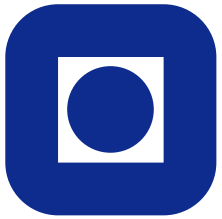


NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (25)

Dimension reduction [ISLR6.3]
Principal component analysis [ISLR10.2]

Mette Langaas

To be lectured: March 31, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

Dimension reduction methods

1. Transform the original predictors, $\boldsymbol{X}$. Make $M \leq p$ new predictors, $\boldsymbol{Z}$.
$$\boldsymbol{Z}_{M\times 1} = \boldsymbol{\phi}_{M\times p}\boldsymbol{X}_{p\times 1}$$
2. Use the transformed predictors to fit a MLR using least squares (LS).
$$\boldsymbol{Y}_{n\times 1} = \theta_0 + \boldsymbol{Z}_{n\times M}\boldsymbol{\theta}_M + \boldsymbol{\varepsilon}_{n\times 1}$$

Comparing this model with a MLR in the original variables $\boldsymbol{Y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ we see that the MLR in the transformed variables also is a MLR in the original variables, but with the following constraint:

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}, \text{ that is } \boldsymbol{\beta}_{p\times 1} = \boldsymbol{\phi}^T_{M\times p}\boldsymbol{\theta}_{M\times 1}$$

Principal components

Q: Is it possible to make a new variable (vector), $\boldsymbol{Z}_{M\times 1} = \phi_{M\times p}\boldsymbol{X}_{p\times 1}$, with $M \leq p$, where
- almost all of the spread in $\boldsymbol{X}$ (variance-covariance) is present in $\boldsymbol{Z}$ and
- $\boldsymbol{Z}$ has an easier/better interpretation than $\boldsymbol{X}$ itself?

Aims: data reduction and interpretation.

PC: Principal components, $\boldsymbol{Z}_{M\times 1} = \phi_{M\times p}\boldsymbol{X}_{p\times 1}$, are *uncorrelated* linear combinations of the original variables $\boldsymbol{X}$, whose variances are as large as possible.

Principal component analysis [ISLR 10.2 + some maths]

Notation:

Let $\underset{p\times 1}{X}$ be a random vector with $\underset{p\times 1}{E(X)} = \mu$ and

$\text{Cov}(X) = \underset{p\times p}{\Sigma}$.

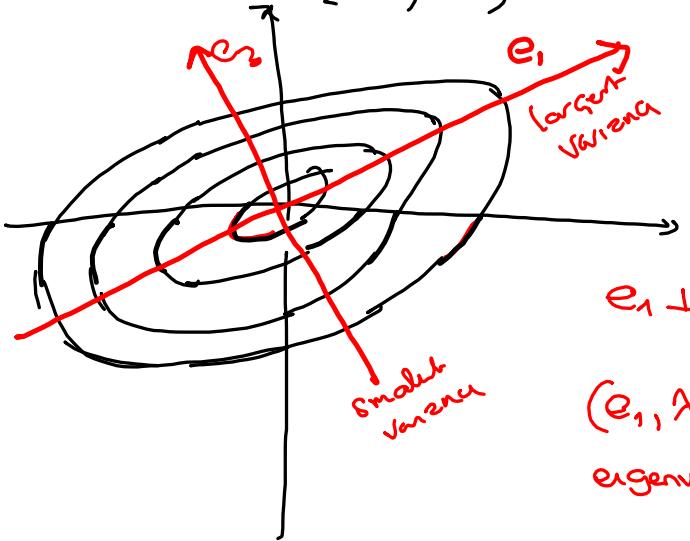
We will consider linear combinations

$$\underset{m\times 1}{Z} = \underset{m\times p}{\Phi}\ \underset{p\times 1}{X} \qquad \underset{m\times p}{\Phi} = \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \\ \vdots \\ \Phi_m^T \end{bmatrix} \leftarrow p\text{-vector}$$

$$E(Z) = \Phi \cdot \mu$$

$$\underset{m\times m}{\text{Cov}(Z)} = \underset{m\times p}{\Phi}\ \underset{p\times p}{\Sigma}\ \underset{p\times m}{\Phi^T} = \begin{bmatrix} \Phi_1^T\Sigma\Phi_1 & & & \\ \Phi_1^T\Sigma\Phi_2 & \ddots & & \\ & & & \Phi_m^T\Sigma\Phi_m \end{bmatrix}$$

What if $X \sim N_2(0, \Sigma)$?



Where is the direction ϕ_1 in which $\mathrm{Var}(\phi_1^T X)$ is the largest?

$e_1 \perp e_2$

$(e_1, \lambda_1), (e_2, \lambda_2)$

eigenvector–value pairs of Σ.

Principal components: idea

1. We choose principal component 1, $\text{PC}_1 = \phi_1^T \boldsymbol{X}$, to have maximal variance

$$\max_{\phi_1 \neq 0, \phi_1^T \phi_1 = 1} \text{Var}(\phi_1^T \boldsymbol{X})$$

2. We choose principal component 2, $\text{PC}_2 = \phi_2^T \boldsymbol{X}$, to have maximal variance and to be uncorrelated with PC_1.

$$\max_{\phi_2 \neq 0, \phi_2^T \phi_2 = 1} \text{Var}(\phi_2^T \boldsymbol{X}) \text{ and } \phi_1^T \boldsymbol{\Sigma} \phi_2 = 0$$

Principal components: idea

3. We choose principal component 3, $\text{PC}_1 = \phi_3^T \boldsymbol{X}$, to have maximal variance and be uncorrelated with PC_1 and PC_2.

$$\max_{\phi_3 \neq 0, \phi_3^T \phi_3 = 1} \text{Var}(\phi_3^T \boldsymbol{X}) \text{ and } \phi_i^T \boldsymbol{\Sigma} \phi_3 = 0$$

   for $i = 1, 2$.

4. and so on.

How can we choose the PCs?
Hint: spectral decomposition of $\boldsymbol{\Sigma}$.

