Principle Component Analysis

December 4, 2023

Outline

Principal Component Analysis, PCA

Principal Component Analysis, PCA

- Principal Component Analysis

Principal Component Analysis, PCA

- Principal Component Analysis
- A “dimension reduction” technique.

Principal Component Analysis, PCA

- Principal Component Analysis
- A “dimension reduction” technique.
- Find (linear) covariation in multidimensional data.

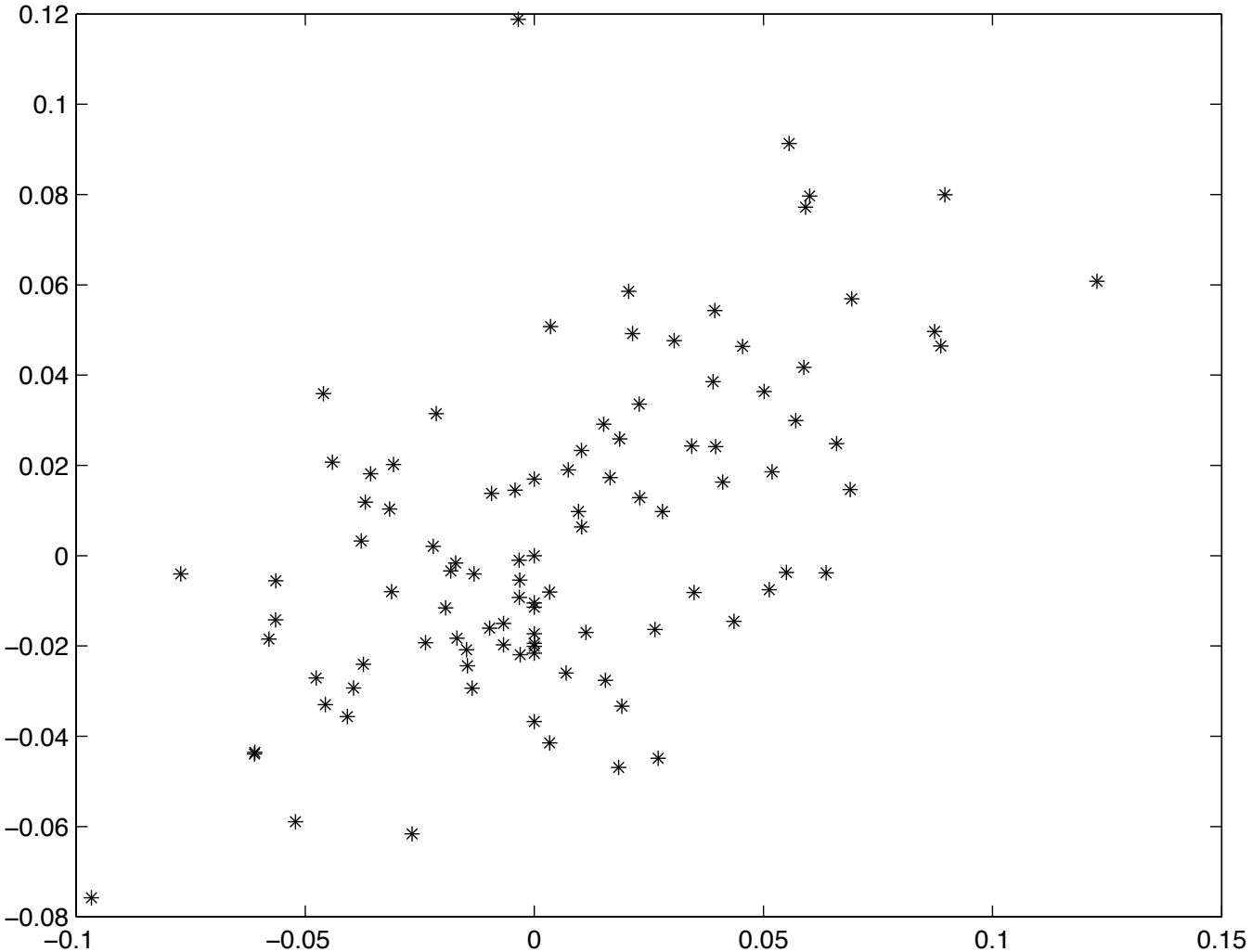
Principal Component Analysis, PCA

- Principal Component Analysis
- A “dimension reduction” technique.
- Find (linear) covariation in multidimensional data.
- Describe this covariation and its strength.

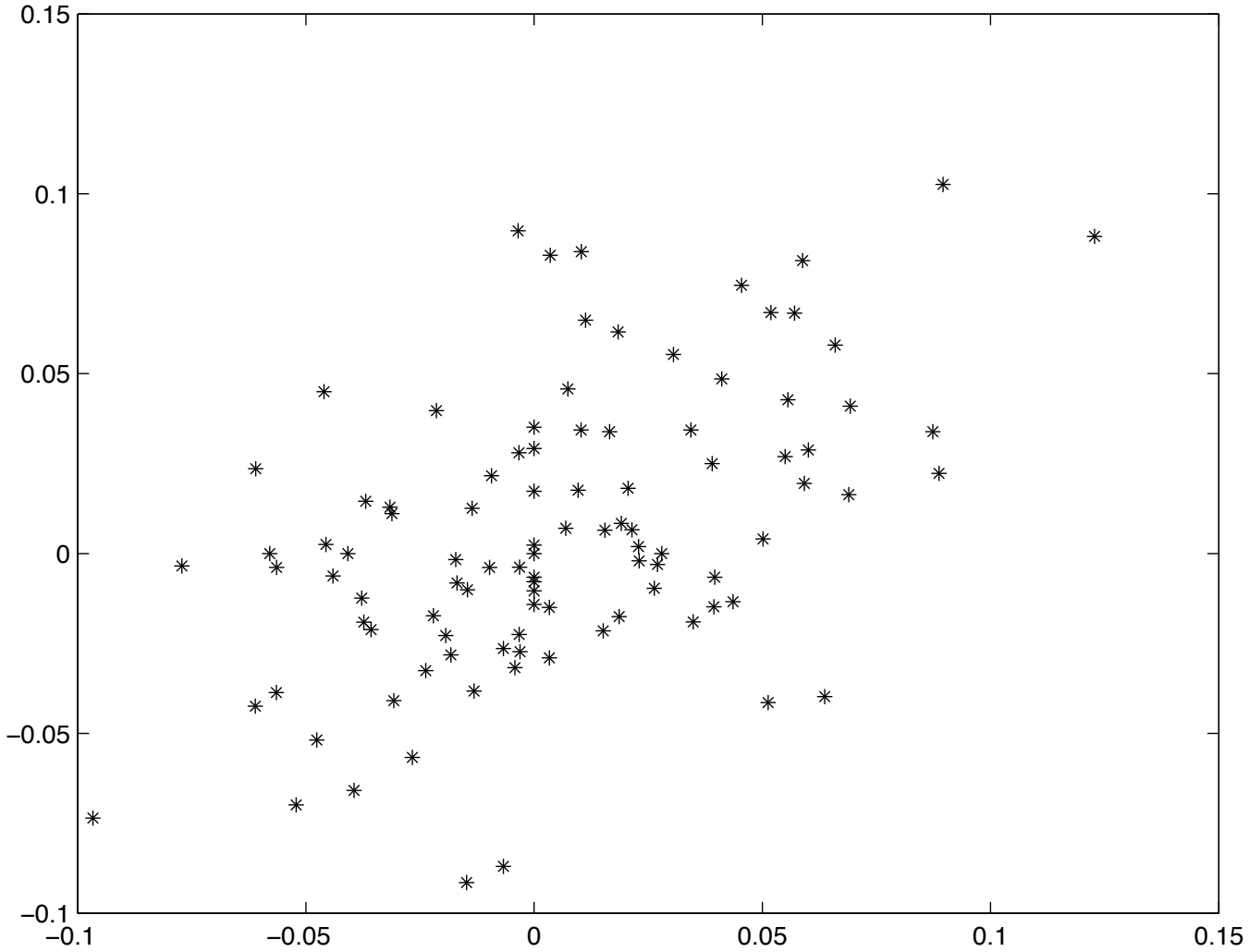
NY Stock Exchange – weakly returns

Week	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	0	0	0	0.0395	0
2	0.0270	-0.0449	-0.0030	-0.0145	0.0435
3	0.1228	0.0608	0.0881	0.0862	0.0781
4	0.0570	0.0299	0.0668	0.0135	0.0195
.	.	.	.	.	
98	0.0103	0.0233	0.0343	-0.0182	-0.0047
99	-0.0441	0.0207	-0.0062	-0.0185	0.0047
100	0.0390	0.0385	0.0250	-0.0283	0.0327

