3. Estimate goodness of fit (chi2, bootstrap)
4. Evaluate model complexity (Exit if enough)
5. C=C+1. Go to Step 2.



10.1 Latent Class Analysis: Introduction

<u>Example:</u> Survey of attitudes towards anti-religionists

Allow anti-religionists to speak (1=Yes, 2=No)

Allow anti-religionists to teach (1=Yes, 2=No)

Keep anti-religionists books in a library (1=Yes, 2=No)

Y_1	Y_2	Y_3	Frequency	$P(X=1\|\mathbf{Y}=\mathbf{y})$	$P(X=2\|\mathbf{Y}=\mathbf{y})$
1	1	1	696	.998	.002
1	1	2	68	.929	.071
1	2	1	275	.876	.124
1	2	2	130	.168	.832
2	1	1	34	.848	.152
2	1	2	19	.138	.862
2	2	1	125	.080	.920
2	2	2	366	.002	.998

(Speak,Teach,Books)=f(cluster)

Model parameters

	$X=1$ (Tolerant)	$X=2$ (Intolerant)
$P(X=x)$	.62	.38
$P(Y_1=1\|X=x)$	.96	.23
$P(Y_2=1\|X=x)$	.74	.04
$P(Y_3=1\|X=x)$	.92	.24

...

Course outline: Session 4

8. Multivariate Analysis of Variance (MANOVA)
    - 8.1. Introduction
    - 8.2. Computations (1-way)
    - 8.3. Computations (2-way)
    - 8.4. Post-hoc tests
    - 8.5. Example
9. Canonical Correlation Analysis (CCA)
    - 9.1. Introduction
    - 9.2. Construction of the canonical variables
    - 9.3. Example
    - 9.4. Extensions
10. Latent Class Analysis (LCA)
    - 10.1. Introduction

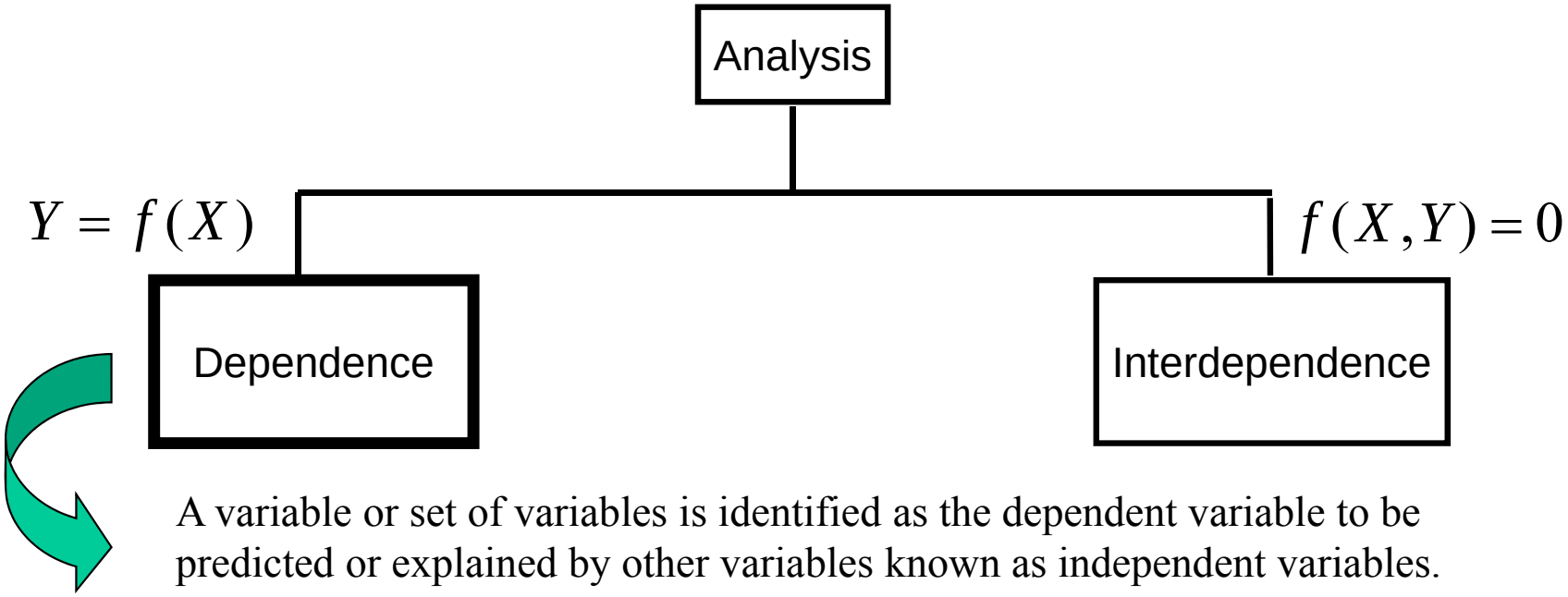
CEU
Universidad
San Pablo

Multivariate Data Analysis

Session 5: Regression based techniques

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid

Introduction: Types of analysis



A variable or set of variables is identified as the dependent variable to be predicted or explained by other variables known as independent variables.

Example:
(No. Sons, House Type)=
f(Income, Social Status, Studies)

- Multiple Discriminant Analysis
- Logit/Logistic Regression
- Multivariate Analysis of Variance (MANOVA) and Covariance
- Conjoint Analysis
- Canonical Correlation
- Multiple Regression
- Structural Equations Modeling (SEM)

Course outline: Session 4

11. Linear Regression
    - 11.1 Introduction
    - 11.2 Calculations
    - 11.3 Correlation coefficients
    - 11.4 Other kinds of regressions
12. Structural Equation Modelling
    - 12.1. Introduction
    - 12.2. Calculations
    - 12.3. Example
13. Conjoint analysis
    - 13.1. Introduction
14. Discriminant Analysis
    - 14.1. Introduction
    - 14.2. Linear Discriminant Analysis

11.1 Linear regression: Introduction

Linear regression

Example: (PriceWheat,SocialWelfare)=f(Year,Rain)

PriceWhear=f_1(Year,Rain)
SocialWelfare=f_2(Year,Rain)

Year	Rain	Price
1500	1010	17
1501	1290	19
1502	985	20
1503	514	15
…	…	…

$$Price = \beta_0 + \beta_1 Year + \beta_2 Rain$$

$$\begin{aligned} 17 &= \beta_0 + 1500\beta_1 + 1010\beta_2 \\ 19 &= \beta_0 + 1501\beta_1 + 1290\beta_2 \\ 20 &= \beta_0 + 1502\beta_1 + 985\beta_2 \\ 15 &= \beta_0 + 1503\beta_1 + 514\beta_2 \\ &\ldots \end{aligned} \longrightarrow \begin{pmatrix} 1 & 1500 & 1010 \\ 1 & 1501 & 1290 \\ 1 & 1502 & 985 \\ 1 & 1503 & 514 \\ 1 & \ldots & \ldots \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 17 \\ 19 \\ 20 \\ 15 \\ \ldots \end{pmatrix}$$

$$\boxed{\mathbf{X\beta} = \mathbf{x}}$$

11.2 Linear regression: Calculations

Linear regression

$$\boxed{\mathbf{X\beta} = \mathbf{x}}$$

$$Price = \beta_0 + \beta_1 Year \longrightarrow \mathbf{x} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \Rightarrow \boldsymbol{\varepsilon} = \mathbf{X\beta} - \mathbf{x} \Rightarrow \sigma_\varepsilon^2 = (\mathbf{X\beta} - \mathbf{x})^t (\mathbf{X\beta} - \mathbf{x})$$

Let's assume $E\{\varepsilon\} = 0$

Homocedasticity

Least Squares Estimate

$$\boxed{\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sigma_\varepsilon^2 = \mathbf{X}^+\mathbf{x} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{x}}$$

Properties:

$$E\{\hat{\boldsymbol{\beta}}\} = \boldsymbol{\beta}$$

$$Cov\{\hat{\boldsymbol{\beta}}\} = \sigma_\varepsilon^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N-k}(\mathbf{X\beta} - \mathbf{x})^t(\mathbf{X\beta} - \mathbf{x})$$

Degree of fit

$$R^2 = 1 - \frac{\|\mathbf{x} - \mathbf{X\beta}\|^2}{\|\mathbf{x} - \bar{\mathbf{x}}\mathbf{1}\|^2}$$

$$R^2_{adjusted} = 1 - (1 - R^2)\frac{N-1}{N-k-1}$$

Linear regression with constraints

$$\boxed{\hat{\boldsymbol{\beta}}_R = \begin{array}{c} \arg\min_{\boldsymbol{\beta}} \sigma_\varepsilon^2 \\ s.t.\ \mathbf{R\beta} = \mathbf{r} \end{array} = \hat{\boldsymbol{\beta}} + \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{R}^t\left(\mathbf{R}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{R}^t\right)(\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\beta}})}$$

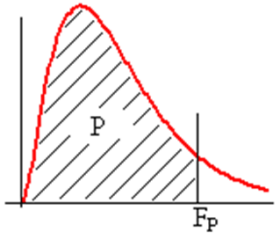
Example:

$$\beta_2 = 2\beta_1 \Rightarrow 2\beta_1 - \beta_2 = 0$$

11.2 Linear regression: Calculations

First test:

$$\boxed{\begin{aligned} H_0 &\equiv \boldsymbol{\beta} = \boldsymbol{\beta}_0 \\ H_1 &\equiv \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \end{aligned}} \longrightarrow F = \frac{1}{\hat{\sigma}_\varepsilon^2 k}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^t \mathbf{X}^t\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{\text{H}_0 \text{ is true}} F(k, N-k)$$



If $F > F_P$, reject H_0

$$\boxed{\mathbf{X}\boldsymbol{\beta} = \mathbf{x}}$$

$$Price = \beta_0 + \beta_1 Year + \beta_2 Year^2 \longrightarrow \begin{cases} 17 = \beta_0 + 1500\beta_1 + 1500^2\beta_2 \\ 19 = \beta_0 + 1501\beta_1 + 1501^2\beta_2 \\ 20 = \beta_0 + 1502\beta_1 + 1502^2\beta_2 \\ 15 = \beta_0 + 1503\beta_1 + 1503^2\beta_2 \\ \dots \end{cases} \longrightarrow \begin{pmatrix} 1 & 1500 & 1500^2 \\ 1 & 1501 & 1501^2 \\ 1 & 1502 & 1502^2 \\ 1 & 1503 & 1503^2 \\ \dots & \dots & \dots \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 17 \\ 19 \\ 20 \\ 15 \\ \dots \end{pmatrix}$$

$$\boxed{\begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \mathbf{x}}$$

$$\mathbf{P}_1 = \mathbf{X}_1\left(\mathbf{X}_1^t\mathbf{X}_1\right)^{-1}\mathbf{X}_1^t$$

Second test:

$$\boxed{\begin{aligned} H_0 &\equiv \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20} \\ H_1 &\equiv \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_{20} \end{aligned}} \longrightarrow F = \frac{1}{\hat{\sigma}_\varepsilon^2 k_2}\left(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20}\right)^t \mathbf{X}_2^t(\mathbf{I} - \mathbf{P}_1)\mathbf{X}_2\left(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20}\right) \xrightarrow{\text{H}_0 \text{ is true}} F(k_2, N-k)$$