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

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

— and has

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

Mathematical derivation of the PCs

1) Let $\phi_1 = \frac{a_1}{\sqrt{a_1^T a_1}}$, so that $\phi_1^T \phi_1 = \frac{a_1^T a_1}{a_1^T a_1} = 1$

Then $M_1 = \max_{\substack{\phi_1 \neq 0 \\ \phi_1^T \phi_1 = 1}} \underbrace{Var(\phi_1^T X)}_{\phi_1^T \Sigma \phi_1}$

$$= \max_{a_1 \neq 0} \frac{a_1^T \Sigma a_1}{a_1^T a_1}$$

2) $\Sigma = P \Lambda P^T$ where $\underset{p \times p}{P} = [e_1\ e_2\ \dots\ e_p]$

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots\ \lambda_p \end{bmatrix} = diag(\lambda_j)$$

where $\Sigma e_j = \lambda_j e_j$ $j = 1, \dots, p$ and

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ ← assume semi positive definite Σ

3)

$$M_1 = \max_{a_1 \neq 0} \frac{\overbrace{a_1^T}^{w^T} P \Lambda P^T \overbrace{a_1}^{w}}{a_1^T \underbrace{PP^T}_{I} a_1} = \max_{w \neq 0} \frac{w^T \Lambda w}{w^T w}$$

$$= \max_{w \neq 0} \frac{\sum_{j=1}^{p} \lambda_j w_j^2}{\sum_{j=1}^{p} w_j^2}$$

$1 \times p \quad p \times p \quad p \times 1$

1×1

$$\leq \lambda_1 \frac{\sum w_j^2}{\sum w_j^2} = \lambda_1$$

$\lambda_1 \geq \lambda_2 \geq \ldots$

The largest M_1 possible is λ_1. This is achieved when $a_1 = e_1$ and $\phi_1 = e_1$ (since $e_1^T e_1 = 1$)

Check: $Z_1 = \phi_1^T X = e_1^T X$

$$Var(Z_1) = e_1^T \Sigma e_1 = \underbrace{e_1^T}_{e_1^T [e_1\ e_2 \ \ldots\ e_p]} P \Lambda P^T e_1$$

$$\underbrace{e_1^T e_1}_{1} \quad \underbrace{e_1^T e_2}_{0} \quad \cdots \quad \underbrace{e_1^T e_p}_{0}$$

$$= [1\ 0 \ \cdots\ 0] \Lambda \begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} = \lambda_1$$

PC_1: $Z_1 = e_1^T X$ with $M_1 = Var(Z_1) = \lambda_1$

4) Further: $\phi_2^T \phi_2 = 1$ and $\phi_1^T \Sigma \phi_2 = 0$ ($\phi_1^T = e_1^T$)

$$\phi_2 = \frac{a_2}{\sqrt{a_2^T a_2}}$$

$$M_2 = \max_{\substack{\phi_2 \neq 0 \\ \phi_2^T\phi_2 = 1 \\ e_1^T \Sigma \phi_2 = 0}} \underbrace{Var(\phi_2^T X)}_{\phi_2^T \Sigma \phi_2} = \max_{\substack{a_2 \neq 0 \\ e_1^T \Sigma a_2 = 0}} \frac{a_2^T \Sigma a_2}{a_2^T a_2}$$

($\Sigma = P \Lambda P^T$)

$$\max_{a_2 \neq 0} \frac{\overbrace{a_2^T P}^{w^T} \Lambda \overbrace{P^T a_2}^{w}}{a_2^T P P^T a_2} = \max_{w \neq 0} \frac{w^T \Lambda w}{w^T w} = \max_{w \neq 0} \frac{\sum_{j=1}^{p} \lambda_j w_j^2}{\sum_{j=1}^{p} w_j^2}$$

Where $w^T = a_2^T P$

5) The additional constraint

$$e_1^T \Sigma a_2 = \underbrace{e_1^T P}_{[1\ 0 \cdots 0]} \Lambda \underbrace{P^T a_2}_{w} = [\lambda_1\ 0\ 0 \cdots 0] \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} = w_1 \cdot \lambda_1$$

$$e_1^T \Sigma a_2 = 0 \iff w_1 \lambda_1 = 0$$

$w_1 = 0$ ← first element of $P^T a_2$

6) $M_2 = \max_{\substack{a_2 \neq 0 \\ e_1^T \Sigma a_2 = 0}} \frac{\sum_{j=1}^{p} \lambda_j w_j^2}{\sum_{j=1}^{p} w_j^2} \overset{w_1 = 0}{=\!=} \max_{a_2 \neq 0} \frac{\sum_{j=2}^{p} \lambda_j w_j^2}{\sum_{j=2}^{p} w_j^2}$

$$\leq \lambda_2 \frac{\sum_{j=2}^{p} w_j^2}{\sum_{j=2}^{p} w_j^2} = \lambda_2$$

PC_2 must be $Z_2 = e_2^T X$ and

$$Var(Z_2) = Var(e_2^T X) = e_2^T \underbrace{\Sigma e_2}_{\lambda_2 e_2} = \lambda_2 e_2^T e_2 = \lambda_2$$

7) Repeat for $i = 3, \ldots, p$

$\Rightarrow$ RESULT:

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
Belgium	7.870000	0.17	5.54	NA	NA	NA
France	5.490000	0.20	1.18	2.5	73.8	40.5
Greece	1.420000	0.05	0.35	NA	30.8	39.1
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
Australia	2.520000	1.16	NA	1.3	18.3	111.0
New Zealand	1.920000	1.46	0.03	1.4	14.6	114.2

Example: Drinking habits

$$\underline{X} = \begin{bmatrix} \text{Coffee} \\ \text{Tea} \\ \text{Cocoa} \\ \text{Liquer} \\ \text{Wine} \\ \text{Beer} \end{bmatrix}$$ 6 variables

