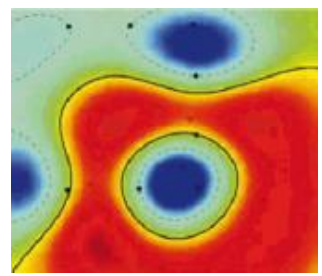
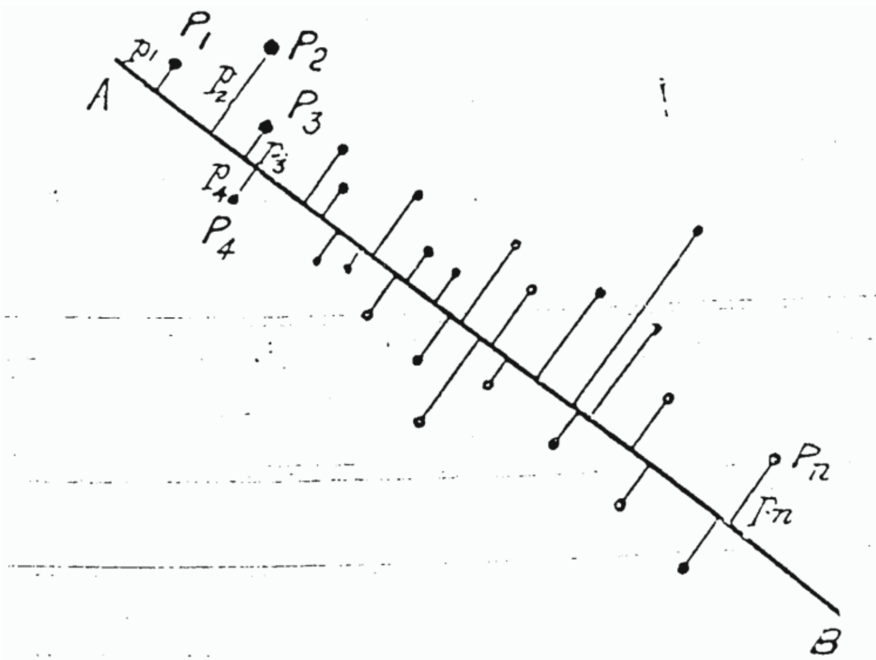
Principal Component Analysis (PCA)



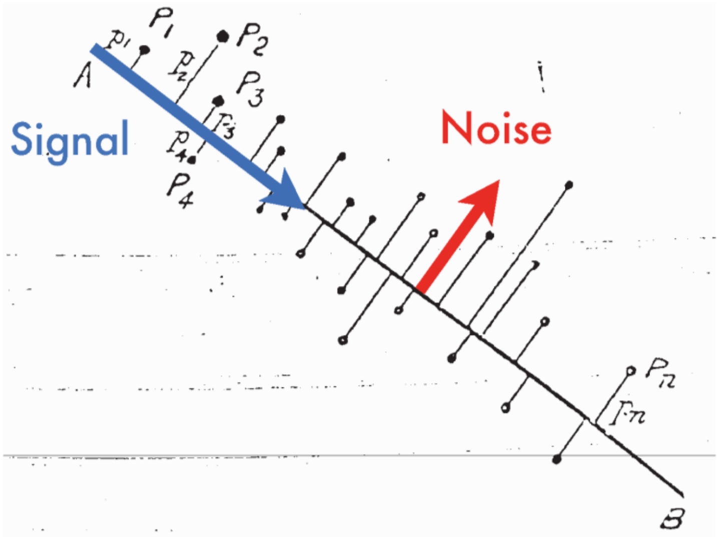
Lecture by Klaus-Robert Müller, TU Berlin 2013

Principal Components Analysis (PCA)



Which line fits data best?

