Machine Learning

Principal Component Analysis (PCA)

Hiroshi Shimodaira

March 2026

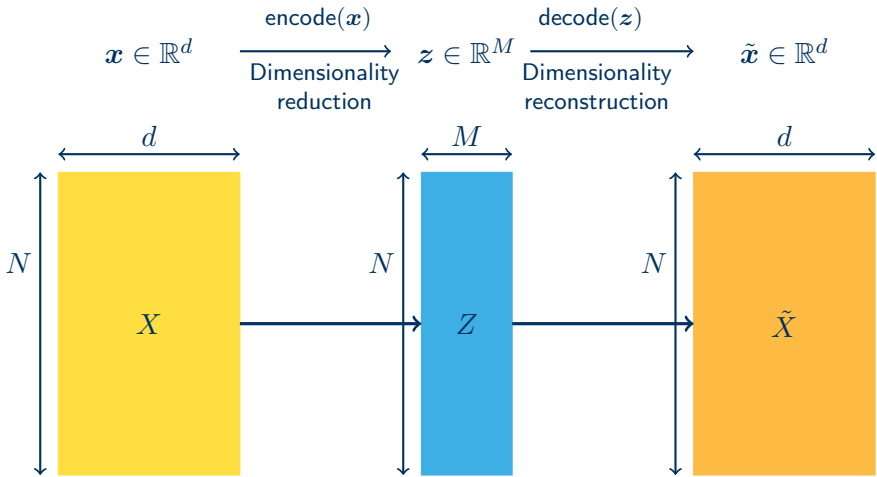
Ver. 1.0.1

partially based on Kia Nazarpour's slides

Topics

- Motivation behind the use of PCA
- Geometrical interpretation of PCA
- PCA with maximum variance
- PCA with minimum reconstruction error
- Orthogonal transformation
- Practical tips
  - Data standardisation

Dimensionality reduction and reconstruction



Why we use PCA?

- Visualisation
- Exploration
- Compression

Note that

- Dimensionality reduction entails a loss of information.
- We'd like to reduce the dimensionality while minimising information loss.

Dimensionality reduction and data visualisation

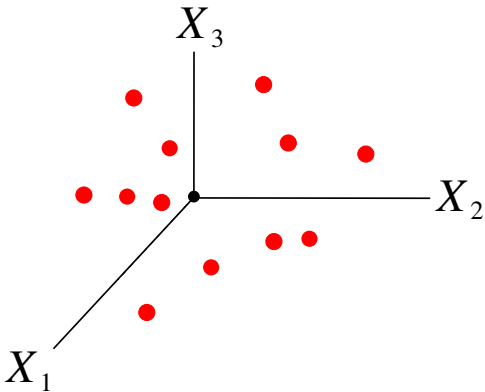
- High-dimensional data are difficult to understand and visualise.

Dimensionality reduction and data visualisation

- High-dimensional data are difficult to understand and visualise.
- Consider dimensionality reduction of data for visualisation

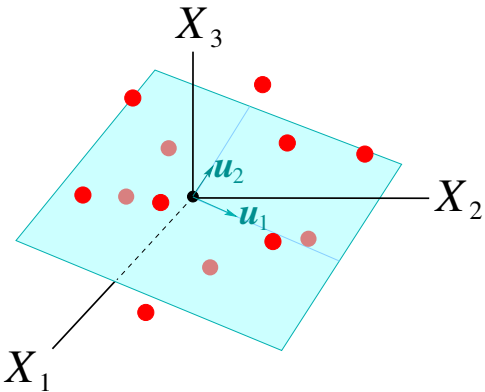
Dimensionality reduction and data visualisation

- High-dimensional data are difficult to understand and visualise.
- Consider dimensionality reduction of data for visualisation



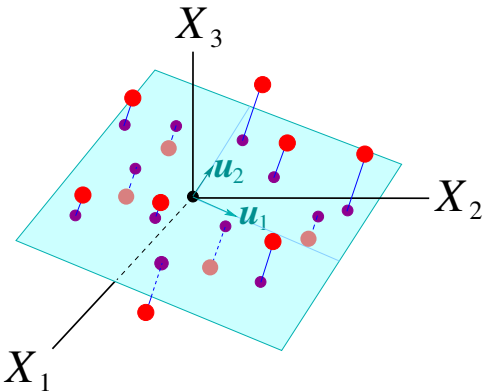
Dimensionality reduction and data visualisation

- High-dimensional data are difficult to understand and visualise.
- Consider dimensionality reduction of data for visualisation



Dimensionality reduction and data visualisation

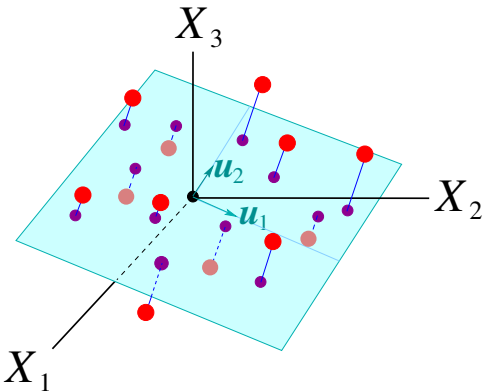
- High-dimensional data are difficult to understand and visualise.
- Consider dimensionality reduction of data for visualisation

