



CEU

*Universidad
San Pablo*

Multivariate Data Analysis

Session 1: Introduction and data examination

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid, July 23rd 2007

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

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. Am J Epidemiol 1988;128:64-74.

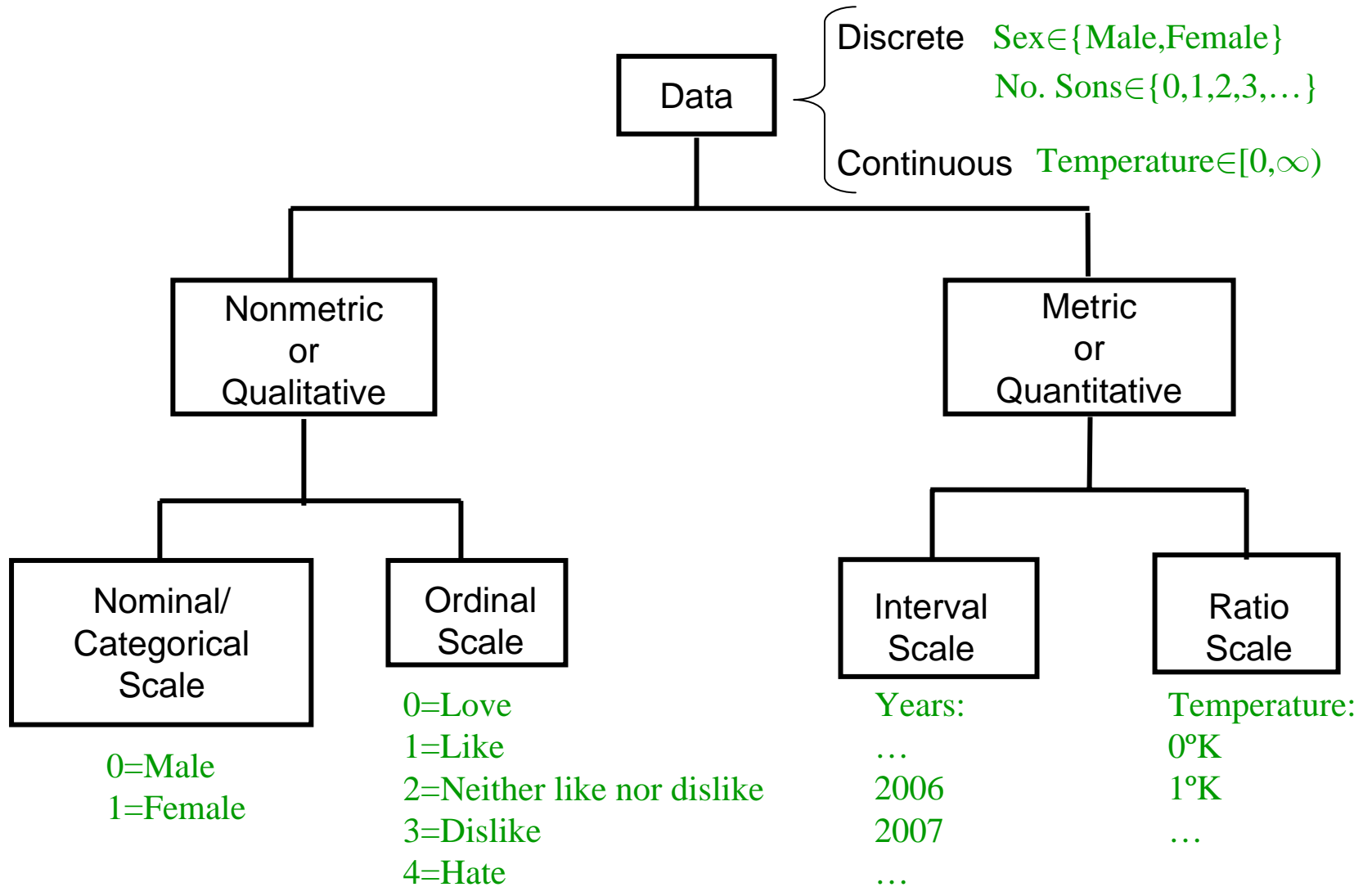
Description: This datafile contains 315 observations on 14 variables. This data set can be used to demonstrate multiple regression.