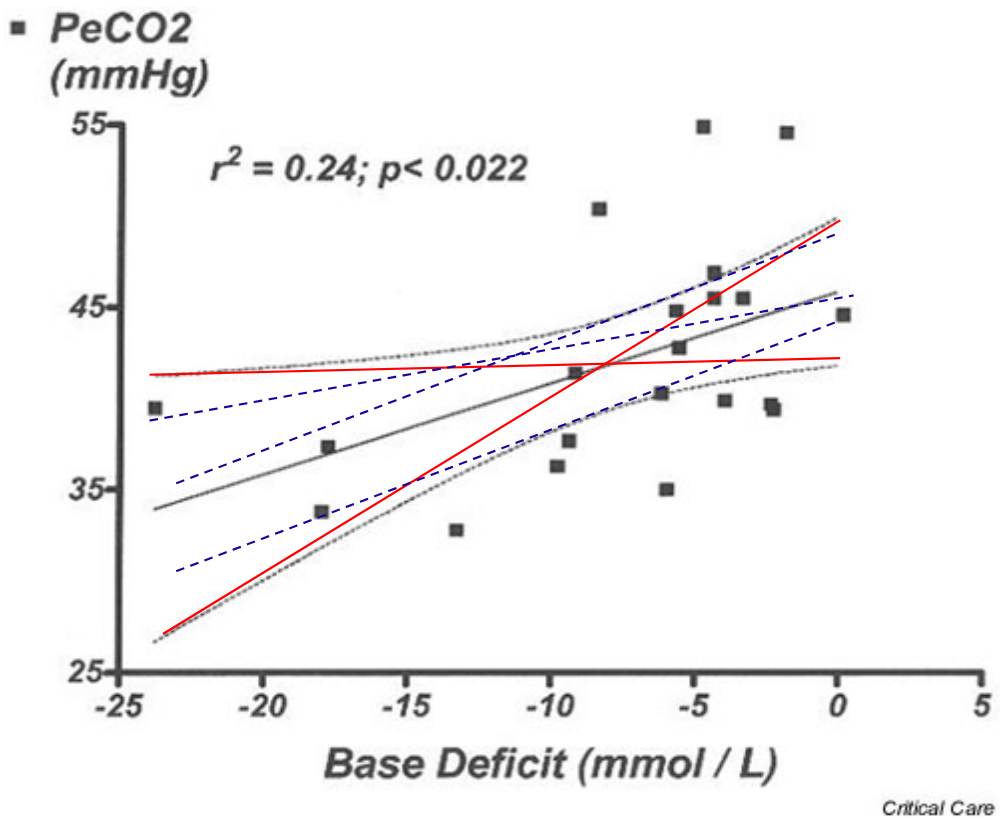
$$\boxed{\begin{aligned} H_0 &\equiv \beta_i = \beta_{i0} \\ H_1 &\equiv \beta_i \neq \beta_{i0} \end{aligned}} \longrightarrow t = \frac{\hat{\beta}_i - \beta_{i0}}{\hat{\sigma}_\varepsilon \sqrt{\omega_{ii}}} \qquad \left[\omega_{ii} = \left(\left(\mathbf{x}^t\mathbf{x}\right)^{-1}\right)_{ii}\right] \xrightarrow{\text{H}_0 \text{ is true}} t(N-k)$$

11.2 Linear regression: Calculations

Confidence intervals for the coefficients

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 \left(\mathbf{X}^t \mathbf{X}\right)^{-1}_{jj}$$ ← Unbiased variance of the j-th regression coefficient

$$\boxed{\beta_j \in \hat{\beta}_j + t_{1-\frac{\alpha}{2},N-k-1}\hat{\sigma}_j^2}$$ ← Confidence interval for the j-th regression coefficient



$$Y = [40, 45] + [0.05, 0.45]X$$

We got a certain regression line but the true regression line lies within this region with a 95% confidence.

11.2 Linear regression: Calculations

<u>Assumptions of regression</u>

- The sample is representative of your population
- The dependent variable is noisy, but the predictors are not!!. **Solution**: Total Least Squares
- Predictors are linearly independent (i.e., no predictor can be expressed as a linear combination of the rest), although they can be correlated. If it happens, this is called multicollinearity. **Solution**: add more samples, remove dependent variable, PCA
- The errors are homoscedastic. **Solution**: Weighted Least Squares
- The errors are uncorrelated to the predictors and to itself. **Solution**: Generalized Least Squares
- The errors follow a normal distribution. **Solution**: Generalized Linear Models

11.3 Correlation coefficients: Pearson correlation coefficient

$$R_{12}^2$$

$$\rho_{12} = \frac{E\left\{(X_1 - \mu_1)(X_2 - \mu_2)\right\}}{\sqrt{E\left\{(X_1 - \mu_1)^2\right\}}\sqrt{E\left\{(X_2 - \mu_2)^2\right\}}}$$

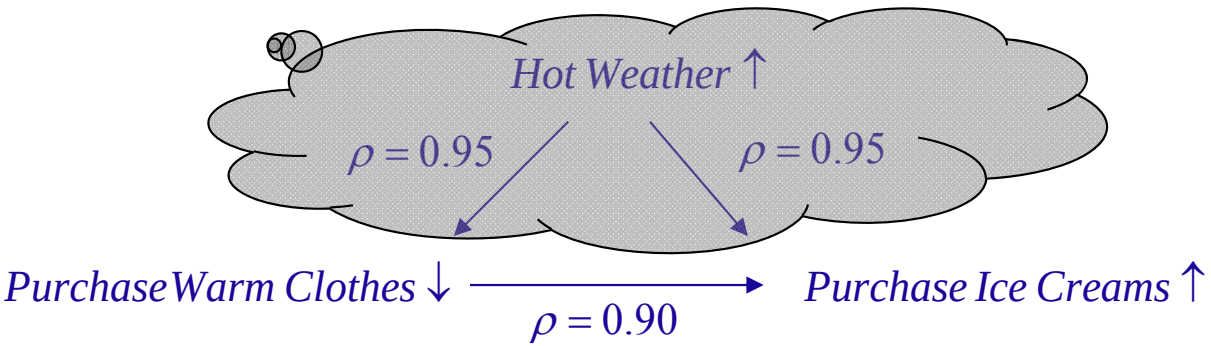
$$R_{12} = r_{12} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_{1i} - \overline{x}_1)(x_{2i} - \overline{x}_2)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{1i} - \overline{x}_1)^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{2i} - \overline{x}_2)^2}} = \frac{\sum_{i=1}^{N}\tilde{x}_{1i}\tilde{x}_{2i}}{\sqrt{\sum_{i=1}^{N}\tilde{x}_{1i}^2}\sqrt{\sum_{i=1}^{N}\tilde{x}_{2i}^2}}$$

$$\hat{\tilde{X}}_1 = \beta_2 \tilde{X}_2 \Rightarrow \hat{X}_1 = \overline{X}_1 + \hat{\tilde{X}}_1$$

$$\boxed{R_{12}^2 = \frac{\sum_{i=1}^{N}(\hat{x}_{1i} - \overline{x}_1)^2}{\sum_{i=1}^{N}(x_{1i} - \overline{x}_1)^2} = \frac{\sum_{i=1}^{N}(\hat{x}_{2i} - \overline{x}_2)^2}{\sum_{i=1}^{N}(x_{2i} - \overline{x}_2)^2}}$$

11.3 Correlation coefficients: Pearson correlation coefficient

Pitfall: Correlation means causation



Correct: Correlation means linear covariation

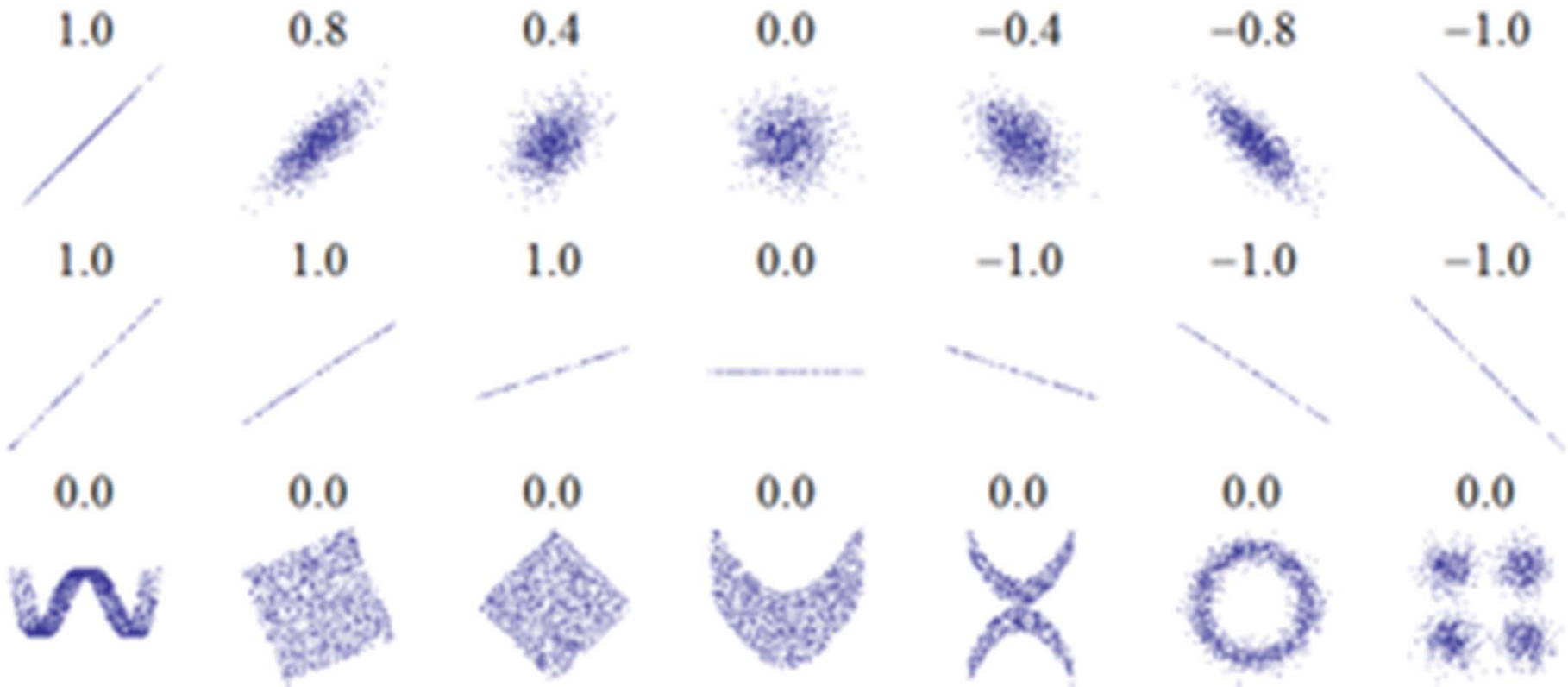
$$\text{Purchase Warm Clothes} \downarrow \xleftrightarrow[\rho = 0.90]{} \text{Purchase Ice Creams} \uparrow$$

$$\left.\begin{array}{l} \text{Salary} \uparrow \Rightarrow \text{FreeTime} \downarrow \\ \text{Education} \uparrow \Rightarrow \text{Salary} \uparrow \end{array}\right\}$$

There is a lot of a priori information!!