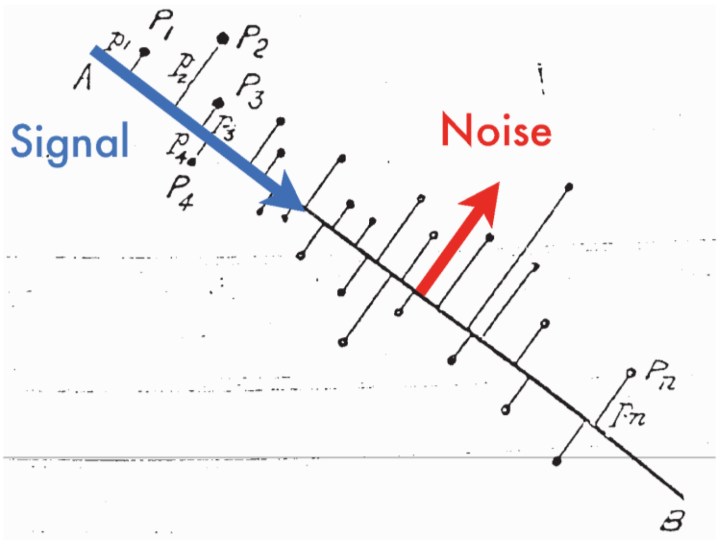
Principal Components Analysis (PCA)



Which line fits data best?

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

Principal Components Analysis (PCA)



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

(the **principal direction**)

Finding the Direction with Largest Variance

We obtained some data $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D \times N}$

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

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

Finding the Direction with Largest Variance

We obtained some data $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D \times N}$

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

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

When optimizing eq. 1 we have to constrain w

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

yielding the Lagrangian

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

Finding the Direction with Largest Variance

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

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

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

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

This is a standard eigenvalue problem.

w is the eigenvector of $XX^\top$ corresponding to the largest eigenvalue

Deflation: finding more informative directions

Idea:

- Repeat analysis in the *(D-1)*-dimensional subspace that is orthogonal to w $(=w_1)$
- Iterate until only 1-dimensional subspace is left

Closed form solution for $W=[w_{1,\ \dots,}\ w_D]$ is obtained by a full **eigendecomposition** of the data covariance matrix:

$$XX^TW = W\Lambda\ ,$$

Where the principal components W are the eigenvectors and Λ contains the corresponding eigenvalues on the diagonal.

Properties (assumptions) of PCA

- Eigenvectors W are **orthogonal**: $W^T W = W W^T = I$.

→ PCA corresponds to generative model

$$x(t) = A\, s(t)$$

with $A = W$ and $s(t) = W^T x(t)$.

- $W^T X X^T W = \Lambda$ is diagonal.

→ Extracted factors s(t) are temporally **uncorrelated.**

- The i-th Eigenvalue is the variance of the i-th factor:

$$\Lambda_{ii} = \mathrm{Var}(s_i) = \mathrm{Var}(w_i^T x) \, .$$

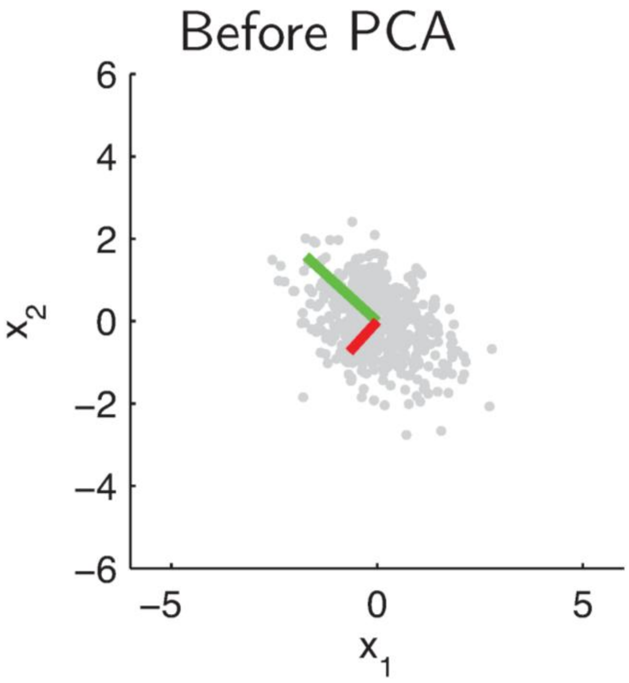
Algorithm

Algorithm 1: Principal Component Analysis

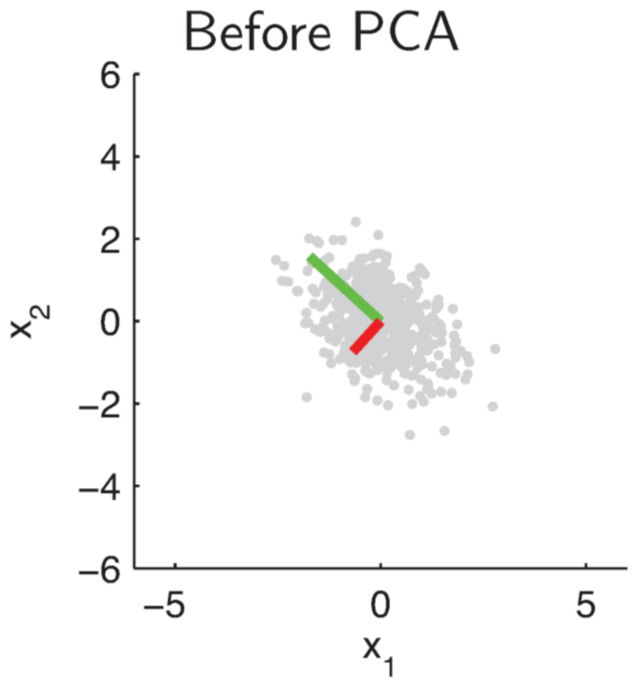
Require: data $x_1, \ldots, x_N \in \mathbb{R}^d$, number of principal components k

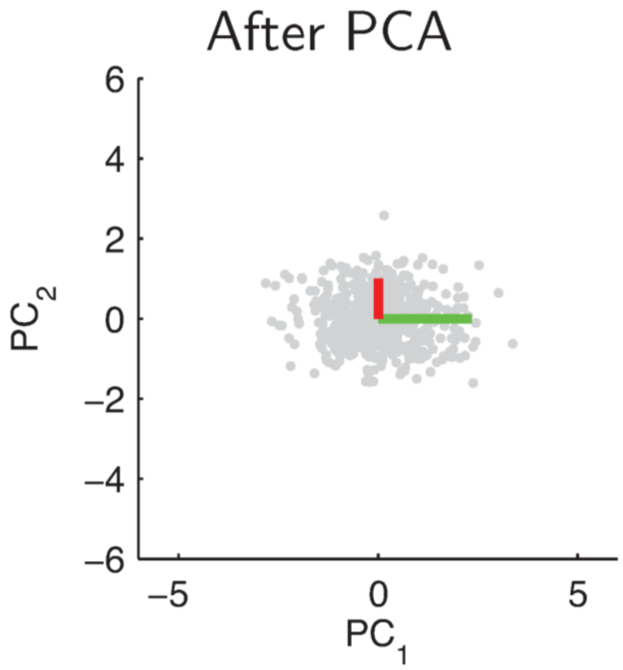
```
# Center Data
X = X − 1/N Σ_i x_i
# Compute Covariance Matrix
C = 1/N XX^⊤
# Compute largest k eigenvectors
W = eig(C)
return W
```

Example



Example





PCA rotates data into new coordinate system with the directions of largest variances being the new coordinate axes.

Optimality for Signal Reconstruction

Assume (WLOG) that the Eigenvalues are ordered:

$$\Lambda_{11} \geq \Lambda_{22} \geq \ldots \Lambda_{DD}$$

(and the Eigenvectors correspondingly).

