PCA and Kernel PCA

Learning Representations. Dimensionality Reduction.

降维方法.

Maria-Florina Balcan

04/08/2015

LDA $\max \frac{w^T \Sigma_B w}{w^T \Sigma_w w}$: w 是投影向量, Σ_B 是分类之间的协方差
Σ_w 是类别之内的协方差.

但这个是有监督的方法.

Big & High-Dimensional Data

- High-Dimensions = Lot of Features

数据是高维的. => $(X^TX)^{-1}$ 无法求 (奇异矩阵)

Document classification => 正则化/降维

Features per document =

thousands of words/unigrams

millions of bigrams, contextual information

推荐系统

Surveys - Netflix

480189 users x 17770 movies

	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6
Tom	5	?	?	1	3	?
George	?	?	3	1	2	5
Susan	4	3	1	?	5	1
Beth	4	3	?	2	4	2

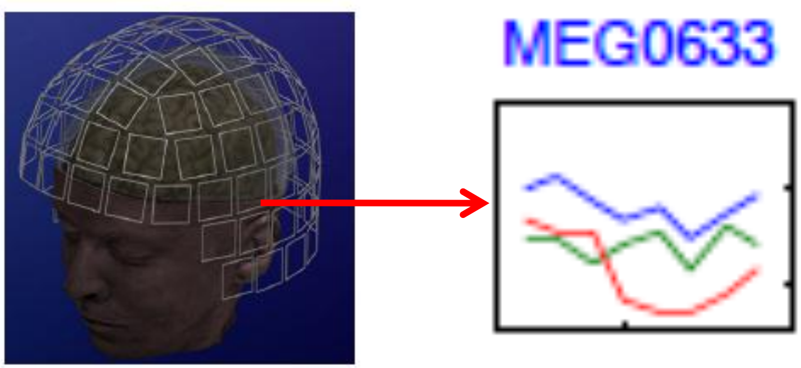
Big & High-Dimensional Data

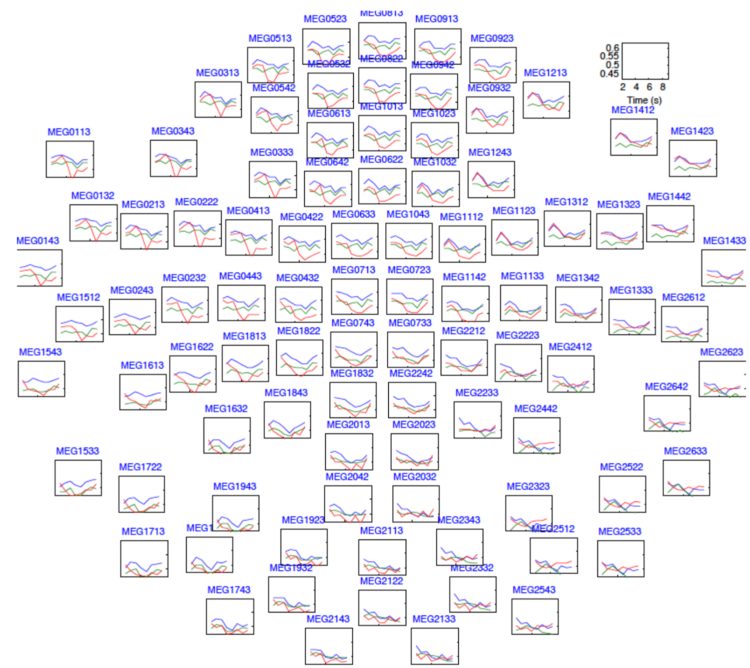
- High-Dimensions = Lot of Features

MEG Brain Imaging

120 locations x 500 time points

x 20 objects





Or any high-dimensional image data

- Big & High-Dimensional Data.

- Useful to learn lower dimensional representations of the data.

Learning Representations

PCA, Kernel PCA, ICA: Powerful unsupervised learning techniques for extracting hidden (potentially lower dimensional) structure from high dimensional datasets.

Useful for:

观测高维的 X，潜在低维隐变量 Z（本质表达）

$\mathbb{R}^D$ $\mathbb{R}^d$

通过解出 Z 然后对 Z 学习以降维.

