CEU
Universidad
San Pablo

# Multivariate Data Analysis

## Session 1: Introduction and data examination

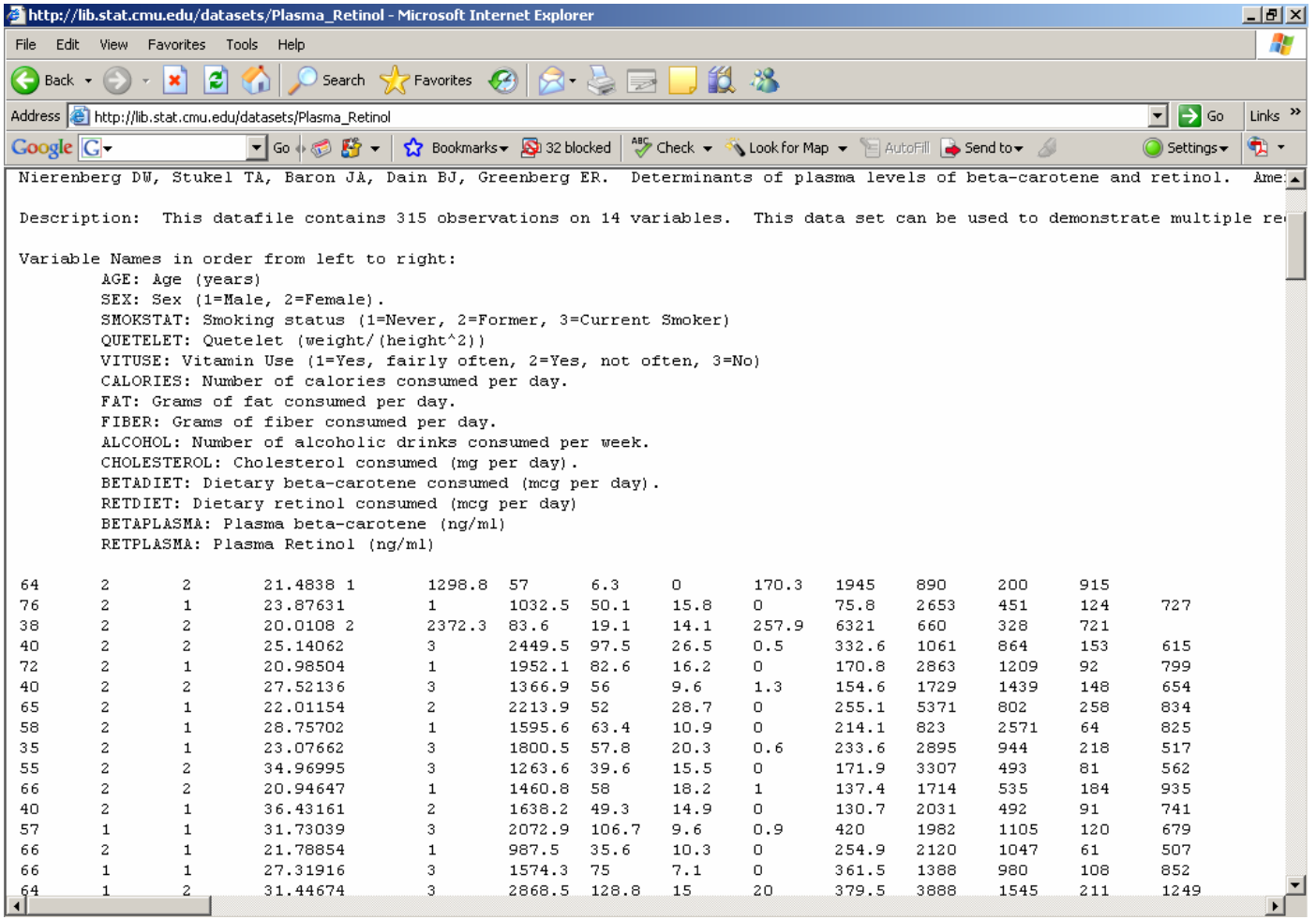
Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid, July 23rd 2007