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

Coding of categorical variables

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

Peter: Black

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

Molly: Blond

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

Charles: Brown

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

Company size
{Small, Medium, Big} $\xrightarrow{\text{Implicit order}}$ $x_{\text{size}} \in \{0,1,2\}$

Company A: Big

Company A: 2

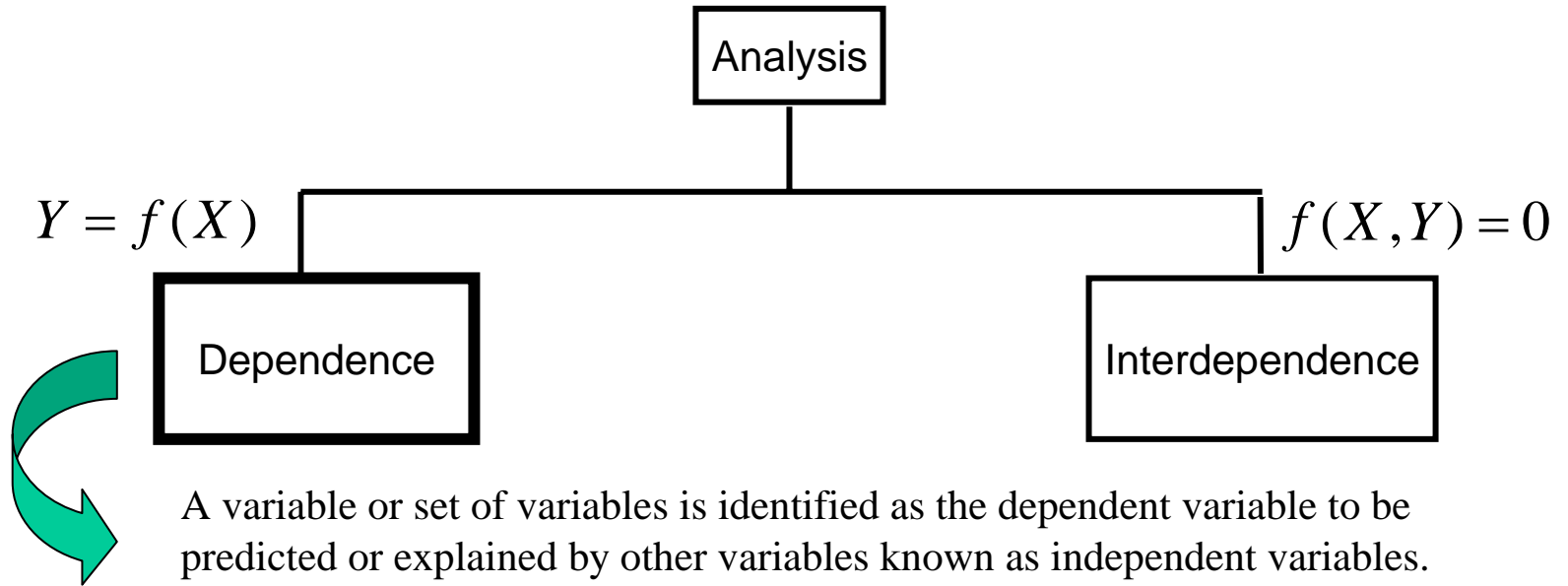
Company B: Small

Company B: 0

Company C: Medium

Company C: 1

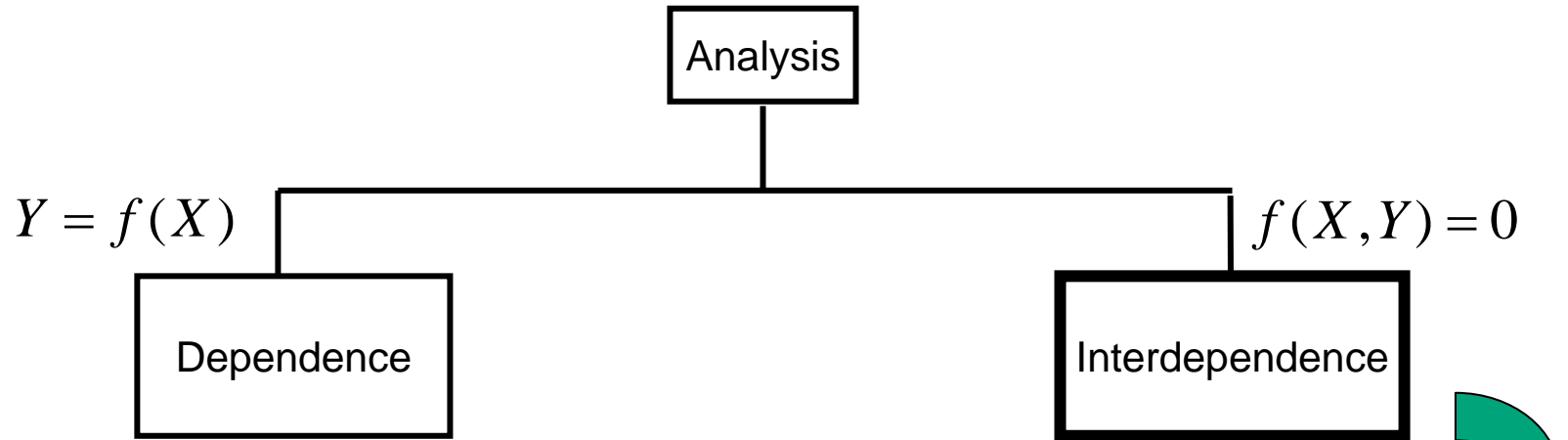
1.2 Introduction: Types of analysis



Example:
(No. Sons, House Type)=
 $f(\text{Income, Social Status, Studies})$

- 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



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

1.2 Introduction: Technique selection

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

Example:

$\text{No. Sons} = f(\text{Income, No. Years working})$

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

Example:

