TMA4267 Linear Statistical Models V2017 [L3]

Part 1: Multivariate RVs and normal distribution (L3)
Covariance and positive definiteness [H:2.2,2.3,3.3],
Principal components [H11.1-11.3]
Quiz with Kahoot!

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: January 17, 2017

Previously

- $\boldsymbol{X}_{(p\times 1)}$: random vector, described by
- joint probability distribution function (pdf) $f(\boldsymbol{x})$ and cumulative distribution function (cdf) $F(\boldsymbol{x})$, or, as we will see (in L4), by the (multivariate) moment generating function (MGF) $M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{X}})$.
- Important aspects: moments.
- $\mathrm{E}(\boldsymbol{X})$: mean of a random vector (or matrix) is found as the mean of each element.
- $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}((\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T)$: $p \times p$ variance-covariance matrix, with variances on the diagonal and covariance off-diagonal, real, symmetric.
- Rules for vector of linear combinations $\boldsymbol{CX}$: $\mathrm{E}(\boldsymbol{CX}) = \boldsymbol{C\mu}$ and $\mathrm{Cov}(\boldsymbol{CX}) = \boldsymbol{C\Sigma C}^T$.

Today!

- Requirements and properties of $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{X})$: symmetric, positive definite (SPD), via spectral decomposition (eigenvalues/eigenvectors).
- The square root matrix.
- Linear combinations with maximal variability: principal components are linear combinations made from eigenvectors.
- PCA-plots.
- Kahoot! on what we have worked with so far.
- Next lecture: move on to multivariate normal data!

Drinking habits data set

	Coffee	Tea	Cocoa	Liquer	Wine	Beer
Norway	9.800000	0.21	0.61	1.1	6.4	52.0
Danmark	10.400001	0.39	0.54	1.4	20.7	123.2
Finland	12.450000	0.17	0.03	3.1	5.4	79.0
Iceland	8.270001	0.23	0.00	2.2	5.2	23.7
Sweden	10.710000	0.32	0.16	1.8	12.3	57.4
France	5.490000	0.20	1.18	2.5	73.8	40.5
Ireland	0.550000	3.14	2.76	1.4	3.9	114.0
Italy	4.670000	0.08	0.97	1.0	67.0	22.7
Jugoslavia	3.100000	0.11	0.59	1.6	20.3	46.8
The Netherlands	10.970000	0.82	15.35	2.0	14.7	87.0
Poland	1.400000	0.54	0.56	4.3	7.6	30.8
Portugal	3.080000	0.03	0.02	0.8	51.5	60.7
Soviet Union	0.300000	0.98	0.56	1.9	6.6	19.3
Spain	4.250000	0.03	1.11	2.8	38.3	70.7
Schweitz	9.400000	0.25	3.06	1.9	49.6	69.3
Great Britain	2.060000	2.62	2.78	1.8	11.5	110.5
Chech Repl	2.200000	0.13	1.21	3.3	13.6	132.9
Germany	8.970000	0.22	3.94	2.0	26.0	143.0
Hungary	2.270000	0.07	0.85	4.6	21.5	103.9
Austria	10.220000	0.16	1.80	1.5	34.8	119.5
New Zealand	1.920000	1.46	0.03	1.4	14.6	114.2

The variance-covariance matrix and positive definiteness

$\boldsymbol{X}_{(p\times 1)}$ with symmetric $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$

- Want variance of linear combination to be positive: $\boldsymbol{c}^T\boldsymbol{\Sigma}\boldsymbol{c}$ for all $\boldsymbol{c} \neq \boldsymbol{0}$,
- which means that $\boldsymbol{\Sigma}$ needs to positive definite. Write $\boldsymbol{\Sigma} > 0$.
- This is true if all eigenvalues of $\boldsymbol{\Sigma}$ are positive (eigenvalues of symmetric matrix are real).
- Spectral decompositions (diagonalization): $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$.
- Square root matrix defined from spectral decomposition: $\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^T$.

