Multivariate Statistics

Principal Component Analysis (PCA)

14.12.2016

Goals of Today's Lecture

- Get familiar with the multivariate counterparts of the expectation and the variance.
- See how principal component analysis (PCA) can be used as a dimension reduction technique.

Introduction

- In your introductory course you started with **univariate statistics**. You had a look at **one** random variable X at a time. E.g., $X =$ "measurement of temperature".

- A random variable can be characterized by its **expectation** μ and the **variance** σ^2 (or standard deviation σ).

$$\mu = E[X], \quad \sigma^2 = \mathrm{Var}(X) = E[(X - \mu)^2].$$

- A **model** that is often used is the **normal distribution**: $X \sim \mathcal{N}(\mu, \sigma^2)$.

- The normal distribution is fully characterized by the expectation and the variance.

- The **unknown parameters** μ and σ can be **estimated** from **data**.
- Say we observe n (independent) realizations of our random variable X: $x_1, \ldots, x_n$.
- You can think of measuring n times a certain quantity, e.g. temperature.
- Usual **parameter estimates** are

$$\widehat{\mu} = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \widehat{\sigma}^2 = \widehat{\text{Var}}(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Multivariate Data

- Now we are going to have a look at the situation where we measure **multiple things** simultaneously.

- Hence, we have a **multivariate random variable (vector)** $\underline{X}$ having m components: $\underline{X} \in \mathbb{R}^m$.

- You can think of measuring temperature at two different locations or measuring temperature and pressure at one location ($m = 2$).

- In that case

$$\underline{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix},$$

where $X^{(1)}$ is temperature and $X^{(2)}$ is pressure.

- A possible data-set now consists of n **vectors** of dimension 2 (or m in the general case):

$$\underline{x}_1, \ldots, \underline{x}_n,$$

where $\underline{x}_i \in \mathbb{R}^2$ (or $\mathbb{R}^m$).

Remark

- In multiple linear regression we already had multiple variables per observation.
- There, we had one **response variable** and many **predictor** variables.
- Here, the situation is more general in the sense that we don't have a response variable but we want to model "relationships" between (any) variables.

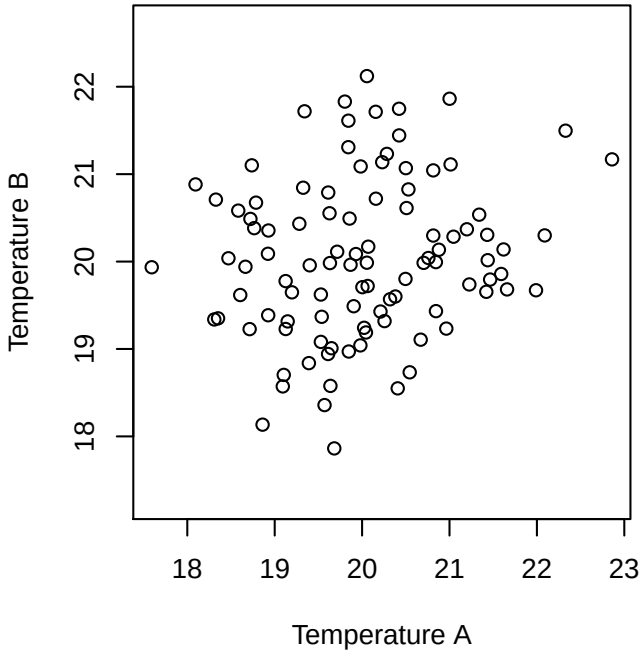
Expectation and Covariance Matrix

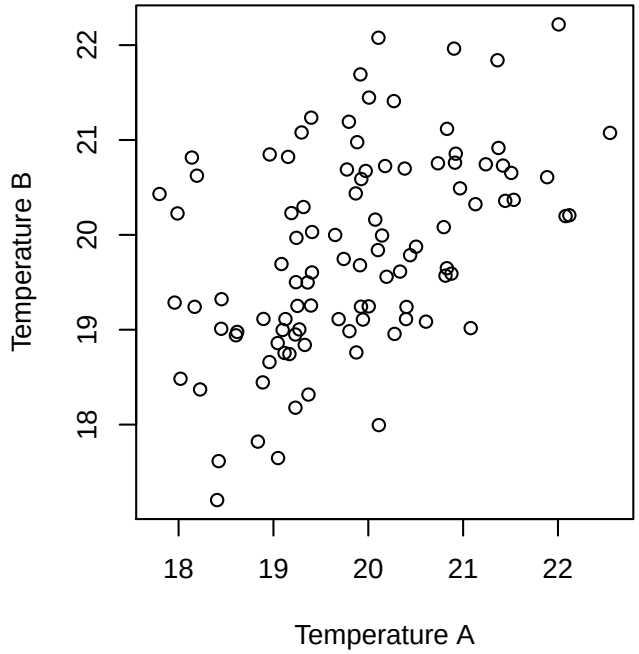
- We need new concepts to model / describe this kind of data.
- We are therefore looking for the multivariate counterparts of the expectation and the variance.
- The (multivariate) **expectation** of $\underline{X}$ is defined as

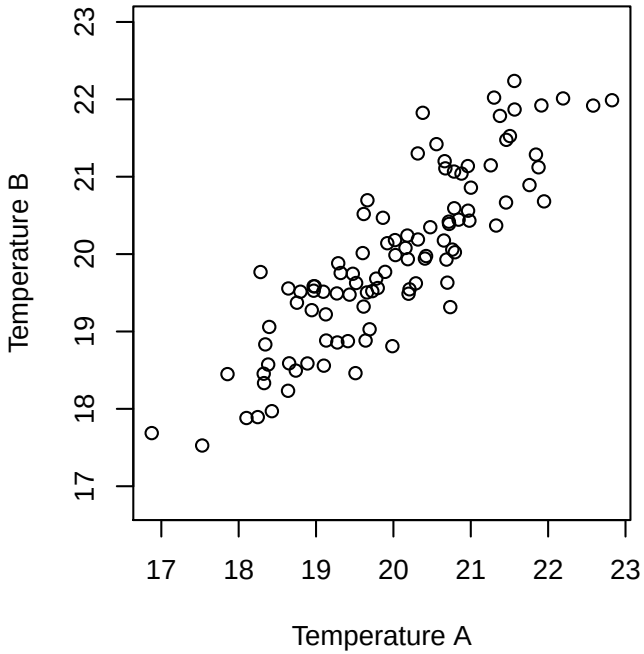
$$E[\underline{X}] = \underline{\mu} = (\mu_1, \ldots, \mu_m)^T = (E[X^{(1)}], \ldots, E[X^{(m)}])^T.$$

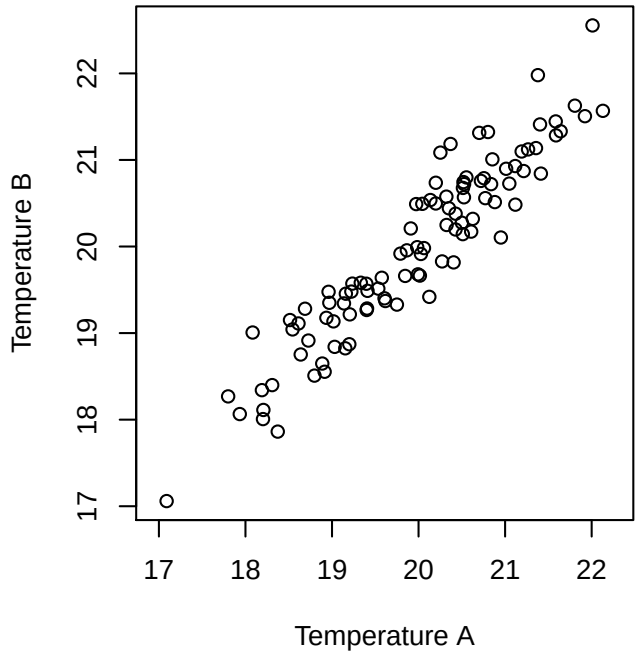
- It's nothing else than the collection of the univariate expectations.

