MULTIVARIATE DATA ANALYSIS

- data $\underset{(n,p)}{\mathbb{X}} = \begin{pmatrix} \underset{\sim}{x}_1 \\ \vdots \\ \underset{\sim}{x}_n \end{pmatrix}$ n measurements on p variables

- the usual model
  Dataset is a realization of i.i.d random vectors $\underset{\sim}{X}_1, \dots, \underset{\sim}{X}_n$

VISUALIZATION METHODS

A BIVARIATE KERNEL DENSITY ESTIMATOR

$(x_{11}, x_{12}), \dots, (x_{n1}, x_{n2})$ i.i.d from F with the density $f(x,y)$

$$\hat{f}(x_1, x_2) = \frac{1}{n h_1 h_2} \sum_{i=1}^{n} K\left(\frac{x_1 - x_{1i}}{h_1}\right) K\left(\frac{x_2 - x_{2i}}{h_2}\right)$$

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

contour ... VERSTEHEN

11. multivariate normal distribution

– slides 95 + 100

– for $p = 2$: $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$ Σ

$\mu_1 = \mathbb{E}X_1$, $\mu_2 = \mathbb{E}X_2$

$\sigma_1^2 = \operatorname{var} X_1$, $\sigma_2^2 = \operatorname{var} X_2$

$\rho\sigma_1\sigma_2 = \operatorname{cov}(X_1, X_2)$

$\rightsquigarrow \rho = \dfrac{\operatorname{cov}(X_1, X_2)}{\sqrt{\sigma_1^2 \sigma_2^2}} = \dfrac{\operatorname{cov}(X_1, X_2)}{\sqrt{\operatorname{var} X_1 \operatorname{var} X_2}}$

$|\Sigma| = \sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$

$\Sigma^{-1} = \dfrac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$

$$f(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} \right) \right\}$$

MATOUCI NA SLIDU 95, 100

– for $p = 1$: $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

III. Testing the multivariate mean

A. ONE-SAMPLE PROBLEM

$\underset{\sim}{X}_1, \dots, \underset{\sim}{X}_n$ i.i.d $N_p(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$ ← UNKNOWN

$H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \qquad H_1: \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$

note that $\operatorname{var}(\bar{\underset{\sim}{X}}) = \frac{1}{n}\underset{\sim}{\Sigma} \approx \frac{1}{n}\hat{\underset{\sim}{\Sigma}}$

$$T_n = \left(\frac{n-p}{p}\right)(\bar{\underset{\sim}{X}}_n - \underset{\sim}{\mu}_0)^T \hat{\underset{\sim}{\Sigma}}_n^{-1} (\bar{\underset{\sim}{X}}_n - \underset{\sim}{\mu}_0)$$

where $\hat{\underset{\sim}{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n (\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})$, $\bar{\underset{\sim}{X}} = \frac{1}{n}\sum_{i=1}^n \underset{\sim}{X}_i$

under H_0: $T_n \overset{a}{\sim} F_{p, n-p}$

BOOTSTRAP $\underset{\sim}{X}_1^*, \dots, \underset{\sim}{X}_n^*$ i.i.d from $\underset{\sim}{X}_1 - \bar{\underset{\sim}{X}}, \dots, \underset{\sim}{X}_n - \bar{\underset{\sim}{X}}$

B. TWO-SAMPLE PROBLEM

$\underset{\sim}{X}_1^{(1)}, \dots, \underset{\sim}{X}_{n_1}^{(1)}$ i.i.d $N_p(\underset{\sim}{\mu}_1, \underset{\sim}{\Sigma})$ } under

$\underset{\sim}{X}_1^{(2)}, \dots, \underset{\sim}{X}_{n_2}^{(2)}$ — " — $N_p(\underset{\sim}{\mu}_2, \underset{\sim}{\Sigma})$ }