$(\text{No. Sons, House m}^2) = f(\text{Income, No. Years working, (No. Years Married)})$

1.2 Introduction: Technique selection

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

Example:

$(\text{Ability in Math, Ability in Physics})=f(\text{Math textbook, Physics textbook, College})$

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

Example:

$\text{Purchaser (or non purchaser)}=f(\text{Income, No. Years working})$

1.2 Introduction: Technique selection

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

Example:

$(\text{Grade Chemistry, Grade Physics}) = f(\text{Grade Math, Grade Latin})$

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

Example:

$\text{TV utility} = f(\text{Screen format, Screen size, Brand, Price})$

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(gen A, gen B)

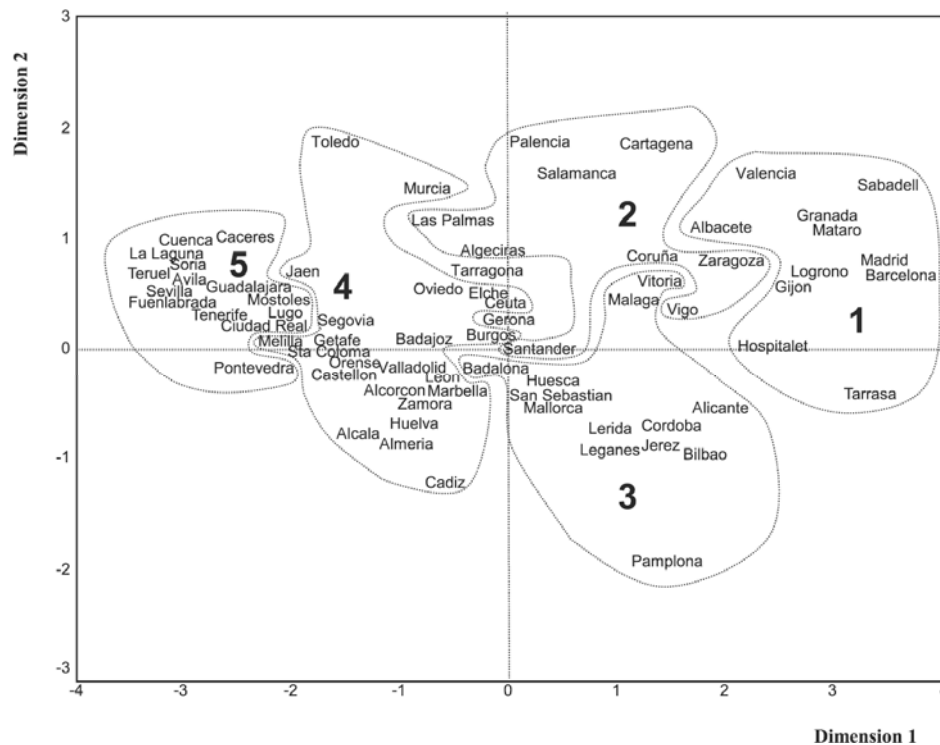
- Cluster analysis: try to group individuals according to similar characteristics

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:

$(x,y)=f(\text{City gross income, health indexes, population, political stability, ...})$

(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.