- Visualization

  可以作可视化（二维，三维）

- More efficient use of resources (e.g., time, memory, communication)

- Statistical: fewer dimensions → better generalization

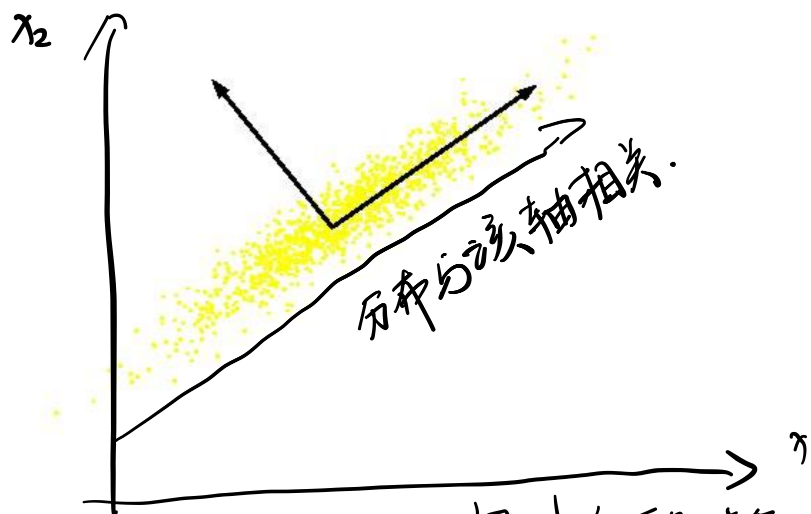
  异常点，（Z 会有较大不同）

- Noise removal (improving data quality)

- Further processing by machine learning algorithms

Principal Component Analysis (PCA)

What is PCA: Unsupervised technique for extracting variance structure from high dimensional datasets.



但未知的时候，与 $x_1 x_2$ 两轴有关

- PCA is an orthogonal projection or transformation of the data into a (possibly lower dimensional) subspace so that the variance (最大化方差) of the projected data is maximized.

$x \leftarrow \frac{x-\mu}{\sigma}$

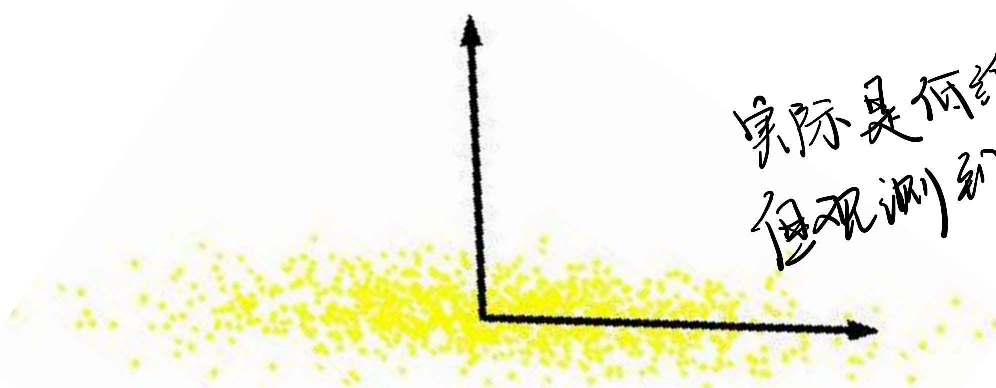
数据预处理

这里具体操作是旋转.

降维后可能减少相关轴能够简化.

Principal Component Analysis (PCA)

Intrinsically lower dimensional than the dimension of the ambient space.



Only one relevant feature

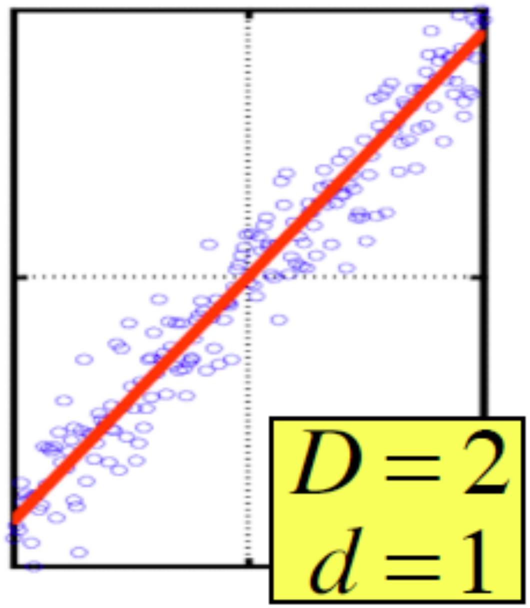
If we rotate data, again only one coordinate is more important.

