Machine Learning

Lecture 9
Principal Component Analysis and Extensions

Felix Bießmann

Beuth University & Einstein Center for Digital Future

June 10, 2019

Supervised vs Unsupervised Algorithms

Often there is no label information available

- Ongoing neural activity
- Mixtures of different speakers in a audio recording
- Complex artefacts in experimental recordings

Unsupervised algorithms

- Find structure in data sets
- Allow partitioning of data in *meaningful* parts
- Allow to remove unwanted aspects (e.g. noise)

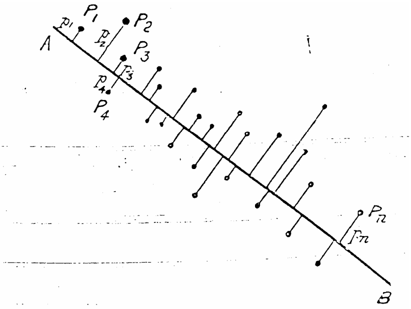
Principal Component Analysis

Principal Component Analysis (PCA):

Popular dimensionality reduction technique

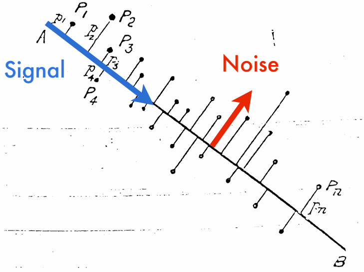
Easy to implement

Principal Component Analysis



Which line fits data best?

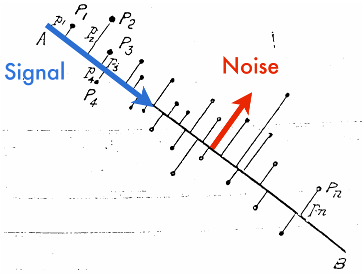
Principal Component Analysis



Which line fits data best?

The line w that minimizes the noise and maximizes the signal [Pearson, 1901]

Principal Component Analysis



Which line fits data best?

The line w that minimizes the noise and maximizes the signal [Pearson, 1901]

Or equivalently:

The line w that maximizes the variance within the data set

Maximizing variance in a data set

We obtained some data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$

PCA finds a direction $\mathbf{w}^* \in \mathbb{R}^D$ such that

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} \tag{1}$$

Maximizing variance in a data set

We obtained some data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$

PCA finds a direction $\mathbf{w}^* \in \mathbb{R}^D$ such that

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} \tag{1}$$

When optimizing eq. 1 we have to constrain $\mathbf{w}$

$$\|\mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{w} = 1 \tag{2}$$

yielding the Lagrangian

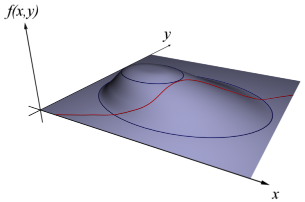
$$\mathcal{L} = \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} + \lambda(1 - \mathbf{w}^\top \mathbf{w}) \tag{3}$$

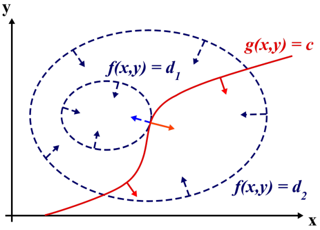
Short excursion: Lagrangians and Optimization

Optimizing a function subject to some constraint

$$\text{maximize } f(x,y)$$
$$\text{subject to the constraint } g(x,y) = c$$
$$\text{Lagrangian: } \mathcal{L}(x,y,\lambda) = f(x,y) + \lambda(g(x,y) - c)$$

where λ is called a *Lagrangian Multiplier*





Source: `http://en.wikipedia.org/wiki/Lagrange_multipliers`

Maximizing variance in a data set

$$\mathcal{L} = \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} + \lambda(1 - \mathbf{w}^\top \mathbf{w})$$

Setting the derivative w.r.t. w to zero yields

$$\frac{\partial \mathcal{L}}{\partial w} = 2\mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\lambda \mathbf{w} = 0$$

$$\Rightarrow \mathbf{X}\mathbf{X}^\top \mathbf{w} = \lambda \mathbf{w} \tag{4}$$

This is a standard eigenvalue problem.