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

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

X^t

64	2	2	21.4838	1	1298.8	57	6.3	0	170.3	1945	890	200	915		
76	2	1	23.07631	1	1032.5	58.1	15.0	0	75.0	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}

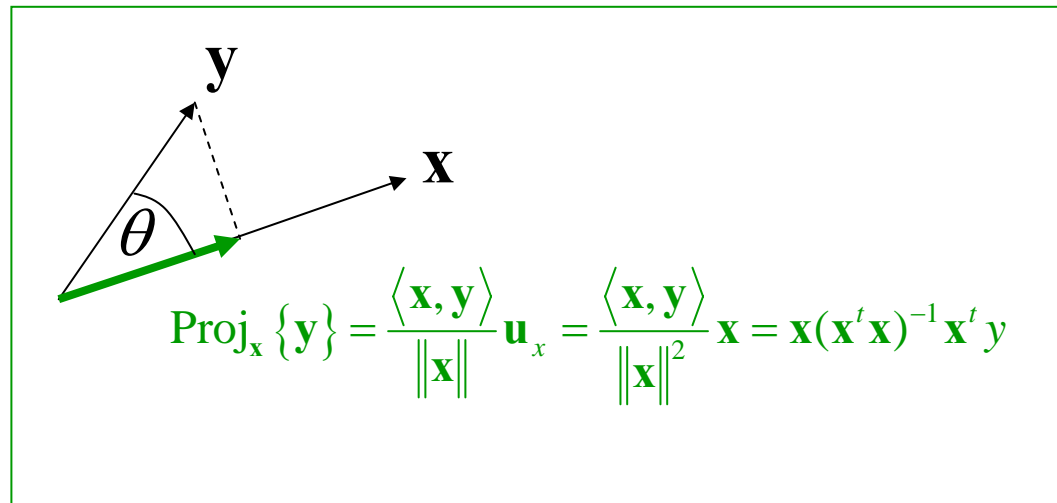
Norm

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

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



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

(Basic vector and matrix algebra)