Then, the projection onto the first *k* principal directions

$$[s_1, \ldots, s_k] = [w_1, \ldots, w_k]^T X$$

preserves $\sum_{i=1}^{k} \Lambda_{ii} / \sum_{i=1}^{D} \Lambda_{ii}$ percent of the data‘s variance (=„information content“), and there is no *k*-dimensional subspace that contains more information about the signal.

Optimality for Signal Reconstruction

Theorem: The minimal reconstruction error using a *k*-dimensional projection is the sum of the *D-k* smallest Eigenvalues of XX^T ,

$$\min_{V=[v_1,\ldots,v_k],\, V^T V = I_k} \left\| X - VV^T X \right\|^2 = \sum_{i=k+1}^{D} \Lambda_{ii}$$

and the minimum is attained at the Eigenvectors $V = [w_1, \ldots, w_k]$ corresponding to the k largest Eigenvalues.

Optimality for Signal Reconstruction

Theorem: The minimal reconstruction error using a *k*-dimensional projection is the sum of the *D-k* smallest Eigenvalues of XX^T ,

$$\min_{V=[v_1,\dots,v_k],\, V^TV=I_k} \left\|X - VV^TX\right\|^2 = \sum_{i=k+1}^{D} \Lambda_{ii}$$

and the minimum is attained at the Eigenvectors $V = [w_1, \dots, w_k]$ corresponding to the k largest Eigenvalues.

Proof:

- Simplify using $\left\|X - VV^TX\right\|^2 = Tr\{XX^T\} - Tr\{V^TXX^TV\}$
- Use result by Ky-Fan (1949):

$$\max_{V=[v_1,\dots,v_k],\, V^TV=I_k} Tr\{V^TXX^TV\} = \sum_{i=1}^{k} \Lambda_{ii}$$

„Kernel Trick“

We get a data set $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D \times N}$ where $N \ll D$

$\rightarrow$ Covariance matrix $XX^\top$ will be very large (D-by-D)

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

We know that w must lie in the span of the data

$$w = X\alpha \tag{7}$$

where α is a weighting of each data point

„Kernel Trick“

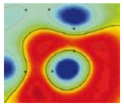
We can plug $w = X\alpha$ in the PCA objective and obtain

$$X \underbrace{X^\top X}_{\text{Kernel } K_X} \alpha = \lambda X \alpha$$

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

$$K_X \alpha = \lambda \alpha. \qquad (8)$$

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



Relation to Singular Value Decomposition (SVD)

By SVD we can decompose any matrix X into

$$X = ESF\ ,$$

where *E* and *F* are orthogonal (containing the *singular vectors*), and *S* is diagonal with positive entries (the *singular values*).

Relation to Singular Value Decomposition (SVD)

Now we see that

$$\text{Covariance Matrix } XX^\top = ESF(ESF)^\top = ESFF^\top S^\top E^\top = ES^2E^\top \tag{10}$$

and

$$\text{Kernel Matrix } X^\top X = FSE(FSE)^\top = FSEE^\top S^\top F^\top = FS^2F^\top \tag{11}$$

$\rightarrow$ E are the eigenvectors of $XX^\top$

$\rightarrow$ F are the eigenvectors of $X^\top X$

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

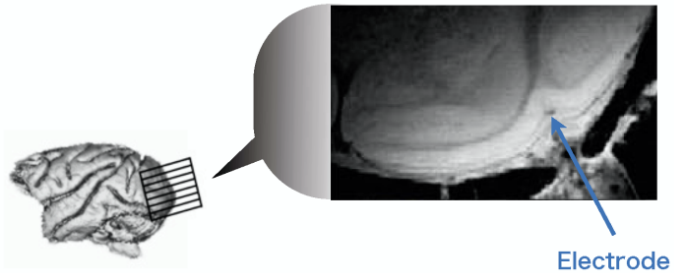
$\rightarrow$ Relation linear kernel PCA and classical PCA: $ES = XF^\top$