$n = 21$ countries

$\rightarrow$ $e_1 \dots e_6$: rotations or loading

PC1: $e_1^T X \leftarrow$ scores

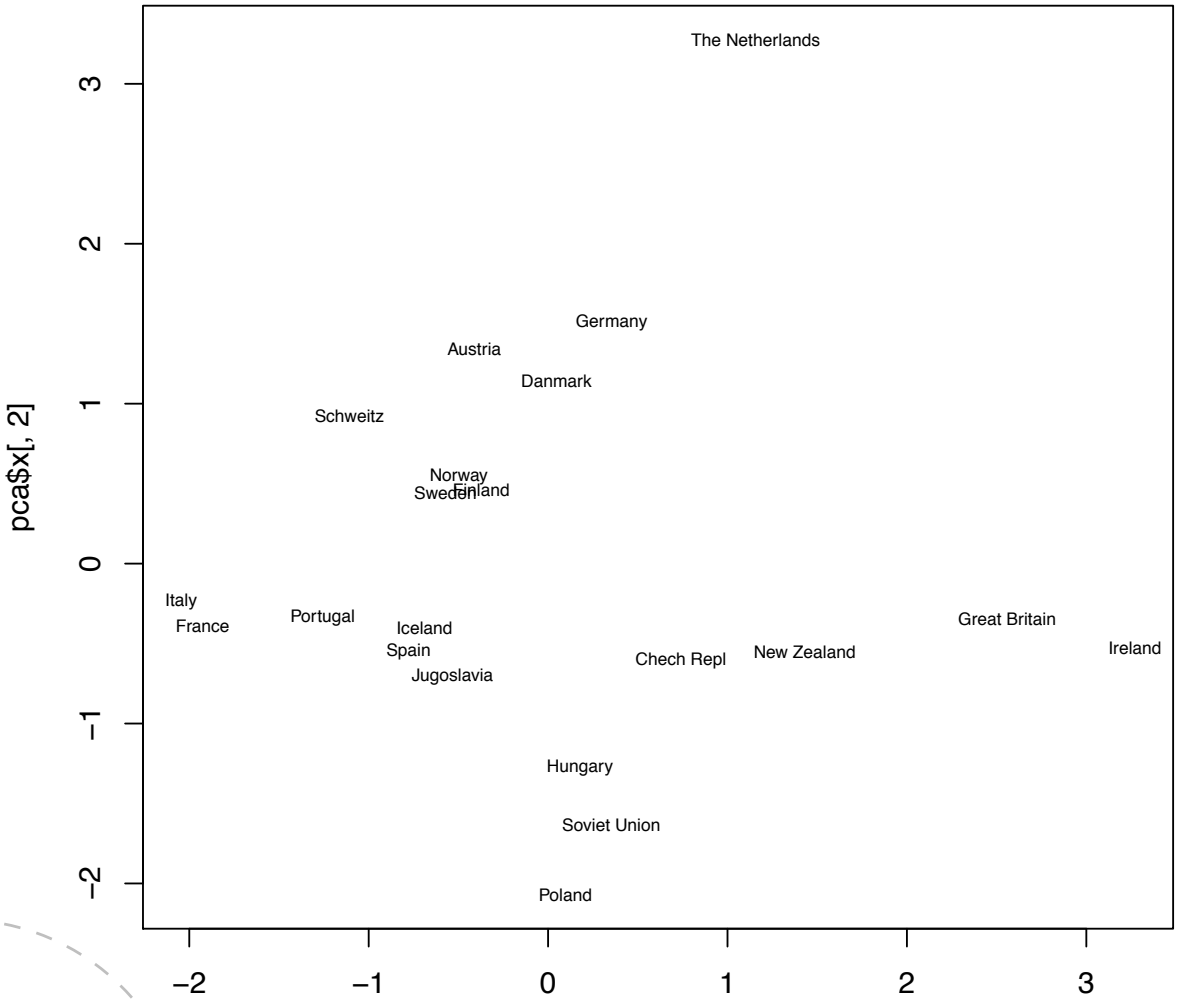
coffee beer

$[-0.26, 0.65, \dots, 0.43]$

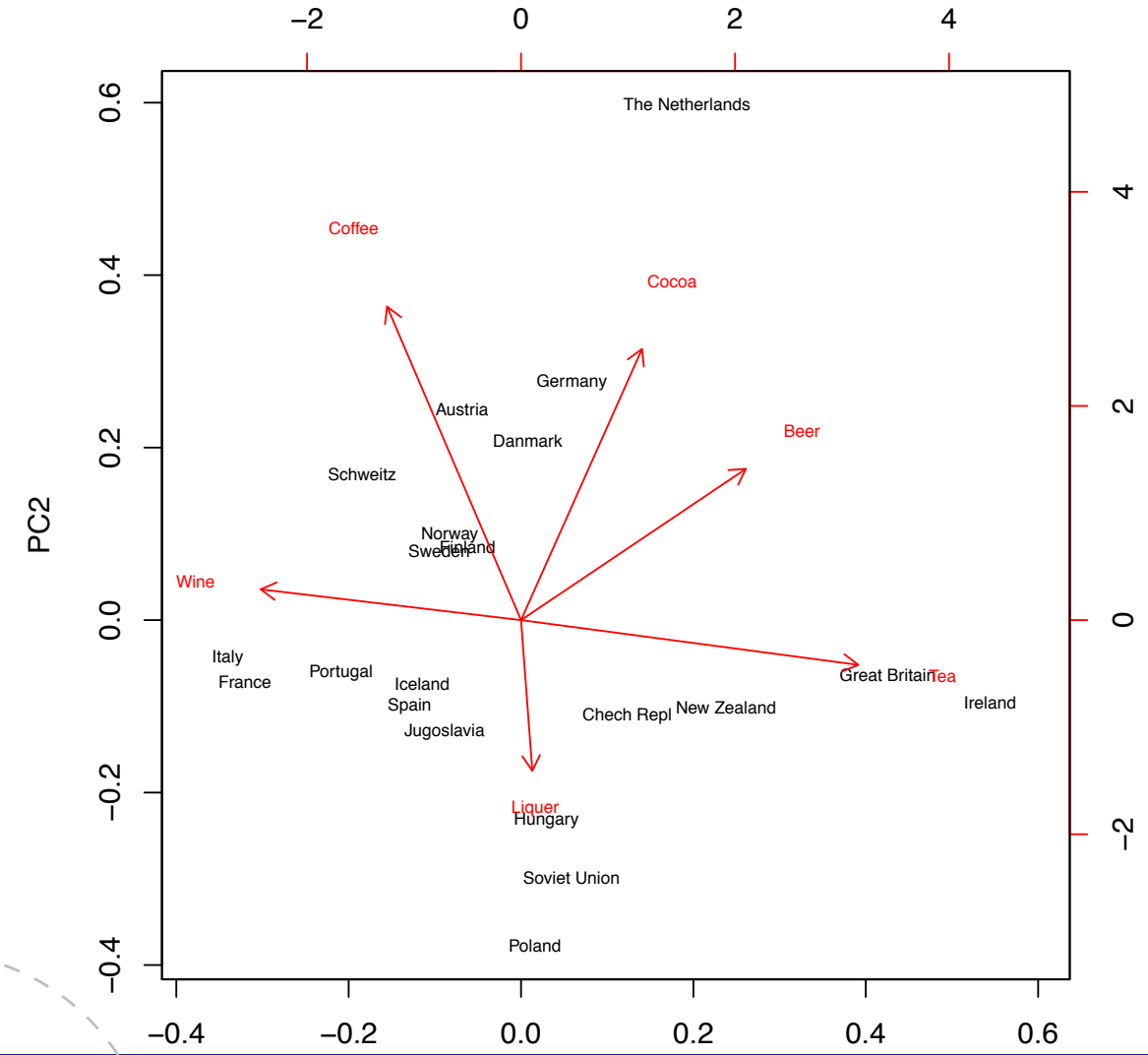
prcomp

```
> pca=prcomp(newds,scale=TRUE)

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

> pca$rotation
               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
```