What about dependency?







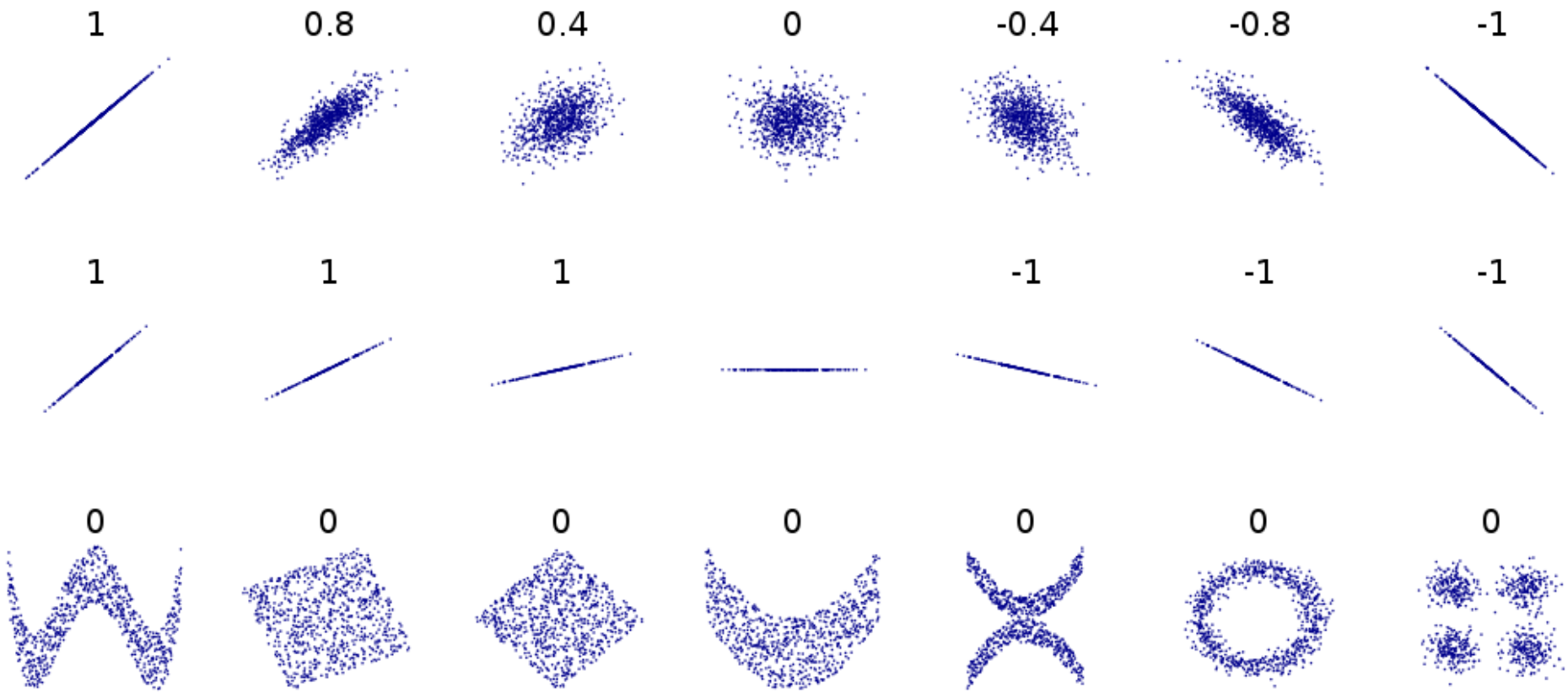


- We need a measure to **characterize the dependency between the different components**.
- The simplest thing one can think of is **linear dependency** between two components.
- The corresponding measure is the **correlation** ρ.
- ρ is dimensionless and it always holds that

$$-1 \leq \rho \leq 1$$

- $|\rho|$ measures the **strength** of the linear relationship.
- The sign of ρ indicates the **direction** of the linear relationship.

Illustration: Empirical Correlation



Source: Wikipedia

- The formal definition of the correlation is based on the covariance.

- The **covariance** is an **unstandardized** version of the correlation. It is defined as

$$\text{Cov}(X^{(j)}, X^{(k)}) = E[(X^{(j)} - \mu_j)(X^{(k)} - \mu_k)].$$

- The correlation between $X^{(j)}$ and $X^{(k)}$ is then

$$\rho_{jk} = \text{Corr}(X^{(j)}, X^{(k)}) = \frac{\text{Cov}(X^{(j)}, X^{(k)})}{\sqrt{\text{Var}(X^{(j)})\,\text{Var}(X^{(k)})}}$$

- You have seen the empirical version in the introductory course.

- The **covariance matrix** $\boldsymbol{\Sigma}$ is an $m \times m$ matrix with elements

$$\boldsymbol{\Sigma}_{jk} = \mathrm{Cov}(X^{(j)}, X^{(k)}) = E[(X^{(j)} - \mu_j)(X^{(k)} - \mu_k)].$$

- We also write $\mathrm{Var}(\underline{X})$ or $\mathrm{Cov}(\underline{X})$ instead of $\boldsymbol{\Sigma}$.

- The special symbol $\boldsymbol{\Sigma}$ is used in order to avoid confusion with the sum sign $\sum$.

- The covariance matrix **contains a lot of information**, e.g.

$$\boldsymbol{\Sigma}_{jj} = \mathrm{Var}(X^{(j)}).$$

  This means that the **diagonal** consists of the **individual variances**.

- We can also compute the **correlations** via

$$\mathrm{Corr}(X^{(j)}, X^{(k)}) = \frac{\mathrm{Cov}(X^{(j)}, X^{(k)})}{\sqrt{\mathrm{Var}(X^{(j)})\,\mathrm{Var}(X^{(k)})}} = \frac{\boldsymbol{\Sigma}_{jk}}{\sqrt{\boldsymbol{\Sigma}_{jj}\boldsymbol{\Sigma}_{kk}}}.$$

- Again, from a real data-set we can estimate these quantities with

$$\begin{aligned} \underline{\widehat{\mu}} &= \left[\overline{x}^{(1)}, \overline{x}^{(2)}, \ldots, \overline{x}^{(m)}\right]^T \\ \widehat{\mathbf{\Sigma}}_{jk} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^{(j)} - \widehat{\mu}_j)(x_i^{(k)} - \widehat{\mu}_k), \end{aligned}$$

- Or more directly the whole matrix

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\underline{x}_i - \underline{\widehat{\mu}})(\underline{x}_i - \underline{\widehat{\mu}})^T.$$

- Remember: (Empirical) correlation only measures strength of **linear relationship** between two variables.

Linear Transformations

The following table illustrates how the expectation and the variance (covariance matrix) change when **linear transformations** are applied to the univariate random variable X or the multivariate random vector $\underline{X}$.