Drinking habits

	Coffee	Tea	Cocoa	Liquer	Wine	Beer
Norway	9.800000	0.21	0.61	1.1	6.4	52.0
Danmark	10.400001	0.39	0.54	1.4	20.7	123.2
Finland	12.450000	0.17	0.03	3.1	5.4	79.0
Iceland	8.270001	0.23	0.00	2.2	5.2	23.7
Sweden	10.710000	0.32	0.16	1.8	12.3	57.4
France	5.490000	0.20	1.18	2.5	73.8	40.5
Ireland	0.550000	3.14	2.76	1.4	3.9	114.0
Italy	4.670000	0.08	0.97	1.0	67.0	22.7
Jugoslavia	3.100000	0.11	0.59	1.6	20.3	46.8
The Netherlands	10.970000	0.82	15.35	2.0	14.7	87.0
Poland	1.400000	0.54	0.56	4.3	7.6	30.8
Portugal	3.080000	0.03	0.02	0.8	51.5	60.7
Soviet Union	0.300000	0.98	0.56	1.9	6.6	19.3
Spain	4.250000	0.03	1.11	2.8	38.3	70.7
Schweitz	9.400000	0.25	3.06	1.9	49.6	69.3
Great Britain	2.060000	2.62	2.78	1.8	11.5	110.5
Chech Repl	2.200000	0.13	1.21	3.3	13.6	132.9
Germany	8.970000	0.22	3.94	2.0	26.0	143.0
Hungary	2.270000	0.07	0.85	4.6	21.5	103.9
Austria	10.220000	0.16	1.80	1.5	34.8	119.5
New Zealand	1.920000	1.46	0.03	1.4	14.6	114.2

Estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ [H3.3]

$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ i.i.d $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X})$.

$$\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{X}_j$$

$$\boldsymbol{S}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^T$$

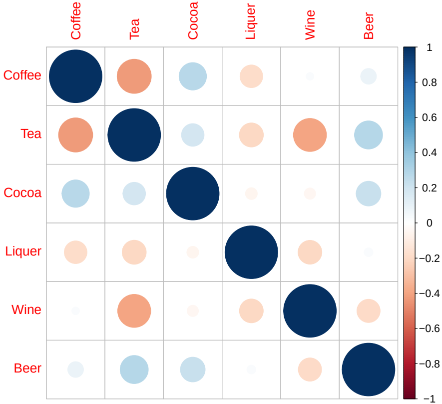
are two commonly used estimators for the mean and covariance matrix.

RecEx1.P7: we may write $\boldsymbol{S}^2$ using a centering matrix.

Drinking habits data set

```
> drinkcov
           Coffee        Tea      Cocoa     Liquer        Wine        Beer
Coffee 16.5471497 -1.4476576  3.6381233 -0.7651859    1.841446   14.386113
Tea    -1.4476576  0.7168891  0.5154248 -0.1786857   -6.939988    9.503595
Cocoa   3.6381233  0.5154248 10.8676967 -0.1700214   -3.191860   28.673374
Liquer -0.7651859 -0.1786857 -0.1700214  1.0222857   -4.307429    1.142071
Wine    1.8414458 -6.9399885 -3.1918596 -4.3074288  432.878505 -159.797675
Beer   14.3861129  9.5035954 28.6733741  1.1420707 -159.797675 1549.731430
```