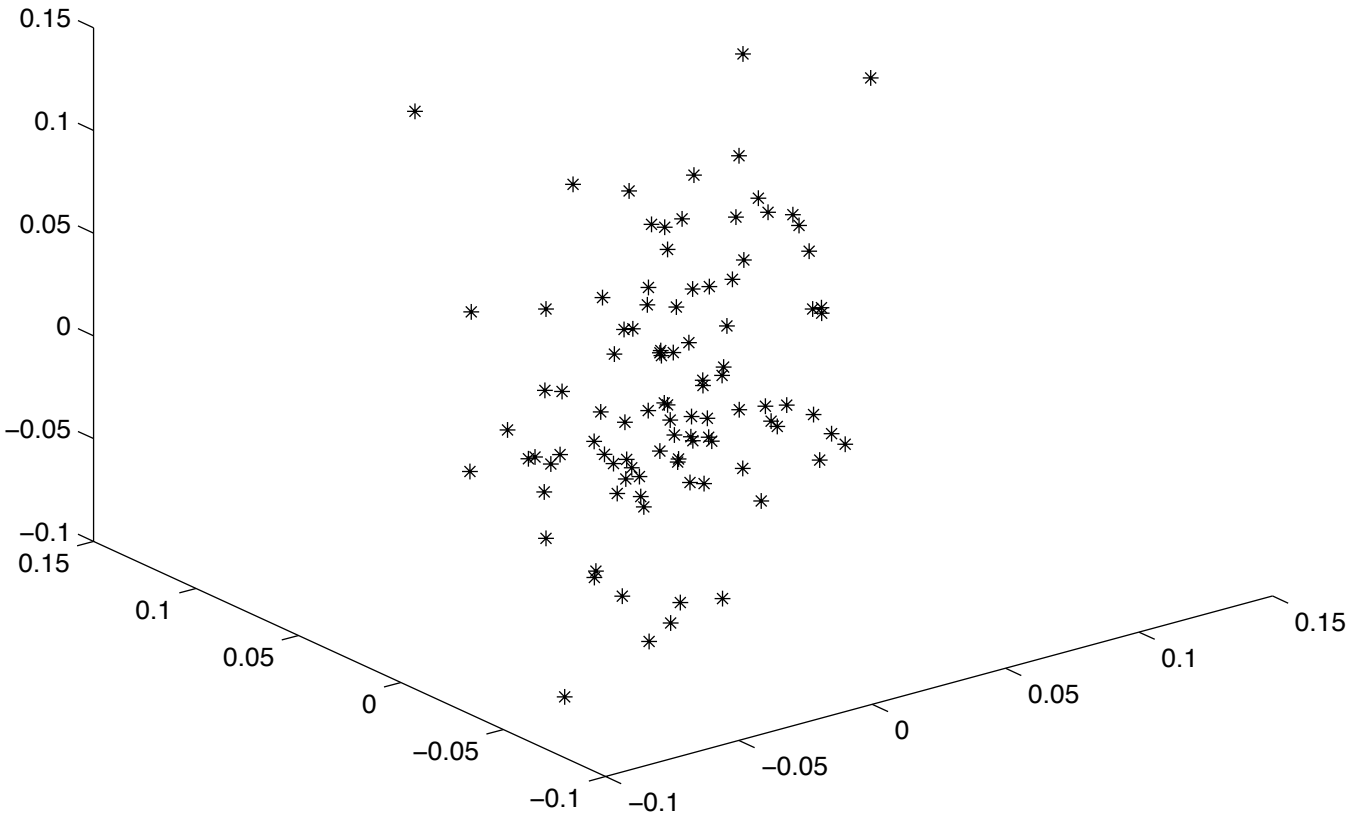
Allied Chemical vs Du Pont



Allied Chemical vs Union Carbide



Allied Chemical, Du Pont and Union Carbide



Linear combinations

Linear combinations

- A random vector $\mathbf{X}$ and a vector $\mathbf{a}$ (each of dimension p) The variance of

$$Y = \mathbf{a}'\mathbf{X}$$

is $\mathbf{a}'\mathbf{\Sigma}\mathbf{a}$, where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$.

Linear combinations

- A random vector $\mathbf{X}$ and a vector $\mathbf{a}$ (each of dimension p) The variance of

$$Y = \mathbf{a}'\mathbf{X}$$

  is $\mathbf{a}'\mathbf{\Sigma}\mathbf{a}$, where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$.
- Important question: Which linear combination represents most variation of the data?

Linear combinations

- A random vector $\mathbf{X}$ and a vector $\mathbf{a}$ (each of dimension p) The variance of

$$Y = \mathbf{a}'\mathbf{X}$$

  is $\mathbf{a}'\mathbf{\Sigma}\mathbf{a}$, where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$.
- Important question: Which linear combination represents most variation of the data?
- **In which direction (represented by a) the variability of the observations is largest?**

Linear combinations

- A random vector $\mathbf{X}$ and a vector $\mathbf{a}$ (each of dimension p) The variance of
$$Y = \mathbf{a}'\mathbf{X}$$
is $\mathbf{a}'\mathbf{\Sigma}\mathbf{a}$, where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$.
- Important question: Which linear combination represents most variation of the data?
- **In which direction (represented by a) the variability of the observations is largest?**
- $\mathbf{a}$ should be normalized i.e. it has to be of the length one.

Outline

First principal component

First principal component

- The linear combination $\mathbf{a}'\mathbf{X}$ with maximal variance under the restriction that $\mathbf{a}'\mathbf{a} = 1$.

First principal component

- The linear combination $\mathbf{a}'\mathbf{X}$ with maximal variance under the restriction that $\mathbf{a}'\mathbf{a} = 1$.
- By the Cauchy-Schwarz inequality

$$\max_{\|\mathbf{a}\|=1} \mathbf{a}'\mathbf{\Sigma}\mathbf{a}$$

is equal to the largest eigenvalue λ_1 of $\mathbf{\Sigma}$ and is obtained for $\mathbf{a} = \mathbf{e}_1$, the eigenvector corresponding to the largest eigenvalue (and standardised to $\mathbf{e}_1'\mathbf{e}_1 = 1$).

First principal component

- The linear combination $\mathbf{a}'\mathbf{X}$ with maximal variance under the restriction that $\mathbf{a}'\mathbf{a} = 1$.
- By the Cauchy-Schwarz inequality

$$\max_{\|\mathbf{a}\|=1} \mathbf{a}'\mathbf{\Sigma}\mathbf{a}$$

  is equal to the largest eigenvalue λ_1 of $\mathbf{\Sigma}$ and is obtained for $\mathbf{a} = \mathbf{e}_1$, the eigenvector corresponding to the largest eigenvalue (and standardised to $\mathbf{e}_1'\mathbf{e}_1 = 1$).
- The Cauchy-Schwarz inequality: $|\mathbf{x}'\mathbf{y}| \leq \|x\|\|y\|$

First principal component

- The linear combination $\mathbf{a}'\mathbf{X}$ with maximal variance under the restriction that $\mathbf{a}'\mathbf{a} = 1$.
- By the Cauchy-Schwarz inequality

$$\max_{\|\mathbf{a}\|=1} \mathbf{a}'\mathbf{\Sigma}\mathbf{a}$$

  is equal to the largest eigenvalue λ_1 of $\mathbf{\Sigma}$ and is obtained for $\mathbf{a} = \mathbf{e}_1$, the eigenvector corresponding to the largest eigenvalue (and standardised to $\mathbf{e}_1'\mathbf{e}_1 = 1$).
- The Cauchy-Schwarz inequality: $|\mathbf{x}'\mathbf{y}| \leq \|x\|\|y\|$
- It simply says that $|\cos\alpha| \leq 1$.

A bit of mathematics

A bit of mathematics

- If $\|\mathbf{a}\| = 1$, then by the Cauchy-Schwartz inequality $0 \leq \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \leq \|\mathbf{\Sigma}\mathbf{a}\|$.

A bit of mathematics

- If $\|\mathbf{a}\| = 1$, then by the Cauchy-Schwartz inequality $0 \leq \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \leq \|\mathbf{\Sigma}\mathbf{a}\|$.
- For which $\mathbf{a}$ the length of $\mathbf{\Sigma}\mathbf{a}$ is the longest

A bit of mathematics

- If $\|\mathbf{a}\| = 1$, then by the Cauchy-Schwartz inequality $0 \leq \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \leq \|\mathbf{\Sigma}\mathbf{a}\|$.
- For which $\mathbf{a}$ the length of $\mathbf{\Sigma}\mathbf{a}$ is the longest
- First note that if λ_1 is the largest (and positive) eigenvalue and $\mathbf{e}_1$ is the corresponding eigenvector then $\|\mathbf{\Sigma}\mathbf{e}_1\| = \lambda_1$.

A bit of mathematics

- If $\|\mathbf{a}\| = 1$, then by the Cauchy-Schwartz inequality $0 \leq \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \leq \|\mathbf{\Sigma}\mathbf{a}\|$.
- For which $\mathbf{a}$ the length of $\mathbf{\Sigma}\mathbf{a}$ is the longest
- First note that if λ_1 is the largest (and positive) eigenvalue and $\mathbf{e}_1$ is the corresponding eigenvector then $\|\mathbf{\Sigma}\mathbf{e}_1\| = \lambda_1$.
- Let $\mathbf{b} = \mathbf{P}'\mathbf{a}$. Then $\|\mathbf{b}\| = 1$.