Univariate			**Multivariate**		
Y	$=$	$a + bX$	$\underline{Y}$	$=$	$\underline{a} + \mathbf{B}\underline{X}$
$E[Y]$	$=$	$a + bE[X]$	$E[\underline{Y}]$	$=$	$\underline{a} + \mathbf{B}E[\underline{X}]$
$\mathrm{Var}(Y)$	$=$	$b^2\,\mathrm{Var}(X)$	$\mathrm{Var}(\underline{Y})$	$=$	$\mathbf{B}\boldsymbol{\Sigma}_X\mathbf{B}^T$

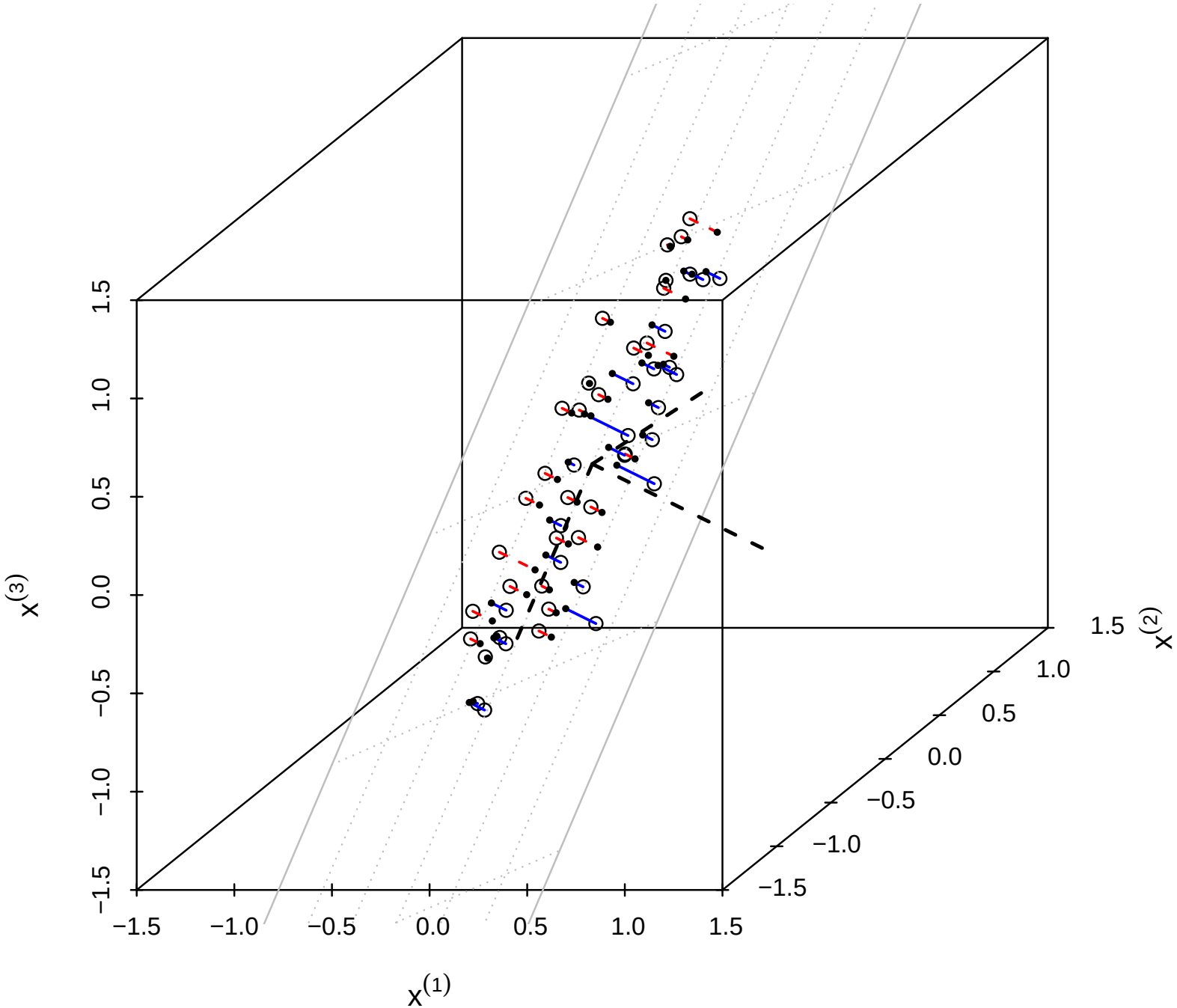
where $a, b \in \mathbb{R}$ and $\underline{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{m \times m}$.

Principal Component Analysis (PCA)

Goal: Dimensionality reduction.

- We have m different dimensions (variables) but we would like to find "a few specific dimensions (projections) of the data that contain most variation".
- If two specific dimensions of the data-set contain most variation, visualizations will be easy (plot these two!).
- Such a plot then can be used to check for any "structure".

Illustration of Artificial 3-Dim Data-Set



- We have to be more precise with what we mean with "variation".

- We **define** the **total variation** in the data as the **sum of all individual empirical variances**

$$\sum_{j=1}^{m} \widehat{\mathrm{Var}}(X^{(j)}) = \sum_{j=1}^{m} \widehat{\sigma}^2(X^{(j)}).$$

- How can we now find projections that contain most variation?

- Conceptually, we are looking for a **new coordinate system** with basis vectors $\underline{b}_1, \ldots, \underline{b}_m \in \mathbb{R}^m$.

- Of course, our data-points $\underline{x}_i \in \mathbb{R}^m$ will then have **new coordinates**

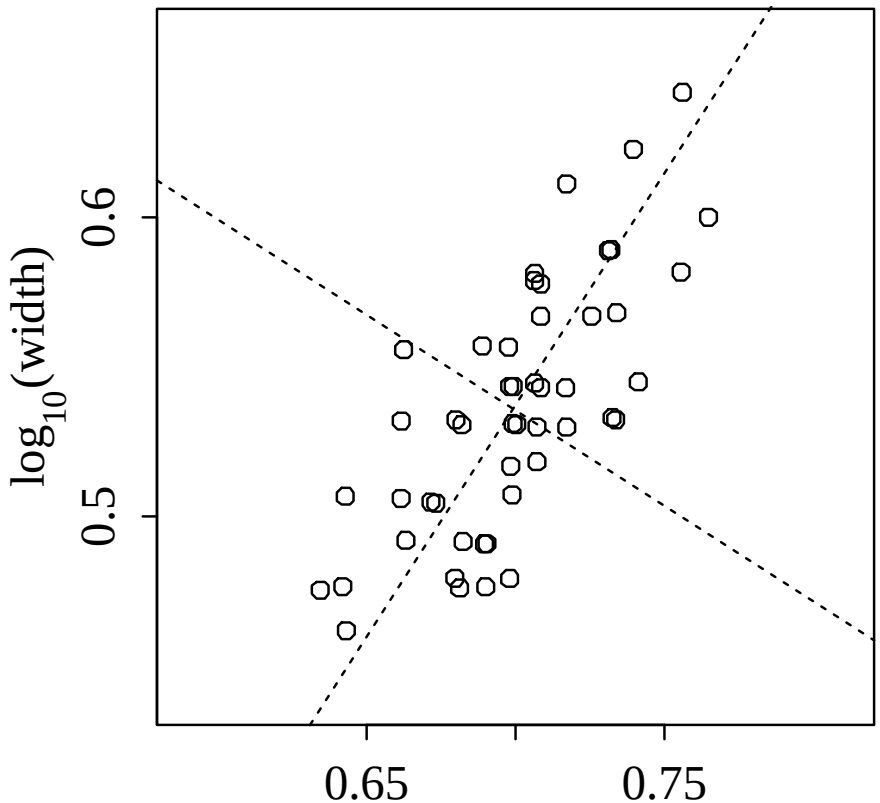
$$z_i^{(k)} = \underline{x}_i^T \underline{b}_k,\ k = 1, \ldots, m$$