Correlation plot



prcomp

```
drink <- read.csv("drikke.TXT",sep=",",header=TRUE)
drink <- na.omit(drink) # remove missing data

# now for PCA
pca <- prcomp(drink,scale=TRUE) # scale: variables are scaled, and automatically center=TRUE

> names(pca)
[1] "sdev"     "rotation" "center"   "scale"    "x"

> pca$rotation # the loadings
              PC1         PC2         PC3           PC4         PC5        PC6
Coffee -0.26029733  0.66788815 -0.22475187  0.4132467433  0.07431918  0.5092751
Tea     0.65540048 -0.09539757  0.36756357 -0.0002927055 -0.12503940  0.6407898
Cocoa   0.23510209  0.57754726 -0.06603093 -0.4200858712 -0.61199325 -0.2362164
Liquer  0.02190508 -0.32118904 -0.79997824 -0.3292322714 -0.12307455  0.3644878
Wine   -0.50599685  0.06551597  0.37109534 -0.6765579799  0.15862233  0.3450672
Beer    0.43693234  0.32219426 -0.17985159 -0.2943302832  0.75099779 -0.1493533

> s <- cor(drink) # cor, not cov, since covariates are scaled
> eigen(s) # same as pca$rotations - opposite sign of some vectors
$values
[1] 1.7204307 1.4295795 1.1408597 0.7731249 0.7354586 0.2005467
$vectors
            [,1]        [,2]        [,3]          [,4]        [,5]       [,6]
[1,]  0.26029733  0.66788815 -0.22475187  0.4132467433  0.07431918 -0.5092751
[2,] -0.65540048 -0.09539757  0.36756357 -0.0002927055 -0.12503940 -0.6407898
[3,] -0.23510209  0.57754726 -0.06603093 -0.4200858712 -0.61199325  0.2362164
[4,] -0.02190508 -0.32118904 -0.79997824 -0.3292322714 -0.12307455 -0.3644878
[5,]  0.50599685  0.06551597  0.37109534 -0.6765579799  0.15862233 -0.3450672
[6,] -0.43693234  0.32219426 -0.17985159 -0.2943302832  0.75099779  0.1493533
```

Principal components

- Let $\boldsymbol{\Sigma}$ be the covariance matrix associated with the random vector $\boldsymbol{X}_{p\times 1}$. The covariance matrix has the eigenvalue-vector pairs $(\lambda_j, \boldsymbol{e}_j)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
- The mth principal component is given by

$$Y_m = \boldsymbol{e}_m^T \boldsymbol{X} = e_{m1}X_1 + e_{m2}X_2 + \cdots + e_{mp}X_p$$

- and has

$$\begin{aligned} \mathrm{Var}(Y_m) &= \boldsymbol{e}_m^T \boldsymbol{\Sigma} \boldsymbol{e}_m = \lambda_m, \quad i = 1, 2, ..., p \\ \mathrm{Cov}(Y_i, Y_m) &= \boldsymbol{e}_i^T \boldsymbol{\Sigma} \boldsymbol{e}_m = 0 \quad i \neq m \end{aligned}$$

Principal components: idea

1. We choose principal component 1, $\mathrm{PC}_1 = \boldsymbol{c}_1^T \boldsymbol{X}$, to have maximal variance

$$\max_{\boldsymbol{c}_1 \neq 0, \boldsymbol{c}_1^T \boldsymbol{c}_1 = 1} \mathrm{Var}(\boldsymbol{c}_1^T \boldsymbol{X})$$

2. We choose principal component 2, $\mathrm{PC}_2 = \boldsymbol{c}_2^T \boldsymbol{X}$, to have maximal variance and to be uncorrelated with PC_1.

$$\max_{\boldsymbol{c}_2 \neq 0, \boldsymbol{c}_2^T \boldsymbol{c}_2 = 1} \mathrm{Var}(\boldsymbol{c}_2^T \boldsymbol{X}) \text{ and } \boldsymbol{c}_1^T \boldsymbol{\Sigma} \boldsymbol{c}_2 = 0$$

Principal components: idea

3. We choose principal component 3, $\mathrm{PC}_1 = \boldsymbol{c}_3^T \boldsymbol{X}$, to have maximal variance and be uncorrelated with PC_1 and PC_2.

$$\max_{\boldsymbol{c}_3 \neq 0, \boldsymbol{c}_3^T \boldsymbol{c}_3 = 1} \mathrm{Var}(\boldsymbol{c}_3^T \boldsymbol{X}) \text{ and } \boldsymbol{c}_i^T \boldsymbol{\Sigma} \boldsymbol{c}_3 = 0$$