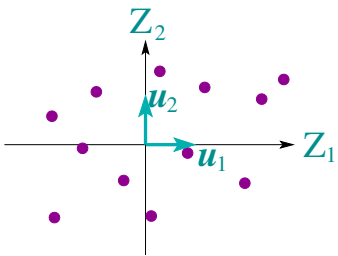


Project each sample in 3D onto a 2D plane

Dimensionality reduction and data visualisation

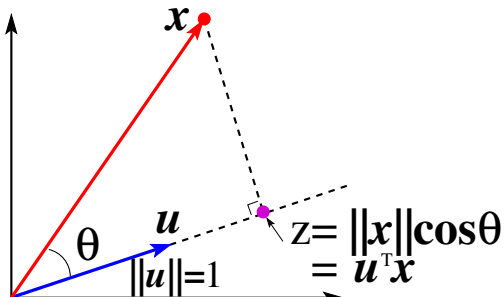
- High-dimensional data are difficult to understand and visualise.
- Consider dimensionality reduction of data for visualisation



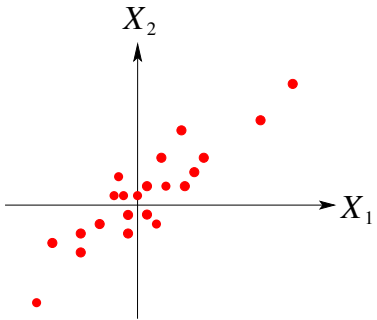


Project each sample in 3D onto a 2D plane

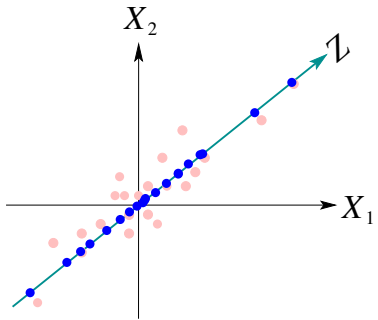
Orthogonal projection of data onto an axis



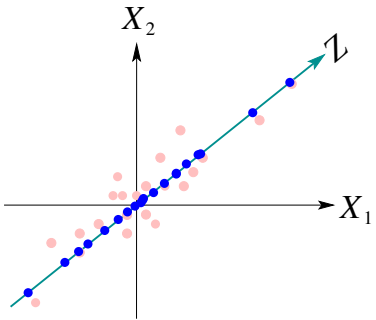
Optimal projection of 2D data onto 1D

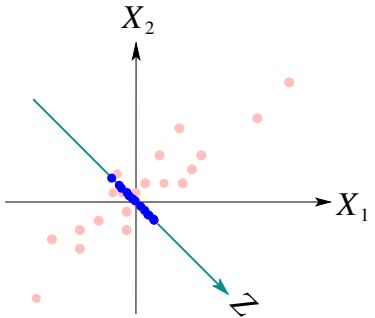


Optimal projection of 2D data onto 1D

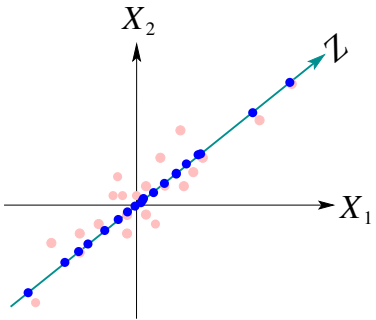


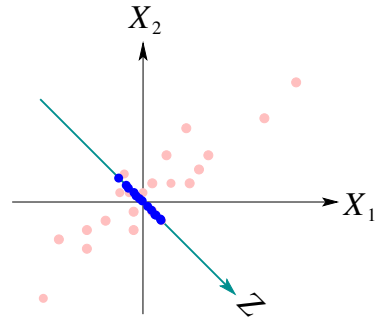
Optimal projection of 2D data onto 1D





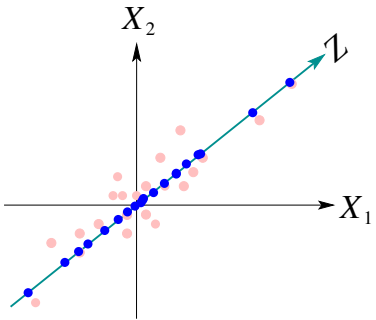
Optimal projection of 2D data onto 1D

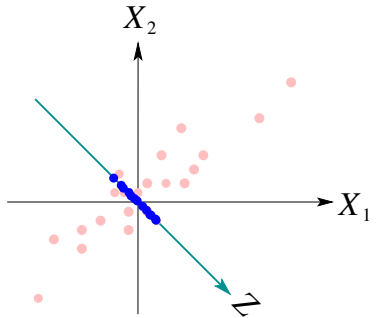




- Mapping 2D to 1D: $z_n = \mathbf{u}^T \mathbf{x}_n = u_1 x_{n1} + u_2 x_{n2}$

Optimal projection of 2D data onto 1D

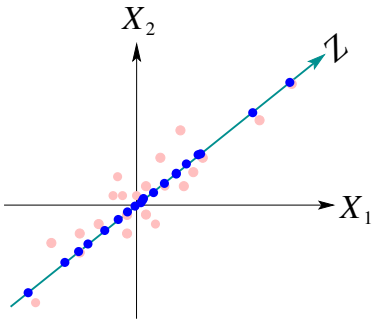


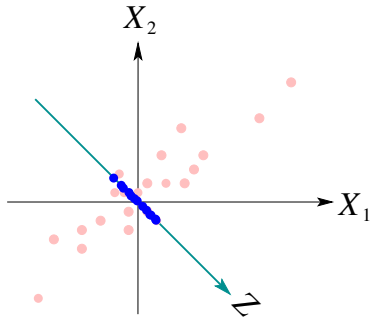


- Mapping 2D to 1D: $z_n = \mathbf{u}^T \mathbf{x}_n = u_1 x_{n1} + u_2 x_{n2}$
- Optimal mapping: $\max_{\mathbf{u}} \mathrm{Var}\,(z)$

$$\mathrm{Var}\,(z) = \frac{1}{N} \sum_{n=1}^{N} (z_n - \bar{z})^2$$

Optimal projection of 2D data onto 1D





- Mapping 2D to 1D: $z_n = \mathbf{u}^T \mathbf{x}_n = u_1 x_{n1} + u_2 x_{n2}$
- Optimal mapping: $\max_{\mathbf{u}} \mathrm{Var}\,(z)$

$$\mathrm{Var}\,(z) = \frac{1}{N} \sum_{n=1}^{N} (z_n - \bar{z})^2$$

- cf. least squares fitting (linear regression)