11.3 Correlation coefficients: Pearson correlation coefficient

Pitfall: Correlation measures all possible associations

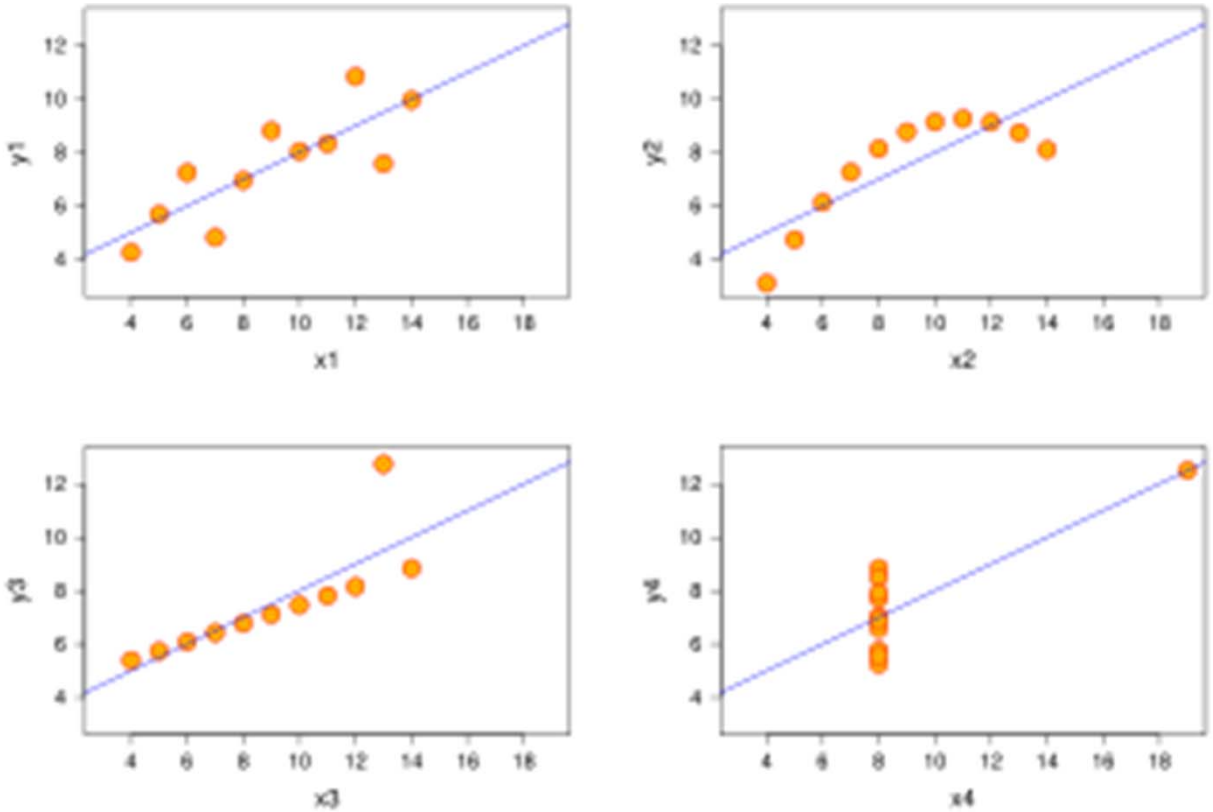


Correct: Correlation measures only linear associations

To measure non-linear associations the coefficient of determination is used (R^2)

11.3 Correlation coefficients: Pearson correlation coefficient

Pitfall: Correlation summarizes well the relationship between two variables



$$\bar{y} = 7.5$$
$$s_Y = 4.12$$
$$y = 3 + 0.5x$$
$$r = 0.81$$

Correct: Visual inspection of the data structure is always needed

11.3 Correlation coefficients: Pearson correlation coefficient

Is there any relationship between education and salary?

Person	Education	Salary $
A	3 (High)	70K
B	3 (High)	60K
C	2 (Medium)	40K
D	1 (Low)	20K

Pitfall: Compute the correlation between a categorical/ordinal variable and an interval variable.

Correct:

- Use ANOVA and the coefficient of determination
- Use Kendall or Spearman's rank correlation coefficient (valid only for ordinal, not categorical, variables)

Is there any relationship between education and salary?

Person	Education	Salary
A	3 (High)	3 (High)
B	3 (High)	3 (High)
C	2 (Medium)	2 (Medium)
D	1 (Low)	1 (Low)

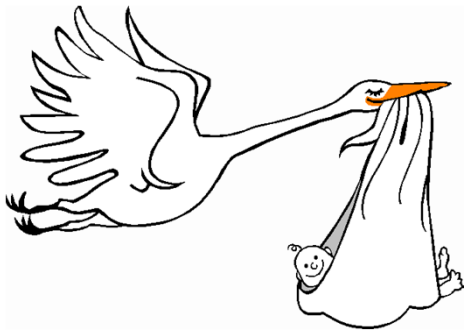
Pitfall: Compute the correlation between a two ordinal variables.

Correct:

Use Kendall or Spearman's rank correlation coefficient

11.3 Correlation coefficients: Pearson correlation coefficient

Pitfall: Correlation between combinations with common variables



Village	#Women	#Babies	#Storks	#Babies/#Women	#Storks/#Women
VillageA	…				
VillageB	…				
VillageC	…				

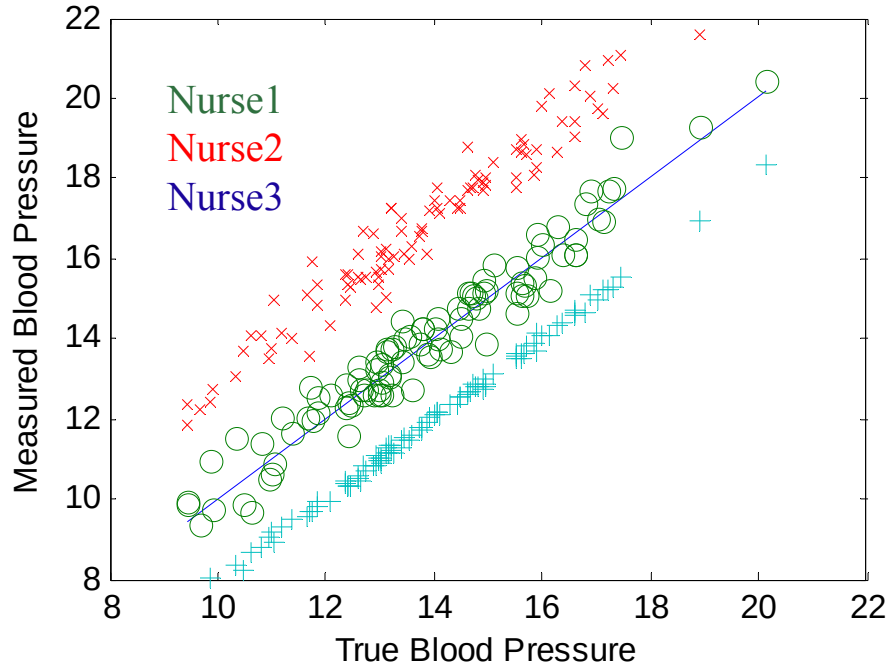
$$r_{BabiesPerWoman,StorkPerWoman} = 0.63!! \quad (p < 0.00001)$$

11.3 Correlation coefficients: Pearson correlation coefficient

Pitfall: Correlation is invariant to changes in mean and variance

Three nurses take blood pressure from the same pool of patients:

- Nurse 1 takes the true value with some variance.
- Nurse 2 takes consistently larger values with the same variance as nurse 1.
- Nurse 3 takes consistently smaller values with much less variance than the other 2.



$$r_{Nurse1,Nurse2} = 0.95$$

$$r_{Nurse1,Nurse3} = 0.97$$

$$r_{Nurse2,Nurse3} = 0.97$$

All correlations are rather high (meaning high agreement) although the data is quite different

11.3 Correlation coefficients: Multiple correlation coefficient $R^2_{1.23\ldots p}$

```
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

64   2   2   21.4838 1   1298.8  57      6.3   0     170.3  1945  890   200   915
76   2   1   23.87631    1       1032.5  50.1  15.8  0      75.8   2653  451   124   727
38   2   2   20.0108 2   2372.3  83.6    19.1  14.1  257.9  6321  660   328   721
40   2   2   25.14062    3       2449.5  97.5  26.5  0.5    332.6  1061  864   153   615
72   2   1   20.98504    1       1952.1  82.6  16.2  0      170.8  2863  1209  92    799
40   2   2   27.52136    3       1366.9  56    9.6   1.3    154.6  1729  1439  148   654
65   2   1   22.01154    2       2213.9  52    28.7  0      255.1  5371  802   258   834
58   2   1   28.75702    1       1595.6  63.4  10.9  0      214.1  823   2571  64    825
```

$$\hat{\tilde{X}}_1 = \beta_2 \tilde{X}_2 + \beta_3 \tilde{X}_3 + \ldots + \beta_p \tilde{X}_p = \boldsymbol{\beta}\tilde{\mathbf{X}}_{2,3,\ldots,p} \longrightarrow \boxed{\boldsymbol{\beta} = S^{-1}_{2,3,\ldots,p} S_1}$$

$$\hat{X}_1 = \bar{X}_1 + \hat{\tilde{X}}_1$$

Multiple correlation coefficient

$$\boxed{R^2_{1.2,3,\ldots p} = \frac{\sum_{i=1}^{n} (\hat{x}_{1i} - \bar{x}_1)^2}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}}$$

Variance of X_1 explained by a linear prediction

Total variance of X_1

$$S = \begin{pmatrix} s_1^2 & s_{12} & \ldots & s_{1p} \\ s_{21} & s_2^2 & \ldots & s_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ s_{p1} & s_{p2} & \ldots & s_p^2 \end{pmatrix}$$

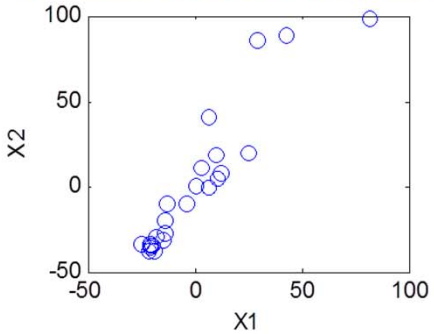
11.3 Correlation coefficients: Partial correlation coefficient

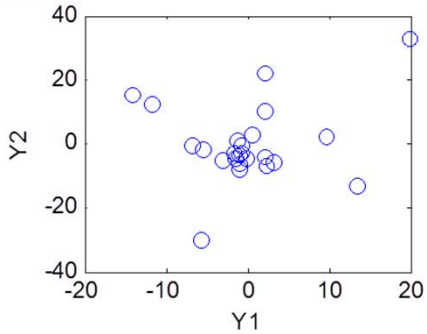
$$R^2_{12.3\ldots p}$$

<u>Partial correlation coefficient</u>

The partial correlation coefficient of Y and X removing the effect of $(Z_1,\ldots,Z_p)$ is the correlation of the residuals of Y after linear multiple regression with $(Z_1,\ldots,Z_p)$ and the residuals of X after linear multiple regression with $(Z_1,\ldots,Z_p)$

<u>Example:</u>





$R_{1.3} = 0.9599 \longrightarrow$ As seen before

$R_{1.2,3} = 0.9615$

Does X2 provide useful information on X1 once the influence of X3 is removed?

$$\hat{\tilde{X}}_1 = \beta_{13}\tilde{X}_3 \rightarrow Y_1 = \tilde{X}_1 - \hat{\tilde{X}}_1$$

$$\hat{\tilde{X}}_2 = \beta_{23}\tilde{X}_3 \rightarrow Y_2 = \tilde{X}_2 - \hat{\tilde{X}}_2$$

$R_{Y_1.Y_2} = 0.1943 \quad p_{value} = 0.363$

No!