# Course outline: Session 1

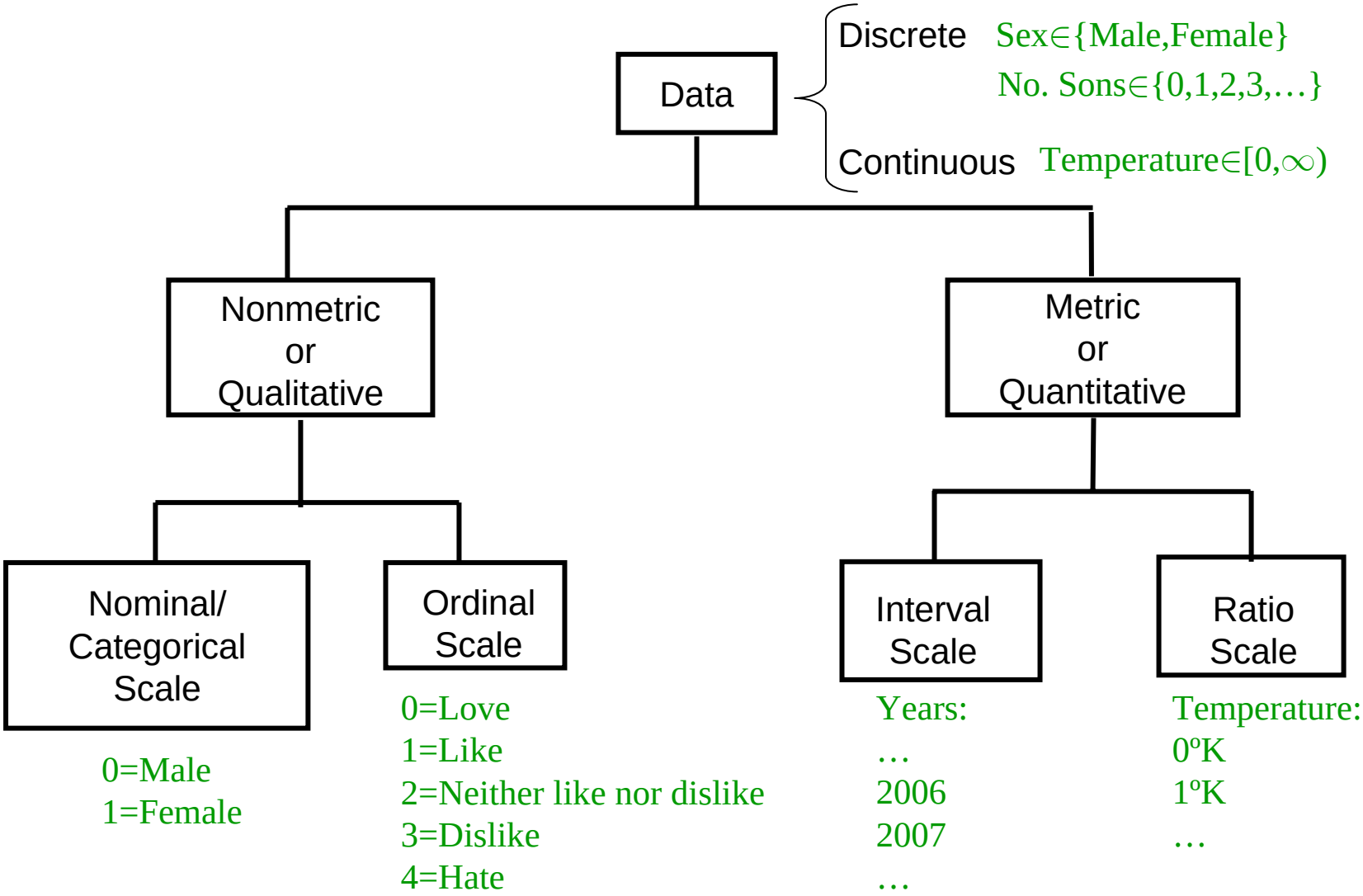
1. Introduction
    - 1.1. Types of variables
    - 1.2. Types of analysis and technique selection
    - 1.3. Descriptors (mean, covariance matrix)
    - 1.4. Variability and distance
    - 1.5. Linear dependence