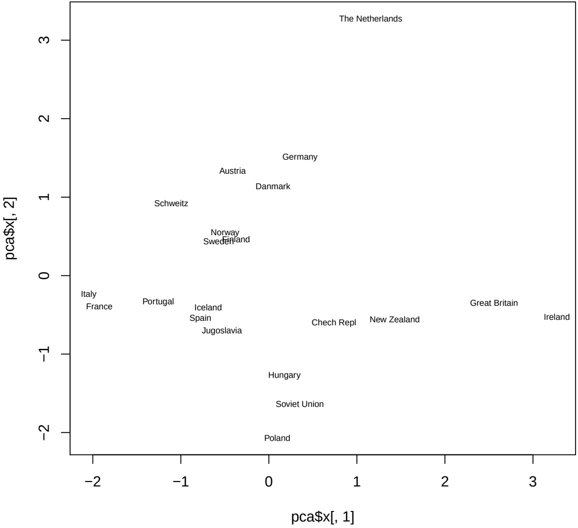
   for $i = 1, 2$.
4. and so on.

It can be shown that choosing $\boldsymbol{c}_i = \boldsymbol{e}_i$ (ith eigenvector of $\boldsymbol{\Sigma}$) fulfills these requirements.

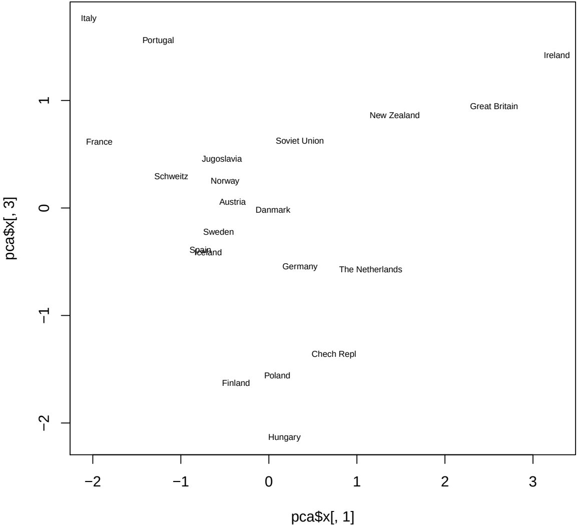
Principal component scores

```
> pca$x
                       PC1        PC2         PC3         PC4          PC5          PC6
Norway         -0.49755266  0.5423321  0.24730564  1.64969605 -0.141194597 -0.259224773
Danmark         0.04736942  1.1407629 -0.01389387  0.62447776  1.286968641  0.032325551
Finland        -0.37212624  0.4628190 -1.62846743  1.17223112  0.284805095  0.685797001
Iceland        -0.68875794 -0.4028320 -0.40970218  1.46441365 -0.741785804  0.092078151
Sweden         -0.57111547  0.4409190 -0.21949104  1.33928137  0.005514957  0.299908368
France         -1.92532968 -0.3909512  0.61668510 -1.42238114 -0.200223837  0.818931395
Ireland         3.27114811 -0.5256052  1.42206819 -0.04493512 -0.014828505  0.477585062
Italy          -2.04540738 -0.2394487  1.76095443 -0.63624103 -0.367324127  0.054512802
Jugoslavia     -0.53258860 -0.7075896  0.45055250  0.39569929 -0.299213104 -0.741778021
The Netherlands 1.15707155  3.2747502 -0.57171870 -0.93460573 -2.324855354 -0.377956367
Poland          0.09667998 -2.0690624 -1.55793948 -0.11990952 -1.118921154  0.196370232
Portugal       -1.25572594 -0.3355743  1.55429933 -0.39165369  0.418471226 -0.587655889
Soviet Union    0.35234451 -1.6341096  0.62745644  0.66818676 -1.138865302 -0.446485485
Spain          -0.77751355 -0.5475718 -0.39572230 -0.70847071  0.084179130 -0.055157063
Schweitz       -1.10755090  0.9248648  0.29621598 -0.49777226 -0.047226891  0.484657020
Great Britain   2.55842016 -0.3473694  0.94755863 -0.24512125  0.028322241  0.555177625
Chech Repl      0.74010278 -0.6054459 -1.36137719 -0.75417409  0.950808503 -0.708681733
Germany         0.35362098  1.5089269 -0.54743939 -0.46975244  0.999859480 -0.289909548
Hungary         0.17766893 -1.2757040 -2.13925810 -1.16454698  0.376532314 -0.009844322
Austria        -0.41110347  1.3402182  0.06024028 -0.01769055  1.108477110  0.029376612
New Zealand     1.43034542 -0.5543289  0.86167320  0.09326849  0.850499977 -0.250026615
```