A bit of mathematics

- If $\|\mathbf{a}\| = 1$, then by the Cauchy-Schwartz inequality $0 \leq \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \leq \|\mathbf{\Sigma}\mathbf{a}\|$.
- For which $\mathbf{a}$ the length of $\mathbf{\Sigma}\mathbf{a}$ is the longest
- First note that if λ_1 is the largest (and positive) eigenvalue and $\mathbf{e}_1$ is the corresponding eigenvector then $\|\mathbf{\Sigma}\mathbf{e}_1\| = \lambda_1$.
- Let $\mathbf{b} = \mathbf{P}'\mathbf{a}$. Then $\|\mathbf{b}\| = 1$. Why?

A bit of mathematics

- If $\|\mathbf{a}\| = 1$, then by the Cauchy-Schwartz inequality $0 \leq \mathbf{a}'\mathbf{\Sigma}\mathbf{a} \leq \|\mathbf{\Sigma}\mathbf{a}\|$.
- For which $\mathbf{a}$ the length of $\mathbf{\Sigma}\mathbf{a}$ is the longest
- First note that if λ_1 is the largest (and positive) eigenvalue and $\mathbf{e}_1$ is the corresponding eigenvector then $\|\mathbf{\Sigma}\mathbf{e}_1\| = \lambda_1$.
- Let $\mathbf{b} = \mathbf{P}'\mathbf{a}$. Then $\|\mathbf{b}\| = 1$. Why? Then

$$\|\mathbf{\Sigma}\mathbf{a}\|^2 = \mathbf{a}'\mathbf{P}\mathbf{\Lambda}\mathbf{P}'\mathbf{P}\mathbf{\Lambda}\mathbf{P}'\mathbf{a} = \mathbf{b}'\mathbf{\Lambda}^2\mathbf{b}$$

- But

$$\mathbf{b}'\mathbf{\Lambda}^2\mathbf{b} \leq \|\mathbf{b}\|^2\lambda_1^2 = \lambda_1^2.$$

Second principal component

Second principal component

- Find another linear combination which is linearly independent of first and maximizes variance

Second principal component

- Find another linear combination which is linearly independent of first and maximizes variance
- The linear combination $\mathbf{a}_2'\mathbf{X}$ with maximal variance under the restriction that $\mathbf{a}_2'\mathbf{a}_2 = 1$ and

$$\mathbb{C}ov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_2 = 0$$

Second principal component

- Find another linear combination which is linearly independent of first and maximizes variance
- The linear combination $\mathbf{a}_2'\mathbf{X}$ with maximal variance under the restriction that $\mathbf{a}_2'\mathbf{a}_2 = 1$ and

$$\mathbb{C}ov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_2 = 0$$

- By the same arguments it follows from Cauchy-Schwarz inequality it follows that $\mathbf{a}_2 = \mathbf{e}_2$, the second normalized eigenvector of $\mathbf{\Sigma}$.

The j :th principal component

The j :th principal component

- The linear combination $\mathbf{a}_j'\mathbf{X}$ with maximal variace under the restictions $\mathbf{a}_j'\mathbf{a}_j = 1$ and

$$\mathbb{C}ov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \mathbb{C}ov(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \ldots$$

$$= \mathbb{C}ov(\mathbf{a}_{j-1}'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = 0$$

The j :th principal component

- The linear combination $\mathbf{a}_j'\mathbf{X}$ with maximal variace under the restictions $\mathbf{a}_j'\mathbf{a}_j = 1$ and

$$\mathbb{C}ov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \mathbb{C}ov(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \ldots$$

$$= \mathbb{C}ov(\mathbf{a}_{j-1}'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = 0$$

- From Cauchy-Schwarz inequality follows that $\mathbf{a}_j = \mathbf{e}_j$, the j:th normalized eigenvector of $\mathbf{\Sigma}$.

The j :th principal component

- The linear combination $\mathbf{a}_j'\mathbf{X}$ with maximal variace under the restictions $\mathbf{a}_j'\mathbf{a}_j = 1$ and

$$\mathbb{C}ov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \mathbb{C}ov(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \ldots$$

$$= \mathbb{C}ov(\mathbf{a}_{j-1}'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = 0$$

- From Cauchy-Schwarz inequality follows that $\mathbf{a}_j = \mathbf{e}_j$, the j:th normalized eigenvector of $\mathbf{\Sigma}$.
- **The vectors that maximize the variance under the above iterative procedure are the eigenvectors of $\mathbf{\Sigma}$.**

Summary of definitions

Summary of definitions

- Principal Components: Find $\mathbf{e}_1, \ldots, \mathbf{e}_p$ which are orthonormal eigenvectors of $\mathbf{\Sigma}$. Let λ_i, $i = 1, \ldots, p$ be the corresponding eigenvalues.

Summary of definitions

- Principal Components: Find $\mathbf{e}_1, \ldots, \mathbf{e}_p$ which are orthonormal eigenvectors of $\mathbf{\Sigma}$. Let λ_i, $i = 1, \ldots, p$ be the corresponding eigenvalues.
- We call $\mathbf{e}_1'\mathbf{X}, \ldots, \mathbf{e}_p'\mathbf{X}$ the principal components of $\mathbf{X}$.

Summary of definitions

- Principal Components: Find $\mathbf{e}_1, \ldots, \mathbf{e}_p$ which are orthonormal eigenvectors of $\mathbf{\Sigma}$. Let λ_i, $i = 1, \ldots, p$ be the corresponding eigenvalues.
- We call $\mathbf{e}_1'\mathbf{X}, \ldots, \mathbf{e}_p'\mathbf{X}$ the principal components of $\mathbf{X}$.
- If $\mathbf{P}$ has columns $\mathbf{e}_i$'s, then $\mathbf{P}'(\mathbf{X} - \mu)$ has a $\mathcal{N}_p(0; \mathbf{\Lambda})$ distribution, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues.

Summary of definitions

- Principal Components: Find $\mathbf{e}_1, \ldots, \mathbf{e}_p$ which are orthonormal eigenvectors of $\mathbf{\Sigma}$. Let λ_i, $i = 1, \ldots, p$ be the corresponding eigenvalues.
- We call $\mathbf{e}_1'\mathbf{X}, \ldots, \mathbf{e}_p'\mathbf{X}$ the principal components of $\mathbf{X}$.
- If $\mathbf{P}$ has columns $\mathbf{e}_i$'s, then $\mathbf{P}'(\mathbf{X} - \mu)$ has a $\mathcal{N}_p(0; \mathbf{\Lambda})$ distribution, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues.
- The entries of $\mathbf{P}'\mathbf{X}$ are the components. We can reconstruct the original X from the components $\mathbf{e}_1'\mathbf{X}, \ldots, \mathbf{e}_p'\mathbf{X}$ because $\mathbf{X} = \mathbf{P}\mathbf{P}'\mathbf{X}$ or $\mathbf{X} = \sum_i (\mathbf{e}_i'\mathbf{X})\mathbf{e}_i$

Outline

Total variance – interpretation

Total variance – interpretation

- If $tr(\mathbf{\Sigma})$ stands for the trace of $\mathbf{\Sigma}$, i.e. it is the sum of the entries on the diagonal.

Total variance – interpretation

- If $tr(\mathbf{\Sigma})$ stands for the trace of $\mathbf{\Sigma}$, i.e. it is the sum of the entries on the diagonal.
- The following properties of the trace of a matrix are easy to check:
  - $tr(\mathbf{AB}) = tr(\mathbf{BA})$
  - $tr(\mathbf{B}^{-1}\mathbf{AB}) = tr(\mathbf{A})$.

Total variance – interpretation

- If $tr(\mathbf{\Sigma})$ stands for the trace of $\mathbf{\Sigma}$, i.e. it is the sum of the entries on the diagonal.
- The following properties of the trace of a matrix are easy to check:
  - $tr(\mathbf{AB}) = tr(\mathbf{BA})$
  - $tr(\mathbf{B}^{-1}\mathbf{AB}) = tr(\mathbf{A})$.
- It follows then that

$$\sum_{j=1}^{p} \mathbb{V}ar(Y_j) = \sum_{j=1}^{p} \lambda_j = tr(\mathbf{\Sigma}) = \sum_{j=1}^{n} \mathbb{V}ar(X_j)$$

Contributions to the total variance

Contributions to the total variance

- The proportion of the total variance explaining the jth principal component is thus

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Contributions to the total variance

- The proportion of the total variance explaining the *j*th principal component is thus

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- The *i*th principal component "explains" the proportion

$$\frac{\lambda_i}{\sum_1^p \lambda_j}$$

of the total variance.

Contributions to the total variance

- The proportion of the total variance explaining the jth principal component is thus

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

- The ith principal component "explains" the proportion

$$\frac{\lambda_i}{\sum_1^p \lambda_j}$$

of the total variance.

- The i largest principal components "explains" together the proportion

$$\frac{\sum_{k=1}^i \lambda_k}{\sum_{j=1}^p \lambda_j}$$

of the total variance.

Components are uncorrelated variables

Components are uncorrelated variables

- The vector of components $\mathbf{Y} = (Y_1, \ldots, Y_m)$ can be viewed as the transformed original vector $\mathbf{X}$ (change of coordinate system).