11.4 Other kinds of regression: Partial Least Squares

Multiple Linear Regression:

$$\underset{N \times m}{\underset{\uparrow}{Y}} = \underset{(N \times p) \times (p \times m)}{\underset{\uparrow}{XB}} + E$$

Partial Least Squares:

$$\underset{N \times p}{\underset{\uparrow}{X}} = \underset{(N \times p') \times (p' \times p)}{\underset{\uparrow}{TP^t}}$$

$$\underset{N \times m}{\underset{\uparrow}{Y}} = \underset{(N \times m') \times (m' \times m)}{\underset{\uparrow}{UQ^t}} + F$$

Succesively

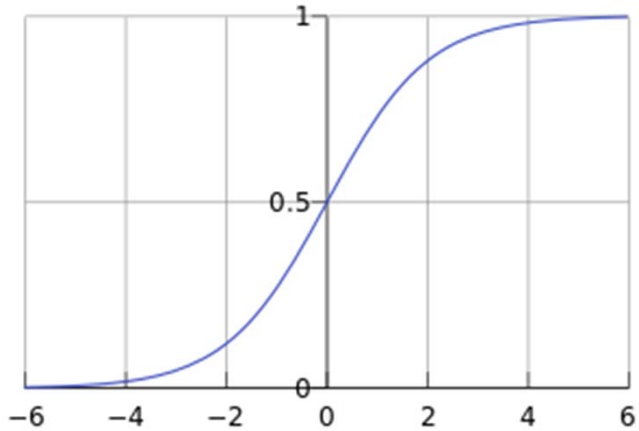
$$\max \operatorname{cov}\{\mathbf{t}_1, \mathbf{u}_1\}$$

$$\max \operatorname{cov}\{\mathbf{t}_2, \mathbf{u}_2\}$$

$$\ldots$$

$$Y = XB' + E$$

11.4 Other kinds of regression: Logistic regression

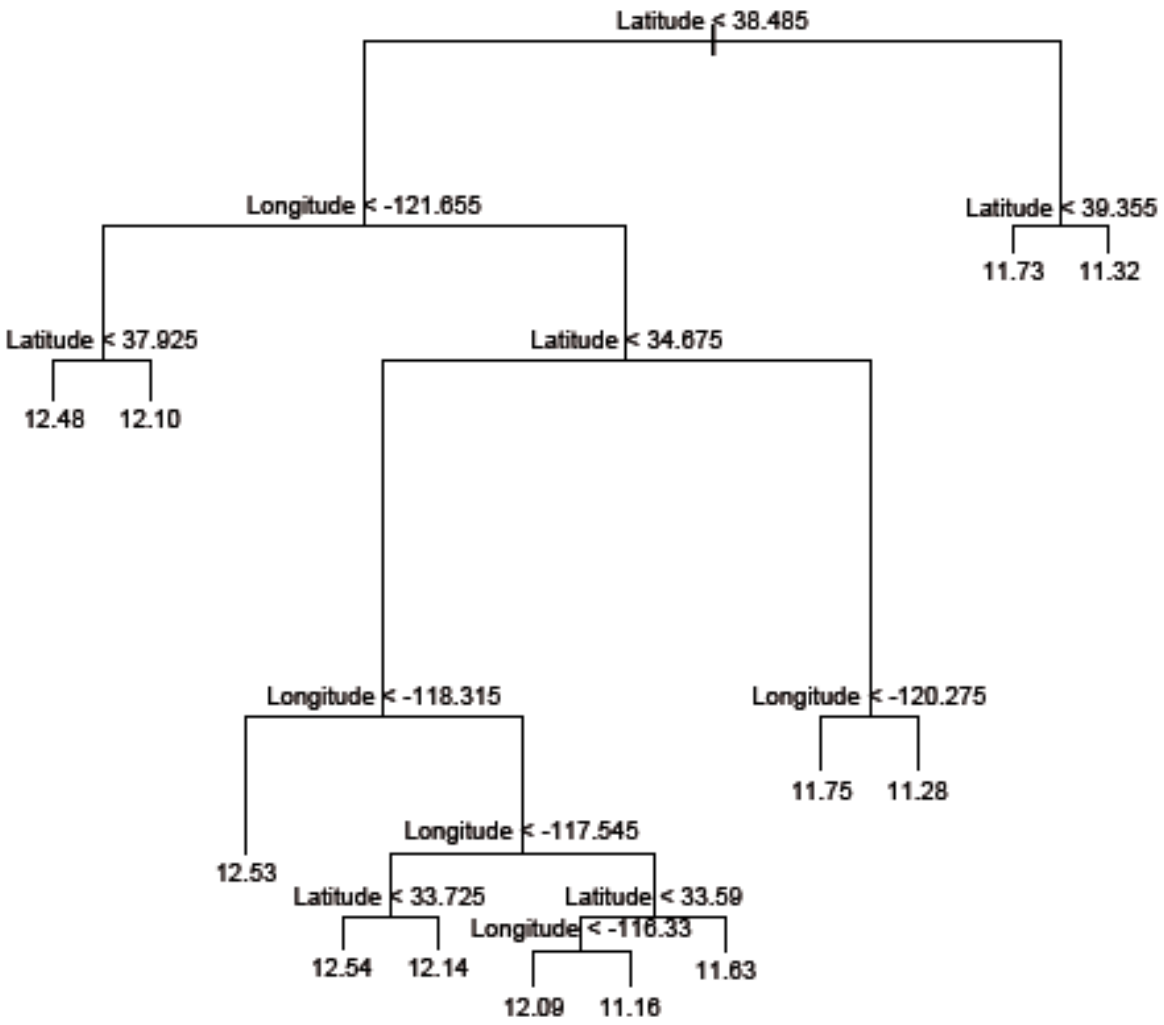


Logistic function

Example: HaveCreditCard=f(Income,Age)

$$E\{Y\} = \Pr\{Y = 1\} = \frac{1}{e^{-\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)}}$$

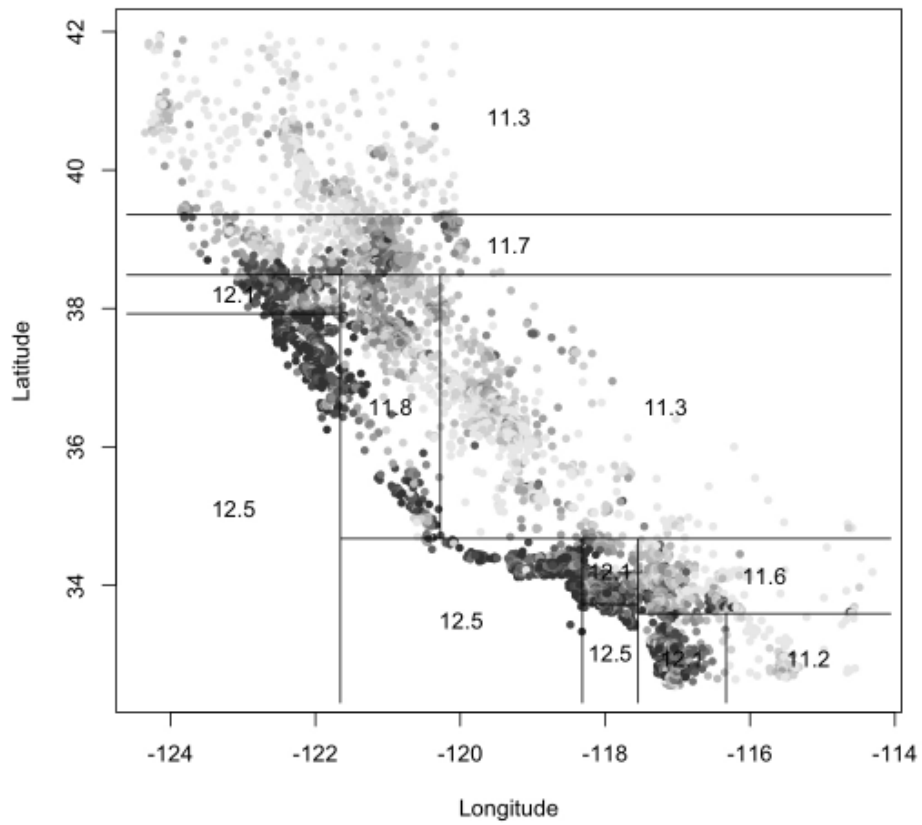
11.4 Other kinds of regressions



<u>Non-parametric regression</u>

<u>Regression tree</u>

The regression function is partitioned into non-overlapping regions.