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

# 1. Introduction



# 1.1 Introduction: Types of variables



# 1.1 Introduction: Types of variables

Coding of categorical variables

Hair Colour
{Brown, Blond, Black, Red}

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

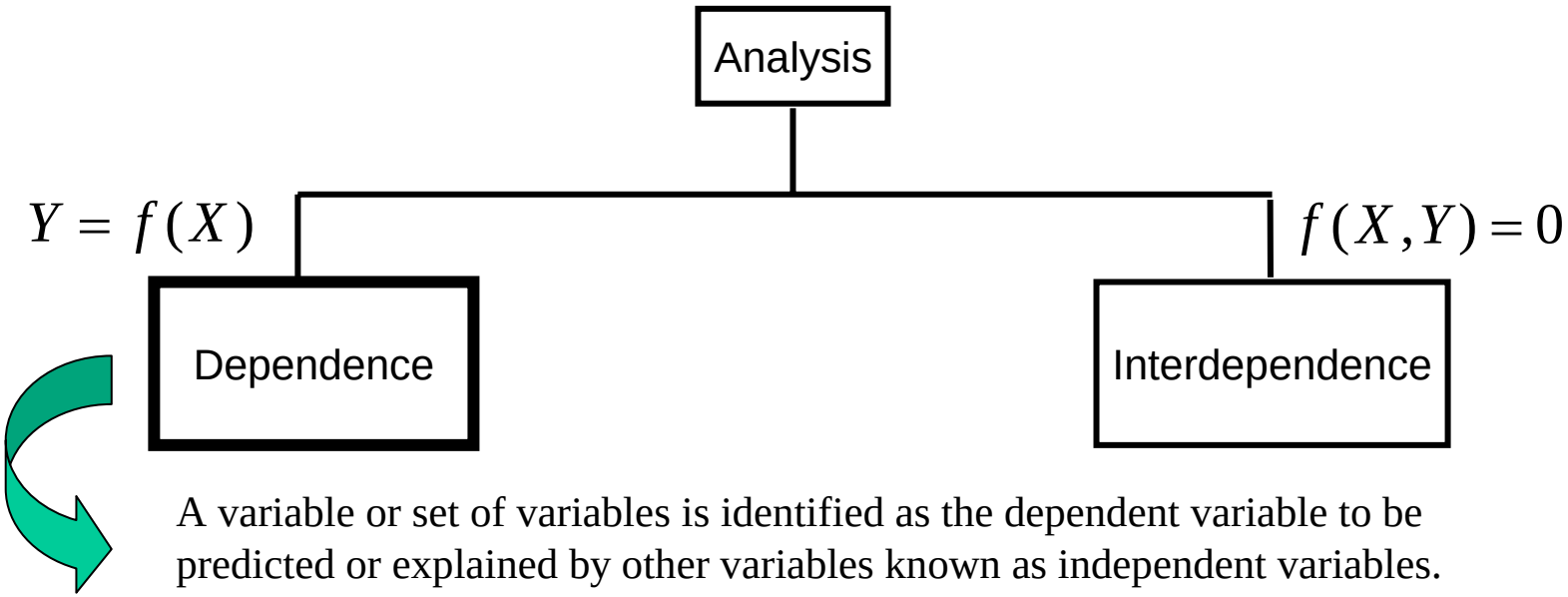
| | |
|---|---|
| Peter: Black | Peter: $\{0,0,1,0\}$ |
| Molly: Blond | Molly: $\{0,1,0,0\}$ |
| Charles: Brown | Charles: $\{1,0,0,0\}$ |

Company size
{Small, Medium, Big}

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

| | |
|---|---|
| Company A: Big | Company A: 2 |
| Company B: Small | Company B: 0 |
| Company C: Medium | Company C: 1 |