Components are uncorrelated variables

- The vector of components $\mathbf{Y} = (Y_1, \ldots, Y_m)$ can be viewed as the transformed original vector $\mathbf{X}$ (change of coordinate system).
- We have one-to-one linear correspondence between $\mathbf{X}$ and $\mathbf{Y}$ as expressed by

$$\mathbf{Y} = \mathbf{P}'\mathbf{X}, \quad \mathbf{X} = \mathbf{P}\mathbf{Y}$$

Components are uncorrelated variables

- The vector of components $\mathbf{Y} = (Y_1, \ldots, Y_m)$ can be viewed as the transformed original vector $\mathbf{X}$ (change of coordinate system).
- We have one-to-one linear correspondence between $\mathbf{X}$ and $\mathbf{Y}$ as expressed by

$$\mathbf{Y} = \mathbf{P}'\mathbf{X}, \quad \mathbf{X} = \mathbf{P}\mathbf{Y}$$

- The sum of variances of the coordinates in $\mathbf{X}$ and in $\mathbf{Y}$ are the same.

Components are uncorrelated variables

- The vector of components $\mathbf{Y} = (Y_1, \ldots, Y_m)$ can be viewed as the transformed original vector $\mathbf{X}$ (change of coordinate system).
- We have one-to-one linear correspondence between $\mathbf{X}$ and $\mathbf{Y}$ as expressed by

$$\mathbf{Y} = \mathbf{P}'\mathbf{X}, \quad \mathbf{X} = \mathbf{P}\mathbf{Y}$$

- The sum of variances of the coordinates in $\mathbf{X}$ and in $\mathbf{Y}$ are the same.
- The coordinates of $\mathbf{Y}$ are uncorrelated: $\mathbb{C}ov(Y_i, Y_j) = 0$, $i \neq j$.

Components are uncorrelated variables

- The vector of components $\mathbf{Y} = (Y_1, \ldots, Y_m)$ can be viewed as the transformed original vector $\mathbf{X}$ (change of coordinate system).
- We have one-to-one linear correspondence between $\mathbf{X}$ and $\mathbf{Y}$ as expressed by

$$\mathbf{Y} = \mathbf{P}'\mathbf{X}, \quad \mathbf{X} = \mathbf{P}\mathbf{Y}$$

- The sum of variances of the coordinates in $\mathbf{X}$ and in $\mathbf{Y}$ are the same.
- The coordinates of $\mathbf{Y}$ are uncorrelated: $\mathbb{C}ov(Y_i, Y_j) = 0$, $i \neq j$.
- The contributions to the total variance of the ith components is decreasing with i and equal to the corresponding eigenvalues.

Components are uncorrelated variables

- The vector of components $\mathbf{Y} = (Y_1, \ldots, Y_m)$ can be viewed as the transformed original vector $\mathbf{X}$ (change of coordinate system).
- We have one-to-one linear correspondence between $\mathbf{X}$ and $\mathbf{Y}$ as expressed by

$$\mathbf{Y} = \mathbf{P}'\mathbf{X}, \quad \mathbf{X} = \mathbf{P}\mathbf{Y}$$

- The sum of variances of the coordinates in $\mathbf{X}$ and in $\mathbf{Y}$ are the same.
- The coordinates of $\mathbf{Y}$ are uncorrelated: $\mathbb{C}ov(Y_i, Y_j) = 0$, $i \neq j$.
- The contributions to the total variance of the ith components is decreasing with i and equal to the corresponding eigenvalues.
- If the distribution of $\mathbf{X}$ is normal, then so is the distribution of $\mathbf{Y}$, i.e. it is $\mathcal{N}(\mathbf{P}\boldsymbol{\mu}, \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}' = \boldsymbol{\Lambda})$.

Component loadings

Component loadings

- The covariance between the ith principal component Y_i and the kth variable X_k in the observation is

$$\mathbb{C}ov(Y_i, X_k) = \lambda_i e_{ik}$$

Component loadings

- The covariance between the ith principal component Y_i and the kth variable X_k in the observation is

$$\mathbb{C}ov(Y_i, X_k) = \lambda_i e_{ik}$$

- Consequently, the correlation between the $i$the component and the kth variable is

$$\rho_{Y_i, X_k} = \sqrt{\lambda_i} e_{ik} / \sigma_k$$

where e_{ik} is the kth element in the ith eigenvector of $\mathbf{\Sigma}$.

Component loadings

- The covariance between the ith principal component Y_i and the kth variable X_k in the observation is

$$\mathbb{C}ov(Y_i, X_k) = \lambda_i e_{ik}$$

- Consequently, the correlation between the $i$the component and the kth variable is

$$\rho_{Y_i, X_k} = \sqrt{\lambda_i} e_{ik} / \sigma_k$$

  where e_{ik} is the kth element in the ith eigenvector of $\mathbf{\Sigma}$.
- For the ith component, the vector of correlations with the variables in observations is called **vector of component loadings**. It is indicative how much particular variables contribute to the given principal component.

The lack of scale invariance

The lack of scale invariance

- Many statistical procedures are expected to be scale invariant – it is expected that the change of measurement units does not affect the conclusion of statistical analysis.

The lack of scale invariance

- Many statistical procedures are expected to be scale invariant – it is expected that the change of measurement units does not affect the conclusion of statistical analysis.
- Unfortunately, this is not the case for the principal component analysis – change of scale results in fundamental change of the principal components

The lack of scale invariance

- Many statistical procedures are expected to be scale invariant – it is expected that the change of measurement units does not affect the conclusion of statistical analysis.
- Unfortunately, this is not the case for the principal component analysis – change of scale results in fundamental change of the principal components
- Generally, we can proceed in two essentially different ways to obtain PCA.

The lack of scale invariance

- Many statistical procedures are expected to be scale invariant – it is expected that the change of measurement units does not affect the conclusion of statistical analysis.
- Unfortunately, this is not the case for the principal component analysis – change of scale results in fundamental change of the principal components
- Generally, we can proceed in two essentially different ways to obtain PCA.
- If the multivariate variables in question are of the similar nature and at the similar numerical scale, so the values between variables can be meaningfully related one to another, then the PCA should be performed on the original scale.

The lack of scale invariance

- Many statistical procedures are expected to be scale invariant – it is expected that the change of measurement units does not affect the conclusion of statistical analysis.
- Unfortunately, this is not the case for the principal component analysis – change of scale results in fundamental change of the principal components
- Generally, we can proceed in two essentially different ways to obtain PCA.
- If the multivariate variables in question are of the similar nature and at the similar numerical scale, so the values between variables can be meaningfully related one to another, then the PCA should be performed on the original scale.
- If the multivariate variables are of different nature and unrelated numerical scale, the PCA should be performed on the standardized data.

Standardized principal components

Standardized principal components

- The components are depending on the scale of the X_i:s. Frequently, it is common to standardize $\mathbf{X}$

$$\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \mu),$$

where $\mathbf{V} = diag(\mathbf{\Sigma})$.

Standardized principal components

- The components are depending on the scale of the X_i:s. Frequently, it is common to standardize $\mathbf{X}$

$$\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \mu),$$

where $\mathbf{V} = diag(\mathbf{\Sigma})$.

- The covariance matrix of $\mathbf{Z}$ is the correlation matrix of $\mathbf{X}$:

$$\mathbb{C}ov(\mathbf{Z}) = \mathbf{V}^{-1/2}\mathbf{\Sigma}\mathbf{V}^{-1/2} = \boldsymbol{\rho}$$

Standardized principal components

- The components are depending on the scale of the X_i:s. Frequently, it is common to standardize $\mathbf{X}$

$$\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \mu),$$

where $\mathbf{V} = diag(\mathbf{\Sigma})$.

- The covariance matrix of $\mathbf{Z}$ is the correlation matrix of $\mathbf{X}$:

$$\mathbb{C}ov(\mathbf{Z}) = \mathbf{V}^{-1/2}\mathbf{\Sigma}\mathbf{V}^{-1/2} = \boldsymbol{\rho}$$

- Let $\lambda_i, \mathbf{e}_i$, $i = 1, \ldots, p$ be eigenvalues and eigenvectors of $\boldsymbol{\rho}$, then the total variance and the component loadings are

$$\sum_{i=1}^{p} \rho_{ii} = \sum_{i=1}^{p} 1 = p$$

$$\rho(Y_i, Z_k) = \sqrt{\lambda_i} e_{ki}$$

Outline

Constant normal density levels

Multivariate normal distribution

- The density of the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-distribution is constant on the ellipsoid

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

This have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$.

Constant normal density levels

Multivariate normal distribution

- The density of the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-distribution is constant on the ellipsoid
$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$
This have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$.
- Thus the principal components are represented by the vectors pointing in the direction of the axes of the ellipsoids.

Constant normal density levels

Multivariate normal distribution

- The density of the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-distribution is constant on the ellipsoid
$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$
This have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$.
- Thus the principal components are represented by the vectors pointing in the direction of the axes of the ellipsoids.
- The first principal component points in the direction of the longest axis, the second in the direction of the second longest axis, etc.