$H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 \qquad H_1: \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$

$$T_n = \frac{n_1+n_2-p-1}{p(n_1+n_2-2)} (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)^T \hat{\underset{\sim}{\Sigma}}_n^{-1} (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2) \overset{H_0}{\sim} F_{p, n_1+n_2-p-1}$$

$$\hat{\underset{\sim}{\mu}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \underset{\sim}{X}_i^{(1)} \qquad \hat{\underset{\sim}{\mu}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \underset{\sim}{X}_i^{(2)}, \qquad \hat{\underset{\sim}{\Sigma}}_n = \frac{1}{n_1+n_2-2}\left((n_1-1)\hat{\underset{\sim}{\Sigma}}_1 + (n_2-1)\hat{\underset{\sim}{\Sigma}}_2\right)$$

$$\hat{\underset{\sim}{\Sigma}}_1 = \frac{1}{n_1-1}\sum_{i=1}^{n_1} (\underset{\sim}{X}_i^{(1)} - \hat{\underset{\sim}{\mu}}_1)(\underset{\sim}{X}_i^{(1)} - \hat{\underset{\sim}{\mu}}_2)^T$$

$$\hat{\underset{\sim}{\Sigma}}_2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2} (\underset{\sim}{X}_i^{(2)} - \hat{\underset{\sim}{\mu}}_2)(\underset{\sim}{X}_i^{(2)} - \hat{\underset{\sim}{\mu}}_2)^T$$

BOOTSTRAP $\underset{\sim}{X}_1^{(1)*}, \dots, \underset{\sim}{X}_{n_1}^{(1)*}, \underset{\sim}{X}_1^{(2)*}, \dots, \underset{\sim}{X}_{n_2}^{(2)*}$ i.i.d

from $\underset{\sim}{X}_1^{(1)} - \hat{\underset{\sim}{\mu}}_1, \dots, \underset{\sim}{X}_{n_1}^{(1)} - \hat{\underset{\sim}{\mu}}_1, \underset{\sim}{X}_1^{(2)} - \hat{\underset{\sim}{\mu}}_2, \dots, \underset{\sim}{X}_{n_2}^{(2)} - \hat{\underset{\sim}{\mu}}_2$

PERMUTATION VERSION OF THE TEST

$\underset{\sim}{X}_1^{(1)*}, \ldots, \underset{\sim}{X}_{n_1}^{(1)*}, \underset{\sim}{X}_1^{(2)*}, \ldots, \underset{\sim}{X}_{n_2}^{(2)*}$ i.i.d from $\underset{\sim}{X}_1^{(1)} - \hat{\underset{\sim}{\mu}}_1, \ldots, \underset{\sim}{X}_{n_2}^{(2)} - \hat{\underset{\sim}{\mu}}_2$

WITHOUT ASSUMPTION OF $\mathrm{var}(\underset{\sim}{X}_1) = \mathrm{var}(\underset{\sim}{X}_2)$

$H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 \qquad H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$

$$\tilde{T}_n = (\hat{\underset{\sim}{\mu}}_1 - \hat{\underset{\sim}{\mu}}_2)^T \left(\frac{\hat{\underset{\sim}{\Sigma}}_1}{n_1} + \frac{\hat{\underset{\sim}{\Sigma}}_2}{n_2} \right)^{-1} (\bar{\underset{\sim}{\mu}}_1 - \bar{\underset{\sim}{\mu}}_2) \underset{H_0}{\overset{as}{\sim}} \chi^2_p$$

PRINCIPAL COMPONENT ANALYSIS (PCA)

- the most common dimension reduction technique

THEOR. INTRODUCTION

- Let $\underset{\sim}{X} = (X_1, \dots, X_n)^T$ be a random vector

LET ~~SEARCH FOR~~ $\underset{\sim}{\alpha} = (\alpha_1, \dots, \alpha_n)^T$ ~~such that~~

AND INTRODUCE $Z(\underset{\sim}{\alpha}) = \underset{\sim}{\alpha}^T \underset{\sim}{X} = \sum_{j=1}^{n} \alpha_j X_j$