# 1.2 Introduction: Types of analysis

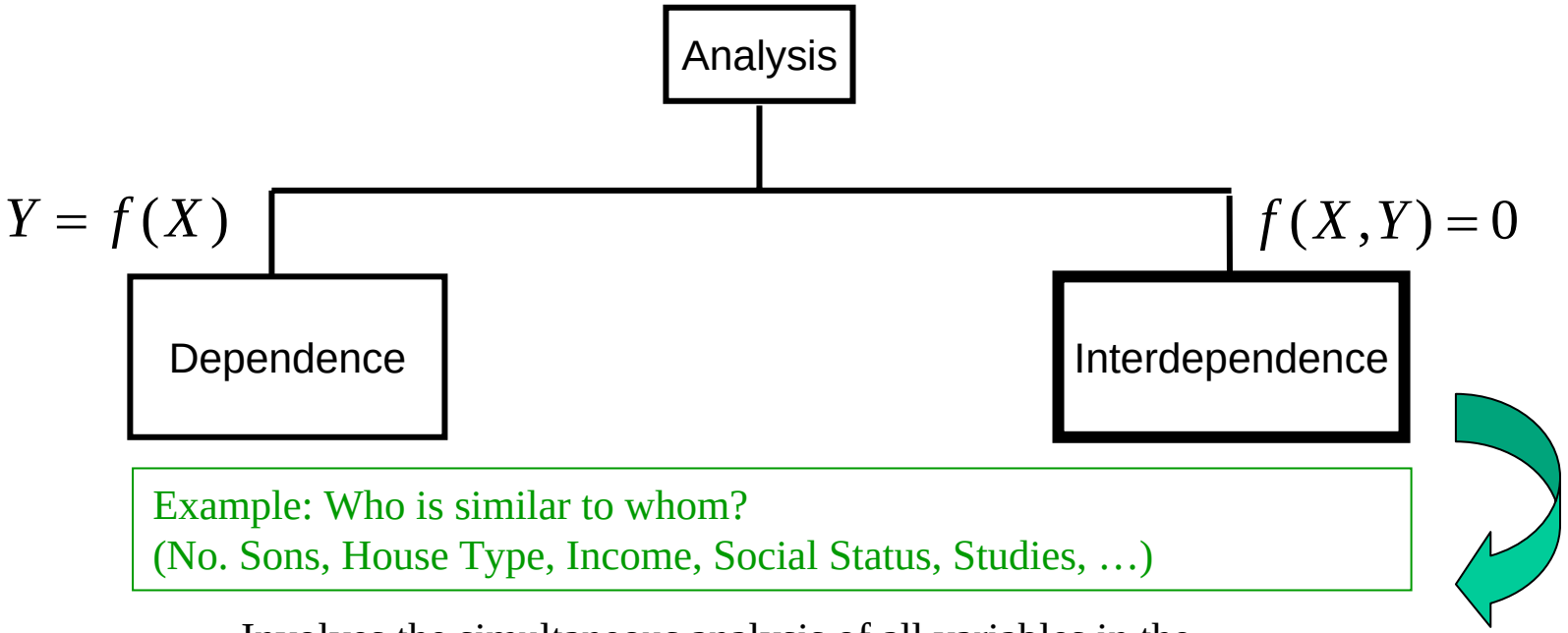



A variable or set of variables is identified as the dependent variable to be predicted or explained by other variables known as independent variables.

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




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:
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)

# 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.

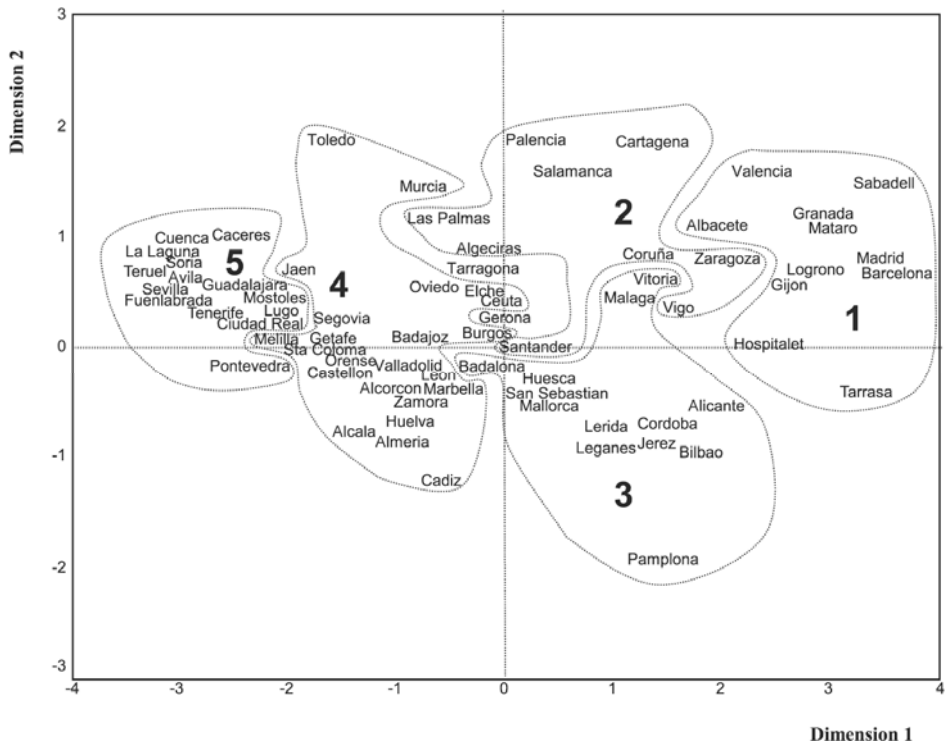
Example:
(Eye colour, Hair colour, Skin colour)=f(gen A, gen B)