Constant normal density levels

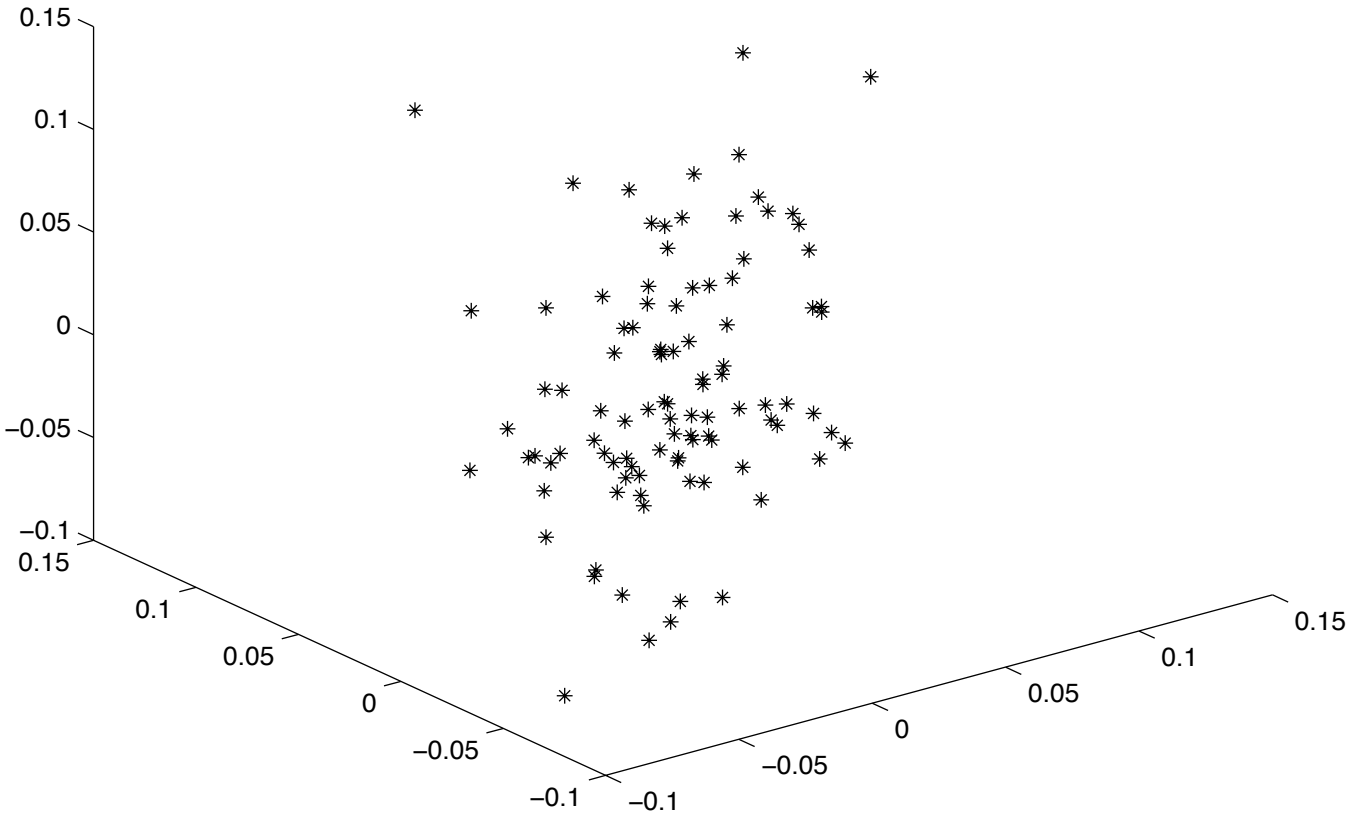
Multivariate normal distribution

- The density of the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-distribution is constant on the ellipsoid
$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$
This have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$.
- Thus the principal components are represented by the vectors pointing in the direction of the axes of the ellipsoids.
- The first principal component points in the direction of the longest axis, the second in the direction of the second longest axis, etc.
- If we deal with multivariate data points they can be interpreted as a set of points (n points) in a p dimensional space.

Constant normal density levels

Multivariate normal distribution

- The density of the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-distribution is constant on the ellipsoid
$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$
This have axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$.
- Thus the principal components are represented by the vectors pointing in the direction of the axes of the ellipsoids.
- The first principal component points in the direction of the longest axis, the second in the direction of the second longest axis, etc.
- If we deal with multivariate data points they can be interpreted as a set of points (n points) in a p dimensional space.
- If the set of points have the shape of an ellipsoid then the normalized directions of the main axes are the principal components.


0.15
0.1
0.05
0
−0.05
−0.1
0.15
0.1
0.05
0
−0.05
−0.1
0.15
0.1
0.05
0
−0.05
−0.1

Component loadings – coordinates of the principal components directions

Component loadings – coordinates of the principal components directions

- The sample covariance between the jth principal component and variable k is equal to

Component loadings – coordinates of the principal components directions

- The sample covariance between the *j*th principal component and variable *k* is equal to
- The score (component loading) can be used as coordinate in the *p* dimensional space spanned by the principal components.

Component loadings – coordinates of the principal components directions

- The sample covariance between the *j*th principal component and variable *k* is equal to
- The score (component loading) can be used as coordinate in the *p* dimensional space spanned by the principal components.
- Make proper scale adjustment depending if the standardized or non-standardized version of the PCA analysis is made.

Outline

Sample based components

Sample based components

- The principal components calculated from **S** or **R** (instead of from $\mathbf{\Sigma}$ and ρ.

Sample based components

- The principal components calculated from **S** or **R** (instead of from $\mathbf{\Sigma}$ and ρ.
- The eigenvalue decomposition of **S** gives eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, and corresponding normalised eigenvectors $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_p$.

Sample based components

- The principal components calculated from **S** or **R** (instead of from $\mathbf{\Sigma}$ and ρ.
- The eigenvalue decomposition of **S** gives eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, and corresponding normalised eigenvectors $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_p$.
- The jth principal component is a variable with values

$$\hat{Y}_{ji} = \hat{\mathbf{e}}_i'(\mathbf{X}_j - \bar{\mathbf{X}})$$

where $\mathbf{X}_j$ is the $j$the row in the data matrix.

Special structure of Σ

Special structure of $\mathbf{\Sigma}$

- $\mathbf{\Sigma}$ diagonal

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{pp} \end{bmatrix}$$

$\mathbf{e}_i' = (0, 0, \dots, 1, 0, \dots, 0)$ och $\lambda_i = \sigma_{ii}$, $i = 1, 2, \dots, p$

Special structure of $\mathbf{\Sigma}$

- $\mathbf{\Sigma}$ diagonal

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{pp} \end{bmatrix}$$

$$\mathbf{e}_i' = (0, 0, \dots, 1, 0, \dots, 0) \text{ och } \lambda_i = \sigma_{ii},\ i = 1, 2, \dots, p$$

- Equicorrelation matrix

$$\rho = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & & & \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

  - $\lambda_1 = 1 + (p-1)\rho$, $\mathbf{e}_1 = \mathbf{1}_p/\sqrt{p}$, $\lambda_2 = \cdots = \lambda_p = 1 - \rho$.

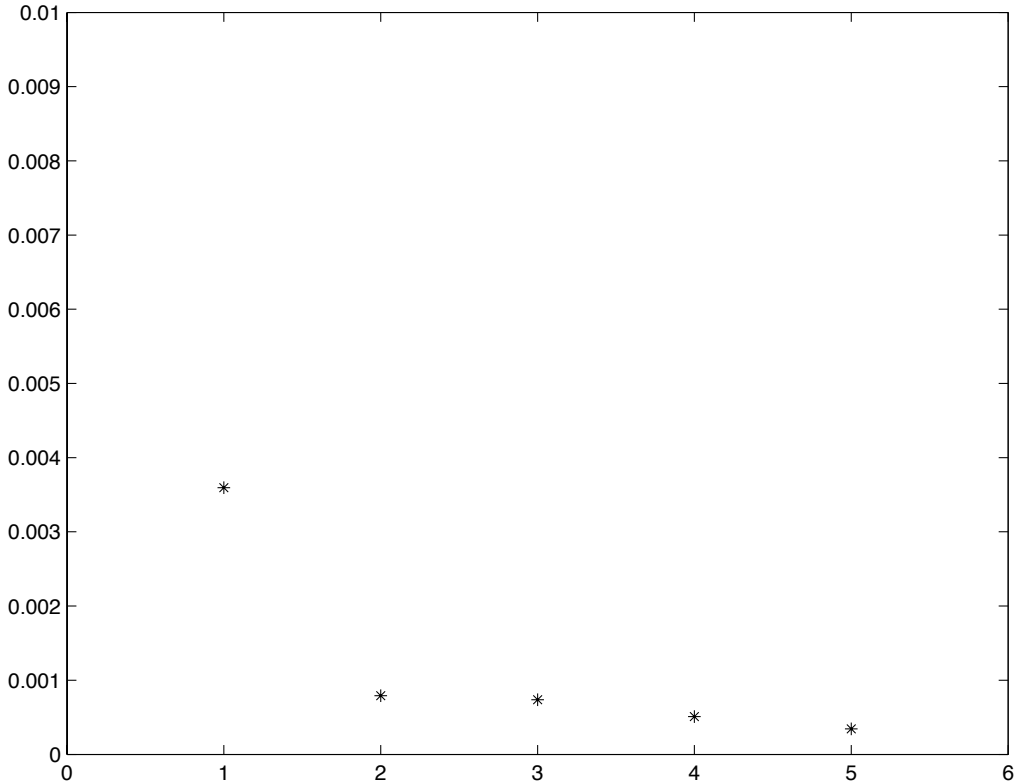
How many principal components?