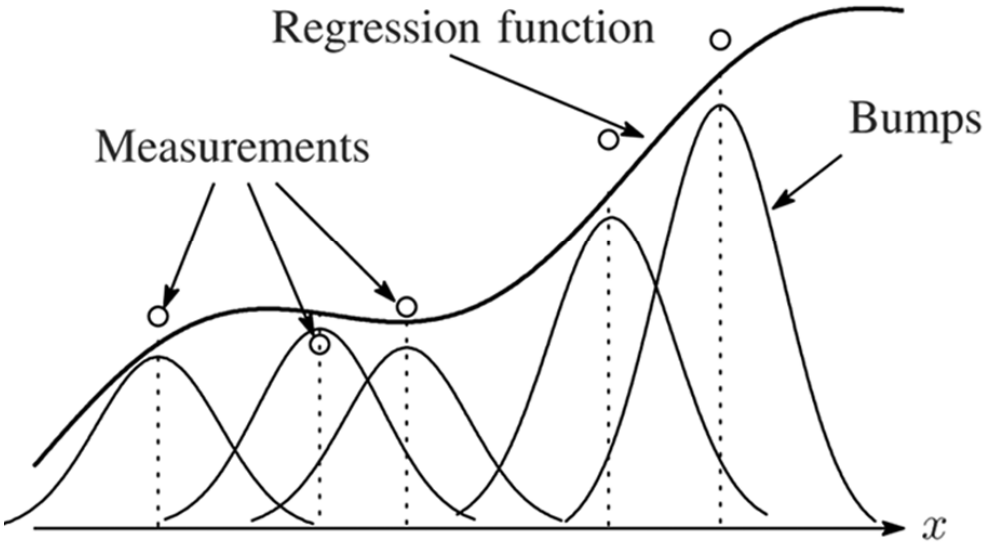
11.4 Other kinds of regressions

<u>Non-parametric regression</u>

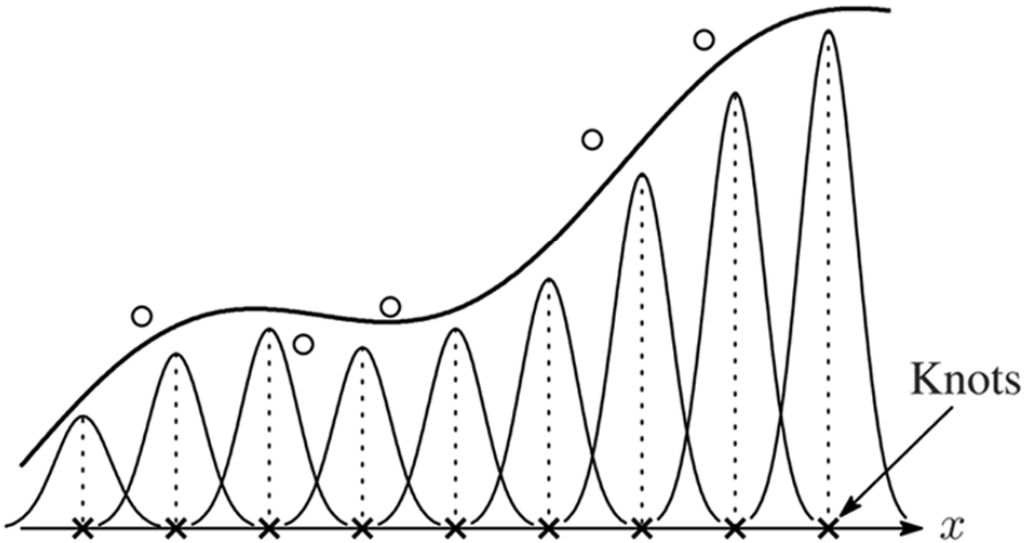
<u>Kernel regression</u>

The regression function is implicitly formed by "convolution" of the data with a kernel

$$\hat{Y} = E(Y \mid X) \Rightarrow Y = \frac{\sum_{i=1}^{N} K_h\left(\mathbf{X} - \mathbf{X}_i\right) Y_i}{\sum_{i=1}^{N} K_h\left(\mathbf{X} - \mathbf{X}_i\right)}$$



(a) Kernel regression



(b) B-spline

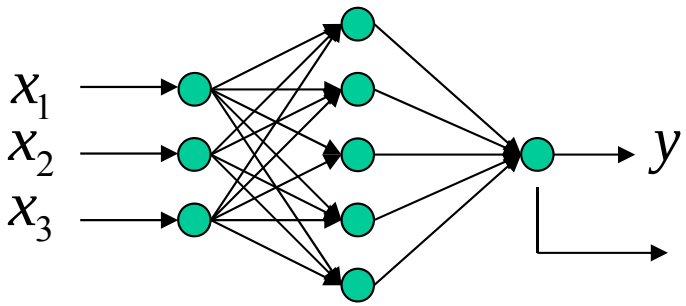
11.4 Other kinds of regressions

<u>Nonlinear regression</u>
The function is nonlinear in β.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - f(\mathbf{X}_i, \boldsymbol{\beta})\right)^2$$

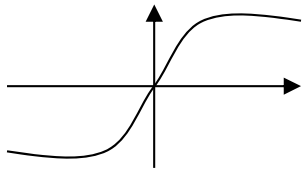
$$Y = \beta_0 + \beta_1 e^{\beta_2 X}$$

<u>Neural networks</u>
Strong nonlinear regression

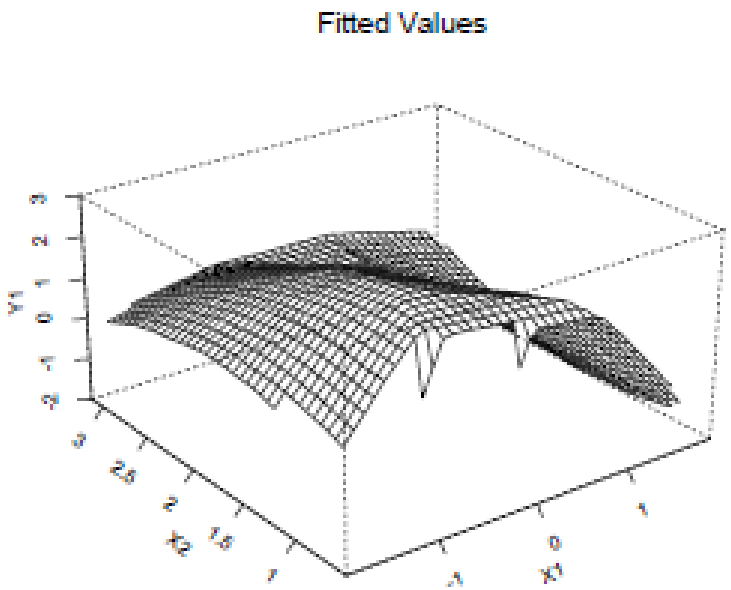


$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - f(\mathbf{X}_i, \boldsymbol{\beta})\right)^2$$

$$Output = f\left(\langle \beta_{node}, Input \rangle\right)$$



$(k+1)H$ Parameters



11.4 Other kinds of regressions

Penalized Least Squares

This is a way of computing the regression coefficients avoiding overfitting. Normally the penalization imposes some kind of smoothness on the solution.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - f(\mathbf{X}_i, \boldsymbol{\beta})\right)^2 + \underbrace{\lambda T(f(\mathbf{X}, \boldsymbol{\beta}))}_{\text{Penalization weight and function}}$$

Bayesian regression

Incorporate a priori information about the distribution of the regression coefficients

$$\max_{\boldsymbol{\beta}} \left(\prod_{i=1}^{N} f_{\varepsilon}\left(Y_i \mid f(\mathbf{X}_i, \boldsymbol{\beta})\right) \right) f_{\boldsymbol{\beta}}(\boldsymbol{\beta})$$

Robust regression

Use M-estimators or least absolute distance instead of least squares

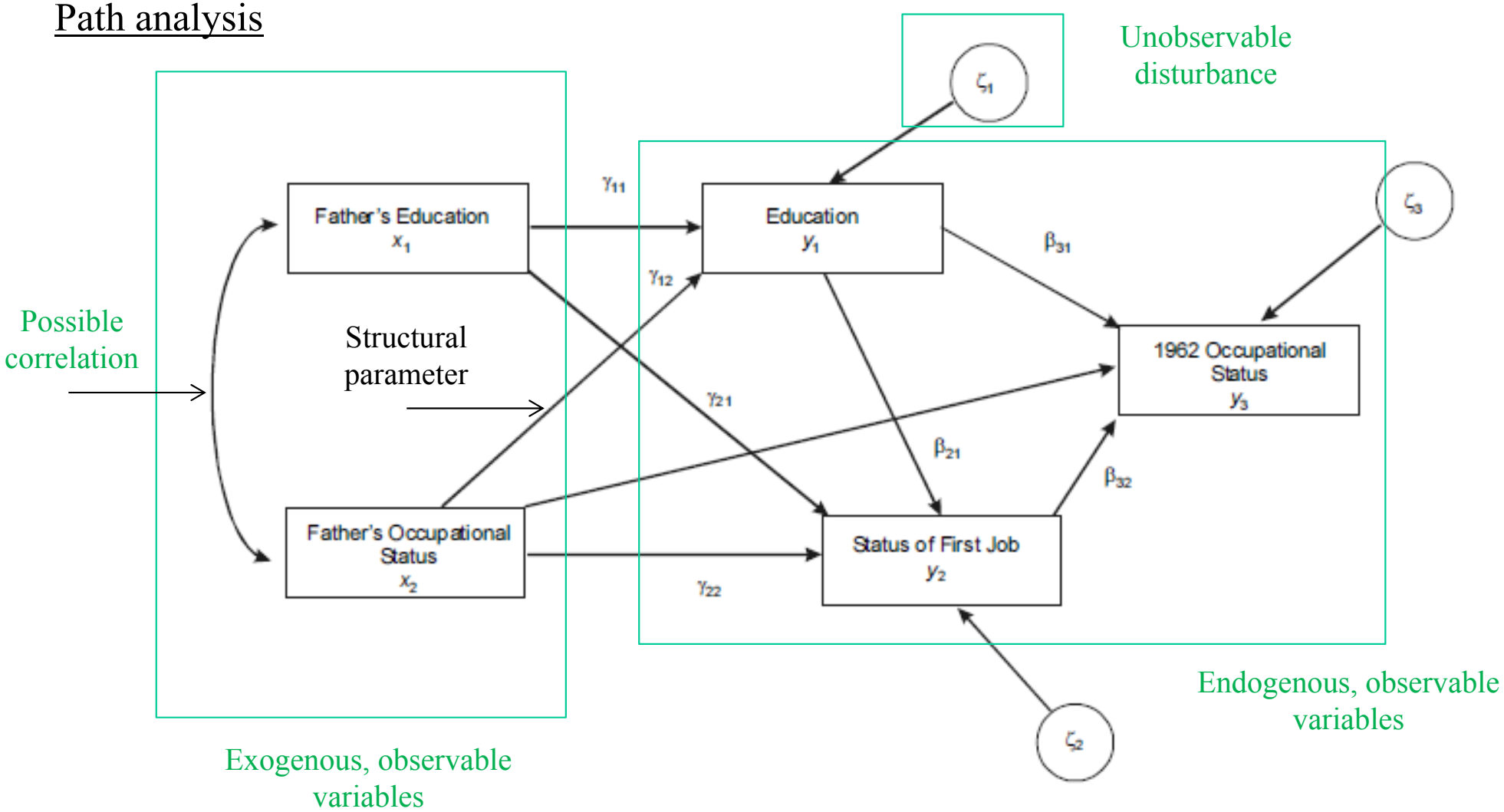
$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \Psi\left(Y_i - f(\mathbf{X}_i, \boldsymbol{\beta})\right) \qquad \min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left|Y_i - f(\mathbf{X}_i, \boldsymbol{\beta})\right|$$

Course outline: Session 4

11. Linear Regression
    - 11.1 Introduction
    - 11.2 Calculations
    - 11.3 Correlation coefficients
    - 11.4 Other kinds of regressions
12. Structural Equation Modelling
    - 12.1. Introduction
    - 12.2. Calculations
    - 12.3. Example
13. Conjoint analysis
    - 13.1. Introduction
14. Discriminant Analysis
    - 14.1. Introduction
    - 14.2. Linear Discriminant Analysis