Problem formulation

- Data set: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \quad \boldsymbol{x}_i = \begin{bmatrix} x_{i1} & \ldots & x_{id} \end{bmatrix}^\top$

  NB: we assume all sample have been mean-normalised, i.e. $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \bar{\boldsymbol{x}}$

Problem formulation

- Data set: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ $\quad \boldsymbol{x}_i = \begin{bmatrix} x_{i1} & \ldots & x_{id} \end{bmatrix}^\top$

  NB: we assume all sample have been mean-normalised, i.e. $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \bar{\boldsymbol{x}}$

- Projection axes (basis vectors): $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M\}$, $\boldsymbol{u}_i \perp \boldsymbol{u}_j$ for $i \neq j$

$$\boldsymbol{u}_m = \begin{bmatrix} u_{m1} & \ldots & u_{md} \end{bmatrix}^\top \quad \cdots \; m\text{-th principal component}$$

Problem formulation

- Data set: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \quad \boldsymbol{x}_i = \begin{bmatrix} x_{i1} & \ldots & x_{id} \end{bmatrix}^\top$
  NB: we assume all sample have been mean-normalised, i.e. $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \bar{\boldsymbol{x}}$
- Projection axes (basis vectors): $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M\}$, $\boldsymbol{u}_i \perp \boldsymbol{u}_j$ for $i \neq j$

$$\boldsymbol{u}_m = \begin{bmatrix} u_{m1} & \ldots & u_{md} \end{bmatrix}^\top \quad \cdots \textit{ m-th principal component}$$

- Projection of $\boldsymbol{x}_i \in \mathbb{R}^d$ onto the basis vectors results in new coordinates in $\mathbb{R}^M$:

$$\boldsymbol{z}_i = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{iM} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^\top \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{u}_M^\top \boldsymbol{x}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^\top \\ \vdots \\ \boldsymbol{u}_M^\top \end{bmatrix} \boldsymbol{x}_i = U^\top \boldsymbol{x}_i \tag{1}$$

$$z_{im} = \boldsymbol{u}_m^\top \boldsymbol{x}_i = u_{m1} x_{i1} + \cdots + u_{md} x_{id} \tag{2}$$

Problem formulation

- Data set: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \quad \boldsymbol{x}_i = \begin{bmatrix} x_{i1} & \ldots & x_{id} \end{bmatrix}^\top$
  NB: we assume all sample have been mean-normalised, i.e. $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \bar{\boldsymbol{x}}$
- Projection axes (basis vectors): $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M\}$, $\boldsymbol{u}_i \perp \boldsymbol{u}_j$ for $i \neq j$

$$\boldsymbol{u}_m = \begin{bmatrix} u_{m1} & \ldots & u_{md} \end{bmatrix}^\top \quad \cdots \textit{ m-th principal component}$$

- Projection of $\boldsymbol{x}_i \in \mathbb{R}^d$ onto the basis vectors results in new coordinates in $\mathbb{R}^M$:

$$\boldsymbol{z}_i = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{iM} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^\top \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{u}_M^\top \boldsymbol{x}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^\top \\ \vdots \\ \boldsymbol{u}_M^\top \end{bmatrix} \boldsymbol{x}_i = U^\top \boldsymbol{x}_i \tag{1}$$

$$z_{im} = \boldsymbol{u}_m^\top \boldsymbol{x}_i = u_{m1} x_{i1} + \cdots + u_{md} x_{id} \tag{2}$$

- Inverse projection of $\boldsymbol{z}_i \in \mathbb{R}^M$ to $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^d$:

$$\tilde{\boldsymbol{x}}_i = \sum_{m=1}^{M} \boldsymbol{u}_m z_{im} = U \boldsymbol{z}_i = U U^\top \boldsymbol{x}_i = \sum_{m=1}^{M} \boldsymbol{u}_m \boldsymbol{u}_m^\top \boldsymbol{x}_i \tag{3}$$

Problem formulation (*cont.*)

- Variance of projected data with m-th basis $\boldsymbol{u}_m$

$$\text{Var}\,(z_m) = \frac{1}{N}\sum_{i=1}^{N}(z_{im} - \bar{z}_m)^2 \;=\; \frac{1}{N}\sum_{i=1}^{N} z_{im}^2 \tag{4}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{u}_m^\top \boldsymbol{x}_i)^2 \tag{5}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}_m^\top \boldsymbol{x}_i\, \boldsymbol{x}_i^\top \boldsymbol{u}_m \tag{6}$$

$$= \boldsymbol{u}_m^\top \left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\boldsymbol{u}_m \tag{7}$$

$$= \boldsymbol{u}_m^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_m \tag{8}$$

NB: $\boldsymbol{\Sigma}$ is the covariance matrix of $\{\boldsymbol{x}_i\}_{i=1}^N$.