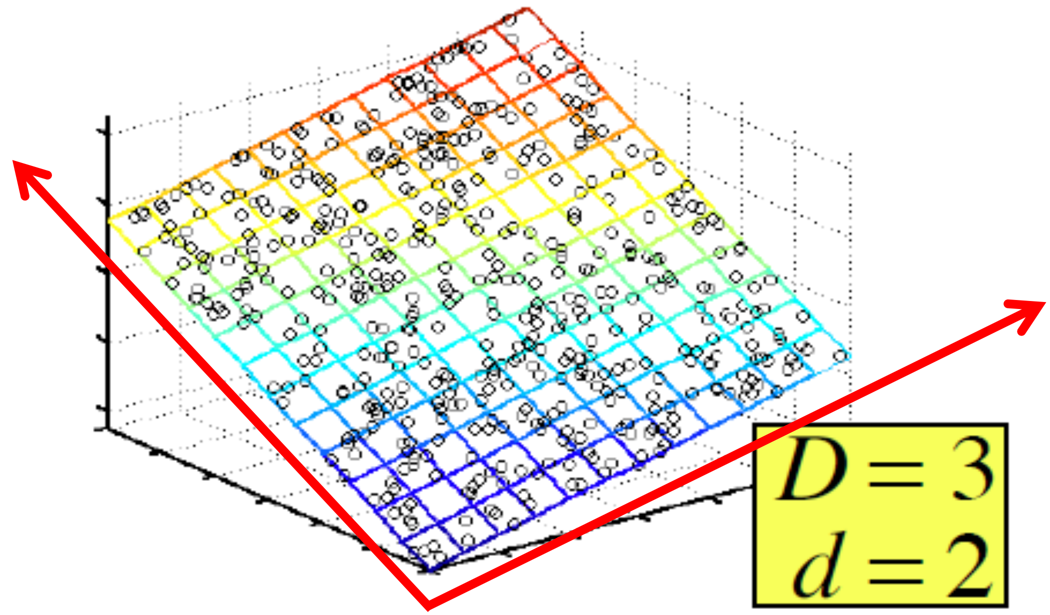
Both features are relevant

实际是低维
但观测到的是一个高维

Question: Can we transform the features so that we only need to preserve one latent feature?

Principal Component Analysis (PCA)





In case where data lies on or near a low d-dimensional linear subspace, axes of this subspace are an effective representation of the data.

Identifying the axes is known as Principal Components Analysis, and can be obtained by using classic matrix computation tools (Eigen or Singular Value Decomposition).

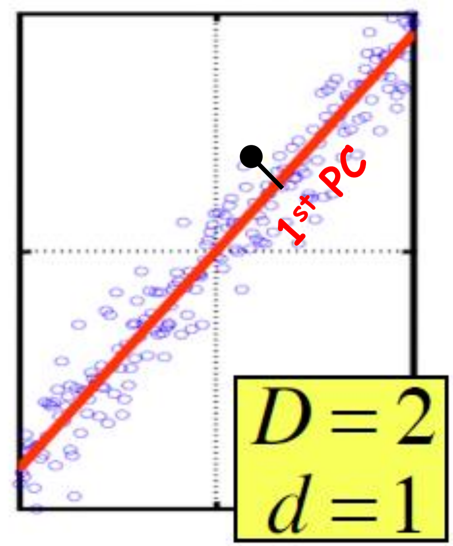
Principal Component Analysis (PCA)

Principal Components (PC) are orthogonal directions that capture most of the variance in the data.

- First PC – direction of greatest variability in data.

↳ 第一主成分方向：保留最大方差的方向.

- Projection of data points along first PC discriminates data most along any one direction (pts are the most spread out when we project the data on that direction compared to any other directions).