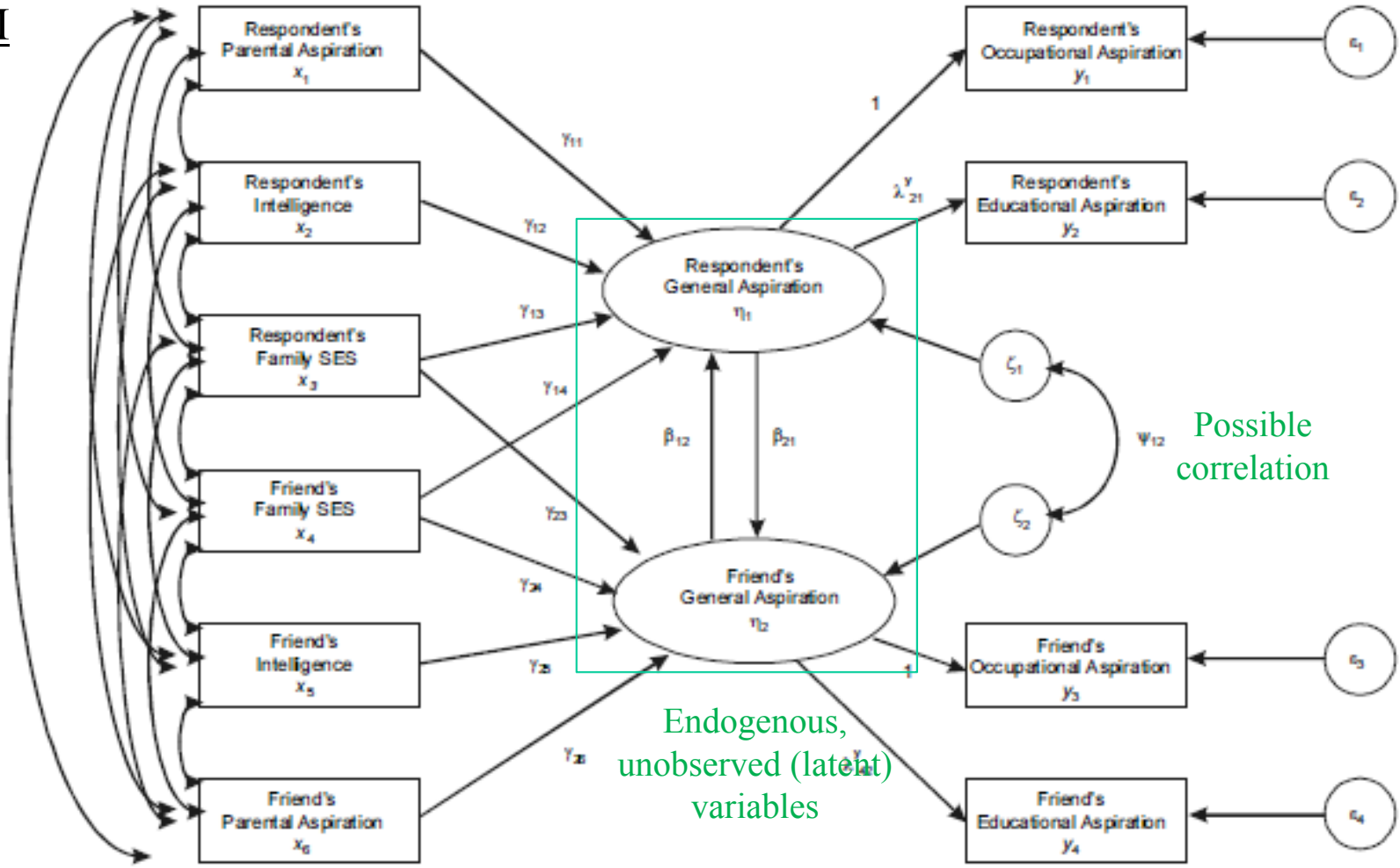
12.1 Structural Equation Modelling: Introduction

Path analysis



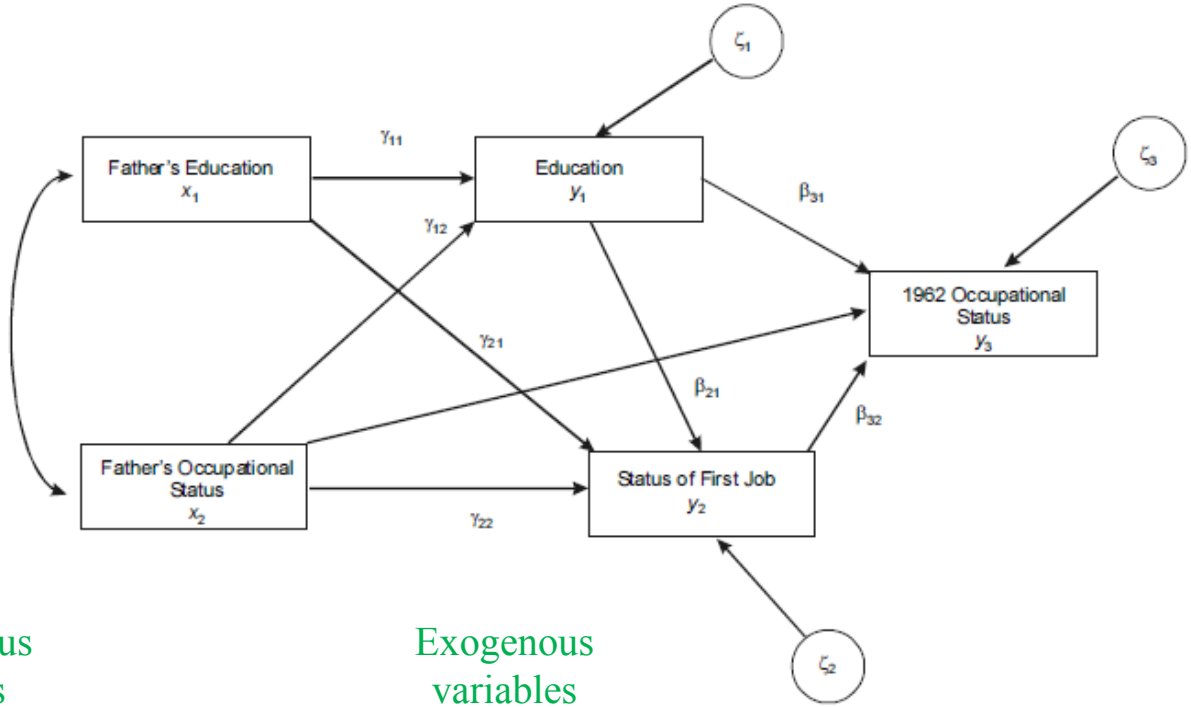
12.1 Structural Equation Modelling: Introduction

SEM



12.2 Structural Equation Modelling: Calculations

Path analysis



Endogenous variables

Exogenous variables

$$
\begin{aligned}
y_{1i} &= \gamma_{10} + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \zeta_{1i} \\
y_{2i} &= \gamma_{20} + \gamma_{21}x_{1i} + \gamma_{22}x_{2i} + \beta_{21}y_{1i} + \zeta_{2i} \\
y_{3i} &= \gamma_{30} + \gamma_{32}x_{2i} + \beta_{31}y_{1i} + \beta_{32}y_{2i} + \zeta_{2i}
\end{aligned}
$$

Endogenous variables

12.2 Structural Equation Modelling: Calculations

<u>SEM</u>

Model variables

Exogenous variables

Endogenous variables

Latent variables

$$
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ \eta_1 \\ \eta_2 \end{bmatrix}
=
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^y_{21} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^y_{42} \\
\gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{12} \\
0 & 0 & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26} & 0 & 0 & 0 & 0 & \beta_{21} & 0
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ \eta_1 \\ \eta_2 \end{bmatrix}
+
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \zeta_1 \\ \zeta_2 \end{bmatrix}
$$

Exogenous variables

Disturbances on observed vars.

Disturbances on unobserved vars.

Exogenous→Latent

Latent →Endogenous

Latent→Latent

$$\mathbf{v} = \mathbf{A}\mathbf{v} + \mathbf{u}$$

12.2 Structural Equation Modelling: Calculations

$$\mathbf{P} = E(\mathbf{u}\mathbf{u}')$$

$$\mathbf{P} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} & \sigma_{36} & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} & \sigma_{46} & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} & \sigma_{56} & 0 & 0 & 0 & 0 & 0 & 0 \\ \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_{66} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta^{\varepsilon}_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta^{\varepsilon}_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta^{\varepsilon}_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta^{\varepsilon}_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \psi_{11} & \psi_{12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \psi_{21} & \psi_{22} \end{bmatrix}$$

Exogenous covariance (estimated from data)

Endogenous, observed disturbance covariance (to be estimated)

Endogenous, unobserved disturbance covariance to be estimated)

12.2 Structural Equation Modelling: Calculations

$$\mathbf{J} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

n variables

m-n variables

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{21}^{y} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{42}^{y} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{12} \\ 0 & 0 & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26} & 0 & 0 & 0 & 0 & \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \zeta_1 \\ \zeta_2 \end{bmatrix}$$

Model→ $\mathbf{C} = E(\mathbf{J}\mathbf{v}\mathbf{v}'\mathbf{J}') = \mathbf{J}(\mathbf{I}_m - \mathbf{A})^{-1}\mathbf{P}(\mathbf{I}_m - \mathbf{A})^{-1'}\mathbf{J}'$

