Lecture 5: A first look at dimension reduction

Felix Held, Mathematical Sciences

MSA220/MVE441 Statistical Learning for Big Data

15th April 2021

CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

Principal Component Analysis

Projection onto a subspace

Assume $\mathbf{x} \in \mathbb{R}^p$. Given **orthonormal vectors** $\mathbf{b}_1, \dots, \mathbf{b}_m$, i.e.

$$\|\mathbf{b}_j\| = 1 \quad \text{and} \quad \mathbf{b}_j^\top \mathbf{b}_k = 0 \text{ for } j \neq k$$

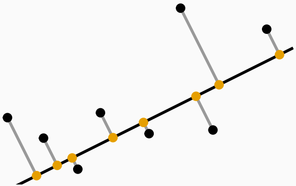
where $m < p$, the projection of $\mathbf{x}$ onto the m-dimensional linear subspace $V_m = \mathrm{span}(\mathbf{b}_1, \dots, \mathbf{b}_m)$ is

$$\hat{\mathbf{x}} = \sum_{j=1}^{m} (\mathbf{x}^\top \mathbf{b}_j)\mathbf{b}_j = \underbrace{\left(\sum_{j=1}^{m} \mathbf{b}_j \mathbf{b}_j^\top \right)}_{\text{Projection matrix}} \mathbf{x}$$

The projection is **orthogonal**, i.e.

$$(\mathbf{x} - \hat{\mathbf{x}})^\top \mathbf{b}_j = 0$$

for all $\mathbf{b}_j$.



Rayleigh Quotient

Let $\mathbf{A} \in \mathbb{R}^{k \times k}$ be a symmetric matrix. For $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^k$ define

$$J(\mathbf{x}) = \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

$J(\mathbf{x})$ is called the **Rayleigh Quotient** for $\mathbf{A}$.

Maximizing the Rayleigh Quotient

The maximization problem

$$\max_{\mathbf{x}} J(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x}^\top \mathbf{x} = 1$$

is solved by a **unit eigenvector** $\mathbf{x}$ of $\mathbf{A}$ corresponding to the **largest eigenvalue** λ of $\mathbf{A}$.

Note: $-\mathbf{x}$ is also a solution.

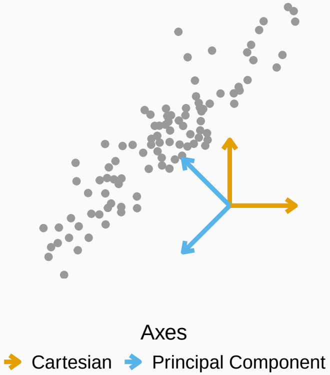
Principal Component Analysis (PCA) (I)

Goal: Given continuous data, find an orthogonal coordinate system such that the variance of the data is maximal along each direction.

Given data points $\mathbf{x}_1, \dots, \mathbf{x}_n$ and a unit vector $\mathbf{r}$, the **variance of the data along $\mathbf{r}$** is

$$S(\mathbf{r}) = \sum_{l=1}^{n} (\mathbf{r}^\top(\mathbf{x}_l - \overline{\mathbf{x}}))^2 = (n-1)\mathbf{r}^\top \widehat{\boldsymbol{\Sigma}} \mathbf{r}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix.



Principal Component Analysis (PCA) (II)

Direction with maximal variance: Find $\mathbf{r}$ such that

$$\max_{\mathbf{r}} S(\mathbf{r}) \quad \text{subject to} \quad ||\mathbf{r}||^2 = \mathbf{r}^\top \mathbf{r} = 1$$

- This is the same problem as maximizing the **Rayleigh Quotient** for the matrix $\widehat{\boldsymbol{\Sigma}}$.
- The **solution** is the eigenvector $\mathbf{r}_1$ of $\widehat{\boldsymbol{\Sigma}}$ corresponding to the largest eigenvalue λ_1.

How do we find the other directions?

Project data on orthogonal complement of $\mathbf{r}_1$, i.e.

$$\hat{\mathbf{x}}_l = \left(\mathbf{I}_p - \mathbf{r}_1 \mathbf{r}_1^\top\right) \mathbf{x}_l$$

and repeat the procedure above.

Intermezzo: Pre-processing

Data is often pre-processed before it is used in computational methods.

Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, let

- $\mathbf{m}_r \in \mathbb{R}^n$ be the vector of row-means,
- $\mathbf{m}_c \in \mathbb{R}^p$ be the vector of column-means, and
- $\mathbf{s} \in \mathbb{R}^p$ be the vector of per-column standard deviations.