只关注主成分方向

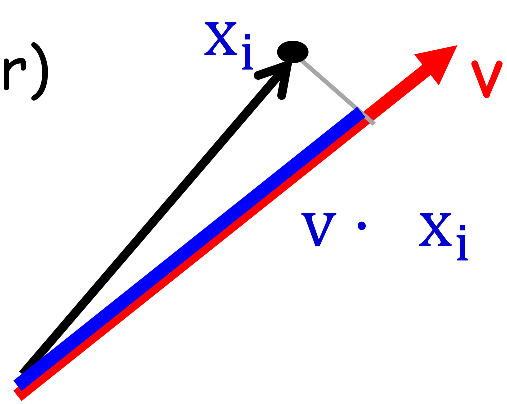
故将主成分向量模长置为1

Quick reminder:

$||v||=1$, Point x_i (D-dimensional vector)

Projection of x_i onto v is $v \cdot x_i$

利用投影去找



每个点都有自己的投影，构成新数据集

$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

是标量，故很好算

预处理

$$\Rightarrow \tilde{S} = \{x_i^T v_1\}_{i=1}^{n} \Rightarrow \max_{v_1} Var(\tilde{S}) = \frac{1}{n}\sum_{i=1}^{n}(x_i^T v_1 - \mu^T v_1)^2 \quad s.t. \|v_1\|=1$$

$$\Rightarrow \max_{v_1} Var(\hat{S}) = \frac{1}{n}\sum_{i=1}^{n}(x_i^T v_1 - \mu^T v_1)^2 = \frac{1}{n}\sum_{i=1}^{n} v_1^T (x_i-\mu)(x_i-\mu)^T v_1 = v_1^T \underbrace{\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)}_{\text{协方差矩阵}} v = v_1^T \underbrace{XX^T}_{\text{数据预处理}} v_1 \quad s.t. \|v_1\|=1$$

数据预处理：令 $x_i \leftarrow x_i - \mu$ 则 $\mu = 0$

对有约束问题应用拉格朗日：

$$L(v,\lambda) = v^T XX^T v - \lambda(v^T v - 1) \Rightarrow XX^T v = \lambda v \Rightarrow v^T XX^T v = \lambda \underline{\underline{v^T v}} = \lambda$$

$\underline{\underline{v^T v}} = 1$

令 matrix $V = \begin{bmatrix} v_1 & v_2 & \cdots & v_d \end{bmatrix}_{p\times d}$ $\quad V^T V = I$

(每个 v_i 之间是相互正交的)

对 X 作 SVD 分解：$X = \underbrace{V}_{\text{正交矩阵}} \overbrace{\Sigma}^{\text{对角矩阵}} \underbrace{U^T}_{\text{正交矩阵}}$

$$\Rightarrow \underset{D\times n}{X}\ \underset{n\times D}{X^T} = V\Sigma \underline{U^T U}\, \Sigma^T V^T = V\Sigma^2 V^T$$

$U^T U = 1$

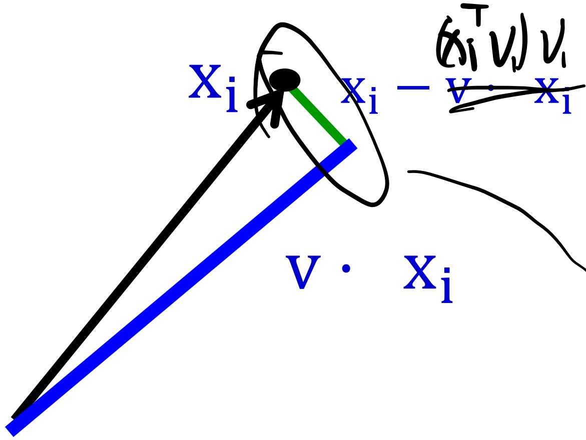
$\hookrightarrow \Sigma\Sigma^T = (\Sigma)^2$ 是对角矩阵

可以用特征值分解加速 XX^T 的计算

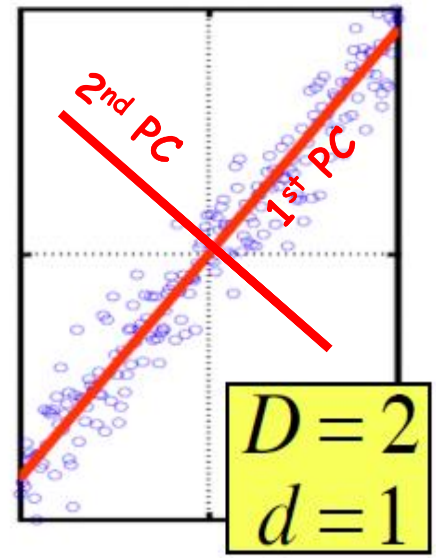
Principal Component Analysis (PCA)

Principal Components (PC) are orthogonal directions that capture most of the variance in the data.