$\mathbf{w}$ is the eigenvector of $\mathbf{X}\mathbf{X}^\top$ corresponding to the largest eigenvalue

Finding k Principal Components

- We found the strongest variance direction
- Now we want to find the second strongest variance direction
- The second direction should not be correlated with the first
- We *project out* the variance explained by the first direction
- And again find the strongest variance direction

Finding k Principal Components

- We found the strongest variance direction
- Now we want to find the second strongest variance direction
- The second direction should not be correlated with the first
- We *project out* the variance explained by the first direction
- And again find the strongest variance direction

Finding k Principal Components

- We found the strongest variance direction
- Now we want to find the second strongest variance direction
- The second direction should not be correlated with the first
- We *project out* the variance explained by the first direction
- And again find the strongest variance direction

Finding k Principal Components

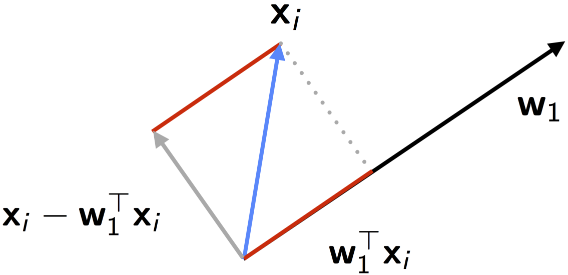
- We found the strongest variance direction
- Now we want to find the second strongest variance direction
- The second direction should not be correlated with the first
- We *project out* the variance explained by the first direction
- And again find the strongest variance direction

Finding k Principal Components

- We found the strongest variance direction
- Now we want to find the second strongest variance direction
- The second direction should not be correlated with the first
- We *project out* the variance explained by the first direction
- And again find the strongest variance direction

Finding k Principal Components

How can we *project out* the variance
along the first principal direction $\mathbf{w}_1$?



Finding k Principal Components

How can we *project out* the variance along the first principal direction $\mathbf{w}_1$?

Project data on $\mathbf{w}_1$ and back through $\mathbf{w}_1$

$$\mathbf{X}_{\mathbf{w}_1} = \mathbf{w}_1 \underbrace{\mathbf{w}_1^\top \mathbf{X}}_{\text{First Principal Component}} \tag{5}$$

Finding k Principal Components

How can we *project out* the variance
along the first principal direction $\mathbf{w}_1$?

Project data on $\mathbf{w}_1$ and back through $\mathbf{w}_1$

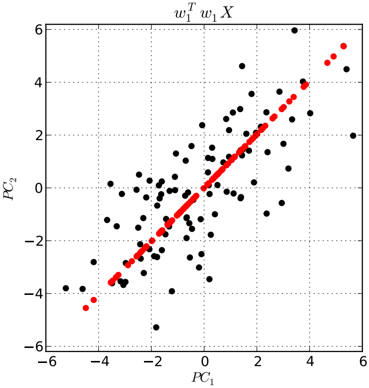
$$\mathbf{X}_{\mathbf{w}_1} = \mathbf{w}_1 \underbrace{\mathbf{w}_1^\top \mathbf{X}}_{\text{First Principal Component}} \tag{5}$$

Now we can subtract $\mathbf{X}_{\mathbf{w}_1}$ from the original data $\mathbf{X}$.

$$\mathbf{X} - \mathbf{X}_{\mathbf{w}_1} = \mathbf{X} - \mathbf{w}_1\mathbf{w}_1^\top\mathbf{X} = (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^\top)\mathbf{X} \tag{6}$$

Finding k Principal Components

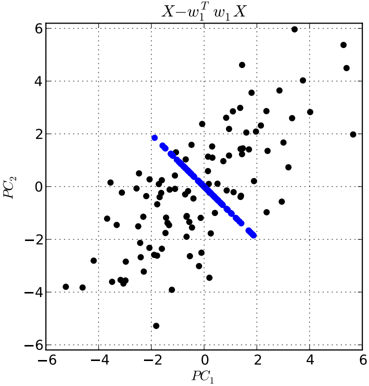
$$\color{red}{\mathbf{X}_{\mathbf{w}_1}} = \mathbf{w}_1 \mathbf{w}_1^\top \mathbf{X}$$



Data projected on $\mathbf{w}_1$ and back through $\mathbf{w}_1$

Finding k Principal Components

$$\mathbf{X} - \mathbf{w}_1\mathbf{w}_1^\top\mathbf{X} = (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^\top)\mathbf{X}$$



Data projected on $\mathbf{w}_1$ and back through $\mathbf{w}_1$ subtracted from $\mathbf{X}$

Finding k Principal Components

Finding the largest eigenvector using gradient descent, projecting it out and finding the next eigenvector is called the **Power Method** for solving eigenvalue equations