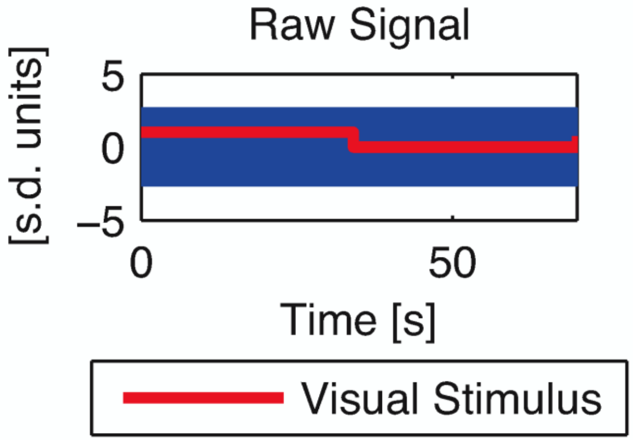
Application: Artifact Reduction

Technical Challenge: fMRI needs strong (>3Tesla) magnetic fields

Multimodal Neuroimaging:

Simultaneous recordings of fMRI and neural activity

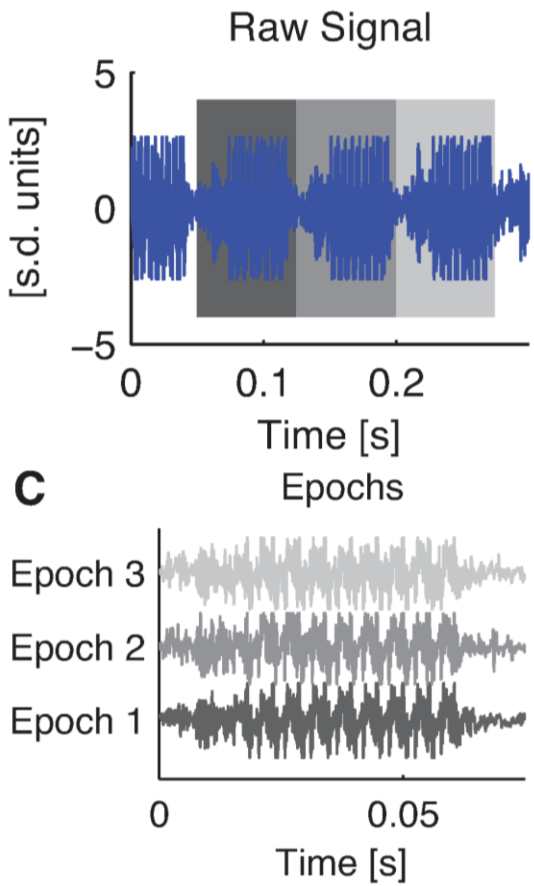




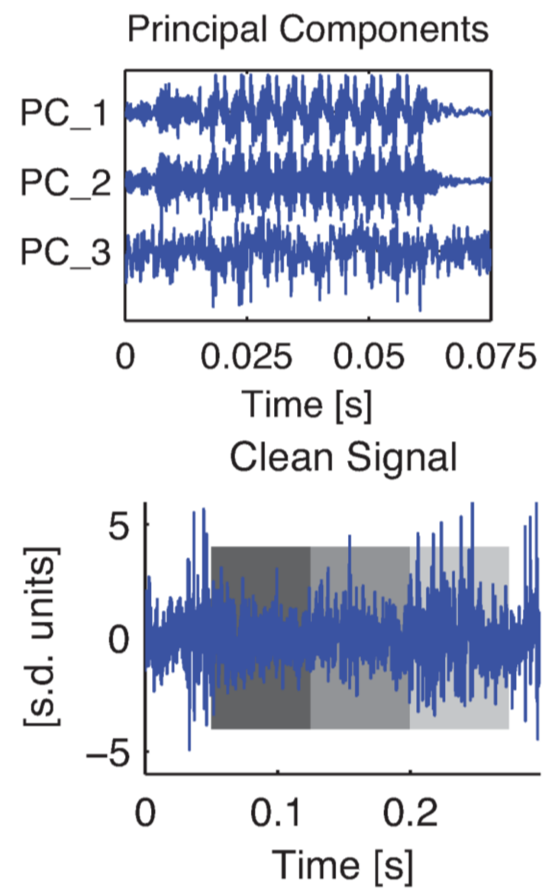
Electrical Artefacts induced by fMRI scanning stronger than neural activity