If we want to "reduce dimension" in the problem we use the k first principal components. How to choose k without loosing to much information in the data?

Screeplot

Plot the eigenvalues $\hat{\lambda}_i$ against i. This will give a tool for determining the number of pricipal components for data reduction.

"Elbow" criteria – Look at the shape created by the points.


0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
0
1
2
3
4
5
6

Other methods

1. With
$$\psi_k = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$
(where the eigenvalues are ordered in decreasing order) choose k such that $\psi_k \geq c$ for some c. (Often $c = 0.9$.)
2. Kaisers rule: Exclude all principal components with eigenvalues less than the mean of the eigenvalues.
(If PCA is based on the correlation matrix this mean is 1.)

Points to ponder:

1: Principal component i has variance λ_i.

2: Can compare λ_i to $\sum_i \lambda_i = trace(\mathbf{\Sigma})$.

3: Try to interpret components; often leading component is some sort of average then others are contrasts.

4: To estimate replace $\mathbf{\Sigma}$ by $\mathbf{S}$.

5: Process can be done on original $\mathbf{X}$ or on standardized variates: $Z_i = (X_i - \mu_i)/\sigma_{ii}$ Covariance matrix of Z_i is $\boldsymbol{\rho}$, correlation matrix.

6: General advice: use $\mathbf{S}$ when different variables in $\mathbf{X}$ are commensurate (same units, comparable). Use $\mathbf{R}$ otherwise.

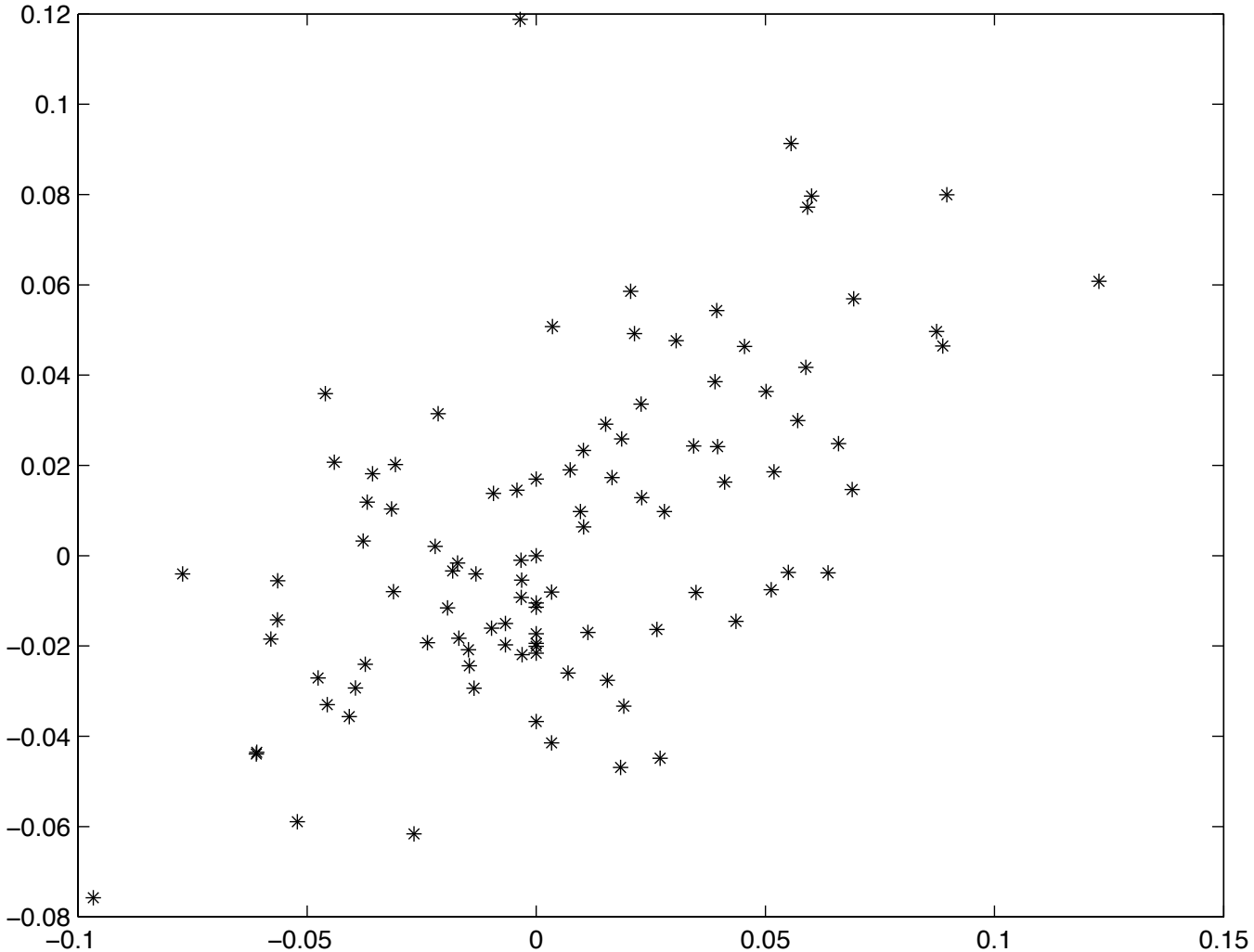
Outline

NY Stock Exchange – Example

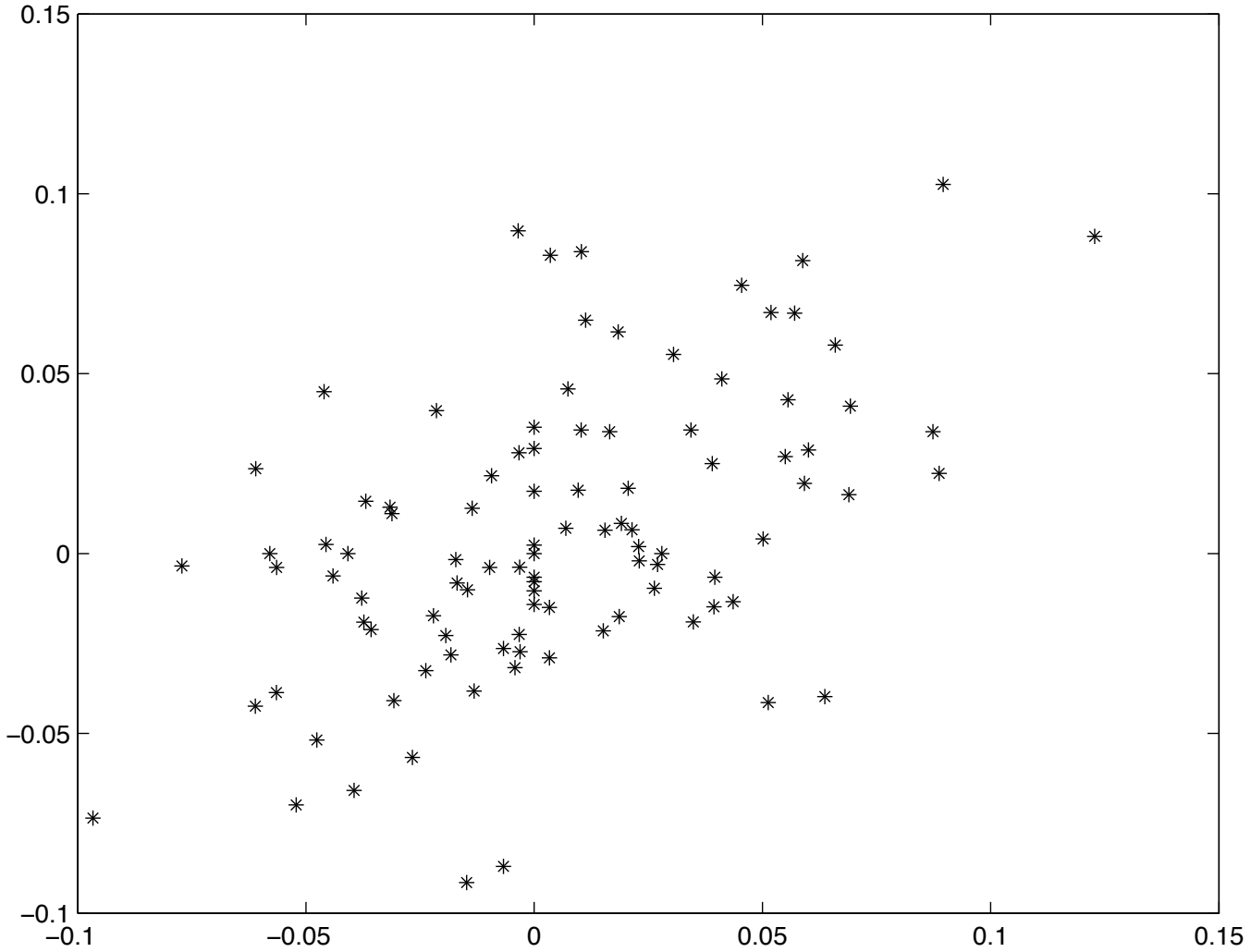
Week	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	0	0	0	0.0395	0
2	0.0270	-0.0449	-0.0030	-0.0145	0.0435
3	0.1228	0.0608	0.0881	0.0862	0.0781
4	0.0570	0.0299	0.0668	0.0135	0.0195
.	.	.	.	.	
98	0.0103	0.0233	0.0343	-0.0182	-0.0047
99	-0.0441	0.0207	-0.0062	-0.0185	0.0047
100	0.0390	0.0385	0.0250	-0.0283	0.0327