Then (with $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$)

- the matrix $\mathbf{X} - \mathbf{m}_r \mathbf{1}_p^\top$ has row means zero (**row-centred**),
- the matrix $\mathbf{X} - \mathbf{1}_n \mathbf{m}_r^\top$ has column means zero (**column-centred**), and
- the matrix $\mathbf{X} \operatorname{diag}(1/\mathbf{s})$ has column standard deviations one (**standardised columns**)

Principal Component Analysis (PCA) (III)

Computational Procedure:

1. **Centre** (and possibly **standardise**) the columns of the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$
2. Calculate the **empirical covariance matrix** $\widehat{\mathbf{\Sigma}} = \dfrac{1}{n-1}\mathbf{X}^\top\mathbf{X}$
3. Determine the **eigenvalues** λ_j and corresponding orthonormal **eigenvectors** $\mathbf{r}_j$ of $\widehat{\mathbf{\Sigma}}$ for $j = 1, \dots, p$ and order them such that
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$
4. The vectors $\mathbf{r}_j$ give the direction of the **principal components (PC)** $\mathbf{r}_j^\top\mathbf{x}$ and the eigenvalues λ_j are the **variances along the PC directions**

Note: Set $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_p)$ and $\mathbf{D} = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ then
$$\widehat{\mathbf{\Sigma}} = \mathbf{R}\mathbf{D}\mathbf{R}^\top \quad \text{and} \quad \mathbf{R}^\top\mathbf{R} = \mathbf{R}\mathbf{R}^\top = \mathbf{I}_p$$

Recall: For a matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$ with eigenvalues $\lambda_1, \dots, \lambda_k$ it holds that

$$\operatorname{tr}(\mathbf{A}) = \sum_{j=1}^{k} \lambda_j$$

For the empirical covariance matrix $\widehat{\mathbf{\Sigma}}$ and the variance of the j-th feature $\operatorname{Var}[x_j]$

$$\operatorname{tr}(\widehat{\mathbf{\Sigma}}) = \sum_{j=1}^{p} \operatorname{Var}[x_j] = \sum_{j=1}^{p} \lambda_j$$

is called the **total variation**.

Using only the first $m < p$ principal components leads to

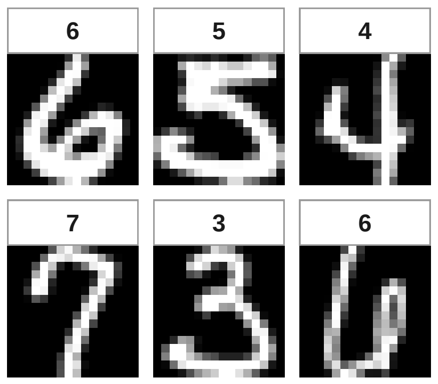
$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p} \cdot 100\% \quad \text{of } \textbf{explained variance}$$

PCA and Dimension Reduction: Example (I)

Variant of the MNIST handwritten digits dataset

($n = 7291$, 16×16 greyscale images, i.e. $p = 256$)

Digit	Frequency
0	0.16
1	0.14
2	0.10
3	0.09
4	0.09
5	0.08
6	0.09
7	0.09
8	0.07
9	0.09



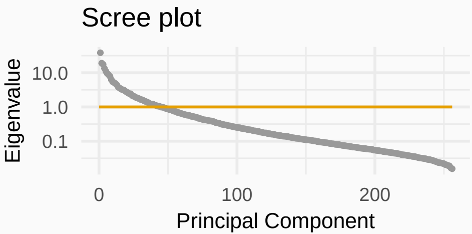
PCA and Dimension Reduction: Example (II)

For standardized variables

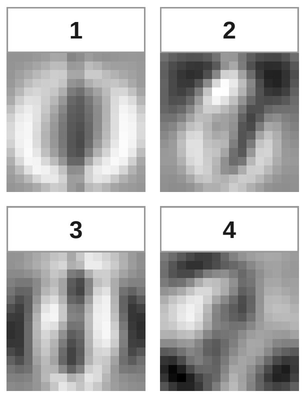
$$\mathrm{tr}(\widehat{\Sigma}) = p$$

Typical selection rule: Components with

$$\lambda_j \geq \frac{1}{p}\,\mathrm{tr}(\widehat{\Sigma}) \quad (= \ 1)$$



Visualisations of the first four principal components

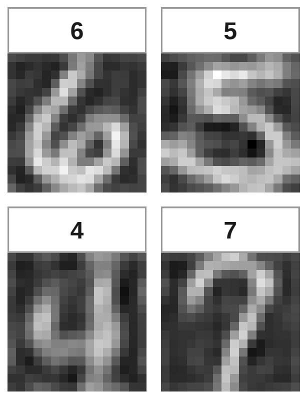


PCA and Dimension Reduction: Example (III)