Problem formulation (*cont.*)

PCA with maximum variance

$$\max_{\{\boldsymbol{u}_m\}} \sum_{m=1}^{M} \mathsf{Var}\,(z_m) \tag{9}$$

PCA with minimum reconstruction error

$$\min_{\{\boldsymbol{u}_m\}} \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2 \tag{10}$$

Reconstruction error:

$$E = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2 = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - UU^\top \boldsymbol{x}_i\|^2 \tag{11}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \sum_{m=1}^{M} \boldsymbol{u}_m \boldsymbol{u}_m^\top \boldsymbol{x}_i\|^2 \tag{12}$$

PCA with maximum variance

Let's consider the variance on the first axis spanned by $\boldsymbol{u}_1$.

$$\text{Var}\,(z_1) = \boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 \tag{13}$$

NB: this is scale dependent.

- Optimisation problem:

$$\max_{\boldsymbol{u}_1} \boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 \tag{14}$$

$$\text{s.t.} \quad \boldsymbol{u}^\top \boldsymbol{u} = 1 \tag{15}$$

- Lagrangian

$$L = \boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 + \lambda_1 (1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1), \quad \lambda_1 \geq 0 \tag{16}$$

PCA with maximum variance (*cont.*)

$$\frac{\partial L}{\partial \boldsymbol{u}_1} = 2\boldsymbol{\Sigma}\,\boldsymbol{u}_1 - 2\lambda_1 \boldsymbol{u}_1 = \mathbf{0} \tag{17}$$

$$\rightarrow\ \boldsymbol{\Sigma}\,\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$$

meaning $\boldsymbol{u}_1$ is an eigenvector of $\boldsymbol{\Sigma}$ and λ_1 is the corresponding eigenvalue.

$$\frac{\partial L}{\partial \lambda_1} = 1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1 \tag{18}$$

Using the above gives:

$$\mathrm{Var}\,(z_1) = \boldsymbol{u}_1^\top \boldsymbol{\Sigma}\,\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^\top \boldsymbol{u}_1 = \lambda_1 \tag{19}$$

$\max \mathrm{Var}\,(z_1)$ is achieved by letting λ_1 be the largest eigen value of $\boldsymbol{\Sigma}$ and $\boldsymbol{u}_1$ be the corresponding eigenvector.

PCA with maximum variance (*cont.*)

- How to find second axis spanned by $\boldsymbol{u}_2$?

PCA with maximum variance (*cont.*)

- How to find second axis spanned by $\boldsymbol{u}_2$?
- Let's create a new data matrix X_2 by subtracting the effect of the first principal component.

$$X_2 = \begin{bmatrix} \boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1 \\ \vdots \\ \boldsymbol{x}_N - \tilde{\boldsymbol{x}}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\top \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N - \boldsymbol{u}_1\boldsymbol{u}_1^\top \boldsymbol{x}_N \end{bmatrix} \tag{20}$$

PCA with maximum variance (*cont.*)

- How to find second axis spanned by $\boldsymbol{u}_2$?
- Let's create a new data matrix X_2 by subtracting the effect of the first principal component.

$$X_2 = \begin{bmatrix} \boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1 \\ \vdots \\ \boldsymbol{x}_N - \tilde{\boldsymbol{x}}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\top \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N - \boldsymbol{u}_1\boldsymbol{u}_1^\top \boldsymbol{x}_N \end{bmatrix} \tag{20}$$

- Using the same method as the one for $\boldsymbol{u}_1$, we can find $\boldsymbol{u}_2$.

PCA with maximum variance (*cont.*)

- How to find second axis spanned by $\boldsymbol{u}_2$?
- Let's create a new data matrix X_2 by subtracting the effect of the first principal component.

$$X_2 = \begin{bmatrix} \boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1 \\ \vdots \\ \boldsymbol{x}_N - \tilde{\boldsymbol{x}}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\top \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N - \boldsymbol{u}_1\boldsymbol{u}_1^\top \boldsymbol{x}_N \end{bmatrix} \tag{20}$$

- Using the same method as the one for $\boldsymbol{u}_1$, we can find $\boldsymbol{u}_2$.
- It is known that $\boldsymbol{u}_m,\ m = 1, \ldots, M$, is the eigenvector of the m-th largest eigen value λ_m of $\boldsymbol{\Sigma}$.

PCA with minimum reconstruction error

Assuming we use the first principal component $\boldsymbol{u}_1$ only, the reconstruction error is give as:

$$E_1 = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i - \boldsymbol{u}_1\boldsymbol{u}_1^\top\boldsymbol{x}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{u}_1\boldsymbol{u}_1^\top\boldsymbol{x}_i)^\top(\boldsymbol{x}_i - \boldsymbol{u}_1\boldsymbol{u}_1^\top\boldsymbol{x}_i) \tag{21}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i^\top\boldsymbol{x}_i - 2\boldsymbol{x}_i^\top\boldsymbol{u}_1\boldsymbol{u}_1^\top\boldsymbol{x}_i + \boldsymbol{x}_i^\top\boldsymbol{u}_1^\top\boldsymbol{u}_1\boldsymbol{u}_1\boldsymbol{u}_1^\top\boldsymbol{x}_i) \tag{22}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i^\top\boldsymbol{x}_i - \boldsymbol{u}_1^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{u}_1) \qquad \text{NB: } \boldsymbol{u}_1^\top\boldsymbol{u}_1 = 1 \tag{23}$$

$$= \frac{1}{N}\left(\sum_{i=1}^{N}\boldsymbol{x}_i^\top\boldsymbol{x}_i\right) - \boldsymbol{u}_1^\top\boldsymbol{\Sigma}\,\boldsymbol{u}_1 \tag{24}$$