Data available here

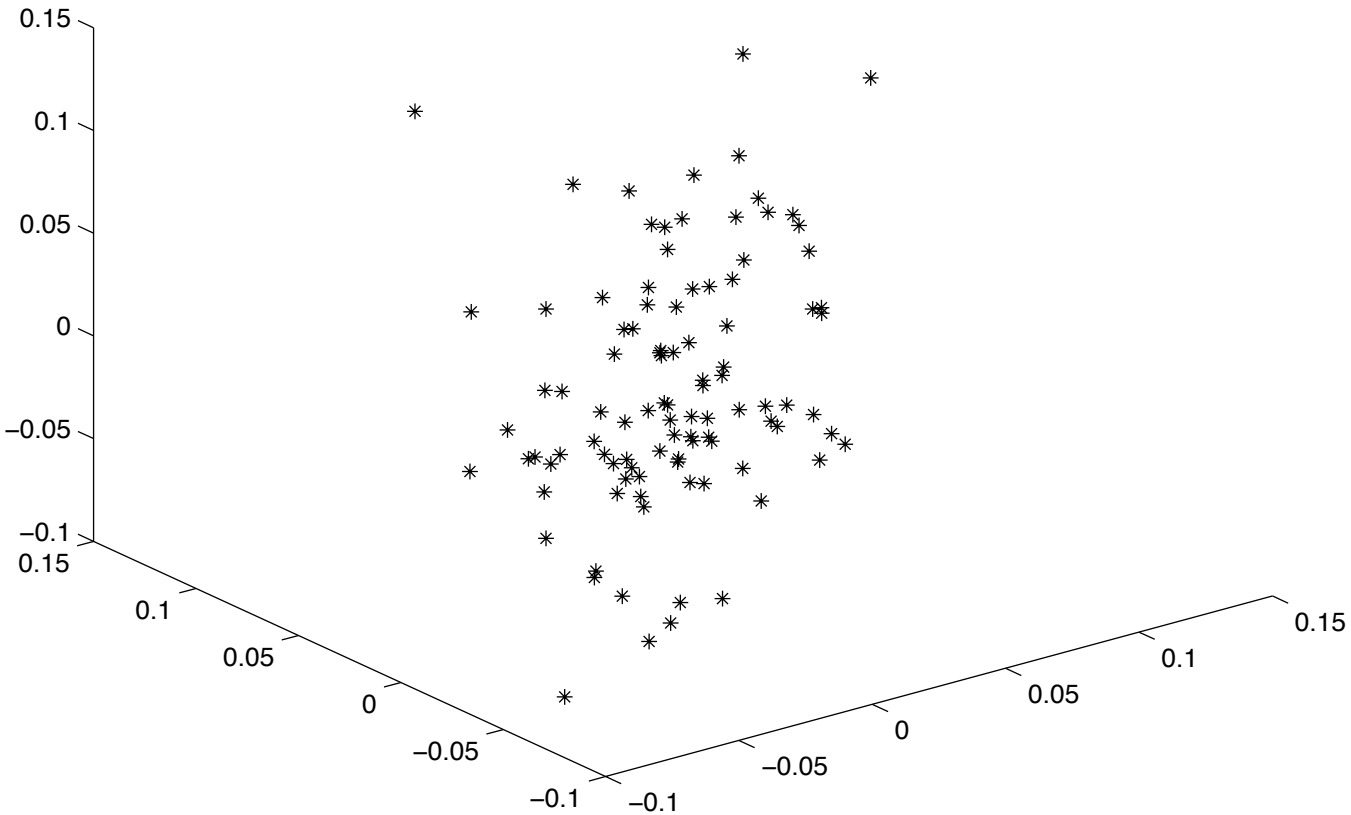
Allied Chemical vs Du Pont



Allied Chemical vs Union Carbide



Allied Chemical, Du Pont and Union Carbide



Code to produce PCA

```
SE=read.table("StockExchange.txt")

S=cov(SE)
R=cor(SE)

EV1=eigen(S)
EV2=eigen(R)

SEV1=sum(EV1$values)
SEV2=sum(EV2$values)

CPCA1=cumsum(EV1$values)/SEV1
CPCA2=cumsum(EV2$values)/SEV2
```

```
plot(EV1$values,col="red",pch="*",
      ylim=c(0,max(EV1$values)))
abline(0,0,col="blue")

plot(EV2$values,col="red",pch="*",
      ylim=c(0,max(EV1$values)))
abline(0,0,col="blue")

plot(CPCA1,col="red",pch="*",cex=3,
      ylim=c(0,max(CPCA1)))
abline(1,0,col="blue")

plot(CPCA2,col="red",pch="*",cex=3,
      ylim=c(0,max(CPCA2)))
abline(1,0,col="blue")
```

NY Stock Exchange – non-normalized data

Covariance matrix

Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
0.0016	0.0008	0.0008	0.0004	0.0005
0.0008	0.0012	0.0008	0.0004	0.0003
0.0008	0.0008	0.0016	0.0005	0.0005
0.0004	0.0004	0.0005	0.0008	0.0004
0.0005	0.0003	0.0005	0.0004	0.0008

NY Stock Exchange – non-normalized data

Covariance matrix

Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
0.0016	0.0008	0.0008	0.0004	0.0005
0.0008	0.0012	0.0008	0.0004	0.0003
0.0008	0.0008	0.0016	0.0005	0.0005
0.0004	0.0004	0.0005	0.0008	0.0004
0.0005	0.0003	0.0005	0.0004	0.0008

- Eigenvalues: 0.00360 0.00079 0.00074 0.00051 0.00034

NY Stock Exchange – non-normalized data

Covariance matrix

Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
0.0016	0.0008	0.0008	0.0004	0.0005
0.0008	0.0012	0.0008	0.0004	0.0003
0.0008	0.0008	0.0016	0.0005	0.0005
0.0004	0.0004	0.0005	0.0008	0.0004
0.0005	0.0003	0.0005	0.0004	0.0008

- Eigenvalues: 0.00360 0.00079 0.00074 0.00051 0.00034
- Percentage of the total variation: 0.602 0.133 0.123 0.085 0.058

NY Stock Exchange – normalized data

Correlation matrix

Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1.000	0.577	0.509	0.387	0.462
0.577	1.000	0.598	0.390	0.322
0.509	0.598	1.000	0.436	0.426
0.387	0.390	0.436	1.000	0.524
0.462	0.322	0.426	0.524	1.000

NY Stock Exchange – normalized data

Correlation matrix

Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1.000	0.577	0.509	0.387	0.462
0.577	1.000	0.598	0.390	0.322
0.509	0.598	1.000	0.436	0.426
0.387	0.390	0.436	1.000	0.524
0.462	0.322	0.426	0.524	1.000