Using the selection rule leads to 44 components. Using the projection

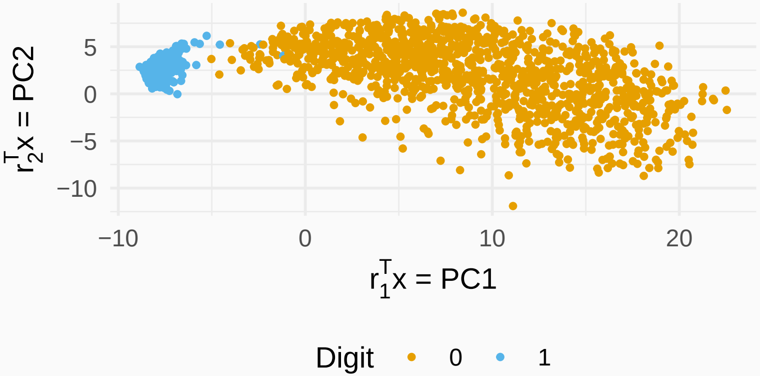
$$\hat{\mathbf{x}} = \left(\sum_{j=1}^{44} \mathbf{r}_j \mathbf{r}_j^\top \right) \mathbf{x}$$

creates a **reconstruction** of $\mathbf{x}$.



PCA and Dimension Reduction: Example (IV)

Projecting the digits onto the first two principal component directions gives a very clear distinction of digits 0 and 1.



Importance of standardisation (I)

The overall issue: **Subjectivity vs Objectivity**

(Co-)variance is scale dependent: If we have a sample (size n) of variables x and y, then their empirical covariance is

$$s_{xy} = \frac{1}{n-1}\sum_{l=1}^{n}(x_l - \bar{x})(y_l - \bar{y})$$

If x is scaled by a factor c, i.e. $z = c \cdot x$, then

$$\begin{aligned} s_{zy} &= \frac{1}{n-1}\sum_{l=1}^{n}(z_l - \bar{z})(y_l - \bar{y}) \\ &= \frac{1}{n-1}\sum_{l=1}^{n}(c \cdot x_l - c \cdot \bar{x})(y_l - \bar{y}) = c \cdot s_{xy} \end{aligned}$$

Importance of standardisation (II)

(Co-)variance is scale dependent: $s_{zy} = c \cdot s_{xy}$ where $z = c \cdot x$

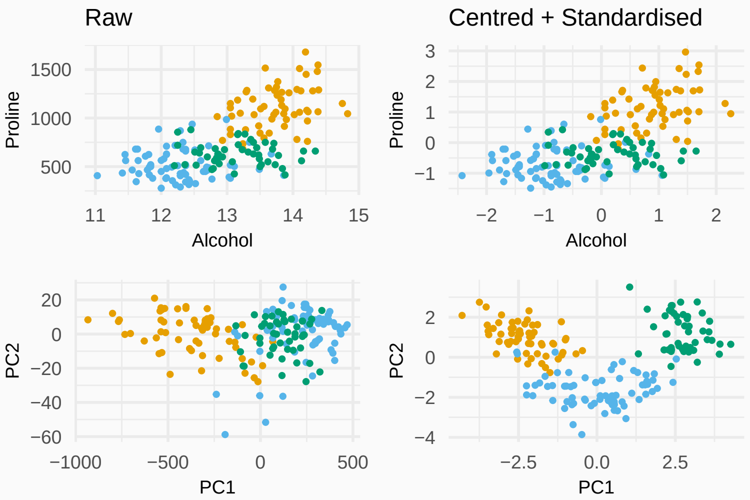
- By scaling variables we can therefore make them as large/influential or small/insignificant as we want, which is a very **subjective** process
- By standardising variables we can get of rid of **scaling** and reach an **objective** point-of-view
- **Do we get rid of information?**
  - The **typical range** of a variable is compressed
  - The overall shape of the data is preserved
  - Outliers will still be outliers

UCI Wine Data Set[1]

- Results of a chemical analysis on multiple samples from three different origins of wine
- $n = 178$ samples (59 origin 1, 71 origin 2, 48 origin 3)
- $p = 13$ features
  - e.g. alcohol in %, ash, colour intensity, magnesium, …

[1] `https://archive.ics.uci.edu/ml/datasets/Wine`

Importance of standardisation (III)



Singular Value Decomposition

Singular Value Decomposition (SVD)

The **singular value decomposition (SVD)** of a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, $n \geq p$, is

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

where $\mathbf{U} \in \mathbb{R}^{n \times p}$ and $\mathbf{V} \in \mathbb{R}^{p \times p}$ with

$$\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p \quad \text{and} \quad \mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$$

and $\mathbf{D} \in \mathbb{R}^{p \times p}$ is diagonal. Usually

$$d_{11} \geq d_{22} \geq \cdots \geq d_{pp}$$

Note: Due to the **orthogonality conditions** on $\mathbf{U}$ and $\mathbf{V}$

$$\mathbf{X}\mathbf{X}^\top\mathbf{U} = \mathbf{U}\mathbf{D}^2$$
$$\mathbf{X}^\top\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{D}^2$$

SVD and PCA

In PCA the empirical covariance matrix $\widehat{\mathbf{\Sigma}}$ is in focus, whereas SVD focuses on the data matrix $\mathbf{X}$ directly.