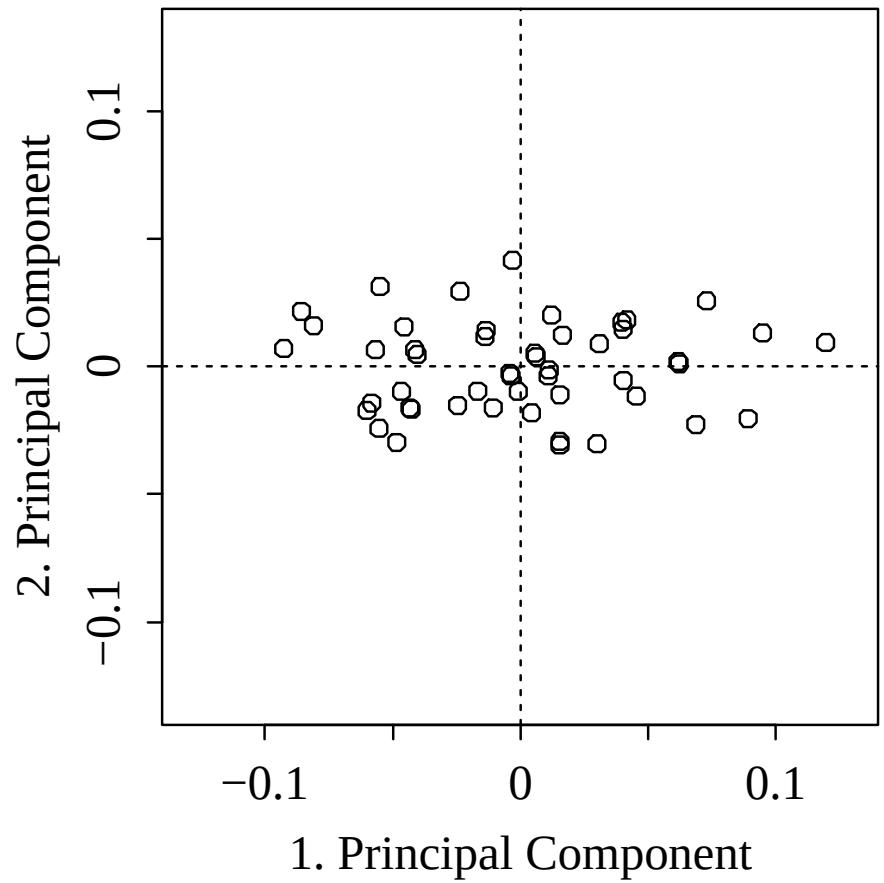
(= projection on new basis vectors).

- How should we choose the new basis?
  - The first basis vector $\underline{b}_1$ should be chosen such that $\widehat{\mathrm{Var}}(Z^{(1)})$ is **maximal** (i.e. it captures most variation).
  - The second basis vector $\underline{b}_2$ should be **orthogonal** to the first one ($\underline{b}_2^T \underline{b}_1 = 0$) such that $\widehat{\mathrm{Var}}(Z^{(2)})$ is **maximized**.
  - And so on...

- The new basis vectors are the so-called **principal components**.
- The individual components of these basis vectors are called **loadings**. The loadings tell us how to interpret the new coordinate system (i.e., how the old variables are weighted to get the new ones).
- The coordinates with respect to the new basis vectors (the transformed variable values) are the so-called **scores** .

PCA: Illustration in Two Dimensions





- We could find the first basis vector $\underline{b}_1$ by solving the following **maximization problem**

$$\max_{\underline{b}:\|b\|=1} \widehat{\text{Var}}(\mathbf{X}\underline{b}),$$

  where $\mathbf{X}$ is the matrix that has different observations in different rows and different variables in different columns (like the design matrix in regression).

- It can be shown that $\underline{b}_1$ is the (standardized) eigenvector of $\widehat{\mathbf{\Sigma}}_X$ that corresponds to the largest eigenvalue.

- Similarly for the other vectors $\underline{b}_2, \ldots, \underline{b}_m$.

- To summarize: We are performing a **transformation to new variables**

$$\underline{z}_i = \mathbf{B}^T(\underline{x}_i - \underline{\widehat{\mu}}),$$

where the transformation matrix $\mathbf{B}$ is orthogonal and contains the $\underline{b}_k$'s as columns.

- In general we also subtract the mean vector to ensure that all components have mean 0.

- $\mathbf{B}$ is the matrix of (standardized) eigenvectors corresponding to the eigenvalues λ_k of $\widehat{\mathbf{\Sigma}}_X$ (in decreasing order).

- Hence, we have

$$\widehat{\mathrm{Var}}(\underline{Z}) \;=\; \widehat{\boldsymbol{\Sigma}}_Z = \mathbf{B}^T \widehat{\boldsymbol{\Sigma}}_X \mathbf{B} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \lambda_m \end{bmatrix}$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0.$$

- Hence, the variance of the different components of $\underline{Z}$ is given by the corresponding **eigenvalue** (values on the diagonal).

- Moreover, the different components are **uncorrelated** (because the off-diagonal elements of the covariance matrix of $\underline{Z}$ are all zero).

Scaling Issues

- The variance is **not** invariant under rescaling.
- If we change the units of a variable, that will change the variance (e.g. when measuring a length in [m] instead of [mm]).
- Therefore, if variables are measured on very different scales, they should first be **standardized** to comparable units.
- This can be done by standardizing each variable to variance 1.
- Otherwise, PCA can be misleading.

PCA and Dimensionality Reduction

- At the beginning we were talking about **dimensionality reduction**.
- We can achieve this by simply looking at the first $p < m$ principal components (and ignoring the remaining components).
- The **proportion** of the variance that is explained by the first p principal components is

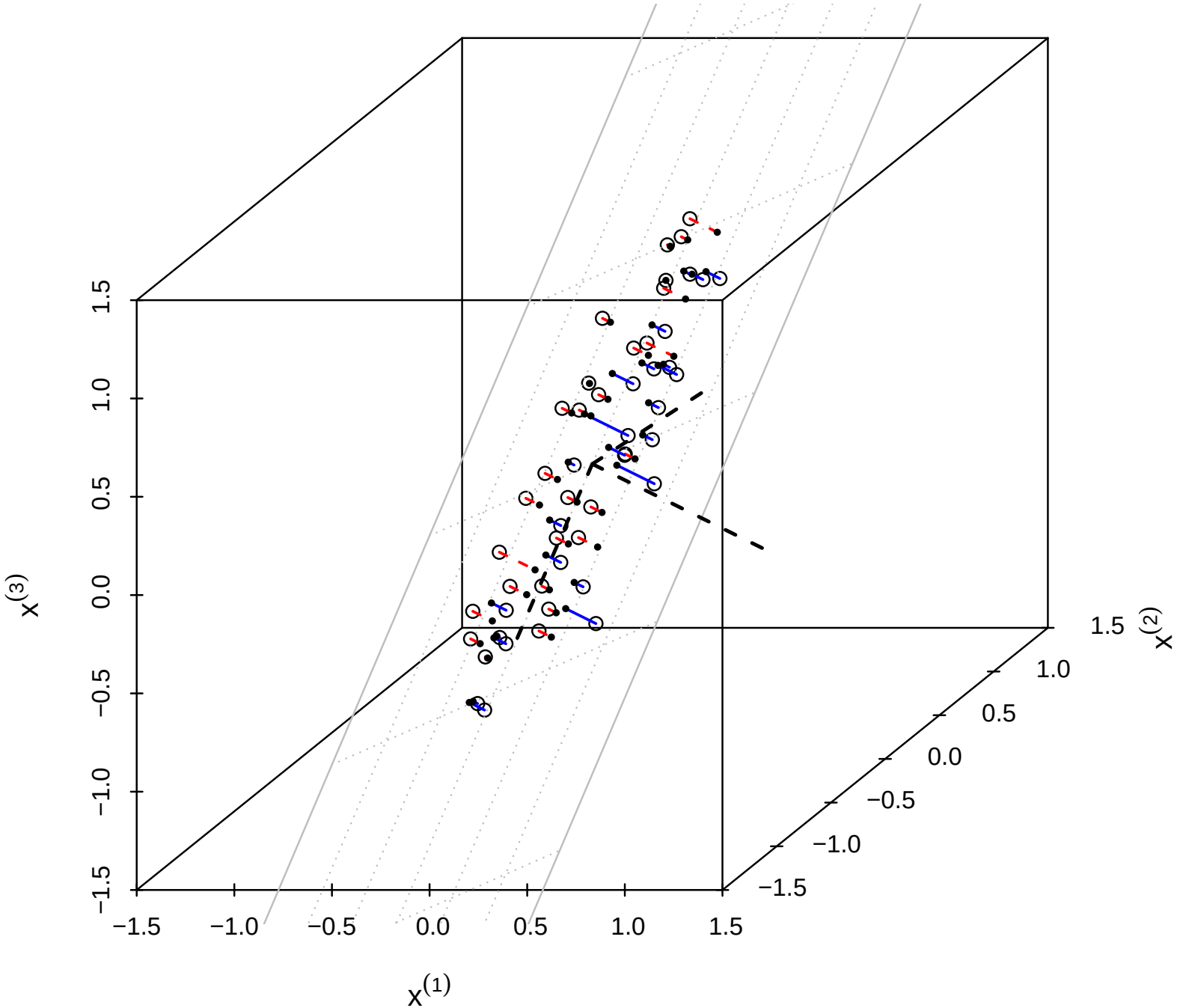
$$\frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{m} \lambda_j}.$$