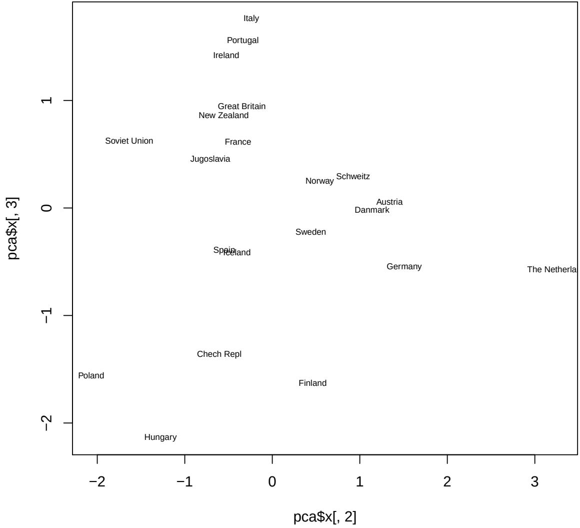
Scores PCA1 vs PCA2



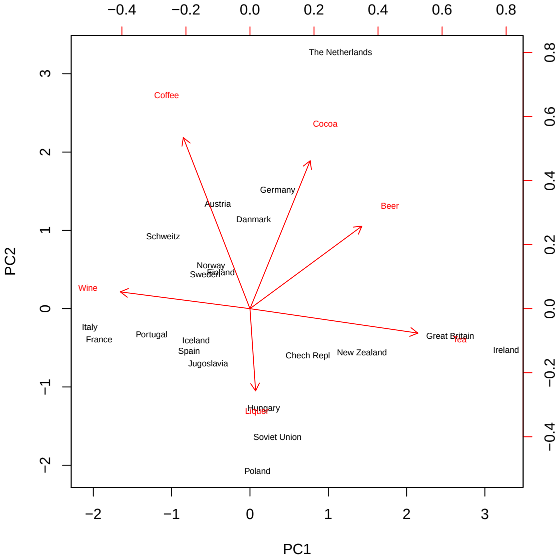
Scores PCA1 vs PCA3



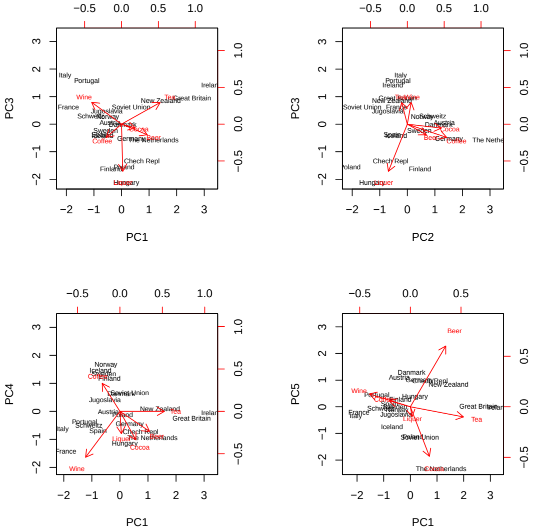
Scores PCA2 vs PCA3



Scores PCA1 vs PCA2 with biplot



Biplots


PC1
PC2
PC3
PC4
PC5
Wine
Tea
Beer
Cocoa
Coffee
Liquer
Italy
Portugal
France
Soviet Union
New Zealand
Great Britain
Jugoslavia
Norway
Sweden
Finland
Poland
Hungary
Chech Repl
Germany
The Netherlands
Austria
Danmark
Spain
Iceland

Proportion of total population variance

- Total population variance:
  $\sum_{j=1}^{p} \mathrm{Var}(X_j) = \mathrm{tr}\boldsymbol{\Sigma} = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \mathrm{Var}(Z_j)$.
- Proportion of total population variance explained by PC m:

$$\frac{\lambda_m}{\sum_{j=1}^{p} \lambda_j}$$

- Proportion of total population variance explained by the first m PCs:

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

How many PCs are needed?