- 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

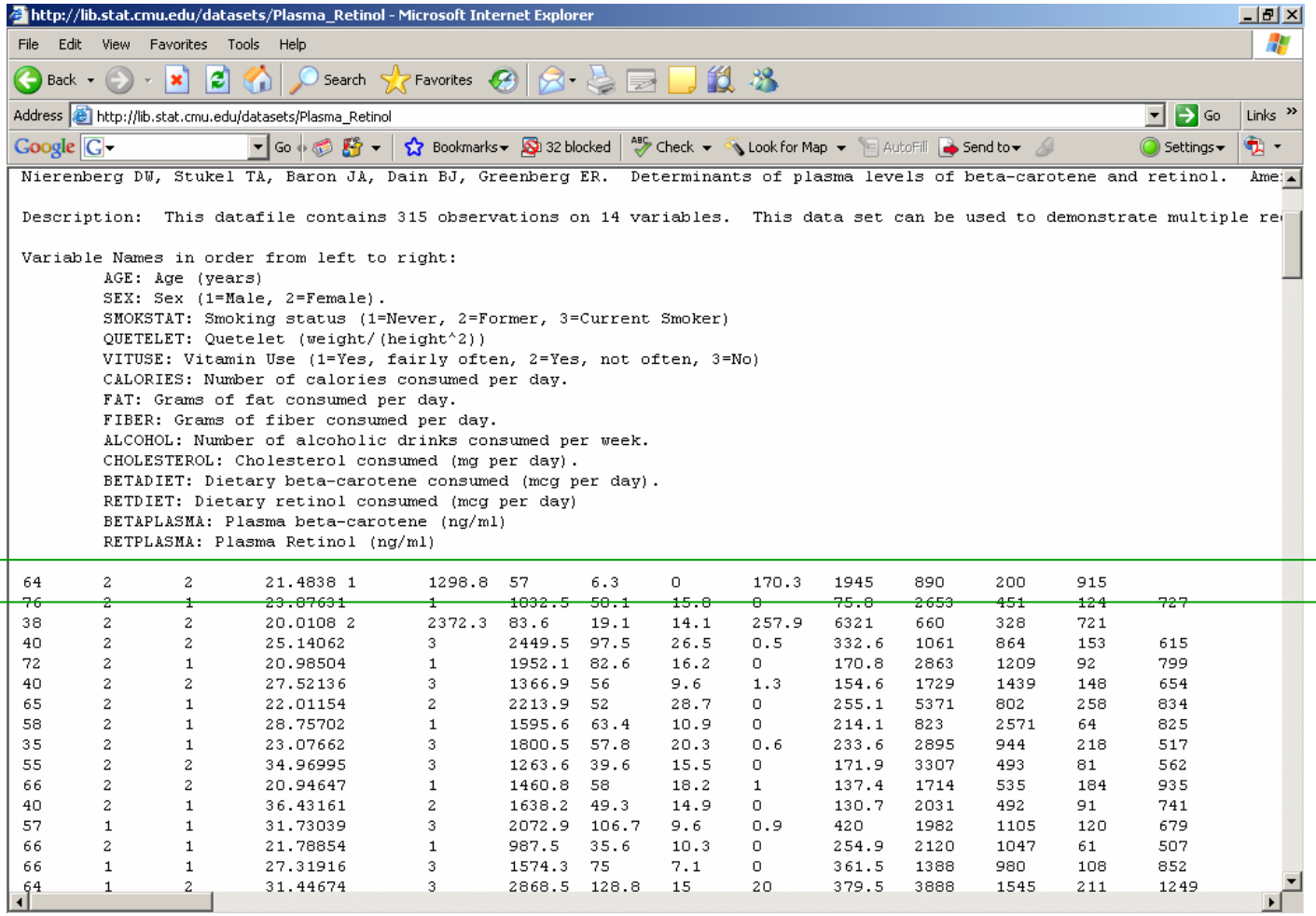
- Multidimensional scaling: Find representative factors so that the relative dissimilarities in the original space are as conserved as possible




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

# (Basic vector and matrix algebra)

$\mathbf{x}^t$

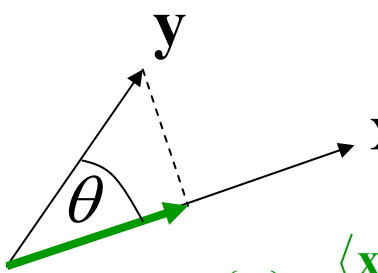


# (Basic vector and matrix algebra)

Vector $\mathbf{x}$

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

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




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

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

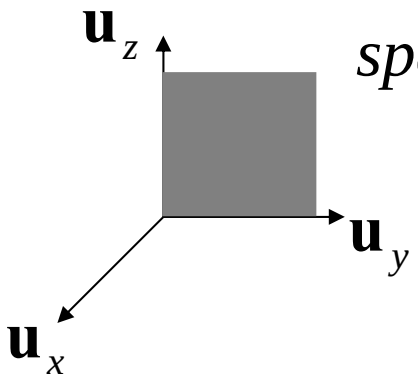
# (Basic vector and matrix algebra)