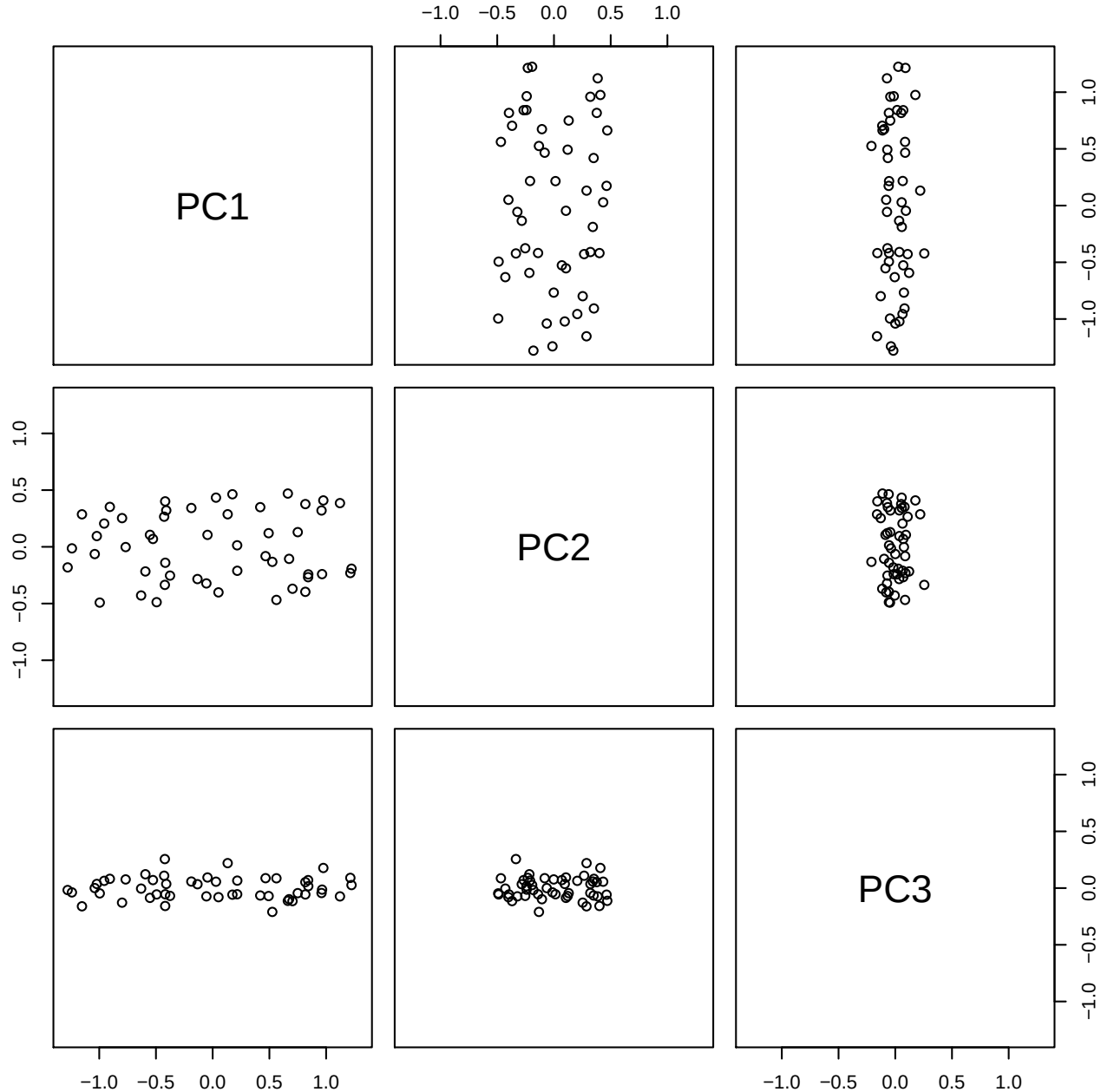
PCA and Dimensionality Reduction

- For visualization of our data, we can for example use a scatterplot of the first two principal components.

- It should show “most of the variation”.