- Eigenvalues: 2.856 0.809 0.540 0.451 0.343

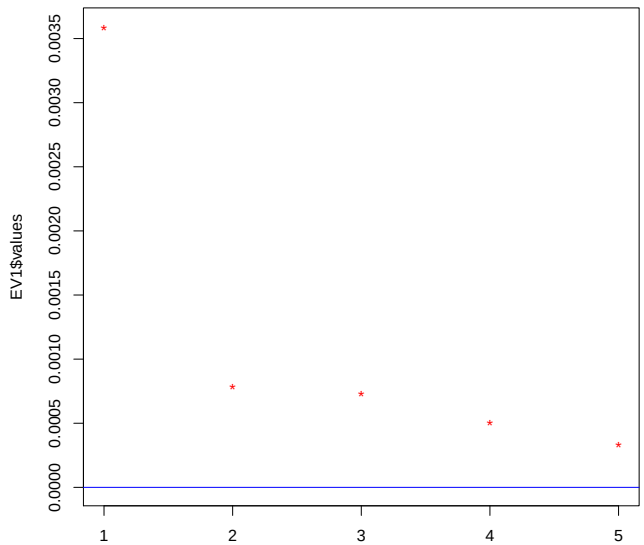
NY Stock Exchange – normalized data

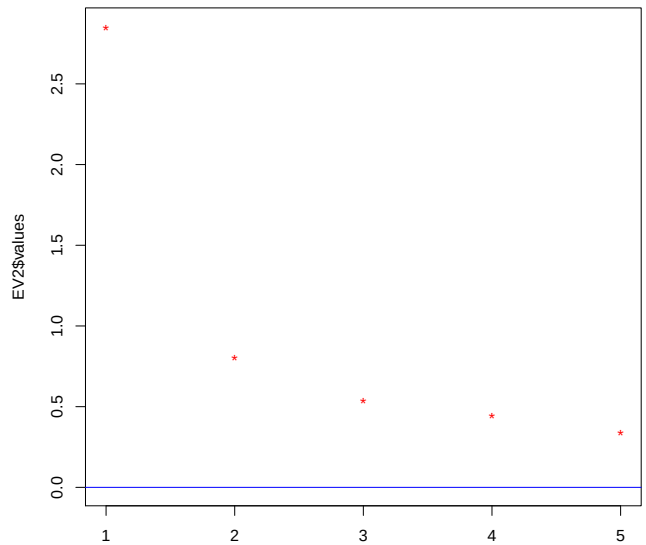
Correlation matrix

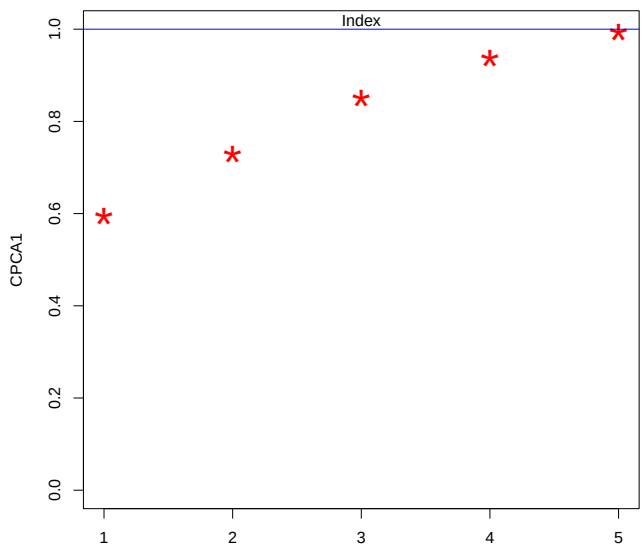
Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1.000	0.577	0.509	0.387	0.462
0.577	1.000	0.598	0.390	0.322
0.509	0.598	1.000	0.436	0.426
0.387	0.390	0.436	1.000	0.524
0.462	0.322	0.426	0.524	1.000

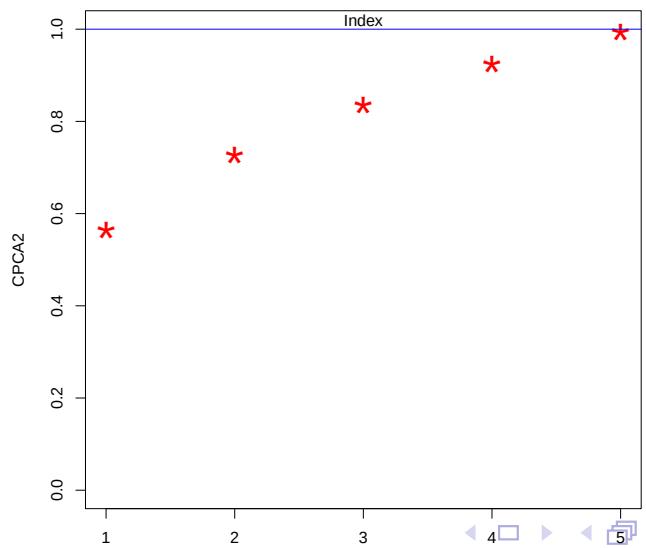
- Eigenvalues: 2.856 0.809 0.540 0.451 0.343
- Percentage of the total variation: 0.571 0.162 0.108 0.090 0.069

How many components? Graphs









Interpretation – not standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.83	-0.80	-0.83	-0.62	-0.62
2	0.52	-0.07	-0.47	-0.11	0.07
3	-0.08	-0.36	-0.08	0.58	0.61
4	0.16	-0.44	0.29	-0.35	0.05
5	-0.10	0.15	-0.05	-0.38	0.49

Interpretation – not standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.83	-0.80	-0.83	-0.62	-0.62
2	0.52	-0.07	-0.47	-0.11	0.07
3	-0.08	-0.36	-0.08	0.58	0.61
4	0.16	-0.44	0.29	-0.35	0.05
5	-0.10	0.15	-0.05	-0.38	0.49

- The first component is roughly equally weighted sum, or "index" of the five stocks: general stock-market component.

Interpretation – not standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.83	-0.80	-0.83	-0.62	-0.62
2	0.52	-0.07	-0.47	-0.11	0.07
3	-0.08	-0.36	-0.08	0.58	0.61
4	0.16	-0.44	0.29	-0.35	0.05
5	-0.10	0.15	-0.05	-0.38	0.49

- The first component is roughly equally weighted sum, or "index" of the five stocks: general stock-market component.
- The second is positively correlated with the 1st chemical company and negatively with the 3rd.

Interpretation – not standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.83	-0.80	-0.83	-0.62	-0.62
2	0.52	-0.07	-0.47	-0.11	0.07
3	-0.08	-0.36	-0.08	0.58	0.61
4	0.16	-0.44	0.29	-0.35	0.05
5	-0.10	0.15	-0.05	-0.38	0.49

- The first component is roughly equally weighted sum, or "index" of the five stocks: general stock-market component.
- The second is positively correlated with the 1st chemical company and negatively with the 3rd.
- The third is positively correlated with the oil companies and negatively correlated with the 2nd chemical companies.

Interpretation – standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.78	-0.77	-0.79	-0.71	-0.71
2	0.22	0.46	0.23	-0.47	-0.52
3	0.45	-0.13	-0.25	-0.40	0.32
4	-0.26	-0.14	0.45	-0.32	0.26
5	-0.27	0.40	-0.23	-0.11	0.23

Interpretation – standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.78	-0.77	-0.79	-0.71	-0.71
2	0.22	0.46	0.23	-0.47	-0.52
3	0.45	-0.13	-0.25	-0.40	0.32
4	-0.26	-0.14	0.45	-0.32	0.26
5	-0.27	0.40	-0.23	-0.11	0.23

- The first component is roughly equally weighted sum, or "index" of the five stocks: general stock-market component.

Interpretation – standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.78	-0.77	-0.79	-0.71	-0.71
2	0.22	0.46	0.23	-0.47	-0.52
3	0.45	-0.13	-0.25	-0.40	0.32
4	-0.26	-0.14	0.45	-0.32	0.26
5	-0.27	0.40	-0.23	-0.11	0.23

- The first component is roughly equally weighted sum, or "index" of the five stocks: general stock-market component.
- The second is positively correlated with the chemical companies and negatively with the oil companies: industry component.

Interpretation – standardized PCA

Matrix of the correlations

Components	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	-0.78	-0.77	-0.79	-0.71	-0.71
2	0.22	0.46	0.23	-0.47	-0.52
3	0.45	-0.13	-0.25	-0.40	0.32
4	-0.26	-0.14	0.45	-0.32	0.26
5	-0.27	0.40	-0.23	-0.11	0.23

- The first component is roughly equally weighted sum, or "index" of the five stocks: general stock-market component.
- The second is positively correlated with the chemical companies and negatively with the oil companies: industry component.
- The remaining components are less clearly interpretable.

Which method to choose?

Which method to choose?

- There is no obvious answer to this question.

Which method to choose?

- There is no obvious answer to this question.
- Typically, choosing the method that gives easier interpretation within smaller number of components would be preferable.

Which method to choose?

- There is no obvious answer to this question.
- Typically, choosing the method that gives easier interpretation within smaller number of components would be preferable.
- With our ‘naive’ insight into stock markets, we would prefer the standardized PCA.

Which method to choose?

- There is no obvious answer to this question.
- Typically, choosing the method that gives easier interpretation within smaller number of components would be preferable.
- With our ‘naive’ insight into stock markets, we would prefer the standardized PCA.
- The standardized method removes the issue of difference in the scale of variability of the stocks.

Formulas and code

Formulas and code

- The correlation between the $i$the component Y_i and the kth variable X_k is

$$\rho_{Y_i,X_k} = \sqrt{\lambda_i} e_{ik} / \sigma_k$$

e_{ik} – the kth element in the ith eigenvector and $\sigma_k = \sqrt{\sigma_{kk}}$

Formulas and code

- The correlation between the $i$the component Y_i and the kth variable X_k is

$$\rho_{Y_i,X_k} = \sqrt{\lambda_i} e_{ik} / \sigma_k$$

  e_{ik} – the kth element in the ith eigenvector and $\sigma_k = \sqrt{\sigma_{kk}}$
- In the matrix notation: let $\mathbf{A} = diag(1/\sigma_k)$ then the correlations of components are row entries of the following matrix

$$\sqrt{\mathbf{\Lambda}}\mathbf{P}'\mathbf{A}$$

Formulas and code

- The correlation between the *i*the component Y_i and the *k*th variable X_k is

$$\rho_{Y_i,X_k} = \sqrt{\lambda_i} e_{ik}/\sigma_k$$

  e_{ik} – the *k*th element in the *i*th eigenvector and $\sigma_k = \sqrt{\sigma_{kk}}$
- In the matrix notation: let $\mathbf{A} = diag(1/\sigma_k)$ then the correlations of components are row entries of the following matrix

$$\sqrt{\mathbf{\Lambda}}\mathbf{P}'\mathbf{A}$$

- Note that for standardized PCA **A** is the identity matrix.

```
L=diag(EV1$values)
P=EV1$vectors
A=diag(1/sqrt(diag(S)))

Crr=sqrt(L)%*%t(P)%*%A
```