Power Method for Eigendecompositions

Algorithm 1 Power Method

Require: Square matrix $\mathbf{A}$, number of eigenvector/-value pairs k

for $k_i = 1$ to k **do**
    # Initialize ith eigenvector b randomly
    while not converged **do**
        $\mathbf{b}_i \leftarrow \frac{\mathbf{A}\mathbf{b}_i}{||\mathbf{A}\mathbf{b}_i||_2}$
    end while
    # Compute ith eigenvalue
    $\lambda_i = \mathbf{A}\mathbf{b}$
    # 'Deflate' $\mathbf{A}$
    $\mathbf{A} \leftarrow \mathbf{A} - \lambda_i \mathbf{b}_i \mathbf{b}_i^\top$
end for
return $[\mathbf{b}_1, \dots, \mathbf{b}_k]$

Principal Directions are Eigenvectors of Covariance Matrix

The k first PCA basis vectors are the eigenvectors corresponding to the largest k eigenvalues

$$\mathbf{X}\mathbf{X}^\top\mathbf{W} = \mathbf{W}\Lambda \tag{7}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k]$ contains the eigenvectors sorted according to their eigenvalues and Λ is a diagonal matrix containing all eigenvalues.

Principal Directions are Eigenvectors of Covariance Matrix

The k first PCA basis vectors are the eigenvectors corresponding to the largest k eigenvalues

$$\mathbf{X}\mathbf{X}^\top \mathbf{W} = \mathbf{W}\Lambda \tag{7}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k]$ contains the eigenvectors sorted according to their eigenvalues and Λ is a diagonal matrix containing all eigenvalues.

A useful property of the new PCA basis:
Eigenvectors $\mathbf{w}_i,\ i \in \{1, 2, \ldots, k\}$ are orthogonal to each other:

$$\mathbf{w}_i^\top \mathbf{w}_j = 0, \forall i \neq j$$

PCA Algorithm

Algorithm 2 Principal Component Analysis

Require: data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^d$, number of principal components k

Ensure: $\mathbf{W}$

1: # Center Data

2: $\mathbf{X} = \mathbf{X} - 1/N \sum_i \mathbf{x}_i$

3: # Compute Covariance Matrix

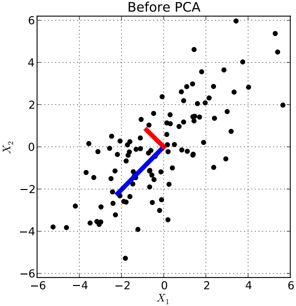
4: $\mathbf{C} = 1/N\ \mathbf{X}\mathbf{X}^\top$

5: # Compute largest k eigenvectors

6: $\mathbf{W} = \mathrm{eig}(\mathbf{C})$

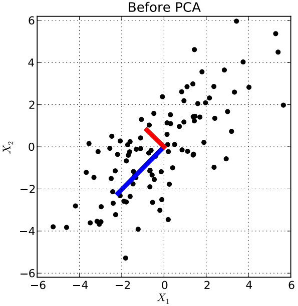
Principal Component Analysis

X

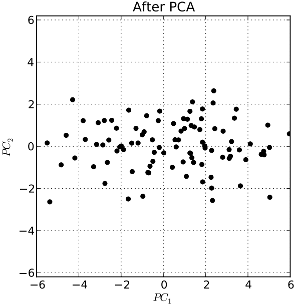


Principal Component Analysis

$\mathbf{X}$



$\mathbf{W}^\top\mathbf{X}$



PCA aligns maximum variance directions with standard basis
$\rightarrow$ Variance along each dimension is **uncorrelated**
$\rightarrow$ Now we can remove each dimension separately

Dimensionality Reduction by PCA

We can reduce the dimensionality of $\mathbf{X}$ from d to k

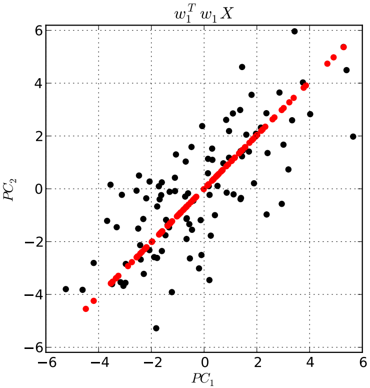
$$\mathbf{X}_{PCA} = \mathbf{W}^\top \mathbf{X} \tag{8}$$

If we want only a set $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ of principal components

$$\mathbf{X}_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \mathbf{w}_i \mathbf{w}_i^\top \mathbf{X}$$