Illustration

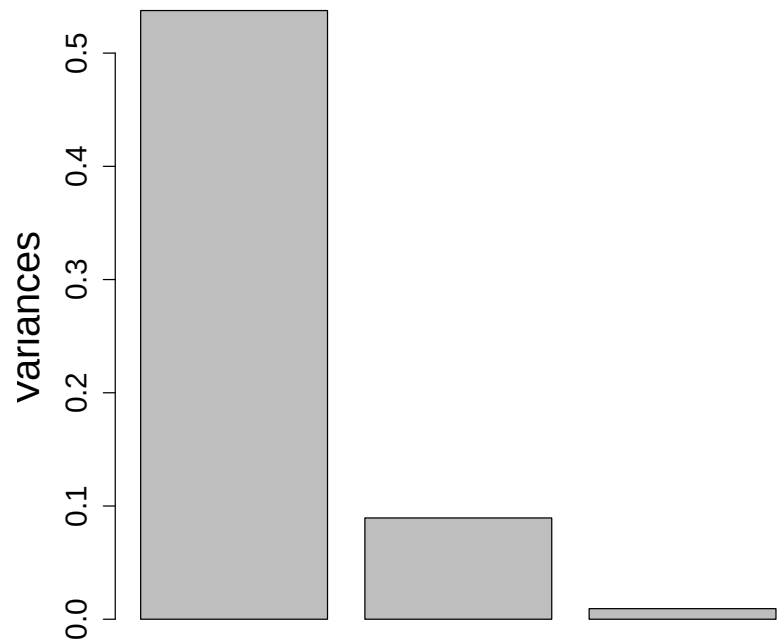



PC1
PC2
PC3
−1.0
−0.5
0.0
0.5
1.0

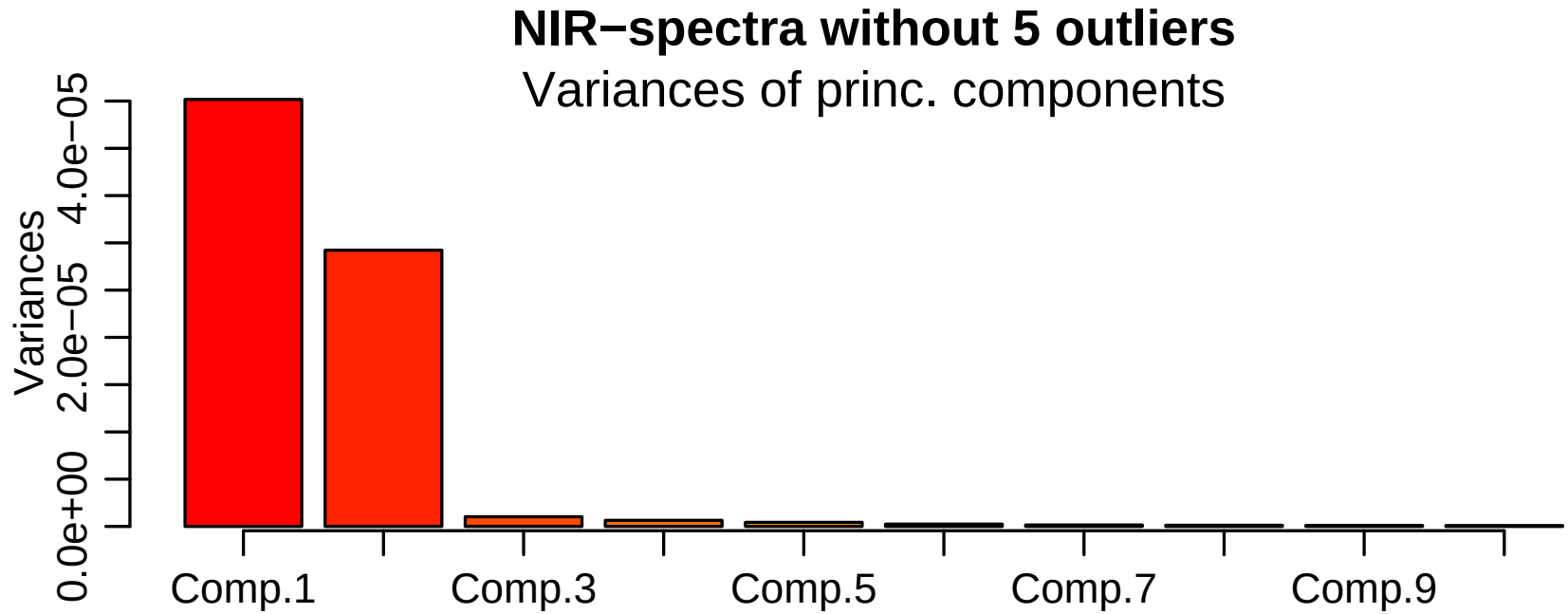
Loadings

```
                PC1           PC2           PC3
x1 -0.395118451  0.06887762  0.91604437
x2 -0.009993467  0.99680385 -0.07926044
x3 -0.918575822 -0.04047172 -0.39316727
```

Screeplot



- Sometimes we see a sharp drop (after component p) when plotting the eigenvalues (in decreasing order).
- $\rightarrow$ Consider only the first p components.

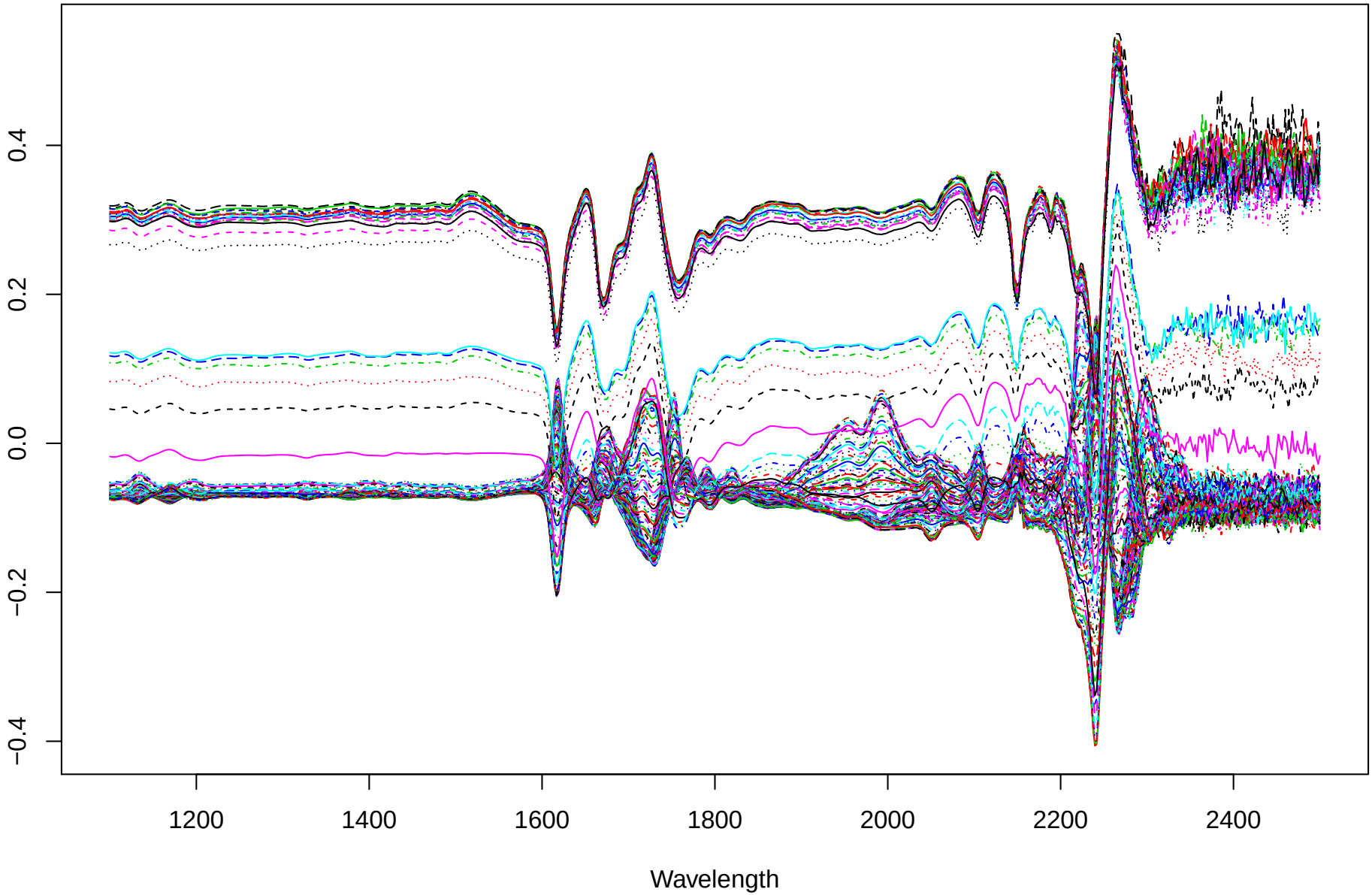


- This plot is also known as the **scree-plot**.