Linear span

field

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

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



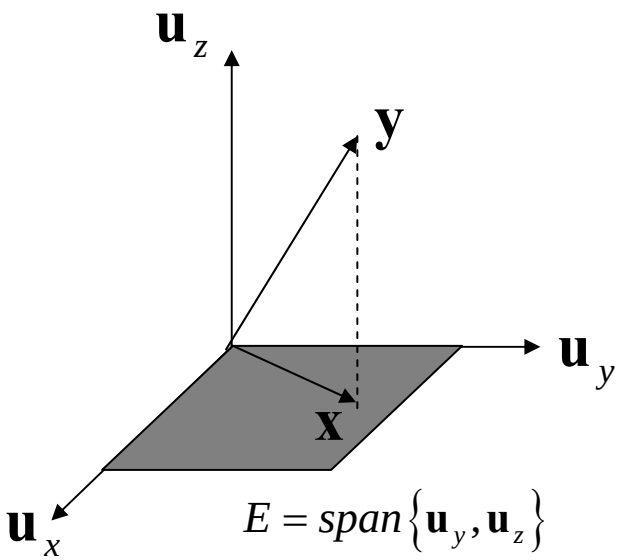

$$span\left\{\mathbf{u}_y, \mathbf{u}_z\right\}$$

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

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

Complementary spaces

# (Basic vector and matrix algebra)




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

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

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

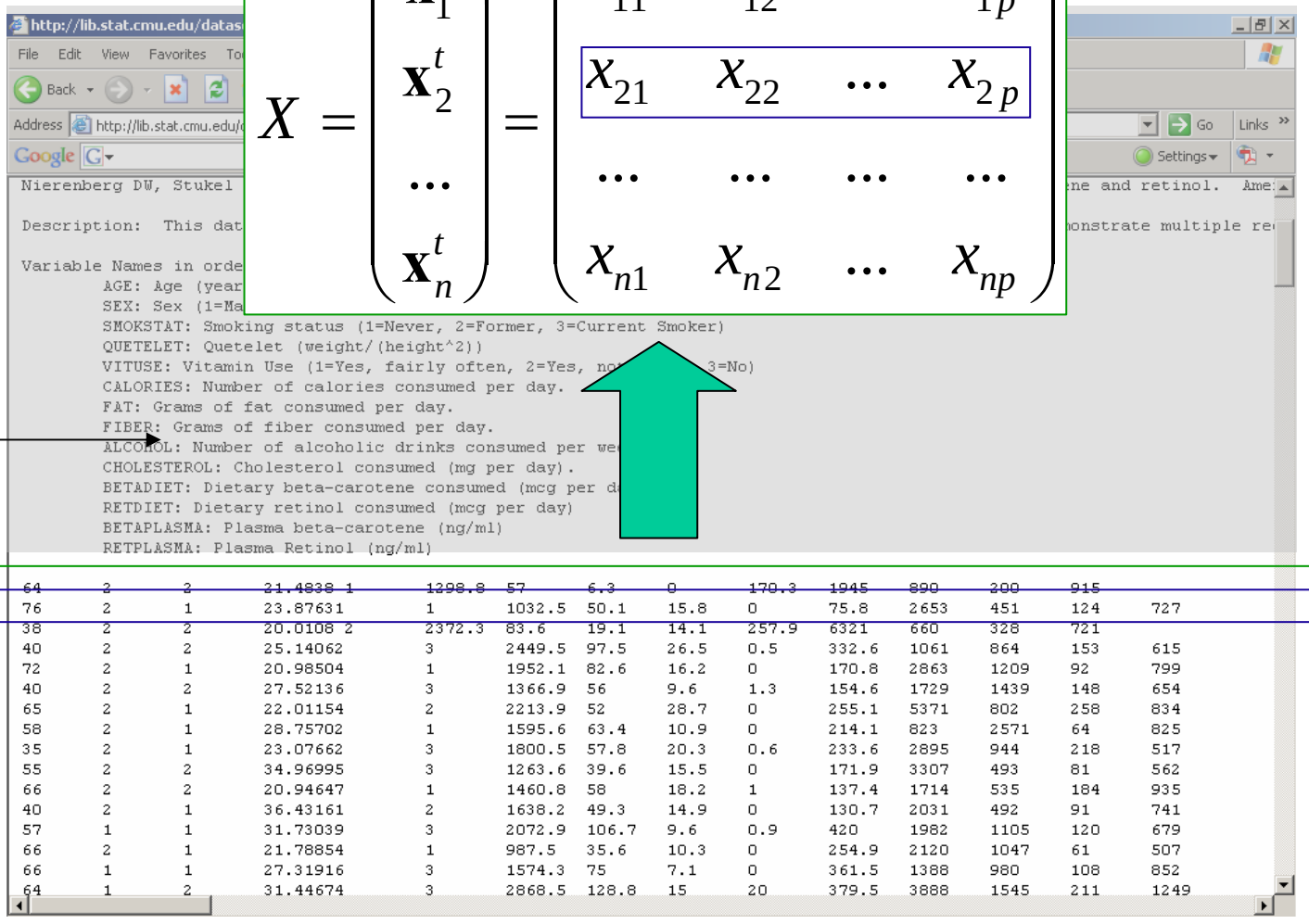
$$X = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \dots \\ \mathbf{x}_n^t \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Features