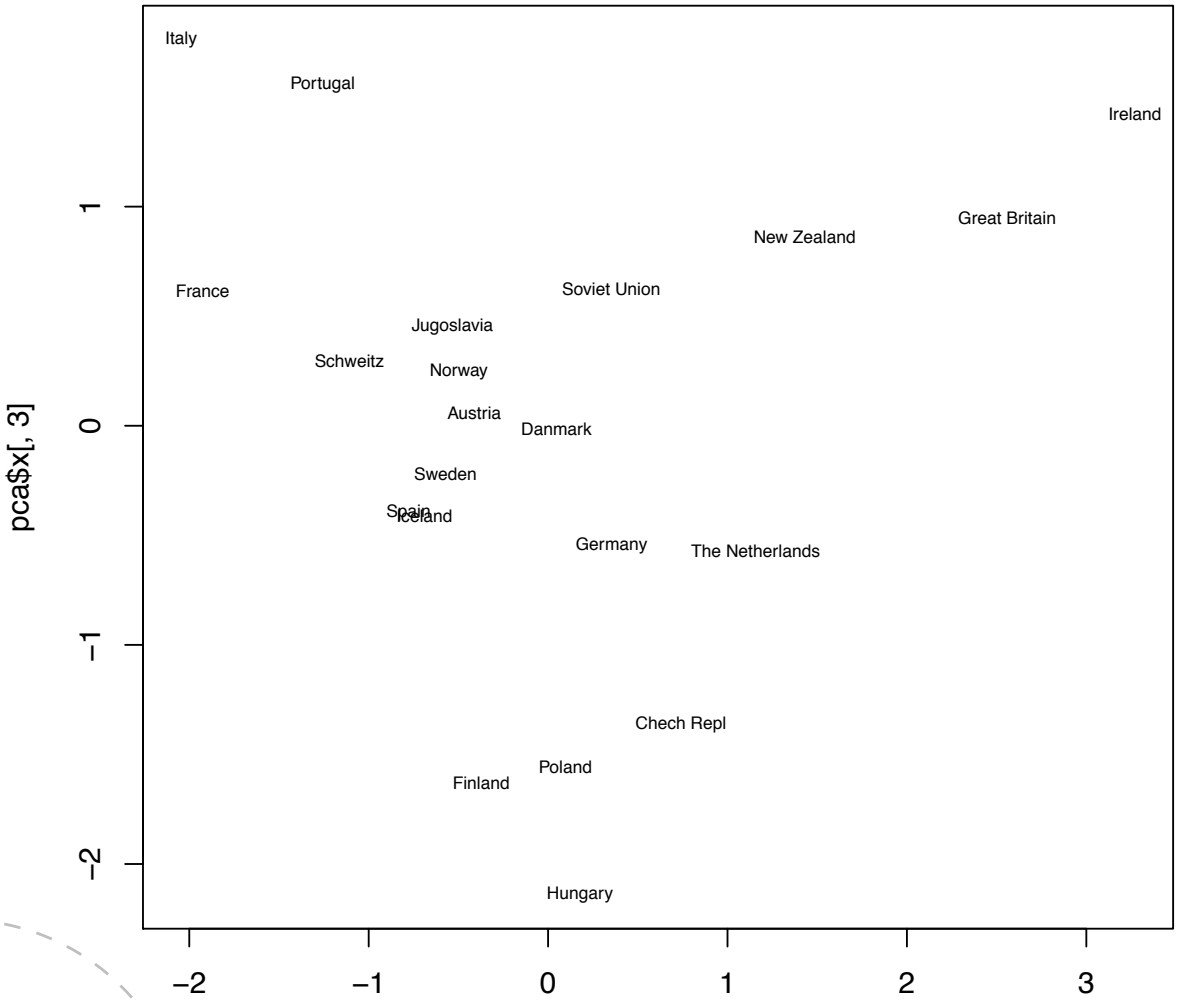
Scores PCA1 vs PCA2



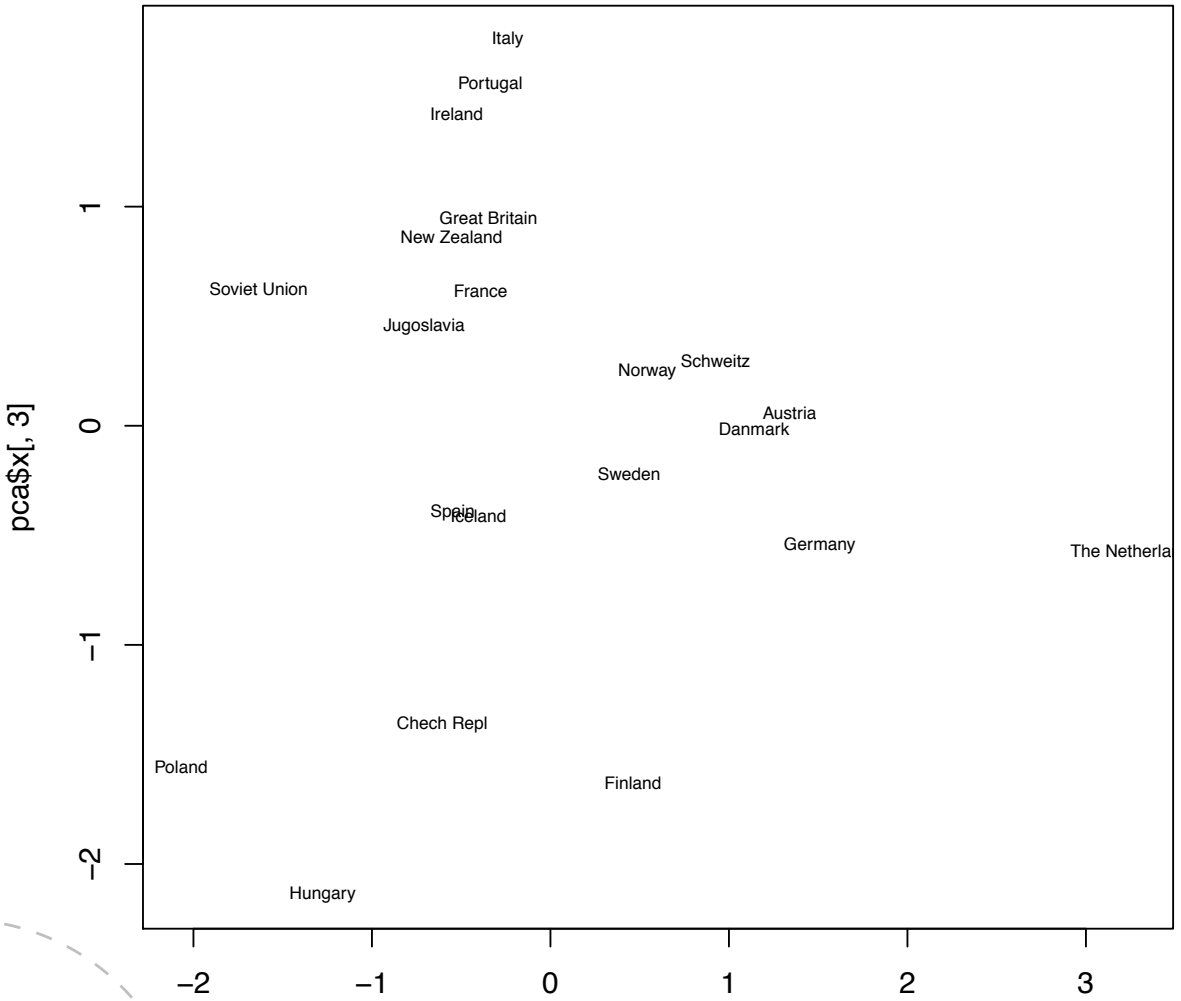
Scores PCA1 vs PCA2 with biplot



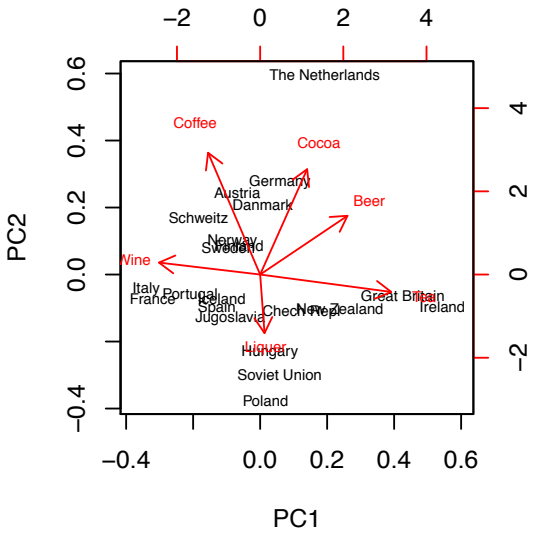
Scores PCA1 vs PCA3

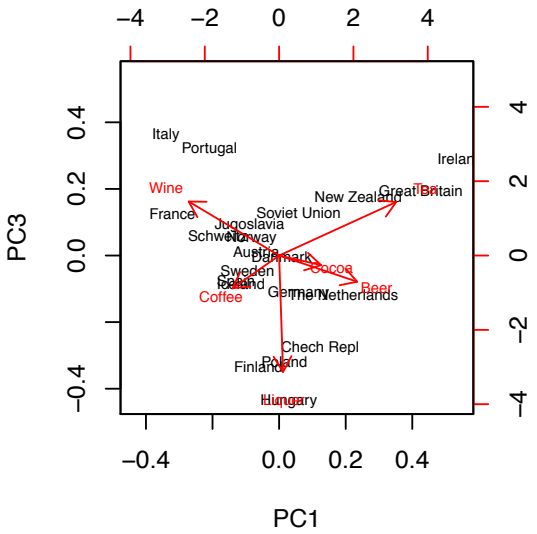


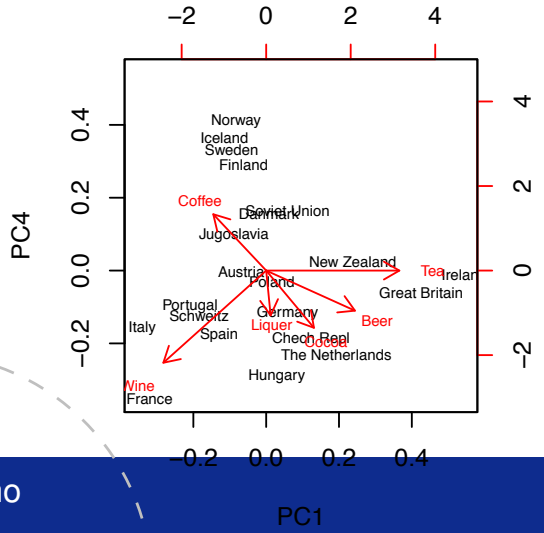
Scores PCA2 vs PCA3

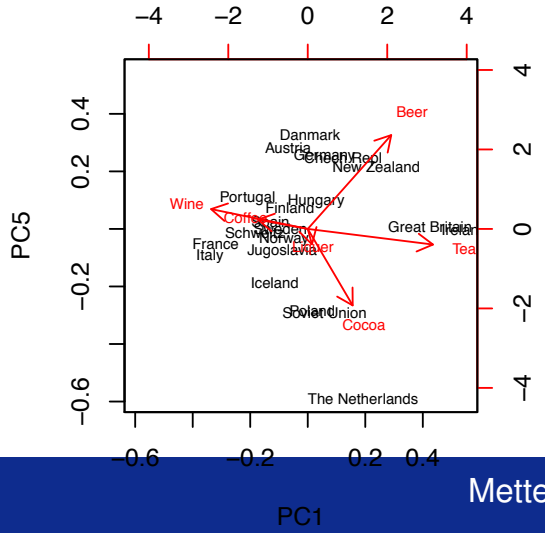


Biplots


PC2
PC1
The Netherlands
Coffee
Cocoa
Germany
Austria
Danmark
Beer
Schweitz
Wine
Great Britain
Ireland
Spain
Jugoslavia
Hungary
Soviet Union
Poland


PC3
PC1
Italy
Portugal
Wine
France
Soviet Union
Jugoslavia
Cocoa
Beer
Coffee
Germany
Chech Repl
Finland


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


PC5
PC1
Beer
Danmark
Austria
New Zealand
Wine
Portugal
Hungary
Finland
Great Britain
France
Italy
Jugoslavia
Tea
Iceland
Soviet Union
Cocoa
The Netherlands

prcomp

```
> pca$sdev
[1] 1.3116519 1.1956502 1.0681103 0.8792752 0.8575888 0.4478244
> 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
```

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 $\boldsymbol{\rho} = \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{V}^{-\frac{1}{2}}$.