Connection: For centred variables

$$\widehat{\mathbf{\Sigma}} = \frac{\mathbf{X}^\top \mathbf{X}}{n-1} = \frac{\mathbf{V}\mathbf{D}\mathbf{U}^\top \mathbf{U}\mathbf{D}\mathbf{V}^\top}{n-1} = \mathbf{V}\left(\frac{\mathbf{D}^2}{n-1}\right)\mathbf{V}^\top$$

The PC directions are in $\mathbf{V}$ and the eigenvalues of $\widehat{\mathbf{\Sigma}}$ are $d_{jj}^2/(n-1)$.

Note: This is how PCA is typically calculated. SVD is a **more general tool** and is used in many other contexts as well.

SVD and best rank-q-approximation / dimension reduction

Write $\mathbf{u}_j$ and $\mathbf{v}_j$ for the columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. Then

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{j=1}^{p} d_{jj} \underbrace{\mathbf{u}_j \mathbf{v}_j^\top}_{\text{rank-1-matrix}}$$

Best rank-q-approximation: For $q < p$

$$\mathbf{X}_q = \sum_{j=1}^{q} d_{jj} \mathbf{u}_j \mathbf{v}_j^\top$$

with **approximation error**

$$\left\| \mathbf{X} - \mathbf{X}_q \right\|_F^2 = \left\| \sum_{j=q+1}^{p} d_{jj} \mathbf{u}_j \mathbf{v}_j^\top \right\|_F^2 = \sum_{j=q+1}^{p} d_j^2$$

Connections to Discriminant Analysis

Discriminant Analysis and the Inverse Covariance Matrix

From PCA or SVD we get $\widehat{\boldsymbol{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ where $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$ and $d_{11} \geq \cdots \geq d_{pp} \geq 0$. Then

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^\top = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{V}^\top = \left(\widehat{\boldsymbol{\Sigma}}^{-1/2}\right)^\top \widehat{\boldsymbol{\Sigma}}^{-1/2}$$

where $(\mathbf{D}^{-1/2})_{jj} := 1/\sqrt{d_{jj}}$ and $\widehat{\boldsymbol{\Sigma}}^{-1/2} := \mathbf{D}^{-1/2}\mathbf{V}^\top$.

In LDA the term involving the inverse covariance matrix is then

$$\begin{aligned}
(\mathbf{x} - \widehat{\boldsymbol{\mu}})^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}) &= (\mathbf{x} - \widehat{\boldsymbol{\mu}})^\top \left(\widehat{\boldsymbol{\Sigma}}^{-1/2}\right)^\top \widehat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x} - \widehat{\boldsymbol{\mu}}) \\
&= \left(\mathbf{V}^\top(\mathbf{x} - \widehat{\boldsymbol{\mu}})\right)^\top \mathbf{D}^{-1} \left(\mathbf{V}^\top(\mathbf{x} - \widehat{\boldsymbol{\mu}})\right) \\
&= \sum_{j=1} \frac{1}{d_{jj}}(\tilde{x}_j - \tilde{\mu}_j)^2
\end{aligned}$$

Inverse of the eigenvalues can lead to **numerical instability**.

Regularised Discriminant Analysis (RDA)

The empirical covariance matrix used by LDA can be **stabilized**:

$$\widehat{\boldsymbol{\Sigma}}_\lambda := \widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}_p = \mathbf{V}(\mathbf{D} + \lambda \mathbf{I}_p)\mathbf{V}^\top$$

where $\lambda > 0$ is a tuning parameter.

- Using $\widehat{\boldsymbol{\Sigma}}_\lambda$ in LDA is called **regularised discriminant analysis (RDA)**.
- Instead of $1/d_{jj}$ the scaling factors are now $1/(d_{jj} + \lambda)$.
- For small d_{jj} this can lead to **numerical stability**, whereas large d_{jj} are not much affected.
- For increasingly large λ the d_{jj} will have diminishing impact and RDA starts to become **nearest centroids**.
- RDA can be used with QDA as well by considering:

$$\widehat{\boldsymbol{\Sigma}}_{i,\lambda} := \underbrace{\widehat{\boldsymbol{\Sigma}}_i}_{\text{QDA}} + \lambda \underbrace{\widehat{\boldsymbol{\Sigma}}}_{\text{LDA}}$$

Take-home message

- Principal component analysis gives a convenient decomposition of the variance of the data
- Pre-processing (centring and standardisation) is important if data is collected on different scales
- Singular value decomposition is a universal workhorse for in numerical methods