- 1^{st} PC – direction of greatest variability in data.



$x_i - (x_i^T v_1) v_1$ ↳作为$\tilde{S}$计算.



- 2^{nd} PC – Next orthogonal (uncorrelated) direction of greatest variability

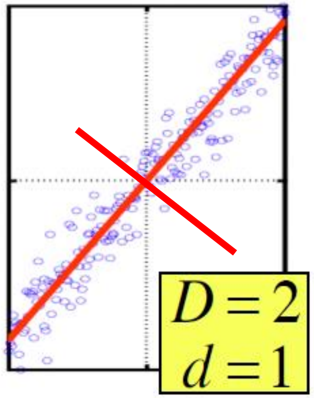
  (remove all variability in first direction, then find next direction of greatest variability)

- And so on …

Principal Component Analysis (PCA)

Let $v_1, v_2, \ldots, v_d$ denote the d principal components.

$v_i \cdot v_j = 0, i \neq j$ and $v_i \cdot v_i = 1, \; i = j$



Assume data is centered (we extracted the sample mean).

Let $X = [x_1, x_2, \ldots, x_n]$ (columns are the datapoints)

Find vector that maximizes sample variance of projected data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

$$\max_{\mathbf{v}} \; \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1$$

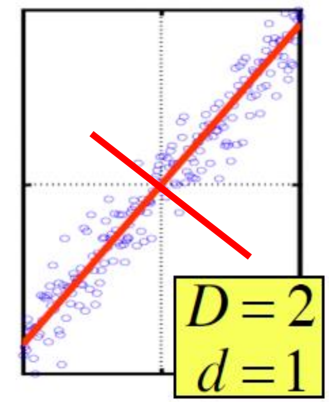
Lagrangian: $\max_{\mathbf{v}} \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} - \lambda\mathbf{v}^T\mathbf{v}$

Wrap constraints into the objective function

$$\partial/\partial\mathbf{v} = 0 \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0 \qquad \Rightarrow \boxed{(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v}}$$

Principal Component Analysis (PCA)

$(X X^T)v = \lambda v$, so v (the first PC) is the eigenvector of sample correlation/covariance matrix $X X^T$



Sample variance of projection $v^T X X^T v = \lambda v^T v = \lambda$

Thus, the eigenvalue λ denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).

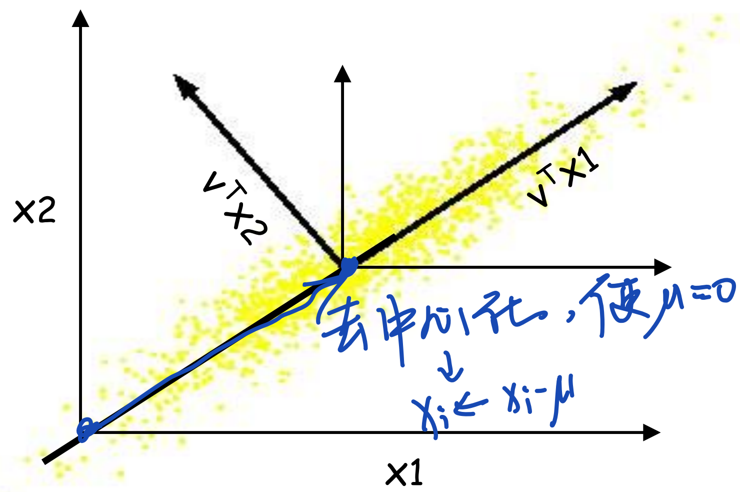
特征值排序,代表从第一主成分到第n主成分

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n$

- The 1st PC v_1 is the the eigenvector of the sample covariance matrix $X X^T$ associated with the largest eigenvalue
- The 2nd PC v_2 is the the eigenvector of the sample covariance matrix $X X^T$ associated with the second largest eigenvalue
- And so on …

Principal Component Analysis (PCA)

- So, the new axes are the eigenvectors of the matrix of sample correlations $X\,X^T$ of the data.
- Transformed features are uncorrelated.