PCA with minimum reconstruction error (*cont.*)

- Optimisation problem

$$\min_{\boldsymbol{u}_1} -\boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 \tag{25}$$

$$\text{s.t.} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1 \tag{26}$$

PCA with minimum reconstruction error (*cont.*)

- Optimisation problem

$$\min_{\boldsymbol{u}_1} -\boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 \tag{25}$$

$$\text{s.t.} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1 \tag{26}$$

- Lagrangian

$$L = -\boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 - \lambda_1(1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1), \quad \lambda_1 \geq 0 \tag{27}$$

PCA with minimum reconstruction error (*cont.*)

- Optimisation problem

$$\min_{\boldsymbol{u}_1} -\boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 \tag{25}$$

$$\text{s.t.} \quad \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1 \tag{26}$$

- Lagrangian

$$L = -\boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 - \lambda_1(1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1), \quad \lambda_1 \geq 0 \tag{27}$$

$$\frac{\partial L}{\partial \boldsymbol{u}_1} = -2\boldsymbol{\Sigma}\boldsymbol{u}_1 + \lambda_1 \boldsymbol{u}_1 = \mathbf{0} \qquad \rightarrow \ \boldsymbol{\Sigma}\, \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$$

$$\frac{\partial L}{\partial \lambda_1} = 1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1 = 0 \tag{28}$$

PCA with minimum reconstruction error (*cont.*)

The reconstruction error E_1 is given as:

$$E_1 = \frac{1}{N}\left(\sum_{i=1}^{N} \boldsymbol{x}_i^\top \boldsymbol{x}_i\right) - \lambda_1 \tag{29}$$

This is minimised by choosing the eigenvector that corresponds to the largest eigenvalue of $\boldsymbol{\Sigma}$.
For the second principal component $\boldsymbol{u}_2$,

$$E_2 = \frac{1}{N}\left(\sum_{i=1}^{N} \boldsymbol{x}_i^\top \boldsymbol{x}_i\right) - \boldsymbol{u}_1^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_1 - \boldsymbol{u}_2^\top \boldsymbol{\Sigma}\, \boldsymbol{u}_2 \tag{30}$$

It can be shown that $\boldsymbol{u}_2$ is the eigenvector that corresponds to the second largest eigenvalue.

PCA - properties

- $\text{Var}\,(z_m) = \text{Var}\left(\boldsymbol{u}_m^\top \boldsymbol{x}\right) = \lambda_m$
- Covariance matrix

$$\boldsymbol{\Sigma} = \frac{1}{N}\mathbf{X}^\top\mathbf{X} = U_d Q_d U_d^\top \tag{31}$$

$$= \begin{bmatrix}\boldsymbol{u}_1 & \dots & \boldsymbol{u}_d\end{bmatrix} \begin{bmatrix}\lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d\end{bmatrix} \begin{bmatrix}\boldsymbol{u}_1 & \dots & \boldsymbol{u}_d\end{bmatrix}^\top \tag{32}$$

- Covariance matrix in the projected/transformed space with PCA

$$\boldsymbol{\Sigma_z} = \begin{bmatrix}\lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M\end{bmatrix} \tag{33}$$

- The Pearson's correlation coefficient $\rho(z_i, z_j) = \delta_{ij}$

PCA - evaluation

- Plotting of cumulative explained variance ratio:

$$\frac{\sum_{i=1}^{m} \mathrm{Var}\left(z_i\right)}{\sum_{i=1}^{d} \mathrm{Var}\left(z_i\right)} = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \tag{34}$$

Example - film review toy data

	Australia	*Body of Lies*	*Burn After*	*Hancock*	*Milk*	*Rev Road*
Denby	3	7	4	9	9	7
McCarthy	7	5	5	3	8	8
M'stern	7	5	5	0	8	4
Puig	5	6	8	5	9	8
Travers	5	8	8	8	10	9
Turan	7	7	8	4	7	8

Example - film review toy data

	Australia	*Body of Lies*	*Burn After*	*Hancock*	*Milk*	*Rev Road*
Denby	3	7	4	9	9	7
McCarthy	7	5	5	3	8	8
M'stern	7	5	5	0	8	4
Puig	5	6	8	5	9	8
Travers	5	8	8	8	10	9
Turan	7	7	8	4	7	8

$$\Sigma = \begin{pmatrix} 2.66 & -1.07 & 0.53 & -4.67 & -1.20 & -0.67 \\ -1.07 & 1.47 & 1.07 & 3.27 & 0.60 & 1.27 \\ 0.53 & 1.07 & 3.47 & 0.67 & 0.20 & 1.87 \\ -4.67 & 3.27 & 0.67 & 10.97 & 2.30 & 3.67 \\ -1.20 & 0.60 & 0.20 & 2.30 & 1.10 & 0.60 \\ -0.67 & 1.27 & 1.87 & 3.67 & 0.60 & 3.07 \end{pmatrix}$$