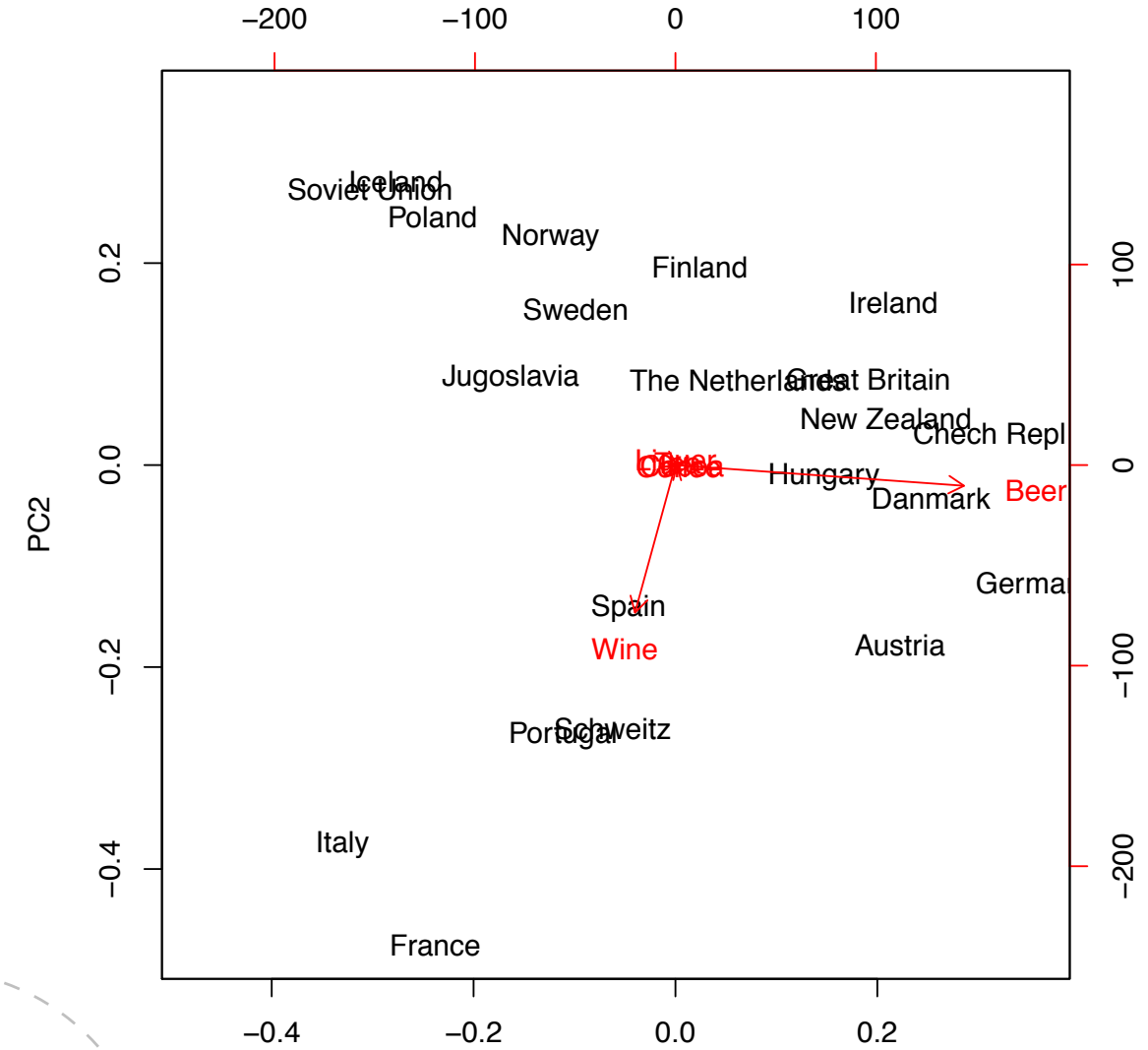
— Achilles heel: Since $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$ do not have the same eigenvectors/eigenvalues, the principal components made from $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$ will not be the same.

— 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: $\frac{\lambda_m}{p}$.

Not standardize



Choosing M

1) Choose M so that 80%, 90%, .. of the total variance of the data is explained.

$$\sum_{j=1}^{p} \text{var}(x_j) = \sum_{j=1}^{p} \text{Var}(z_j) = \sum_{j=1}^{p} \lambda_j$$

Drinking habits : 4 PC explain 84.4%

2) Scree plot : look for elbow.

Drinking habit : 4 PC

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:

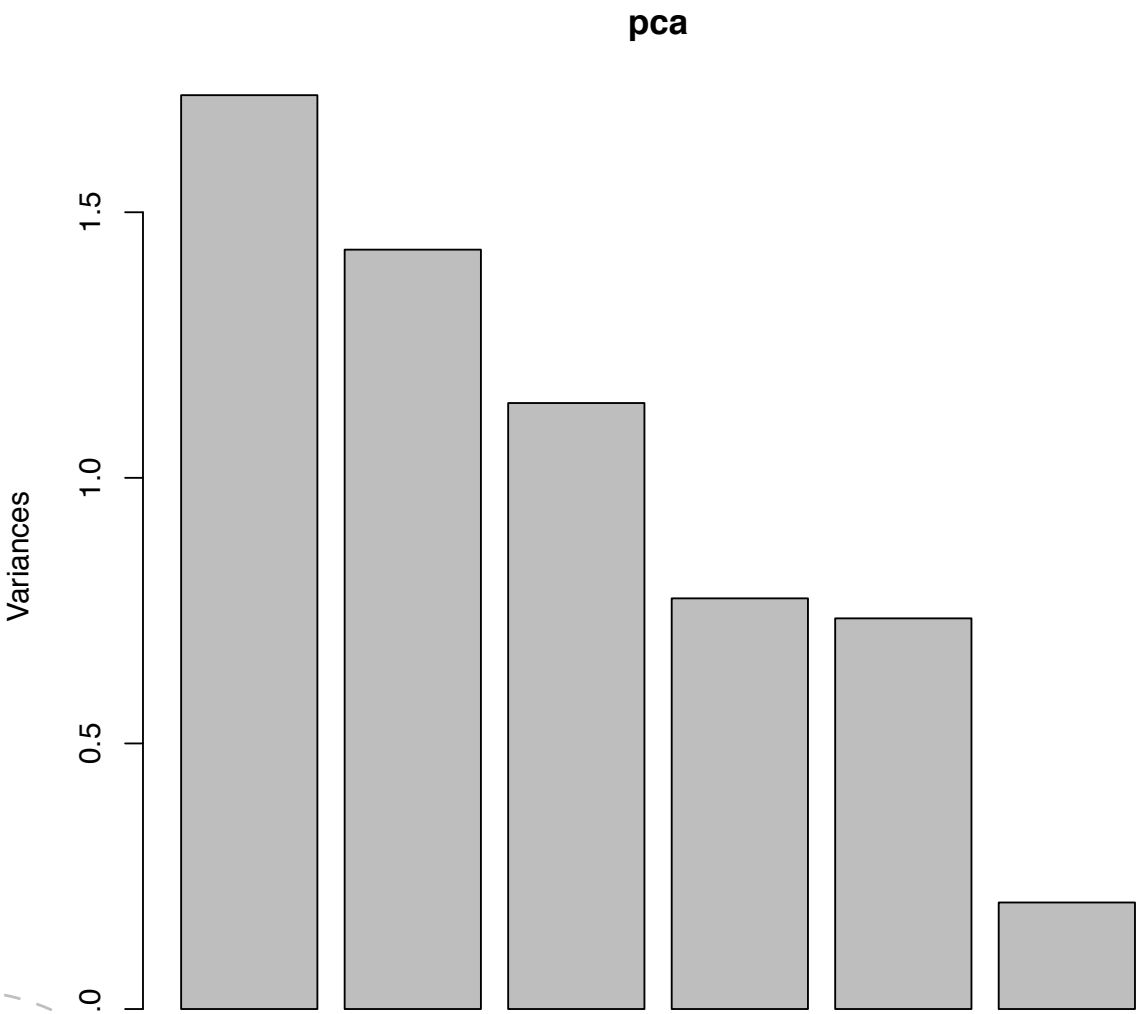
$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