Individuals

$X$

$\mathbf{x}_2^t$



# 1.3 Descriptors: Univariate analysis

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

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

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{pmatrix}^t$$

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

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

Sample mean

↑ Vector of 1s

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

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

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

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

Sample covariance

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

Covariance

---

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

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

---

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

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

# 1.3 Descriptors: Covariance

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

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

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

---

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

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

---

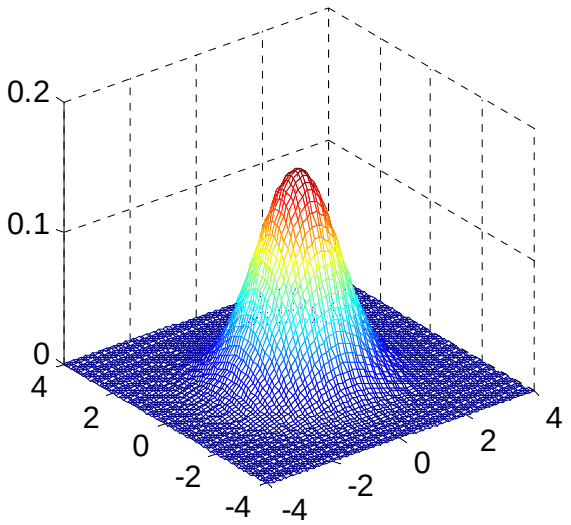
$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\left(2\pi |\Sigma|\right)^{\frac{N}{2}}} \exp\left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

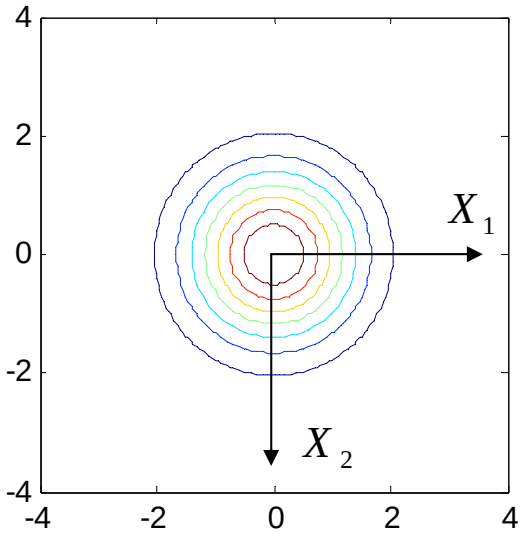
# 1.3 Descriptors: Covariance

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

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

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




$X_1$ and $X_2$ are independent

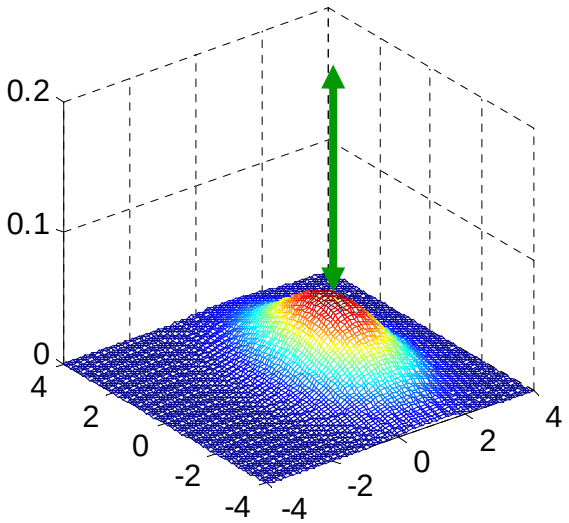
**For multivariate Gaussians, covariance=0 implies independency**