Example - film review toy data

	Australia	*Body of Lies*	*Burn After*	*Hancock*	*Milk*	*Rev Road*
Denby	3	7	4	9	9	7
McCarthy	7	5	5	3	8	8
M'stern	7	5	5	0	8	4
Puig	5	6	8	5	9	8
Travers	5	8	8	8	10	9
Turan	7	7	8	4	7	8

$$\Sigma = \begin{pmatrix} 2.66 & -1.07 & 0.53 & -4.67 & -1.20 & -0.67 \\ -1.07 & 1.47 & 1.07 & 3.27 & 0.60 & 1.27 \\ 0.53 & 1.07 & 3.47 & 0.67 & 0.20 & 1.87 \\ -4.67 & 3.27 & 0.67 & 10.97 & 2.30 & 3.67 \\ -1.20 & 0.60 & 0.20 & 2.30 & 1.10 & 0.60 \\ -0.67 & 1.27 & 1.87 & 3.67 & 0.60 & 3.07 \end{pmatrix} \quad U = \begin{pmatrix} -0.341 & 0.345 & 0.326 & -0.180 & 0.603 & -0.512 \\ 0.255 & 0.151 & -0.240 & -0.548 & 0.496 & 0.554 \\ 0.101 & 0.786 & -0.503 & 0.028 & -0.280 & -0.198 \\ 0.827 & -0.154 & 0.096 & -0.182 & 0.025 & -0.450 \\ 0.181 & -0.065 & -0.341 & 0.733 & 0.556 & 0.015 \\ 0.304 & 0.461 & 0.676 & 0.309 & -0.047 & 0.375 \end{pmatrix}$$

Example - film review toy data

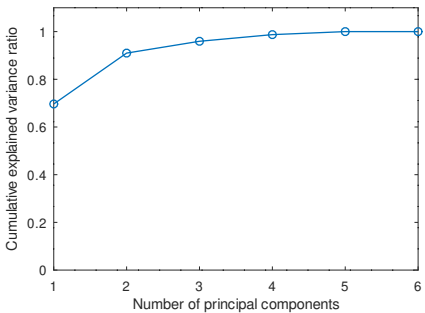
	Australia	*Body of Lies*	*Burn After*	*Hancock*	*Milk*	*Rev Road*
Denby	3	7	4	9	9	7
McCarthy	7	5	5	3	8	8
M'stern	7	5	5	0	8	4
Puig	5	6	8	5	9	8
Travers	5	8	8	8	10	9
Turan	7	7	8	4	7	8

$$\mathbf{\Sigma} = \begin{pmatrix} 2.66 & -1.07 & 0.53 & -4.67 & -1.20 & -0.67 \\ -1.07 & 1.47 & 1.07 & 3.27 & 0.60 & 1.27 \\ 0.53 & 1.07 & 3.47 & 0.67 & 0.20 & 1.87 \\ -4.67 & 3.27 & 0.67 & 10.97 & 2.30 & 3.67 \\ -1.20 & 0.60 & 0.20 & 2.30 & 1.10 & 0.60 \\ -0.67 & 1.27 & 1.87 & 3.67 & 0.60 & 3.07 \end{pmatrix} \quad U = \begin{pmatrix} -0.341 & 0.345 & 0.326 & -0.180 & 0.603 & -0.512 \\ 0.255 & 0.151 & -0.240 & -0.548 & 0.496 & 0.554 \\ 0.101 & 0.786 & -0.503 & 0.028 & -0.280 & -0.198 \\ 0.827 & -0.154 & 0.096 & -0.182 & 0.025 & -0.450 \\ 0.181 & -0.065 & -0.341 & 0.733 & 0.556 & 0.015 \\ 0.304 & 0.461 & 0.676 & 0.309 & -0.047 & 0.375 \end{pmatrix}$$

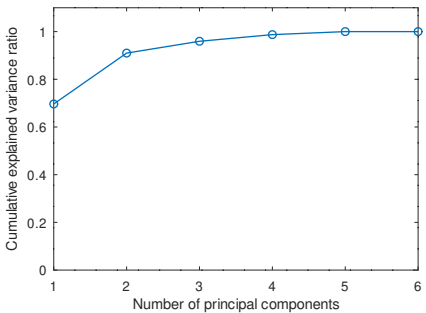
$$Q = \begin{pmatrix} 15.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4.85 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.13 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.634 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.288 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

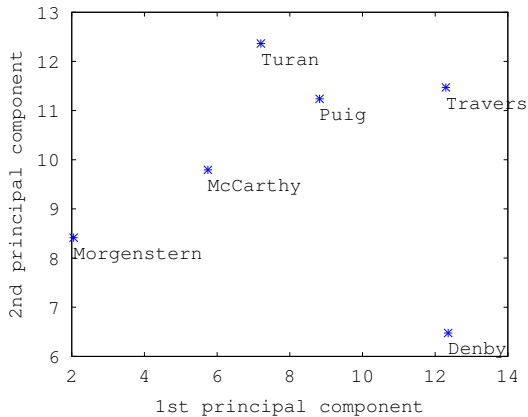
where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_6)$ and $(Q)_{ii} = \lambda_i$ for $i = 1, \ldots, 6$

Example - film review toy data (*cont.*)



Example - film review toy data (*cont.*)

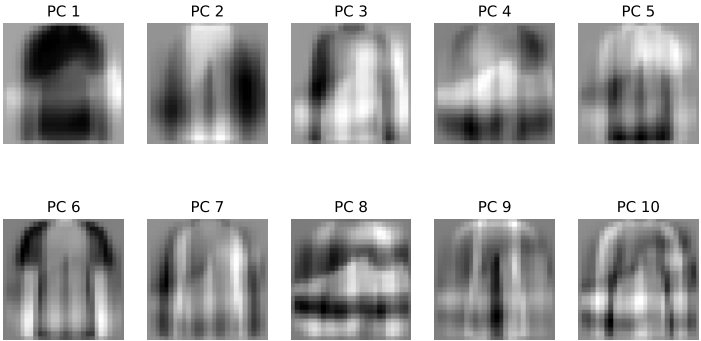


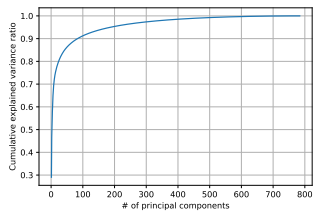


Example - image data (Fashion-MNIST)