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.

Screeplot



Importance of X_j on the mth PC

- $Z_m = \boldsymbol{e}_m^T \boldsymbol{X} = \sum_{j=1}^{p} e_{mj} X_j$, thus e_{mj} will be related to the importance of X_j for Z_m.
- Covariance: $\mathrm{Cov}(Z_m, X_j) = \lambda_m e_{mj}$.
- Correlation: $\rho(Z_m, X_j) = \frac{\sqrt{\lambda_m}}{\sqrt{\sigma_{jj}}} e_{mj}$.
- e_{mj} should be regarded relative to e.g. e_{kj}, since the relative (and not the absolute) importance of a variable on the PC will be most informative.

PCR = principal component regression

1) Find PC's : $Z = \phi X$

2) Fit MLR : $Y = \theta_0 + Z\theta + \varepsilon$

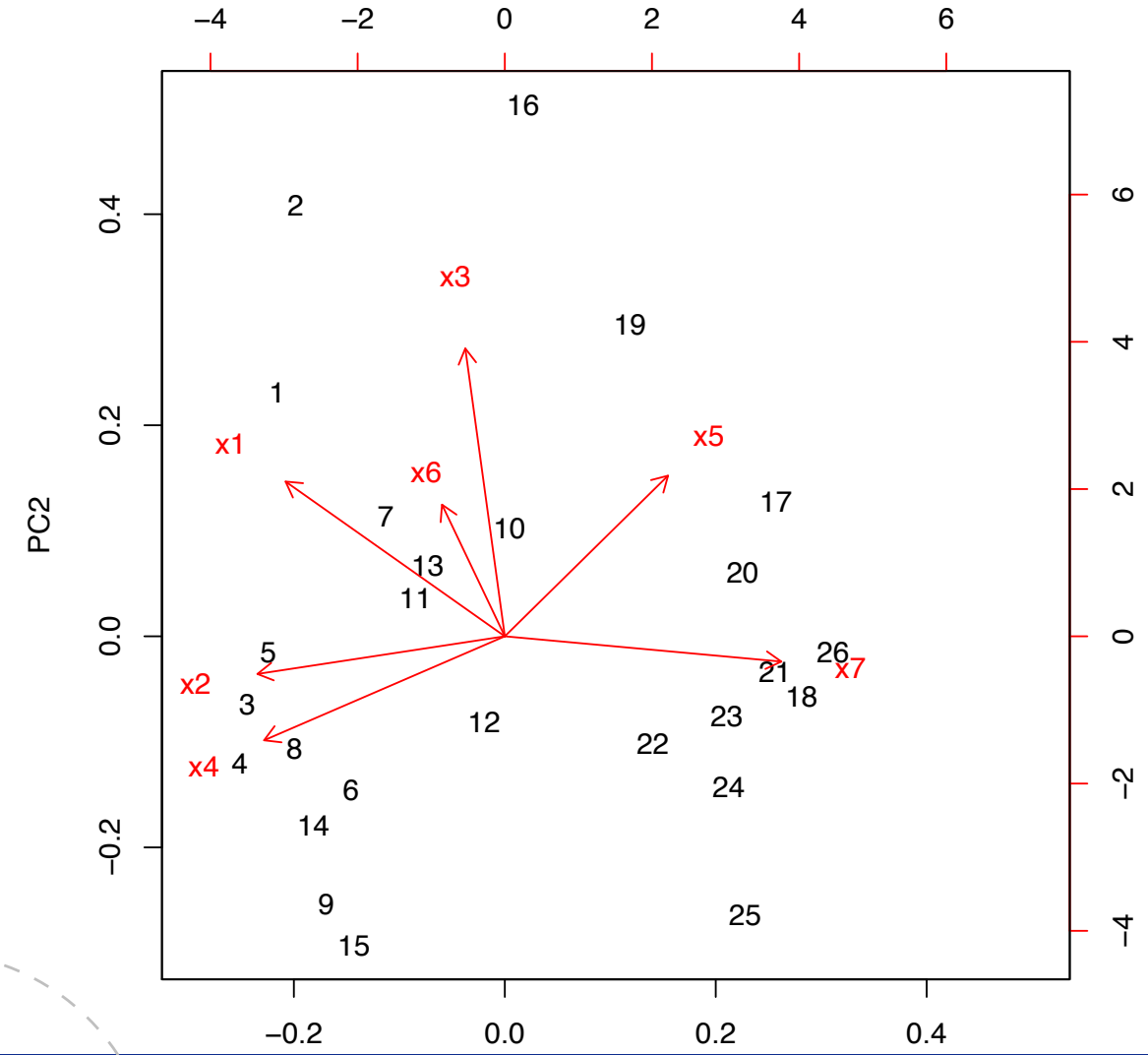
Acid rain → choose 3PC

see slide to compare $\hat{\beta}_{ORG}$ and $\hat{\phi}^T\hat{\theta} = \hat{\beta}_{PC}$