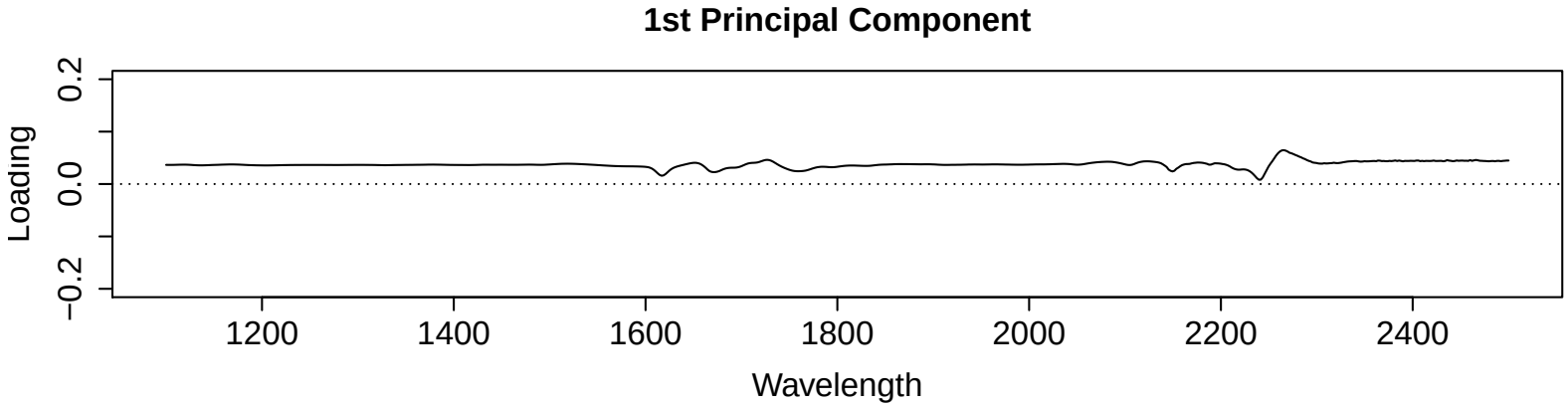
Applying PCA to NIR-Spectra

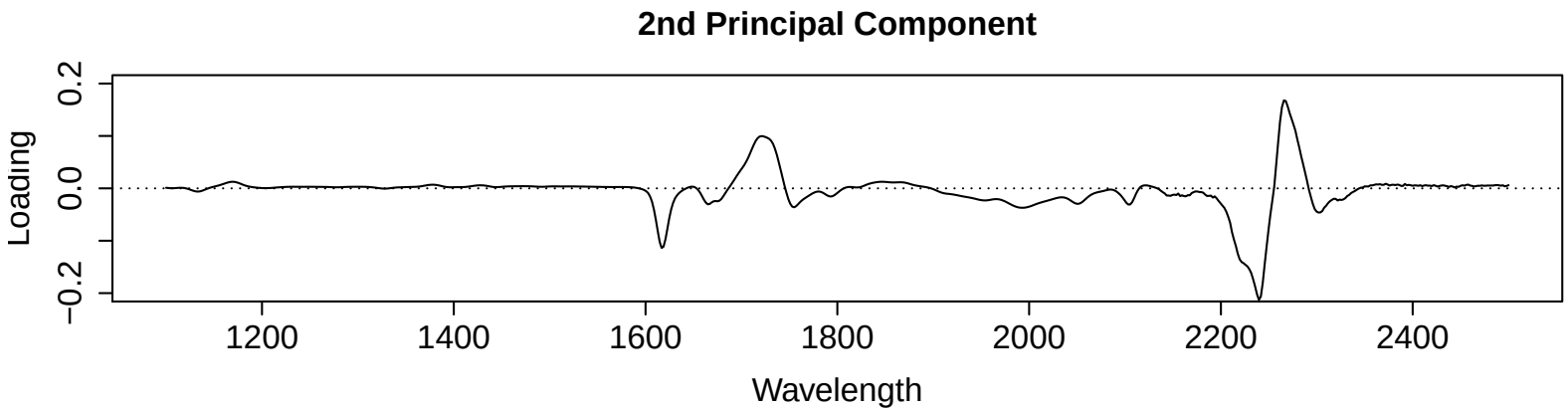
- An **NIR-spectrum** can be thought of as a multivariate observation (the different variables are the measurements at different wavelengths).
- A spectrum has the property that the different variables are "ordered" and we can plot one observation as a "function" (see plot on next slide).
- If we apply PCA to this kind of data, the individual components of the $\underline{b}_k$'s (the so called **loadings**) can again be plotted as spectra.
- As an example we have a look at spectra measured at different time-points of a chemical reaction.

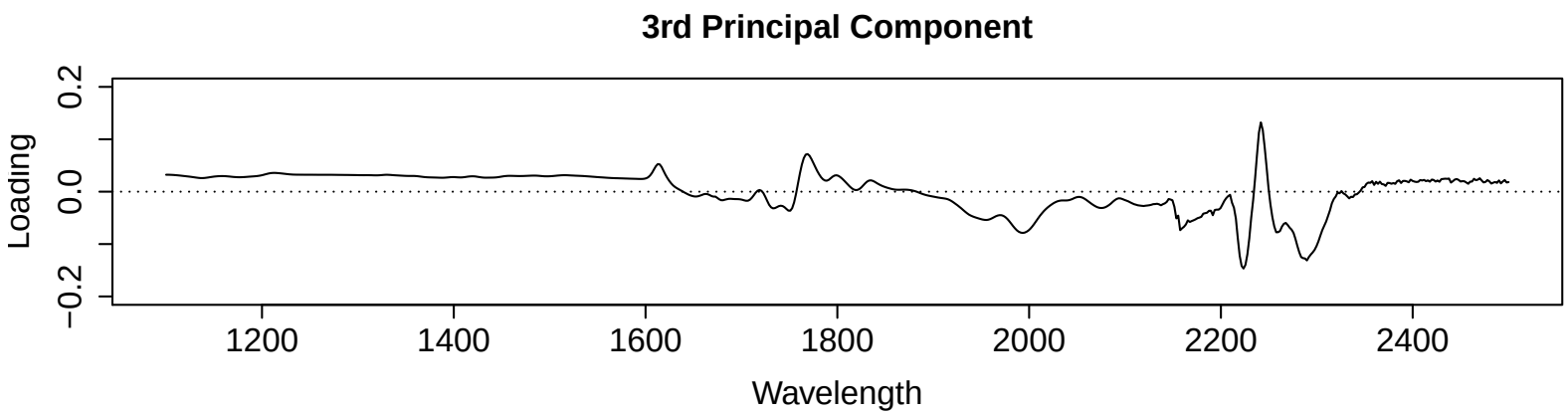
Illustration: Spectra (centered at each wavelength)



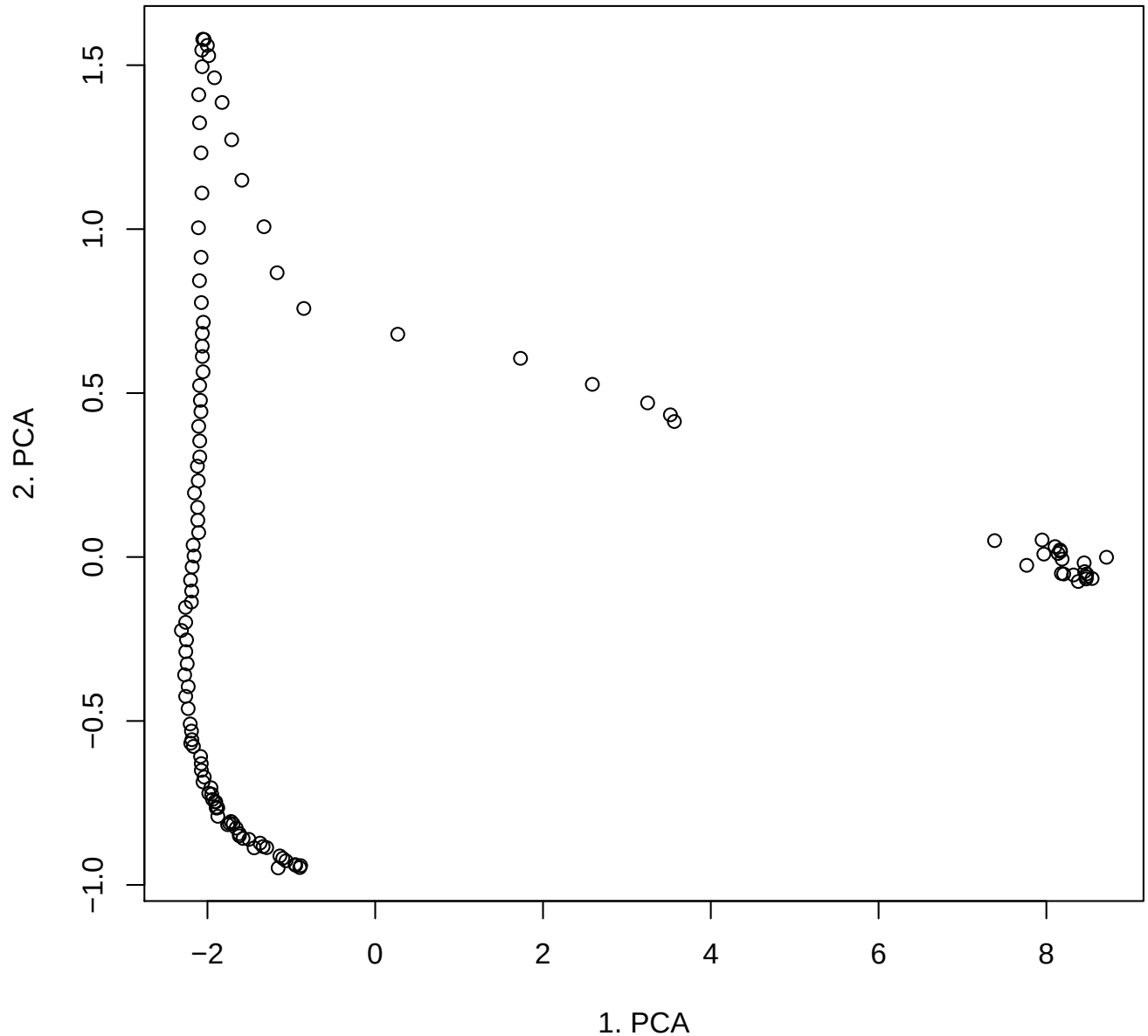
One observation is one spectrum, i.e. a whole "function".