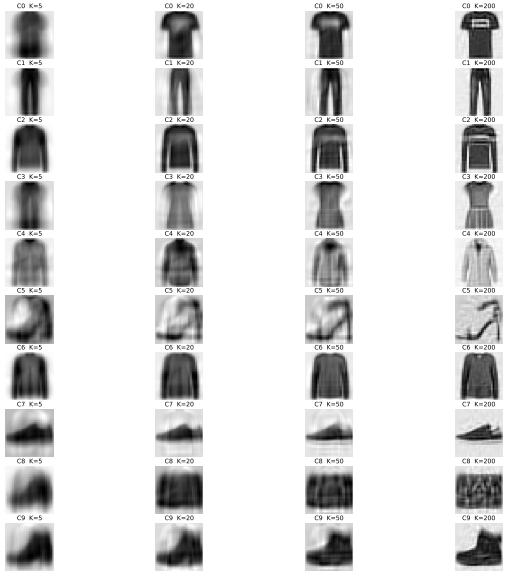
28x28 gray-scale images of 60000 samples
10 classes
https://github.com/zalandoresearch/fashion-mnist

Example - image data (Fashion-MNIST) (*cont.*)

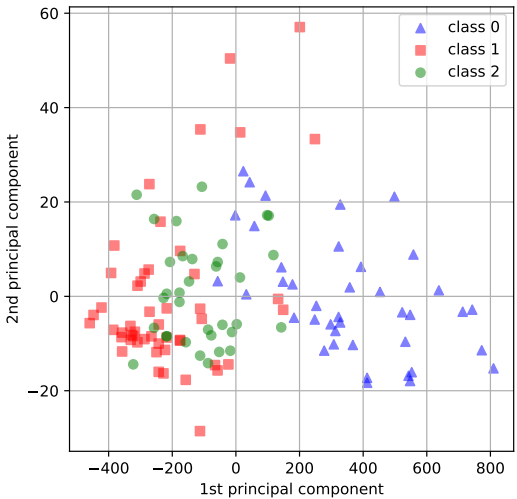




Example - image data (Fashion-MNIST) (*cont.*)



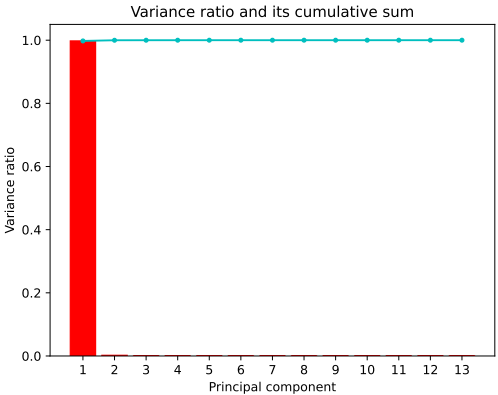
Practical tips - data standardisation



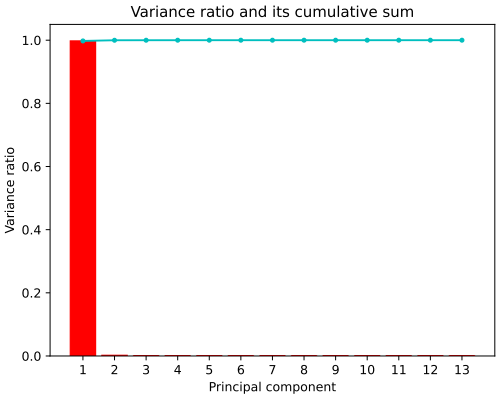
Dataset	UCI Wine data set
# instances	178
# attributes	13
# classes	3

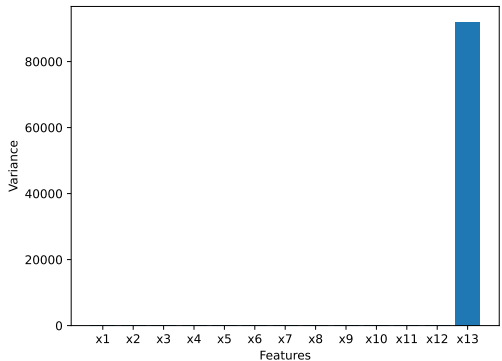
Credit: Sklearn - "importance of feature scaling"

Practical tips - data standardisation (*cont.*)

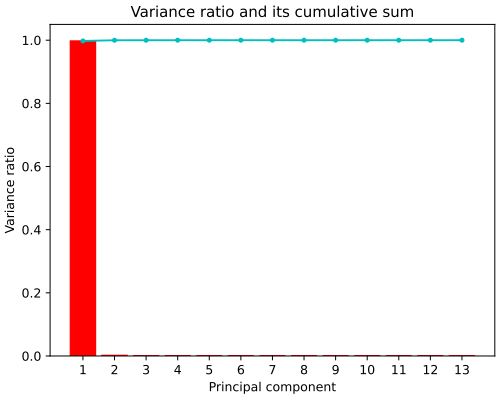


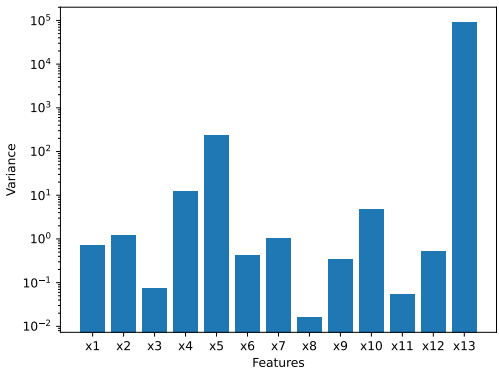
Practical tips - data standardisation (*cont.*)