Application: Artifact Reduction

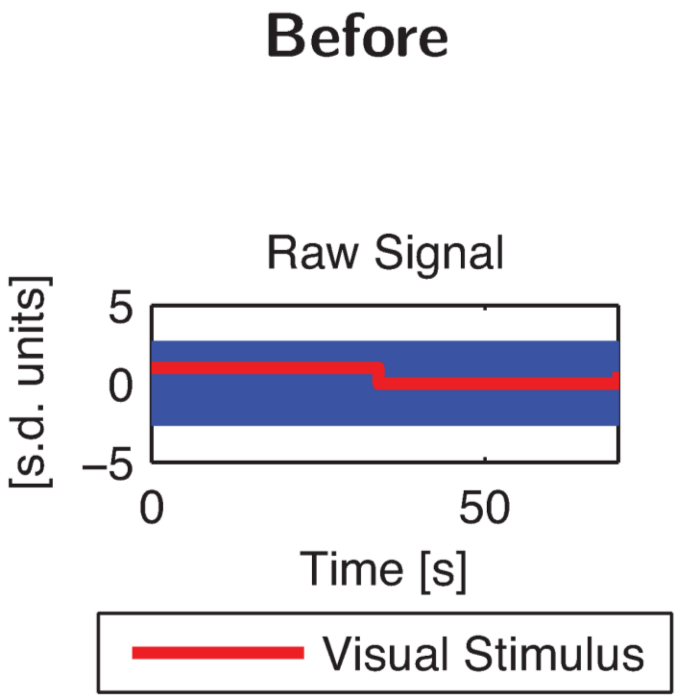
Before

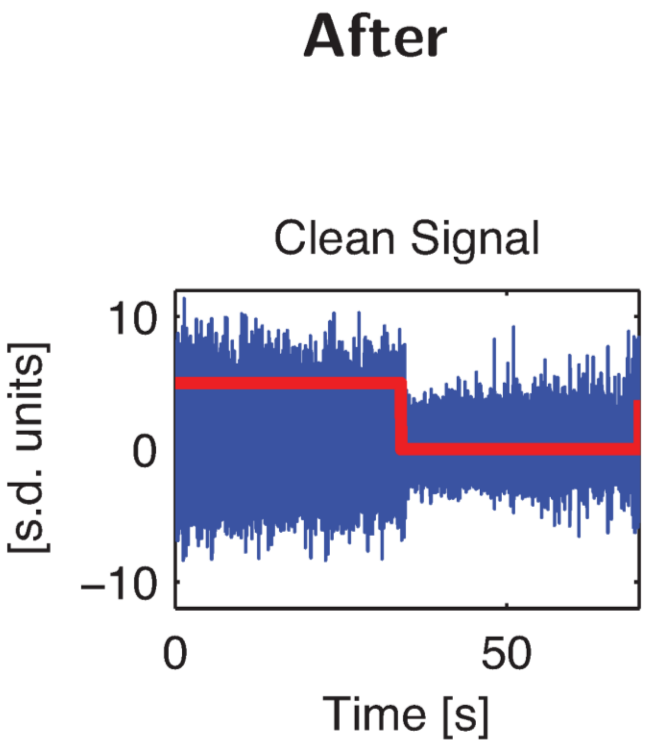


After



Application: Artifact Reduction





Summary

- PCA finds directions of maximal variance in a dataset

- kernel PCA
  - extends PCA to potentially non-linear dependencies
  - Makes PCA applicable to high dimensional data