# The covariance matrix [H.2.2, H:2.3]

TMA4267 L3
17.01.2017

$\underset{p\times 1}{X}$ random vector $\quad E(X) = \mu_{p\times 1}$

$$\mathrm{Cov}(X) = \underset{p\times p}{\Sigma} = E\left[(X-\mu)(X-\mu)^T\right]$$

$$= \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & & \\ & & \ddots & \\ \mathrm{Cov}(X_1, X_p) & \cdots & & \mathrm{Var}(X_p) \end{bmatrix}$$

EX: Drinking habits $X_1$ = coffee, $X_2$ = tea, $X_3$ = cocoa, $X_4$ = liquer, $X_5$ = wine, $X_6$ = beer

i.i.d $n = 21$ countries. $\qquad$ Consumption of.

$\Sigma$ is real and symmetric. Other requirements?
We will need $\Sigma^{-1}$ so we require that $\Sigma$ is invertible, that, $\det(\Sigma) \neq 0$ (or we might use generalized inverses)

But, we looked at $\underset{k\times p}{C}\underset{p\times 1}{X}$ and found $\mathrm{Cov}(CX) = C\Sigma C^T$.

If $k = 1$ then $\underbrace{\underset{1\times p}{c^T}\underset{p\times 1}{X}}_{1\times 1}$ and $\underbrace{\mathrm{Cov}(c^T X)}_{\underbrace{\mathrm{Var}(c^T X)}_{\text{scalar}}} = \underbrace{\boxed{c^T \Sigma c}}_{\text{scalar}}$

We want $\mathrm{Var}(c^T X) = c^T \Sigma c$ to be positive
(because we don't want 0 or negative variances).

So, what do we need?

$$\boxed{\underset{1\times p}{c^T}\underset{p\times p}{\Sigma}\underset{p\times 1}{c} > 0} \quad \text{for all } c \neq 0$$

this is the definition of a S P D matrix.

S — symmetric, P — positive, D — definite

Useful result:

If $\Sigma$ is SPD $\iff$ all $\lambda_i > 0$

$\lambda_i$: eigenvalues of $\Sigma$

Proof via spectral decomposition:

Let $P \Lambda P^T = \Sigma$ where $P = [e_1\ e_2\ \cdots\ e_p]$ is a matrix with the normalized eigenvectors of $\Sigma$ as column vectors and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ is a diagonal matrix with the eigenvalues of $\Sigma$ on the diagonal.

Remember: $(\lambda_i, e_i)$ satisfy $\Sigma e_i = \lambda_i e_i$ and use $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. And that a (real) symmetric matrix has real eigenvalues (and eigenvectors of distinct eigenvalues are orthogonal)

$\Leftarrow$

$$\underset{1\times p}{c^T} \Sigma c > 0 \qquad \text{for all } c \neq 0$$

$$\underbrace{c^T \overset{p\times p}{P}} \Lambda P^T c > 0 \qquad P \text{ is invertible}$$

$$\underset{1\times p}{y^T} \underset{p\times p}{\Lambda} \underset{p\times 1}{y} > 0 \qquad \text{for all } y \neq 0 \text{ since } P^T c = y$$

$$\boxed{\sum_{i=1}^{p} \lambda_i \underbrace{y_i^2} > 0}$$

This is true when all $\lambda_i$ are positive.

$\Rightarrow$

If $y = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \rightarrow \lambda_1$, same for all the eigenvalues

QED

This also implies

RecEx 1.P4

$$\det(\Sigma) = \prod_{i=1}^{p} \lambda_i > 0$$

$$\mathrm{tr}(\Sigma) = \sum_{i=1}^{p} \lambda_i > 0$$

Homework: why is $\det(\Sigma) = \prod \lambda_i$ and
$\mathrm{tr}(\Sigma) = \sum \lambda_i$ ?

## What about $\Sigma^{-1}$

If $\Sigma$ is SPD then $\Sigma^{-1}$ is also SPD, and the eigenvalues of $\Sigma^{-1}$ are the inverses of the eigenvalues of $\Sigma$.

PROOF: $\Sigma^{-1} = (P \Lambda P^T)^{-1} = (P^T)^{-1} \Lambda^{-1} P^{-1}$

$(P^T)^{-1} = P$ and $P^{-1} = P^T$ for orthogonal matrix

$$\Sigma^{-1} = P \Lambda^{-1} P^T = P \begin{bmatrix} \frac{1}{\lambda_1} & & & 0 \\ & \frac{1}{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\lambda_p} \end{bmatrix} P^T$$

eigenvectors

eigenvalues of $\Sigma^{-1}$

Finally: $\Sigma^{\frac{1}{2}} = P \Lambda^{\frac{1}{2}} P^T$

$$[e_1 \cdots e_p] \quad \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_p} \end{bmatrix}$$

RecEx1.P4: $\Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} = \Sigma$

$$\Sigma^{-\frac{1}{2}} = (\Sigma^{\frac{1}{2}})^{-1} = P \Lambda^{-\frac{1}{2}} P^T$$

$$\begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & & 0 \\ & \frac{1}{\sqrt{\lambda_2}} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sqrt{\lambda_p}} \end{bmatrix}$$

$\Sigma^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} = I$ (show)

The transformation $Y = \Sigma^{-\frac{1}{2}}(X - \mu)$

Mahalanobis transform

Pizzavalg: 6 Tiddlibom
9 Sult

Glutenfri =

Vegeter : 2 + 3

# Principal Component Analysis [vtl 11.1- 11.3]

why is $\Sigma$ important and what can $C\underline{X}$ be?