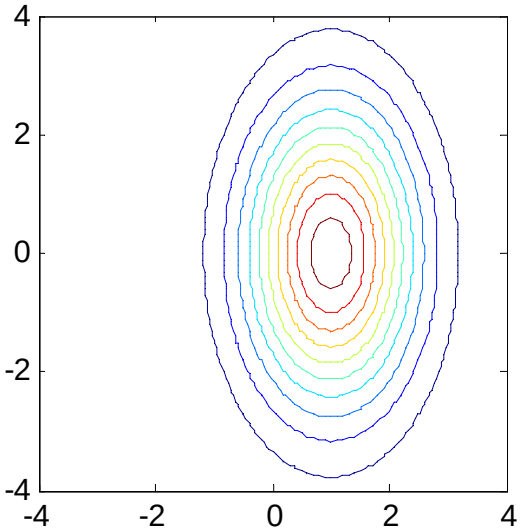
# 1.3 Descriptors: Covariance

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

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

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




$X_1$ and $X_2$ are independent

# 1.3 Descriptors: Covariance

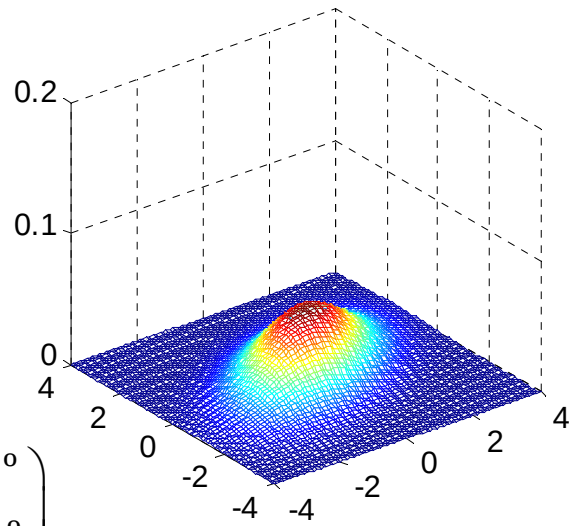
$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \longrightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi|\Sigma|)^{\frac{N}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

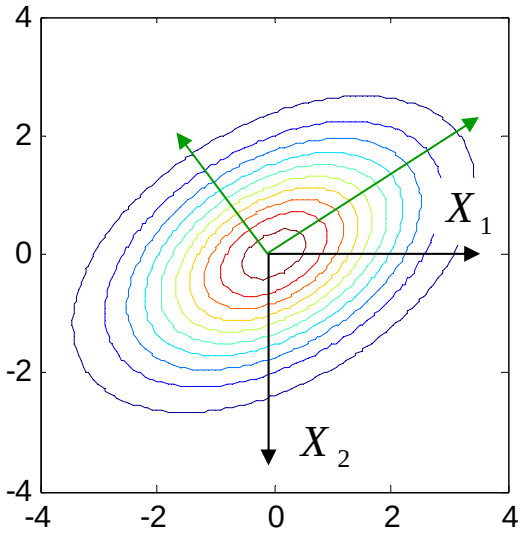
$$\boldsymbol{\mu} = \mathbf{0}$$

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

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

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




$X_1$ and $X_2$ are NOT independent

BUT there exist two independent variables

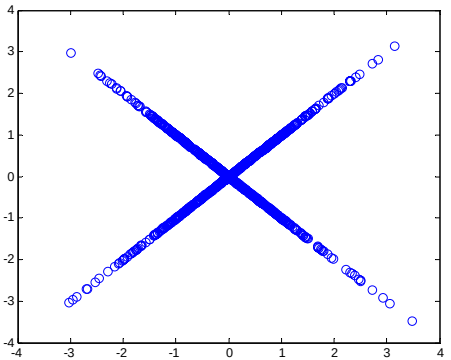
# 1.3 Descriptors: Covariance

Pitfalls of the covariance matrix

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

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

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




---

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

$$X_2 = X_1^2$$

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

$$Cov(X_1, X_2) = 0 \Rightarrow Uncorrelated \not\Rightarrow Independent$$

$$Cov(X_1, X_2) = 0 \wedge Gaussian \Rightarrow Independent$$

# 1.3 Descriptors: Covariance

Redundant variables

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

---

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

---

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

---

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