prcomp acid rain

```
> ds=read.table("http://www.math.ntnu.no/~mettela/TMA4267/
Data/acidrain.txt",header=TRUE)
> res <- prcomp(ds[,-1],scale=TRUE)
> summary(res)
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6    PC7
Standard deviation     1.7497 1.3436 1.0993 0.68695 0.4636 0.40972 0.2646
Proportion of Variance 0.4374 0.2579 0.1726 0.06741 0.0307 0.02398 0.0100
Cumulative Proportion  0.4374 0.6953 0.8679 0.93531 0.9660 0.99000 1.0000
> res$rotation
           PC1         PC2         PC3         PC4         PC5         PC6
x1 -0.41745808  0.38347726 -0.31995472  0.07423069 -0.15826636  0.56360503
x2 -0.47105443 -0.09326022  0.27410932 -0.52032503  0.59996587  0.21657756
x3 -0.07524827  0.71340688 -0.03065928  0.25149764  0.15344298 -0.01435770
x4 -0.45833730 -0.25767224 -0.30813170 -0.27382885 -0.57359528 -0.06648117
x5  0.31079500  0.39811482 -0.37506082 -0.70067633  0.01087656 -0.29511209
x6 -0.11955401  0.32631845  0.75718800 -0.19310266 -0.47900914 -0.12188733
x7  0.52651154 -0.06264353  0.11226592 -0.23933743 -0.18129636  0.72723447
```

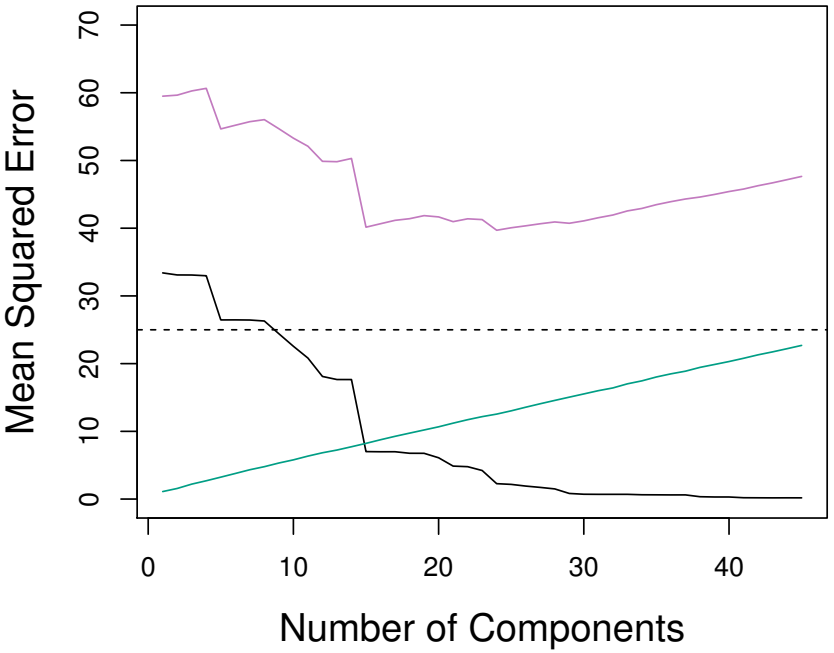
Acid rain biplot

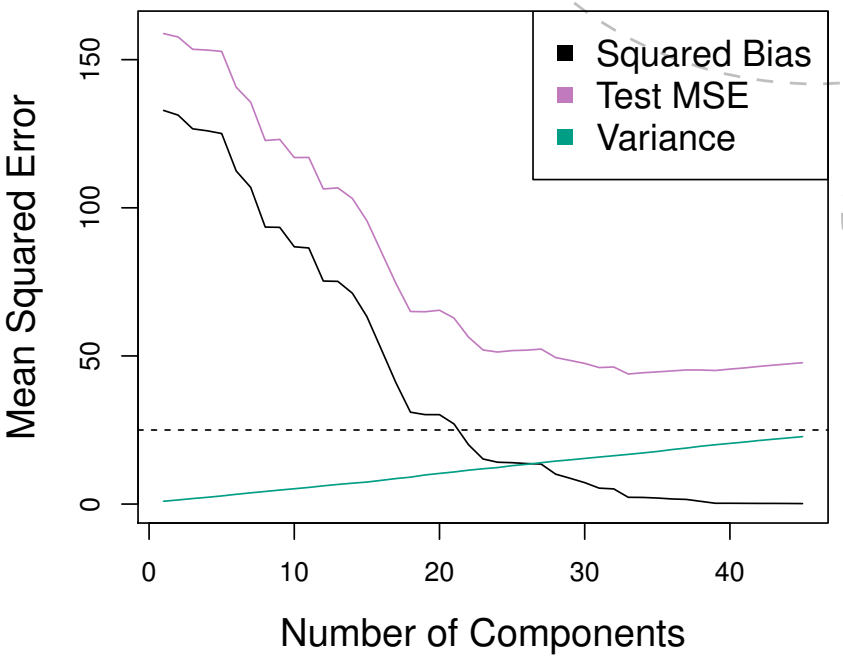


PCR vs full MLR (standardized)

```
> fit <- lm(y~z,data=ds)
> betas <- res$rotation[,1:3]%*%matrix(fit$coeff[2:4],ncol=1)
> x <- apply(ds[,-1],2,scale)
> y <- ds[,1]
> fullfit <- lm(y~x)
> cbind(fullfit$coeff[-1],betas)
             [,1]         [,2]
xx1 -0.292943103 -0.02722807
xx2 -0.058456343 -0.05173126
xx3  0.390470155  0.09491575
xx4 -0.011818319 -0.12845986
xx5 -0.050943297  0.06710810
xx6 -0.001844708  0.10303149
xx7  0.044092942  0.07113932
```

Does PCR work?





Right: simulated data A and left: B.

PCR: summary

- PCA finds linear combinations $\boldsymbol{Z}$ that "best" represents the predictors $\boldsymbol{X}$.
- The PCs are found in an unsupervised way - without the use of the response Y.
- Potential problem: there is no guarantee that the direction that best explain the predictors also will be the best directions to use for predicting the response.
- Often this is no problem.
- Other methods, like partial least squares (PLS), chooses $\boldsymbol{Z}$ with the aid of the response.