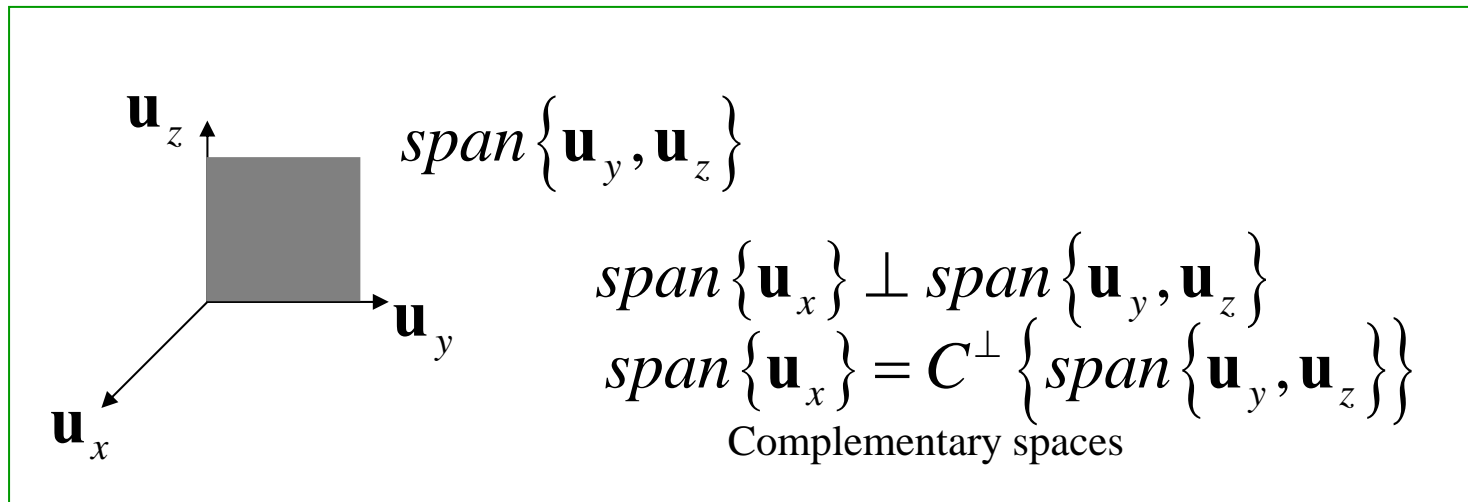
Linear span

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

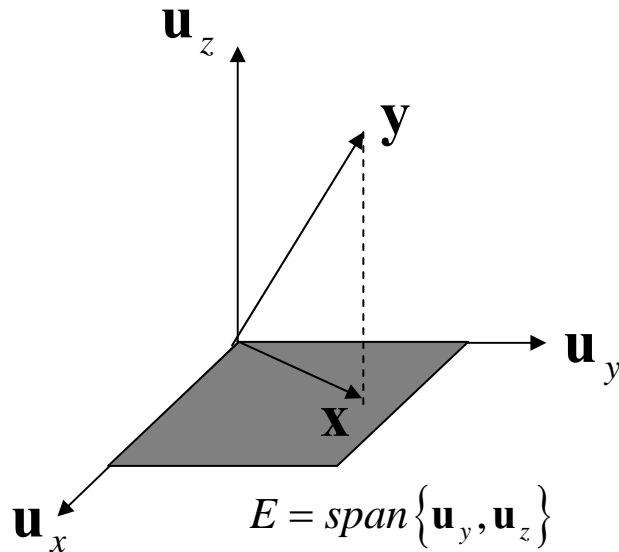
field
↓

Linearly dependent, i.e.,

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



(Basic vector and matrix algebra)



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

$$\begin{aligned} \mathbf{x} &= \text{Proj}_E \{\mathbf{y}\} = \text{Proj}_{\mathbf{u}_y} \{\mathbf{y}\} + \text{Proj}_{\mathbf{u}_z} \{\mathbf{y}\} \\ &= \mathbf{X} (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y} \end{aligned}$$

Basis vectors of E as columns

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

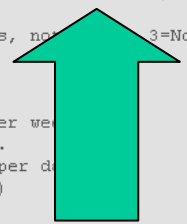
$$\|\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y} - \mathbf{x}\|^2$$

1.3 Descriptors: Data representation

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

```

http://lib.stat.cmu.edu/datas
File Edit View Favorites To
Back
Address http://lib.stat.cmu.edu/
Google
Nierenberg DW, Stukel
Description: This dat
Variable Names in orde
AGE: Age (year)
SEX: Sex (1=Ma
SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
QUETELET: Quetelet (weight/(height^2))
VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, no 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 we
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per d
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)
    
```



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

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

Sample variation coefficient

$$VC_2 = \sqrt{\frac{\overline{x_2^2}}{s_2^2}}$$

m_2 Sample median

$$\Pr\{X_2 \leq m_2\} \geq \frac{1}{2} \leq \Pr\{X_2 \geq m_2\}$$

If outliers

Robust statistics

MAD_2 Sample Median Absolute Deviation

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

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

$$\bar{\mathbf{x}} = (\bar{x}_1 \quad \bar{x}_2 \quad \dots \quad \bar{x}_p)^t$$

$$\tilde{X} = X - \mathbf{1}\mathbf{x}^t$$

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

Sample mean
↑
Vector of 1s

Sample covariance $s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ ← Measures how variables j and k are related

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

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

1.3 Descriptors: Covariance

$X = (x_1 \quad x_2 \quad x_3)$ 3 variables
200 samples

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

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

Sample covariance

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

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

Covariance

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

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

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

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

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

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

1.3 Descriptors: Covariance

$$\begin{aligned} X_1 &\sim N(1, 2) \\ X_2 &\sim N(2, 3) \\ X_3 &= X_1 - X_2 \end{aligned}$$

$$S = \begin{pmatrix} 1.6338 & -0.0970 & 1.5368 \\ -0.0970 & 2.8298 & -2.7329 \\ 1.5368 & -2.7329 & 4.2696 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 3 & -3 \\ 2 & -3 & 5 \end{pmatrix}$$