Note that $\mathcal{I}$ does not need to contain the *strongest* components

Dimensionality Reduction by PCA



Here we assume the relevant signal is along the high variance direction

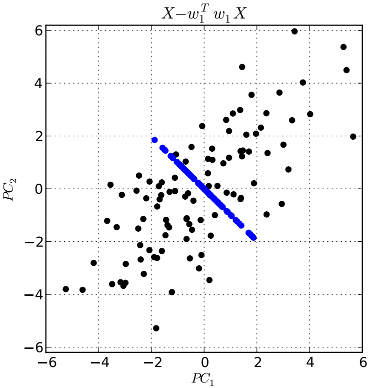
Denoising by PCA

Assuming that noise has high (or low) variance we can remove those components

If we want to project out a set $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ of principal components but we want the data to be in the input space

$$\mathbf{X}_{PCA} = \mathbf{X} - \mathbf{W}_{\mathcal{I}}\mathbf{W}_{\mathcal{I}}^{\top}\mathbf{X} = (\mathbf{I} - \mathbf{W}_{\mathcal{I}}\mathbf{W}_{\mathcal{I}}^{\top})\mathbf{X} \tag{8}$$

Denoising by PCA



Here we assume noise is along the high variance direction

PCA For High-Dimensional Data

We get a data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ where $N \ll D$

$\rightarrow$ Covariance matrix $\mathbf{X}\mathbf{X}^\top$ will be very large (D-by-D)

$\rightarrow$ Too few samples for a robust covariance matrix estimate

We know that $\mathbf{w}$ must lie in the span of the data

$$\mathbf{w} = \mathbf{X}\mathbf{a} \tag{8}$$

where $\mathbf{a}$ is a weighting of each data point

PCA For High-Dimensional Data

We can plug $\mathbf{w} = \mathbf{Xa}$ in the PCA objective and obtain

$$\mathbf{X} \underbrace{\mathbf{X}^\top \mathbf{X}}_{\text{Kernel } \mathbf{K}_X} \mathbf{a} = \lambda \mathbf{Xa}$$

which is equivalent to [Schölkopf et al., 1998]

$$\mathbf{K}_X \mathbf{a} = \lambda \mathbf{a}. \tag{9}$$

Solving PCA via $\mathbf{X}^\top \mathbf{X}$ instead of $\mathbf{XX}^\top$ is called **linear kernel PCA**

Eigenvectors of $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$

By Singular Value Decomposition we can decompose $\mathbf{X}$ into

$$\mathbf{X} = \mathbf{E}\mathbf{S}\mathbf{F}^\top$$

Eigenvectors of $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}^\top\mathbf{X}$

By Singular Value Decomposition we can decompose $\mathbf{X}$ into

$$\mathbf{X} = \mathbf{E}\mathbf{S}\mathbf{F}^\top$$

Now we see that

Covariance Matrix $\mathbf{X}\mathbf{X}^\top = \mathbf{E}\mathbf{S}\mathbf{F}^\top(\mathbf{E}\mathbf{S}\mathbf{F}^\top)^\top = \mathbf{E}\mathbf{S}\mathbf{F}^\top\mathbf{F}\mathbf{S}^\top\mathbf{E}^\top = \mathbf{E}\mathbf{S}^2\mathbf{E}^\top$

and

Kernel Matrix $\mathbf{X}^\top\mathbf{X} = (\mathbf{E}\mathbf{S}\mathbf{F}^\top)^\top\mathbf{E}\mathbf{S}\mathbf{F}^\top = \mathbf{F}\mathbf{S}^\top\mathbf{E}^\top\mathbf{E}\mathbf{S}\mathbf{F}^\top = \mathbf{F}\mathbf{S}^2\mathbf{F}^\top$

$\rightarrow$ $\mathbf{E}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^\top$

$\rightarrow$ $\mathbf{F}$ are the eigenvectors of $\mathbf{X}^\top\mathbf{X}$

$\rightarrow$ $\mathbf{S}$ are the (square root of) the eigenvalues of $\mathbf{X}^\top\mathbf{X}$ **and** $\mathbf{X}\mathbf{X}^\top$

$\rightarrow$ Relation linear kernel PCA and classical PCA: $\mathbf{E}\mathbf{S} = \mathbf{X}\mathbf{F}^\top$

Eigenvectors of $\mathbf{XX}^\top$ and $\mathbf{X}^\top\mathbf{X}$

By Singular Value Decomposition we can decompose $\mathbf{X}$ into

$$\mathbf{X} = \mathbf{ESF}^\top$$

If there are more dimensions than samples ($N \ll D$)

$\rightarrow$ Compute PCA on linear kernel matrix $\mathbf{XX}^\top \in \mathbb{R}^{N\times N}$

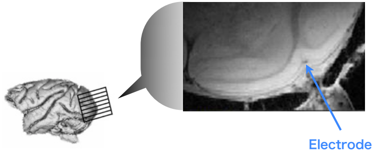
If there are more samples than dimensions ($D \ll N$)