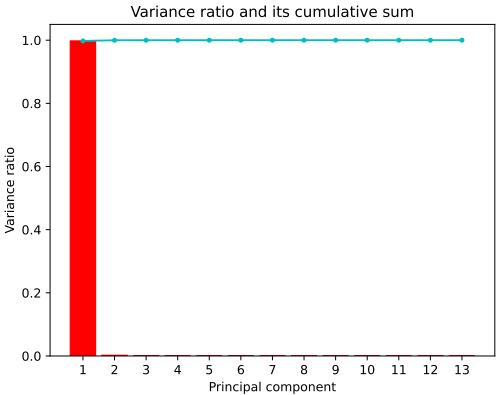


Practical tips - data standardisation (*cont.*)



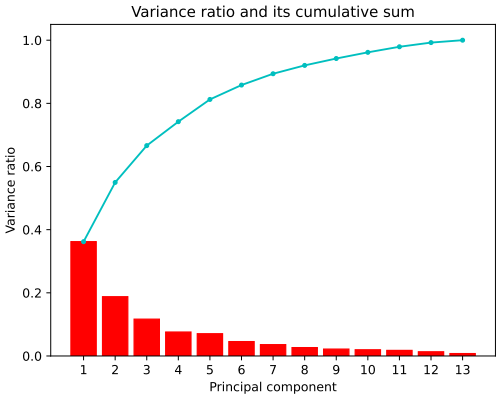


Practical tips - data standardisation (*cont.*)



Attribute	min	max	mean	std
x_1	11.0	14.8	13.0	0.81
x_2	0.74	5.80	2.34	1.12
x_3	1.36	3.23	2.36	0.27
x_4	10.6	30.0	19.5	3.34
x_5	70.0	162.0	99.7	14.3
x_6	0.98	3.88	2.30	0.63
x_7	0.34	5.08	2.03	1.00
x_8	0.13	0.66	0.36	0.12
x_9	0.41	3.58	1.59	0.57
x_{10}	1.28	13.0	5.06	2.32
x_{11}	0.48	1.71	0.96	0.23
x_{12}	1.27	4.00	2.61	0.71
x_{13}	278	1680	746	315

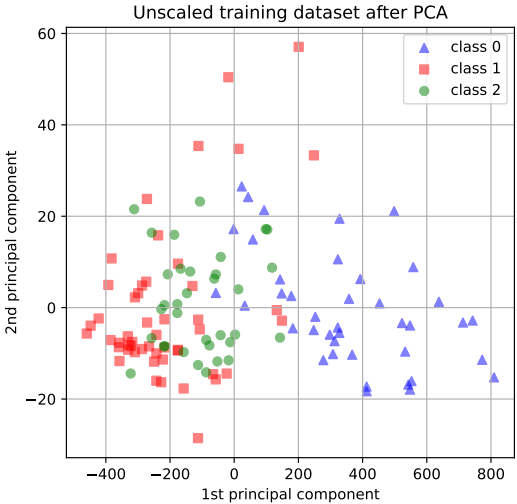
Practical tips – data standardisation (*cont.*)

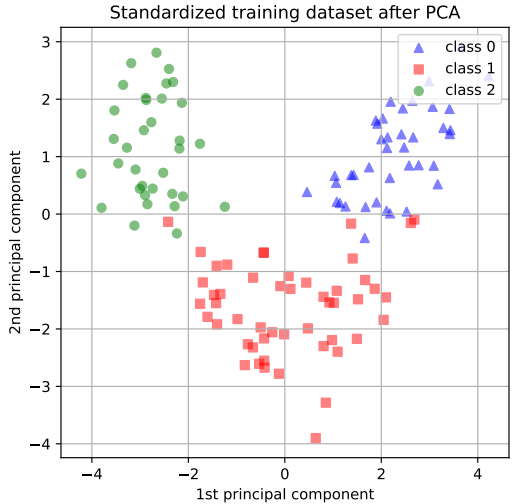


For each dimension (feature),

$$x' = \frac{x - \bar{x}}{\sigma}$$

Practical tips – standardisation (*cont.*)





PCA - Bad Applications

1. Doing PCA to avoid overfitting is a bad idea. Instead use regularisation.
2. Doing PCA to for dimensionality reduction before classification is also a bad idea. Instead use a method called, linear discriminant analysis (LDA).

More tips

- Using singular value decomposition (of $\mathbf{X}$) is more common in practice than eigen value decomposition (of $\mathbf{\Sigma}$).
- PCA is an orthogonal transformation, also known as the Karhunen–Loève transform (KLT).

Related topics

- Kernel PCA
- Discrete Cosine Transform (DCT)
- Multidimensional scaling (MDS)
- Nonnegative matrix factorisation (NMF)
- Independent Component Analysis (ICA)
- Auto-encoder (AE), variational autoencoder (VAE)
- t-distributed stochastic neighbour embedding (t-SNE)
- Uniform manifold approximation and projection (UMAP)

Quizzes

- If you forget to apply mean normalisation to a data set and obtain the eigenvectors and eigenvalues of $\mathbf{S} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ for PCA, what difference will you have compared with the case you employ mean-normalisation?
- Assuming two-dimensional data, explain the similarities and differences between principal components and linear regression lines.