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

$$\begin{aligned} X_1 &\sim N(\mu_1, \sigma_1^2) \\ X_2 &\sim N(\mu_2, \sigma_2^2) \\ X_3 &= a_1 X_1 + a_2 X_2 \end{aligned}$$

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

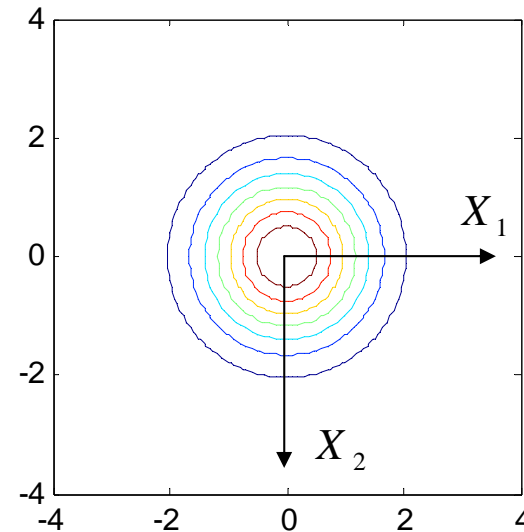
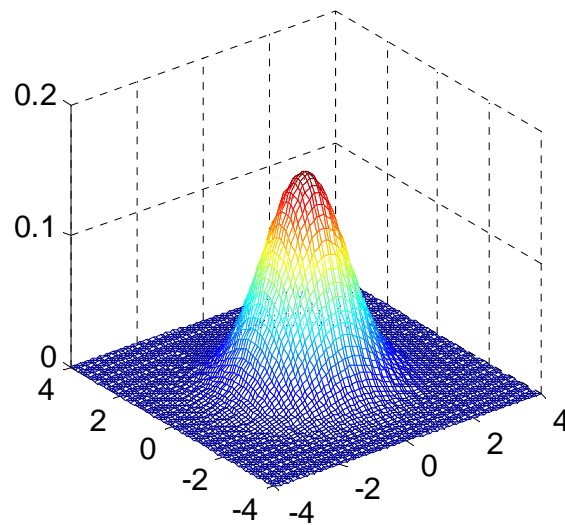
$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi |\Sigma|)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

1.3 Descriptors: Covariance

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

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

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



X_1 and X_2 are independent

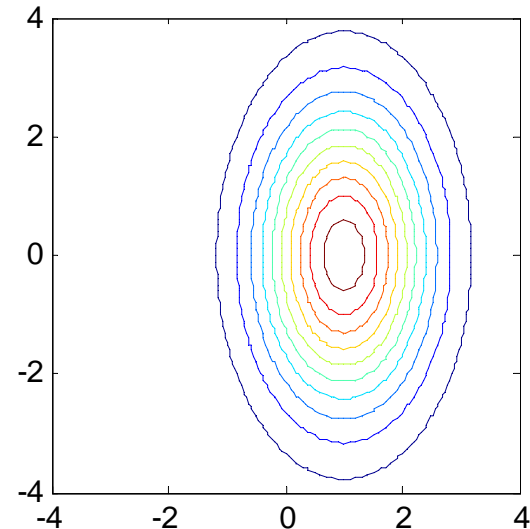
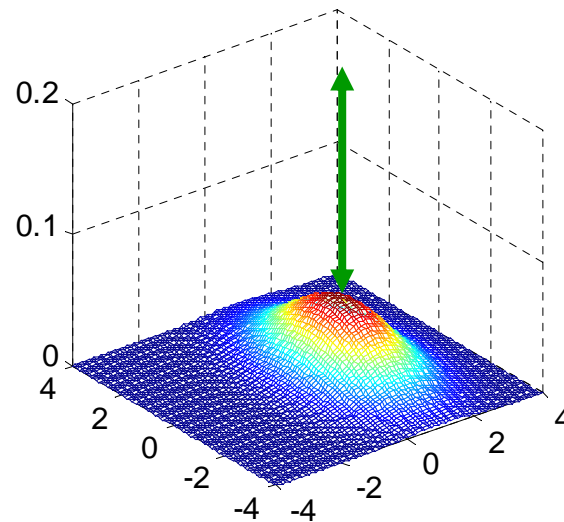
For multivariate Gaussians, covariance=0 implies independency