WE SEARCH FOR $\underset{\sim}{\alpha}$ such that:

~~max~~ $\operatorname{var}(Z(\underset{\sim}{\alpha}))$ is maximized subject to $\|\underset{\sim}{\alpha}\| = 1$

$\|\underset{\sim}{\alpha}\| = \sqrt{\underset{\sim}{\alpha}^T \underset{\sim}{\alpha}} = \sqrt{\sum_{j=1}^{n} \alpha_j^2}$

- note that

$$\operatorname{var}(Z(\underset{\sim}{\alpha})) = \operatorname{var}(\underset{\sim}{\alpha}^T \underset{\sim}{X}) = \underset{\sim}{\alpha}^T \operatorname{var}(\underset{\sim}{X}) \underset{\sim}{\alpha} = \sum_{j=1}^{n} \sum_{i=1}^{n} \operatorname{cov}(X_j, X_i) \alpha_j \alpha_i$$

$$\begin{pmatrix} \operatorname{var}(X_1) & & \\ \operatorname{cov}(X_2, X_1) & \operatorname{var}(X_2) & \\ \operatorname{cov}(X_n, X_1) & & \operatorname{var}(X_n) \end{pmatrix}$$

where $\operatorname{Cov}(X_j, X_i) = \operatorname{var}(X_j)$

$\rightsquigarrow$ our task is:

$$\max \ \underset{\sim}{\alpha}^T \operatorname{var}(\underset{\sim}{X}) \underset{\sim}{\alpha} \quad \text{s.t.} \ \|\underset{\sim}{\alpha}\| = 1$$

Spectral decomposition

- Let A be a symmetric matrix of order (n, n). Then one can write A as

$$A = \underset{(n,n)}{\Gamma} \ \underset{(n,n)}{\Lambda} \ \underset{(n,n)}{\Gamma^T}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{pmatrix}$

with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ being the eigenvalues of A

Further $\Gamma = (\underline{x}_1, \dots, \underline{x}_n)$, where

$\underline{x}_1, \dots, \underline{x}_n$ are eigenvectors corresponding to $\lambda_1, \dots, \lambda_n$.

Note that $\Gamma^T \Gamma = \begin{pmatrix} - \underline{x}_1^T - \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix} \begin{pmatrix} | & & | \\ \underline{x}_1 & \cdots & \underline{x}_n \\ | & & | \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$,

as $\underline{x}_\gamma^T \underline{x}_\ell = 0$ for $\gamma \neq \ell$.

$A \underline{x}_\ell = \lambda_\ell \underline{x}_\ell$, $\ell = 1, \dots, n$

Note that $A = \sum_{\gamma=1}^{n} \lambda_\gamma \underline{x}_\gamma \underline{x}_\gamma^T$

Lemma Let A be a sym. (p,p) matrix and $\underline{d} \in \mathbb{R}^n$ s.t. $\|\underline{d}\| = 1$

Then $\underline{d}^T A \underline{d} \leq \lambda_1$, where λ_1 is the max. eigenvalue of A.

Proof: One can write $\underline{d} = \sum_{\ell=1}^{n} c_\ell \underline{x}_\ell$

note that $1 = \underline{d}^T \underline{d} = \left(\sum_{\ell=1}^{n} c_\ell \underline{x}_\ell^T \right) \sum_{\gamma=1}^{n} c_\gamma \underline{x}_\gamma =$

$= \sum_{\ell=1}^{n} c_\ell^2 \underline{x}_\ell^T \underline{x}_\ell + \sum_{\ell \neq \gamma} c_\ell c_\gamma \underline{x}_\ell^T \underline{x}_\gamma = \sum_{\ell=1}^{n} c_\ell^2$