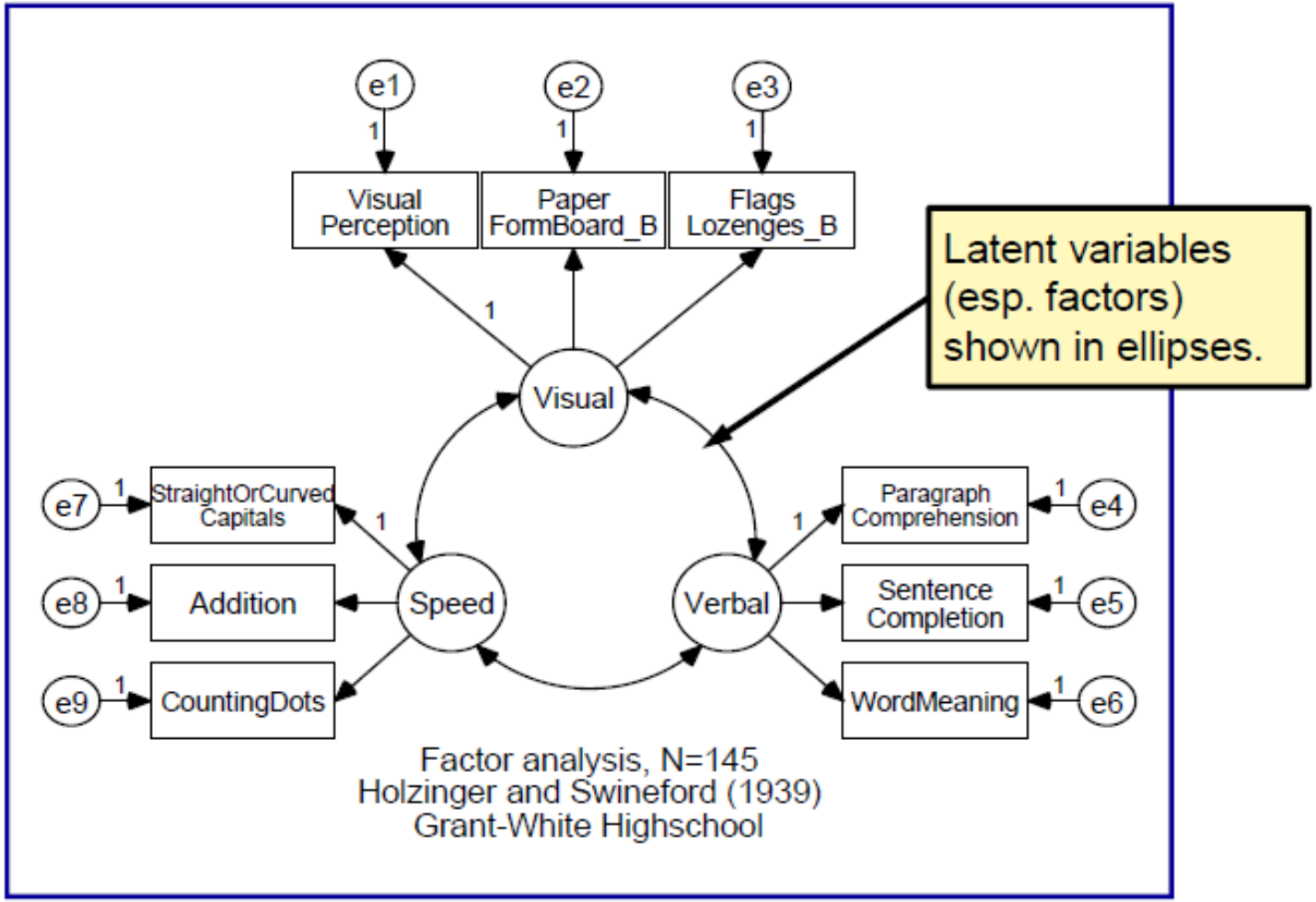
Observed→ $\mathbf{S}$

$$\min_{\mathbf{A},\mathbf{P}} \ \mathrm{trace}(\mathbf{S}\mathbf{C}^{-1}) - n + \log_e \det \mathbf{C} - \log_e \det \mathbf{S}$$

CEU Universidad San Pablo

12.3 Structural Equation Modelling: Example

Confirmatory Factor Analysis



12.3 Structural Equation Modelling: Example

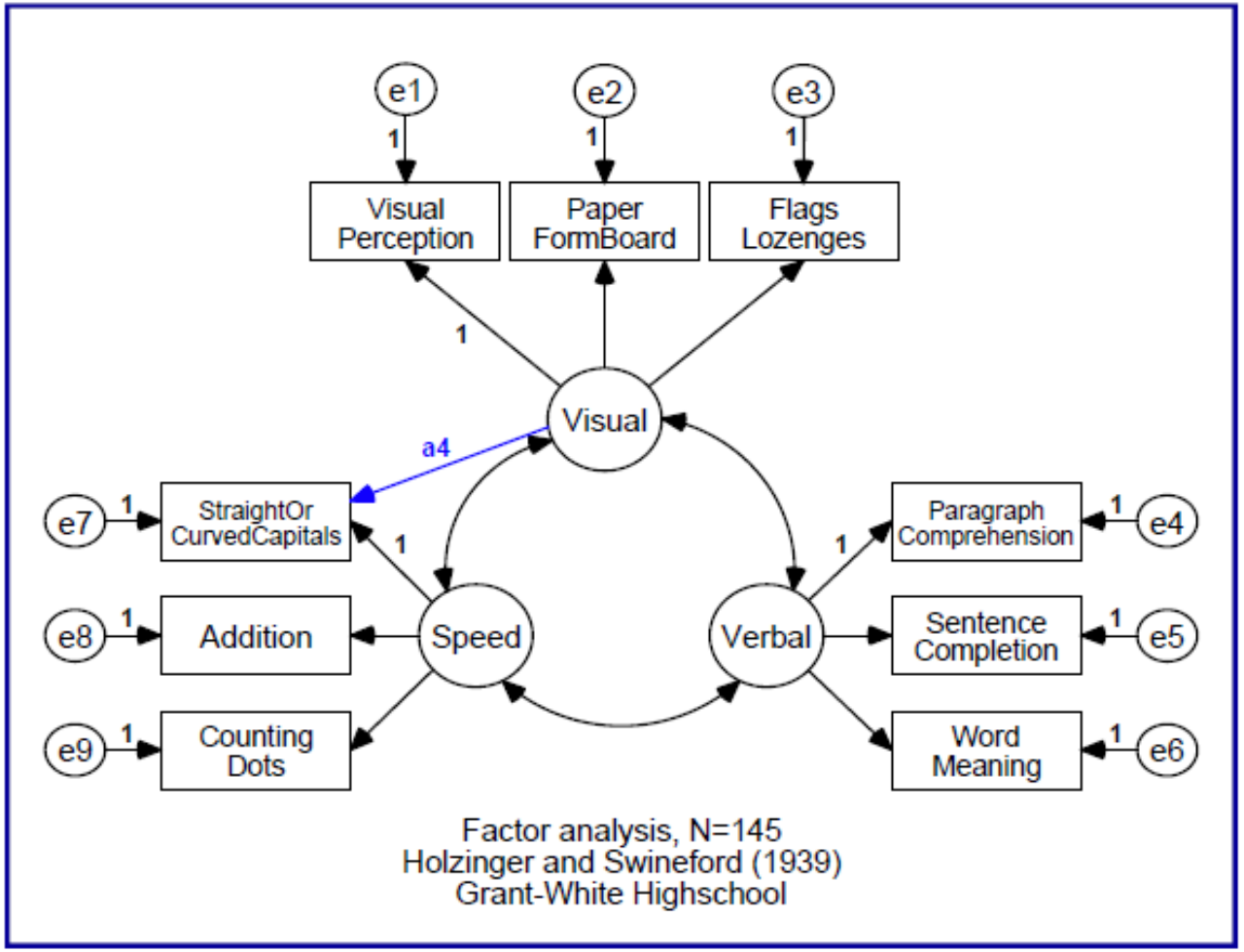
Confirmatory Factor Analysis

Asymptotically Standardized Residual Matrix

	Visual Perception	PaperForm Board	Flags Lozenges_B	Paragraph Comprehension	Sentence Completion
VisualPerc	0.000000000	-0.490645663	0.634454156	-0.376267466	-0.853201760
PaperFormB	-0.490645663	0.000000000	-0.133256120	-0.026665527	0.224463460
FlagsLozen	0.634454156	-0.133256120	0.000000000	0.505250934	0.901260142
ParagraphC	-0.376267466	-0.026665527	0.505250934	0.000000000	-0.303368250
SentenceCo	-0.853201760	0.224463460	0.901260142	-0.303368250	0.000000000
WordMeanin	-0.530010952	0.187307568	0.474116387	0.577008266	-0.268196124
StraightOr	4.098583857	2.825690487	1.450078999	1.811782623	2.670254862
Addition	-3.084483125	-1.069283994	-2.383424431	0.166892980	1.043444072
CountingDo	-0.219601213	-0.619535105	-2.101756596	-2.939679987	-0.642256508

12.3 Structural Equation Modelling: Example

Confirmatory Factor Analysis



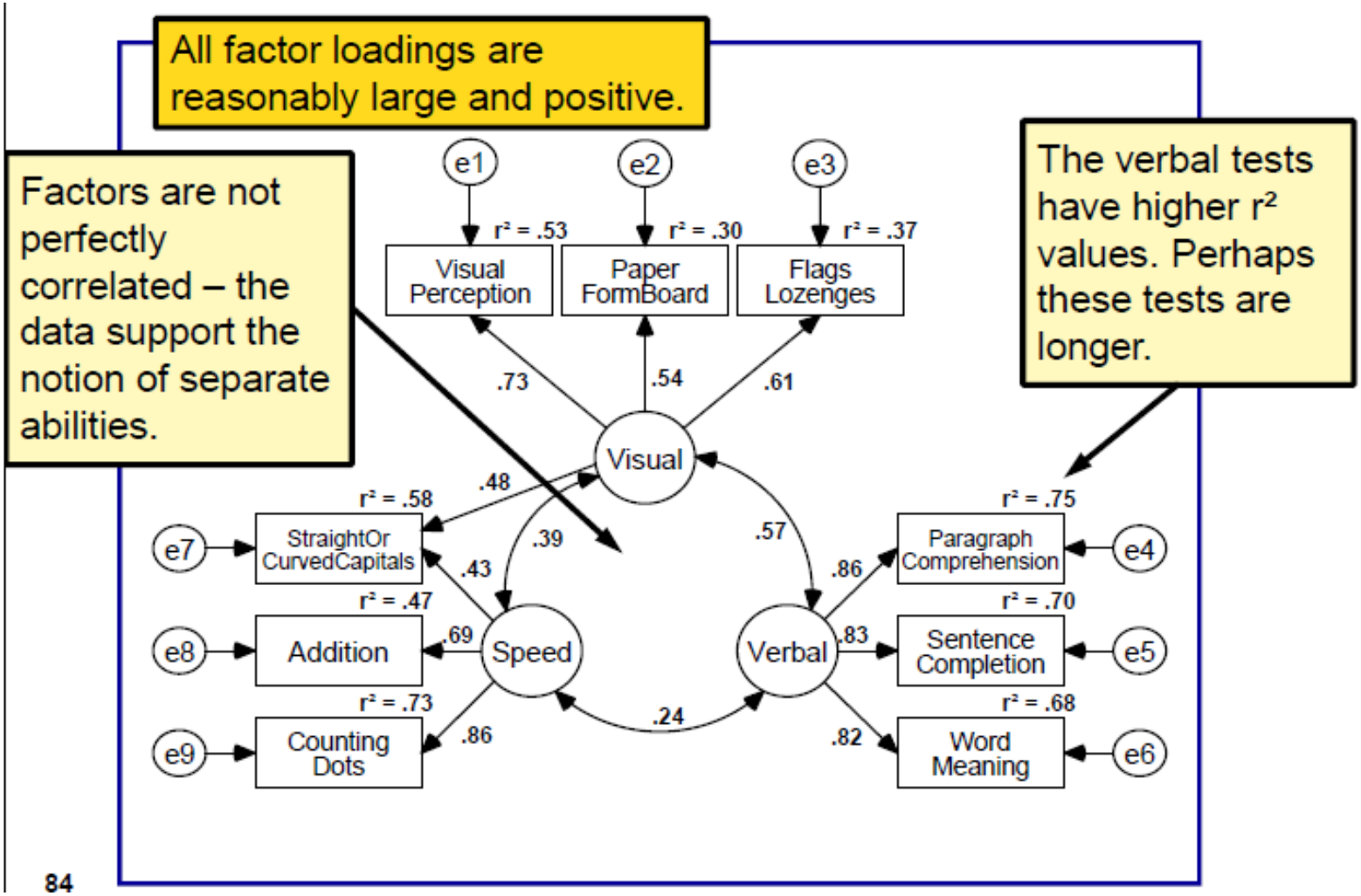
12.3 Structural Equation Modelling: Example

Nested models

Model	Chi-square	DF	P-Value	Comment
Model 1	48.0536	24	0.0025	Base model
Added Path "StraightOrCurvedCapitals <- F_Visual"	20.5494	23	0.6086	More general model
Difference	27.5042	1	0.0000	**"Significance of added parameters"**