$\rightarrow$ Compute PCA on covariance matrix $\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{D\times D}$

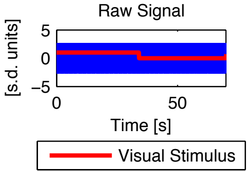
Application PCA: Automatic Artefact rejection

Multimodal Neuroimaging:

Simultaneous recordings of fMRI and neural activity



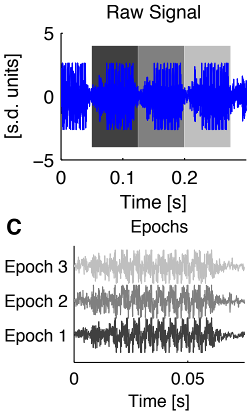
fMRI needs strong (>3Tesla) magnetic fields



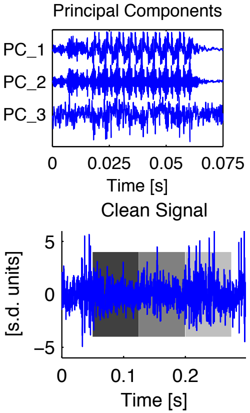
Electrical Artefacts induced by fMRI scanning stronger than neural activity

Application PCA: Automatic Artefact rejection

Before

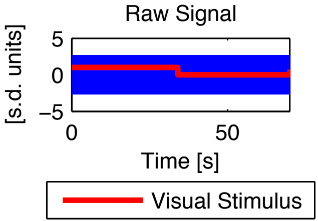


After

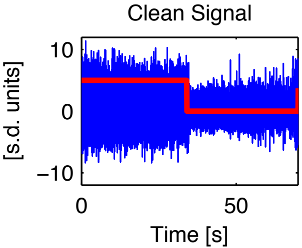


Application PCA: Automatic Artefact rejection

Before

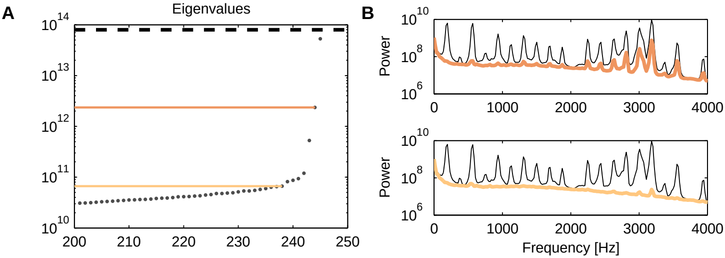


After



Application PCA: Automatic Artefact rejection

How many principal components should be rejected?

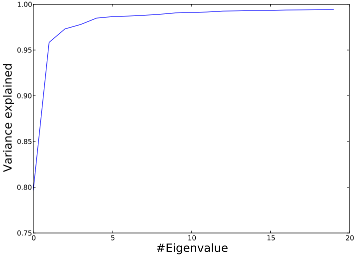


Often empirically defined heuristics have to be used

Application PCA: Dimensionality Reduction of Text Data

We are looking at Bag-Of-Words data from news web pages

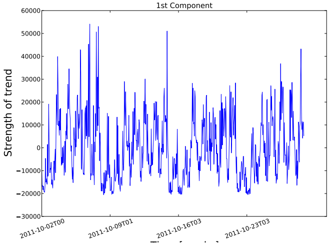
We store the data in a matrix $X \in \mathbb{R}^{W \times T}$
$X_{wt} = 10$ means: word w was counted 10 times in time bin t



We only need 15 principal directions to explain >99% of the data

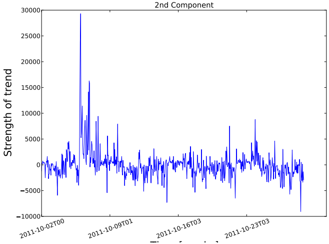
Application PCA: Dimensionality Reduction of Text Data

First Principal Component



Main Variance due to weekly/daily publishing activity

Second Principal Component



Steve Jobs died on Oct 5th

Summary

Unsupervised Data Analysis

- Finds structure in data in explorative fashion
- Can be used for
  - Dimensionality reduction
  - Visualization
  - Denoising

Principal Component Analysis (PCA)

- Finds directions of maximal variance
- Is solved by eigendecomposition of Covariance/Kernel Matrix

Linear PCA

- Finds *linear* subspaces
- If there are more dimensions than data points

$\rightarrow$ Do eigendecomposition on kernel matrix

References

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(6):1299–1319, 1998.