Further

$\underline{d}^T A \underline{d} = \sum_{\ell=1}^{n} c_\ell \underline{x}_\ell^T \left(\sum_{\gamma=1}^{n} \lambda_\gamma \underline{x}_\gamma \underline{x}_\gamma^T \right) \sum_{\ell=1}^{n} c_\ell \underline{x}_\ell$

$= \sum_{\ell=1}^{n} c_\ell \underline{x}_\ell^T \sum_{\gamma=1}^{n} \lambda_\gamma c_\gamma \underline{x}_\gamma = \sum_{\gamma=1}^{n} \lambda_\gamma c_\gamma^2 \leq$

$\leq \lambda_1 \sum_{\gamma=1}^{n} c_\gamma^2 = \lambda_1$ □

Consider now $\underline{d} = \underline{x}_1$. Then $\underline{d}^T \underline{d} = \|\underline{d}\|^2 = 1$

and further $\underline{d}^T A \underline{d} = \underline{x}_1^T A \underline{x}_1 = \underline{x}_1^T \lambda_1 \underline{x}_1 = \lambda_1$

$\rightsquigarrow$ $\underline{d}^T$ var X $\underline{d}$ as maximized (s.t. $\|\underline{d}\| = 1$)

for $\underline{x}_1$ being the eigenvector corresponding to

the maximal eigenvalue λ.

Bsp „Bivariate Verteilung"

$$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad , \quad \mathrm{var}(\underset{\sim}{X}) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) \end{pmatrix}$$

$Z(\underset{\sim}{d}) = d_1 X_1 + d_2 X_2$, $\|\underset{\sim}{d}\|^2 = d_1^2 + d_2^2 = 1 \rightsquigarrow d_2^2 = 1 - d_1^2$

$\mathrm{var}(Z(\underset{\sim}{d})) = d_1^2 \mathrm{var} X_1 + d_2^2 \mathrm{var} X_2 + 2 d_1 d_2 \mathrm{cov}(X_1, X_2)$

a) X_1 ~~and X_2 are independent~~ $\mathrm{cov}(X_1, X_2) = 0$

~~$\rightsquigarrow \mathrm{cov}(X_1, X_2) = 0$~~

$\rightsquigarrow \mathrm{var}(Z(\underset{\sim}{d})) = d_1^2 \mathrm{var} X_1 + d_2^2 \mathrm{var} X_2 = d_1^2 \mathrm{var} X_1 + (1 - d_1^2) \mathrm{var} X_2$

if $\mathrm{var} X_1 > \mathrm{var} X_2 \rightsquigarrow$ $\mathrm{var}(Z(\underset{\sim}{d}))$ is maximized for $d_1 = 1, d_2 = 0$

$\mathrm{var} X_1 < \mathrm{var} X_2 \rightsquigarrow$ —"— $d_1 = 0, d_2 = 1$

$\mathrm{var}(X_1) = \mathrm{var} X_2 \rightsquigarrow \mathrm{var}(Z(\underset{\sim}{d})) = \mathrm{var} X_1 = \mathrm{var} X_2$
CONSTANT IN d_1, d_2

b) $\mathrm{cov}(X_1, X_2) \neq 0$

For simplicity consider $\mathrm{var} X_1 = \mathrm{var} X_2$

$\rightsquigarrow \mathrm{var}(Z(\underset{\sim}{d})) = d_1^2 \mathrm{var} X_1 + (1 - d_1^2) + 2 d_1 d_2 \mathrm{cov}(X_1, X_2)$

($\mathrm{var} X_2$)

"$\mathrm{var} X_1$

maximal < 0

$\Leftrightarrow |d_1| = |d_2|$

$\rightsquigarrow |d_1| = \frac{1}{\sqrt{2}} = d_2$

(b-1) $\mathrm{cov}(X_1, X_2) > 0$

$(d_1, d_2)^T = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T$ oder $(d_1, d_2) = \left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$

(b-2) $\mathrm{cov}(X_1, X_2) < 0$

$(d_1, d_2)^T = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)^T$ oder $(d_1, d_2)^T = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T$