Dependent on:

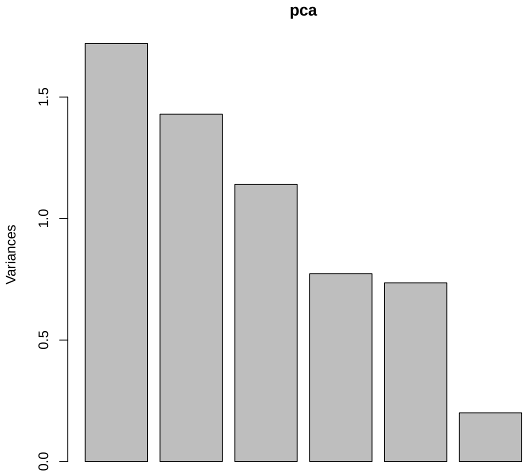
- The proportion of the total sample variance that we would like to explain. 80%? More?
- Look at the eigenvalues; small eigenvalues may be an evidence of collinearity problems.

Importance of components

```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6
Standard deviation     1.3117 1.1957 1.0681 0.8793 0.8576 0.44782
Proportion of Variance 0.2867 0.2383 0.1901 0.1288 0.1226 0.03342
Cumulative Proportion  0.2867 0.5250 0.7151 0.8440 0.9666 1.00000

> eigen(s)
$values
[1] 1.7204307 1.4295795 1.1408597 0.7731249 0.7354586 0.2005467
> sqrt(eigen(s)$values)
[1] 1.3116519 1.1956502 1.0681103 0.8792752 0.8575888 0.4478244
```

Screeplot



PC from standardized variables

- $\boldsymbol{X}$ can be standardized to have mean $\mathbf{0}$ and unit variances.

$$\boldsymbol{X}^* = \boldsymbol{V}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$$

- Principal components made from standardized variables will be based on the eigenvalues and eigenvectors of the correlation matrix $\boldsymbol{\rho} = \boldsymbol{V}^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{V}^{-\frac{1}{2}}$.
- Achilles heel: Since $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$ do not have the same eigenvectors/eigenvalues, the principal components made from $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$ will not be the same.
- Unless we have a good reason to compare the variances for the different X_js we should make PCs from the standardized variables.
- For standardized variables $\sum_{j=1}^{p} \mathrm{Var}(X_j^*) = p$, and
- Proportion of total population variance explained by PC m: $\frac{\lambda_m}{p}$.

Principal components from singular value decomposition

The singular value decomposistion of a (data) matrix $\boldsymbol{X}_{n\times p}$ is given by:

$$\boldsymbol{X}_{n\times p} = \boldsymbol{U}_{n\times p}\boldsymbol{D}_{p\times p}\boldsymbol{V}^T_{p\times p}$$

where

- the columns of $\boldsymbol{U}$ are the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$
- $\boldsymbol{D}$ is a diagonal matrix with singular values on the diagonal, i.e. the square root of the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$ (they have the same eigenvalues).
- the columns of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$.

And, the principal components (scores) of the data are defined as the columns of

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}$$

PCR: summary

- PCA finds linear combinations $\boldsymbol{Y}$ that “best” represents the $\boldsymbol{X}$.
- The PCs are found in an unsupervised way. The "truth" is not known.
- A plot of PC1 vs PC2 is often used to see if there is separation (subgroups in the data).
- The principal component loadings are often given interpretation (overall consumption,
- PCA can be combined with linear regression.

Quiz

with Kahoot! at kahoot.it. - based on what we have gone through so far!