1.3 Descriptors: Covariance

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

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

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



X_1 and X_2 are independent

1.3 Descriptors: Covariance

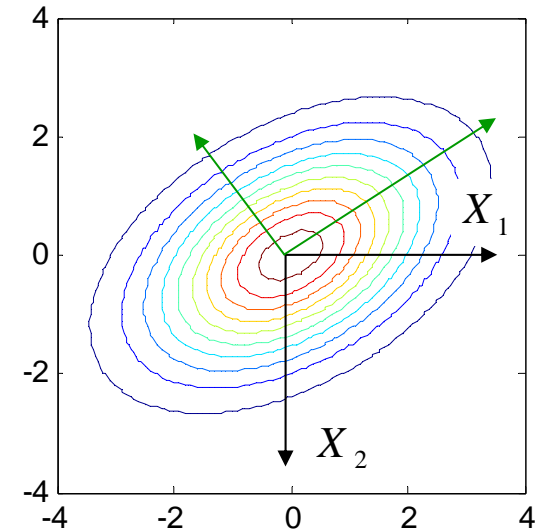
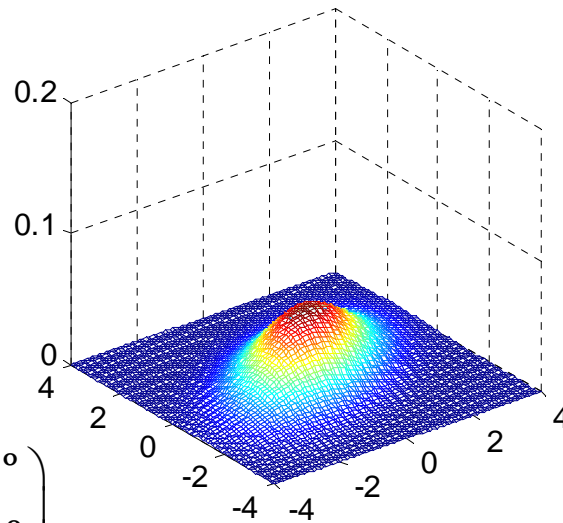
$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi |\Sigma|)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

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

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

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

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



X_1 and X_2 are NOT independent
BUT there exist two independent variables

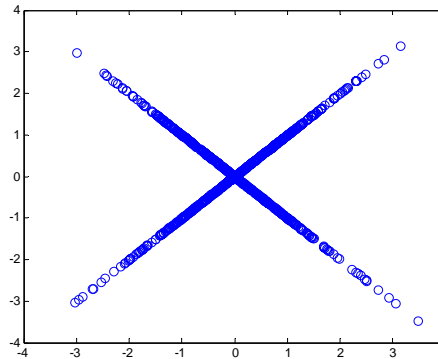
1.3 Descriptors: Covariance

Pitfalls of the covariance matrix

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

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

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



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

$$X_2 = X_1^2$$

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

$Cov(X_1, X_2) = 0 \Rightarrow \text{Uncorrelated} \not\Rightarrow \text{Independent}$
 $Cov(X_1, X_2) = 0 \wedge \text{Gaussian} \Rightarrow \text{Independent}$

1.3 Descriptors: Covariance

Redundant variables

$$\begin{array}{l} X_1 \sim N(0,1) \\ X_2 \sim N(0,1) \\ X_3 \sim N(0,1) \end{array} \quad \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{eig}(\Sigma) = (1,1,1)$$

$$\begin{array}{l} X_1 \sim N(0,1) \\ X_2 = X_1 \\ X_3 = -X_1 \end{array} \quad \Sigma = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \quad \text{eig}(\Sigma) = (1,0,0)$$

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

$$\begin{array}{l} X_1 \sim N(1,2) \\ X_2 \sim N(2,3) \\ X_3 = X_1 - X_2 \end{array} \quad \Sigma = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 3 & -3 \\ 2 & -3 & 5 \end{pmatrix} \quad \text{eig}(\Sigma) = (7.64, 2.35, 0)$$