- Geometrically: centering followed by rotation.
  - Linear transformation

Key computation: eigendecomposition of XX^T (closely related to SVD of X).

Two Interpretations

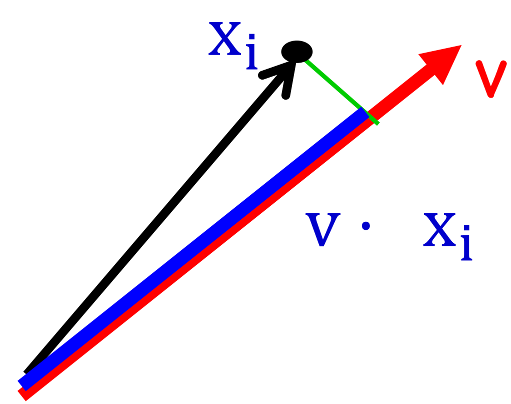
So far: Maximum Variance Subspace. PCA finds vectors v such that projections on to the vectors capture maximum variance in the data

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad s.t. \quad v^Tv=1$$

Alternative viewpoint: Minimum Reconstruction Error. PCA finds vectors v such that projection on to the vectors yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$

Square Loss 描述误差.



Two Interpretations

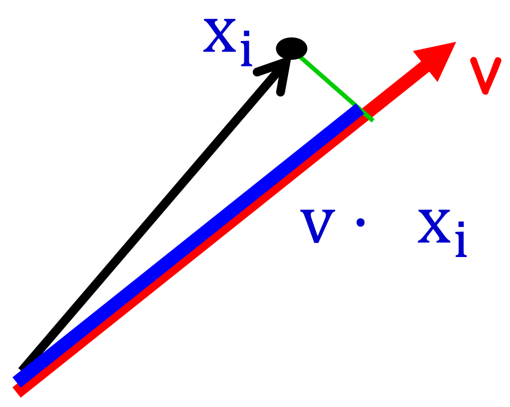
E.g., for the first component.

Maximum Variance Direction: 1st PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

Minimum Reconstruction Error: 1st PC a vector v such that projection on to this vector yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$



Why? Pythagorean Theorem

E.g., for the first component. 最大化方差 $\overset{\text{相同}}{\Longleftrightarrow}$ 最小化误差

Maximum Variance Direction: 1^{st} PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

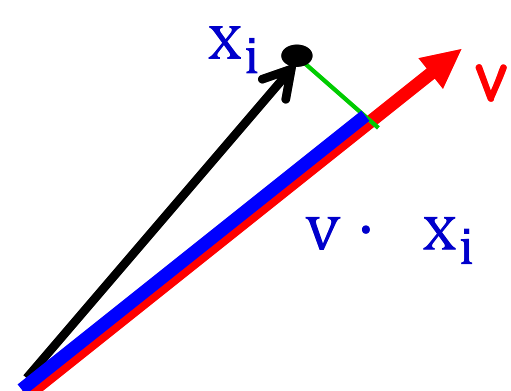
$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \qquad\qquad \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$

Minimum Reconstruction Error: 1^{st} PC a vector v such that projection on to this vector yields minimum MSE reconstruction

blue^2 + green^2 = black^2

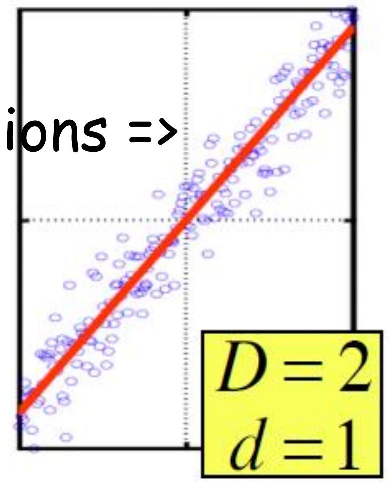
black^2 is fixed (it's just the data)

So, maximizing blue^2 is equivalent to minimizing green^2



Dimensionality Reduction using PCA

The eigenvalue λ denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).