12.3 Structural Equation Modelling: Example

Model result



Course outline: Session 4

11. Linear Regression
    - 11.1 Introduction
    - 11.2 Calculations
    - 11.3 Correlation coefficients
    - 11.4 Other kinds of regressions
12. Structural Equation Modelling
    - 12.1. Introduction
    - 12.2. Calculations
    - 12.3. Example
13. Conjoint analysis
    - 13.1. Introduction
14. Discriminant Analysis
    - 14.1. Introduction
    - 14.2. Linear Discriminant Analysis

13.1 Conjoint Analysis: Introduction

Example: Utility=f(Seat Comfort, Price, Duration)

Choice	Seat Comfort	Price	Duration
1 (u_{111})	extra-wide	$700	5 hours
2 (u_{112})	extra-wide	$700	3 hours
3 (u_{121})	extra-wide	$400	5 hours
4 (u_{122})	extra-wide	$400	3 hours
5 (u_{211})	regular	$700	5 hours
6 (u_{212})	regular	$700	3 hours
7 (u_{221})	regular	$400	5 hours
8 (u_{222})	regular	$400	3 hours

$$u_{ijkn} = u_0 + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \eta_{ik} + \nu_{jk} + \omega_{ijk} + \varepsilon_{ijkn}$$

13.1 Conjoint Analysis: Discrete Choice Experimentation

Choice	Seat Comfort	Price	Duration
3 (u_{121})	extra-wide	$400	5 hours
4 (u_{122})	extra-wide	$400	3 hours
7 (u_{221})	regular	$400	5 hours

$u_{122} > u_{121}$

$u_{122} > u_{221}$

Choice	Seat Comfort	Price	Duration
2 (u_{112})	extra-wide	$700	3 hours
6 (u_{212})	regular	$700	3 hours
7 (u_{221})	regular	$400	5 hours

$u_{221} > u_{112}$

$u_{221} > u_{212}$

Choice	Seat Comfort	Price	Duration
1 (u_{111})	extra-wide	$700	5 hours
5 (u_{211})	regular	$700	5 hours
8 (u_{222})	regular	$400	3 hours

$u_{222} > u_{111}$

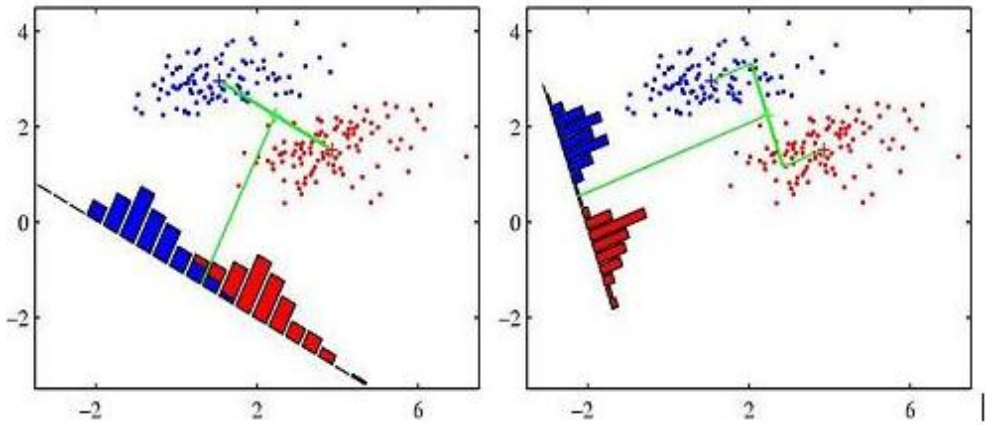
$u_{222} > u_{211}$

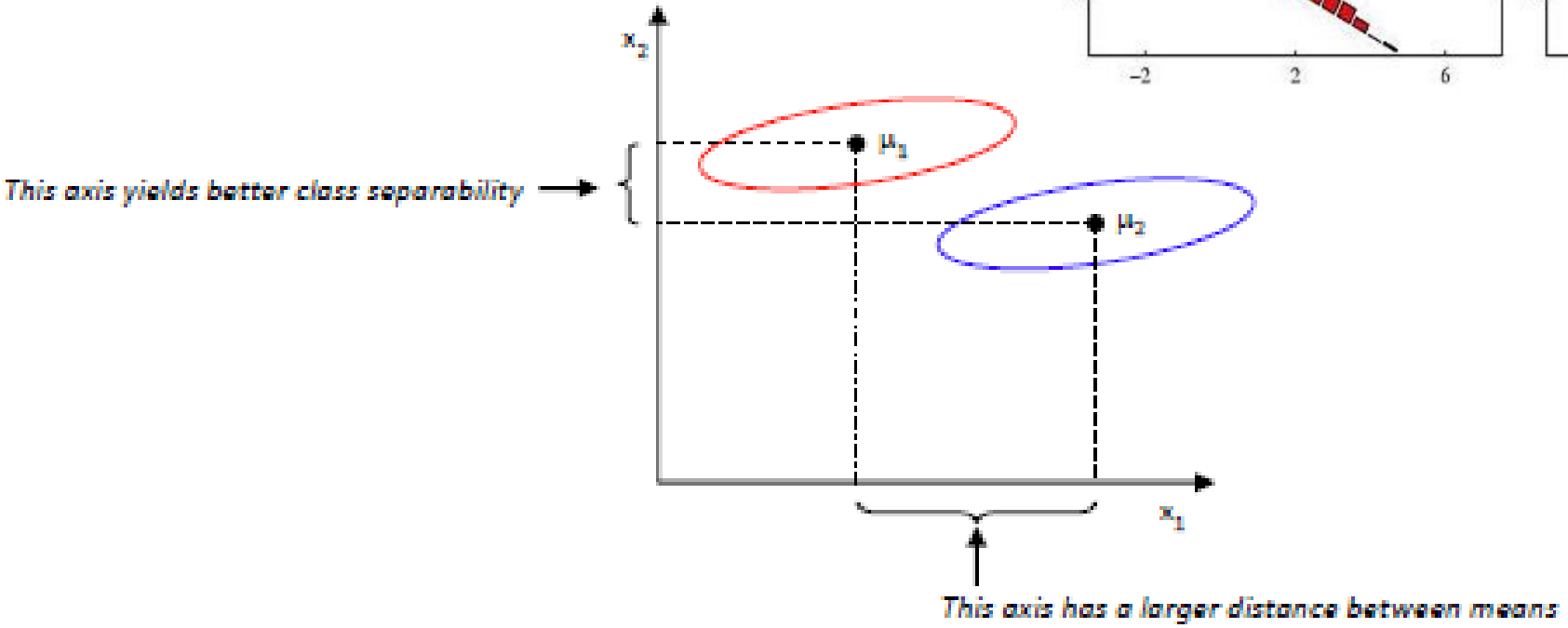
Course outline: Session 4

11. Linear Regression
    - 11.1 Introduction
    - 11.2 Calculations
    - 11.3 Correlation coefficients
    - 11.4 Other kinds of regressions
12. Structural Equation Modelling
    - 12.1. Introduction
    - 12.2. Calculations
    - 12.3. Example
13. Conjoint analysis
    - 13.1. Introduction
14. Discriminant Analysis
    - 14.1. Introduction
    - 14.2. Linear Discriminant Analysis

14.1 Discriminant Analysis: Introduction

HouseType=f(Income,Age)





14.2 Linear Discriminant Analysis

Maximize $J(\mathbf{w}) = \dfrac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$

Overall mean

$$S_B = \sum_c N_c(\boldsymbol{\mu}_c - \bar{\mathbf{x}})(\boldsymbol{\mu}_c - \bar{\mathbf{x}})^T$$ Between classes covariance

$$S_W = \sum_c \sum_{i \in c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T$$ Within classes covariance

Class mean

Solution $S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}$

2-class classification: $\begin{cases} \mathbf{w}^t \mathbf{x} < c & Class1 \\ \mathbf{w}^t \mathbf{x} > c & Class2 \end{cases}$

K-class classification: Construct K classifiers (a class vs the rest)

14.2 Linear Discriminant Analysis

Assumptions: Data within classes is normally distributed.

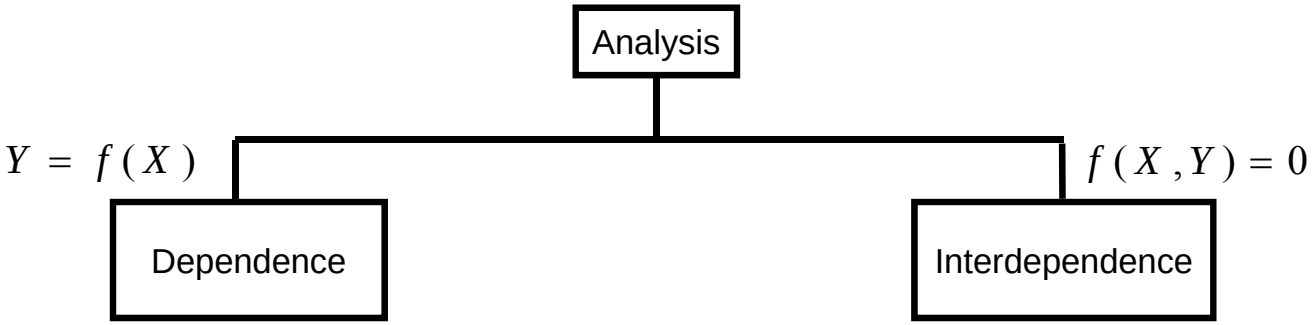
Limitations: LDA can only compute up to C-1 projection directions.
LDA fails if the differences between group is not in the mean, but in the variances.

Extension: PLS-DA: PLS with a categorical variable

Course outline: Session 4

11. Linear Regression
    - 11.1 Introduction
    - 11.2 Calculations
    - 11.3 Correlation coefficients
    - 11.4 Other kinds of regressions
12. Structural Equation Modelling
    - 12.1. Introduction
    - 12.2. Calculations
    - 12.3. Example
13. Conjoint analysis
    - 13.1. Introduction
14. Discriminant Analysis
    - 14.1. Introduction
    - 14.2. Linear Discriminant Analysis

Conclusions



- Discriminant Analysis
  HouseType=f(Income,Age)
- Logit/Logistic Regression
  HaveCreditCard=f(Income,Age)
- Multiple Regression
  (PriceWheat,SocialWelfare)=f(Year,Rain)
- Multivariate Analysis of Variance (MANOVA) and Covariance
  (Ability in Math, Ability in Physics)=f(Math textbook, Physics textbook, College)
- Conjoint Analysis
  Utility=f(Seat Comfort, Price, Duration)
- Structural Equations Modeling (SEM)
  OccupationalAspiration=f(…)

- Principal Components, Factor Analysis
  (Grade Math, Grade Latin, Grade Physics)=f(Intelligence, Maturity)
- Multidimensional Scaling (perceptual mapping)
  (x,y)=f(City gross income, health indexes, population, political stability, …)
- Correspondence Analysis
  (Eye colour, Hair colour, Skin colour)=f(gen A, gen B)
- Canonical Correlation
  (Grade Chemistry, Grade Physics)=f(Grade Math, Grade Latin)
- Latent Class Analysis
  (Speak,Teach,Books)=f(cluster)
- Cluster Analysis (Course)