$\underset{p\times 1}{\underline{X}}$ RV and $E(\underline{X}) = \mu$, $Cov(\underline{X}) = \Sigma$

$\underset{p\times 1}{\underline{X}_1}, \underset{p\times 1}{\underline{X}_2}, \dots, \underset{p\times 1}{\underline{X}_n}$ i.i.d

identically distributed

independent

$\mu = E(\underline{X}_j)$ and $\Sigma = Cov(\underline{X}_j)$ $j = 1, \dots, n.$

Estimators for $\mu$ and $\Sigma$?

$$\underline{\hat{\mu}} = \frac{1}{n} \sum_{j=1}^{n} \underline{X}_j = \bar{\underline{X}}$$ p-dim vector

$$\left[ Cov(\underline{X}) = E\left( (\underline{X} - \mu)(\underline{X} - \mu)^T \right) = \Sigma \right]$$

$$\underset{p\times p}{\underline{S^2}} = \frac{1}{n-1} \sum_{j=1}^{n} \underset{p\times 1}{(\underline{X}_j - \bar{\underline{X}})} \underset{1\times p}{(\underline{X}_j - \bar{\underline{X}})^T}$$ pxp matrix

We will work with scaled variables so $\Sigma = \rho$.

Q : When faced with a large set of correlated variables – is it possible to define a set of linear combinations of the original variables that capture a large part of the variability of the data?

$$\left.\begin{array}{l} \text{– interpretability} \\ \text{– data reduction} \\ \text{– visualization} \end{array}\right\} \text{Aims!}$$

A: Yes. Let $(\lambda_i, e_i)$ be eigenvalue/vector pairs of $\Sigma$ where $\lambda_1 \geq \dots \geq \lambda_p$. Then

construct $Y_i = e_i^T \underline{X}$

$Y_i$: principal components

$e_i$: loadings or rotations

<u>Important property:</u>

$\underset{p\times 1}{\underline{X}}$, $Cov(\underline{X}) = \underset{p\times p}{\Sigma}$ and $(\lambda_i, e_i)$ eigenvalue/vector pairs of $\Sigma$, $i = 1, \dots, p$

Let $Y_i = e_i^T \underline{X}$ and $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} = P^T \underline{X}$

Then $\text{Cov}(Y) = \text{Cov}(P^T \underline{X})$

$$= P^T \,\text{Cov}(\underline{X})\, P = P^T \Sigma P = \underbrace{P^T P}_{I} \Lambda \underbrace{P^T P}_{I}$$

$$= \Lambda$$

so $\text{Var}(e_i^T \underline{X}) = \lambda_i$ and

$\text{Cov}(e_i^T \underline{X}, e_k^T \underline{X}) = 0$

$\Rightarrow$ the PCs $Y_i$ have $\text{Var}(Y_i) = \lambda_i$

and are uncorrelated.

<u>PC scores</u>

$$\left.\begin{aligned} Y_1 &= e_1^T \underline{X} \\ Y_2 &= e_2^T \underline{X} \\ &\vdots \\ Y_p &= e_p^T \underline{X} \end{aligned}\right\} \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} \quad \text{new random vector}$$

I have a data set of $n$ $\underline{X}$'s

Ex: $n = 21$

For each we may calculate $Y$ = PC scores $y_1, \dots, y_p$

Ex: Norway:

data

$x_1 = 9.8$   $e_1^T x \rightarrow y_1 = -0.497$

$x_2 = 0.21$   $e_2^T x \rightarrow y_2 = 0.54$

$x_3 = 0.61$   $y_3 =$

$x_4 = 1.1$   $\vdots$

$x_5 = 6.4$   $y_6 =$

$x_6 = 52$

$$y_1 = x_1 \cdot e_{11} + x_2 e_{12} + \dots + x_6 \cdot e_{16}$$
$$= 9.8 \cdot -0.26 + 0.21 \cdot 0.66 + 0.61 \cdot 0.24 + \dots = -0.497$$

The same for each country

PCA plot

x-axis: PCA1 score
y-axis: PCA2 score $\Big\}$ n points

Biplot: - display graphically the loadings in the PC-space:

1) Coffee arrow: PC1 loading: $-0.26$
  2   $0.67$

Tea arrow   PC1   $0.65$
  PC2   $-0.09$

$\Rightarrow$ sum up next time!