1st Principal Component
Loading
0.2
0.0
−0.2
1200
1400
1600
1800
2000
2200
2400
Wavelength

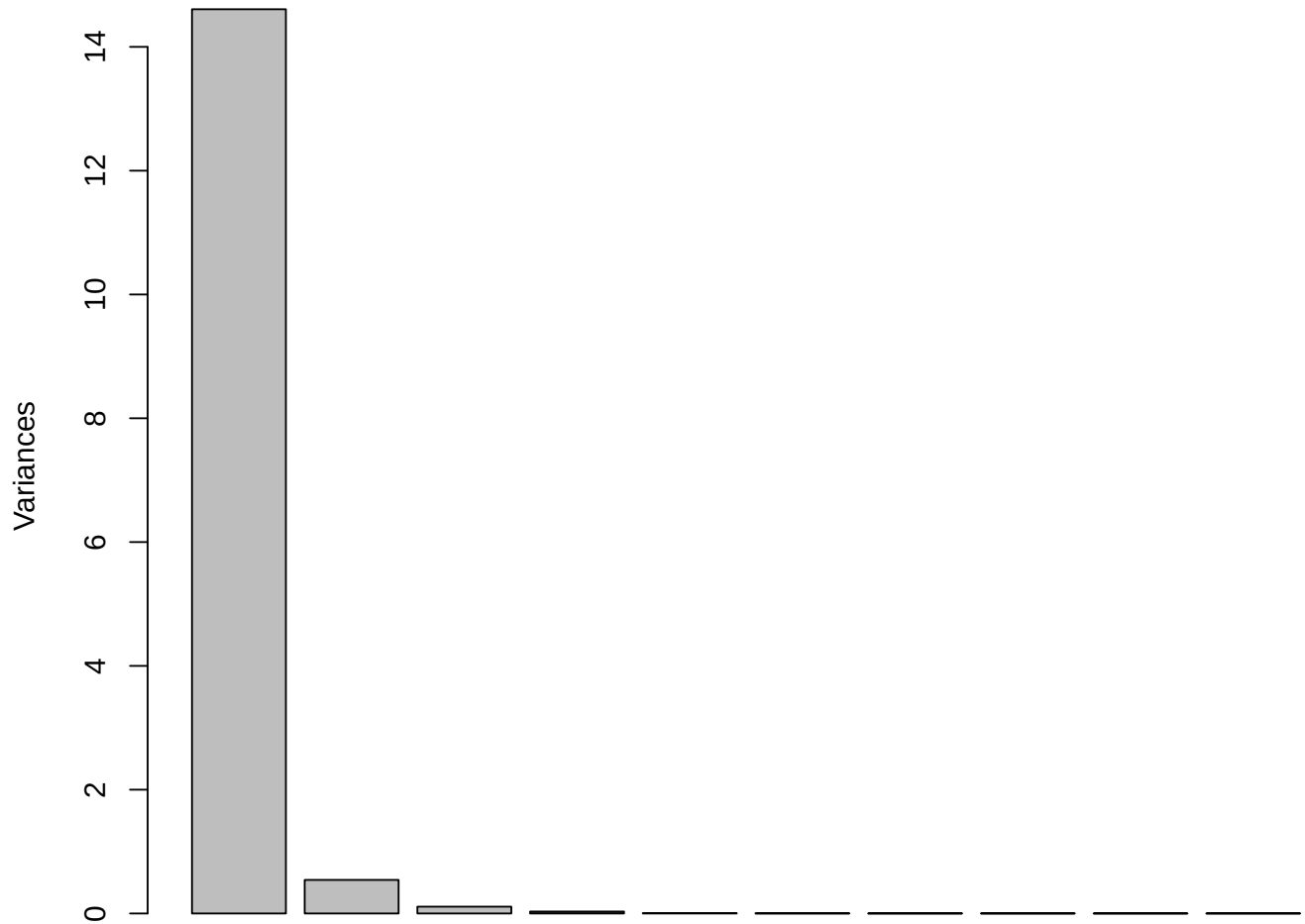

2nd Principal Component
Loading
0.2
0.0
−0.2
1200
1400
1600
1800
2000
2200
2400
Wavelength


3rd Principal Component
Loading
0.2
0.0
−0.2
1200
1400
1600
1800
2000
2200
2400
Wavelength

Scatterplot of the First Two Principal Components



Scree-Plot



Alternative Interpretations

If we restrict ourselves to the first $p < m$ principal components we have

$$\underline{x}_i - \underline{\widehat{\mu}} = \underline{\widehat{x}}_i + \underline{e}_i$$

where

$$\underline{\widehat{x}}_i = \sum_{k=1}^{p} z_i^{(k)} \underline{b}^{(k)}, \quad \underline{e}_i = \sum_{k=p+1}^{m} z_i^{(k)} \underline{b}^{(k)}.$$

Linear Algebra

It can be shown that the data matrix consisting of the $\widehat{\underline{x}_i}$ is the **best approximation of our original (centered) data matrix** if we restrict ourselves to matrices of **rank** p (with respect to the Frobenius norm), i.e. it has smallest

$$\sum_{i,j} \left(e_i^{(j)} \right)^2 .$$

Statistics

It's the best approximation in the sense that it has the smallest sum of variances

$$\sum_{j=1}^{m} \widehat{\mathrm{Var}}(E^{(j)}).$$

PCA via Singular Value Decomposition (SVD)

- We can also get the principal components from the singular value decomposition (SVD) of the data matrix $\mathbf{X}$.
- For that reason we require $\mathbf{X}$ to have **centered columns**! 
- Why does this work?
  SVD of $\mathbf{X}$ yields the decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

  where
  - $\mathbf{U}$ is $n \times n$ and orthogonal
  - $\mathbf{D}$ is $n \times m$ and generalized diagonal (containing the so-called singular values in **descending** order)
  - $\mathbf{V}$ is $m \times m$ and orthogonal

Properties of SVD

- The (standardized) eigenvectors of $\mathbf{X}^T\mathbf{X}$ make up the columns of $\mathbf{V}$.
- The singular values are the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$.

But what is $\mathbf{X}^T\mathbf{X}$? If the columns of $\mathbf{X}$ are **centered**, this is the rescaled (empirical) covariance matrix $\widehat{\mathbf{\Sigma}}_X$, because

$$(\mathbf{X}^T\mathbf{X})_{jk} = \sum_{i=1}^{n} (\underline{x}_i \underline{x}_i^T)_{jk} = \sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} = (n-1)\widehat{\mathbf{\Sigma}}_{jk}.$$

Hence, the singular value decomposition of the **centered** data-matrix automatically gives us the principal components (in $\mathbf{V}$).

The data-matrix in new coordinates is given in $\mathbf{UD}$.

Summary

- PCA is a useful tool for dimension reduction.
- New basis system is given by (standardized) eigenvectors of covariance matrix.
- Eigenvalues are the variances of the new coordinates.
- In the case of spectra, the loadings can again be plotted as spectra.