Zero eigenvalues indicate no variability along those directions => data lies exactly on a linear subspace

Only keep data projections onto principal components with non-zero eigenvalues, say $\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{k}$, where k=rank($X\,X^T$)

Original representation

Data point

$$x_i = (x_i^1, \ldots, x_i^D)$$

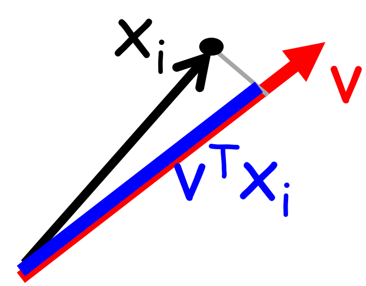
D-dimensional vector

Transformed representation

projection

$$(v_1 \cdot x^i, \ldots, v_d \cdot x^i)$$

d-dimensional vector



Dimensionality Reduction using PCA

Original representation

Data point

$$x_i = (x_i^1, \ldots, x_i^D)$$

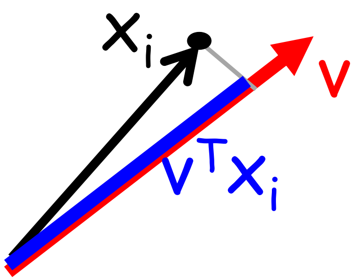
D-dimensional vector

Transformed representation

projection

$$(v_1 \cdot x^i, \ldots, v_d \cdot x^i)$$

d-dimensional vector

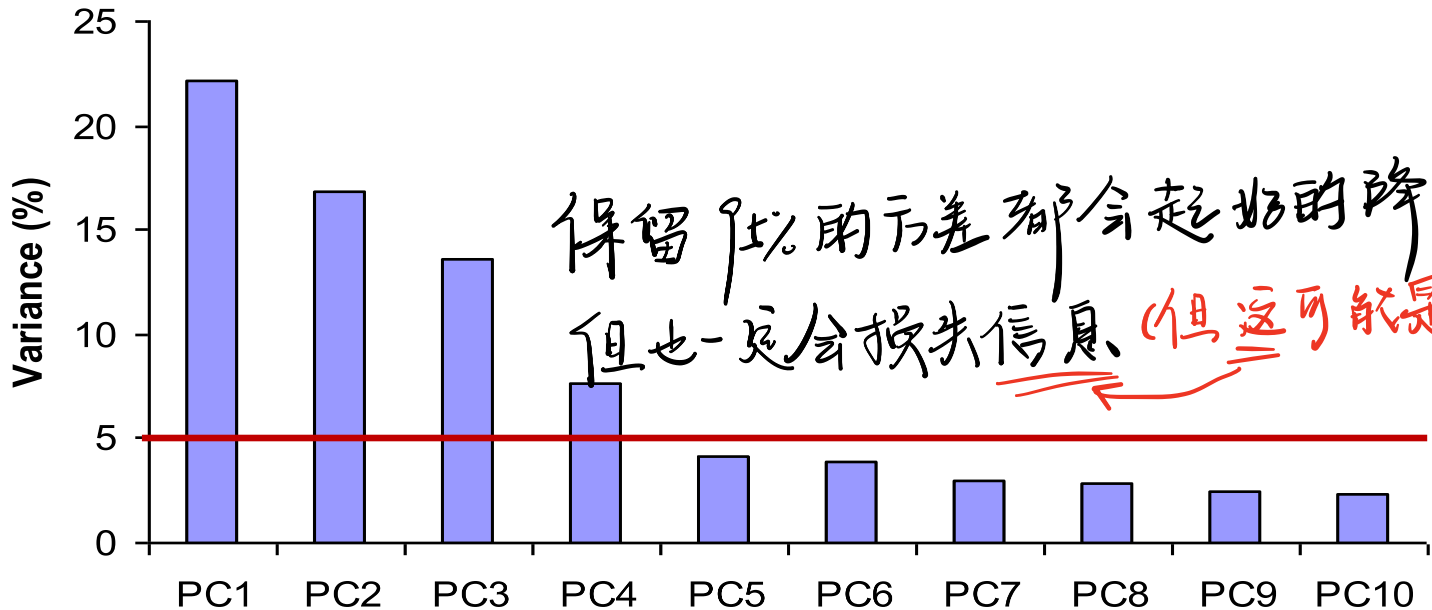


Dimensionality Reduction using PCA

In high-dimensional problems, data sometimes lies near a linear subspace, as noise introduces small variability

Only keep data projections onto principal components with **large** eigenvalues

Can *ignore* the components of smaller significance.

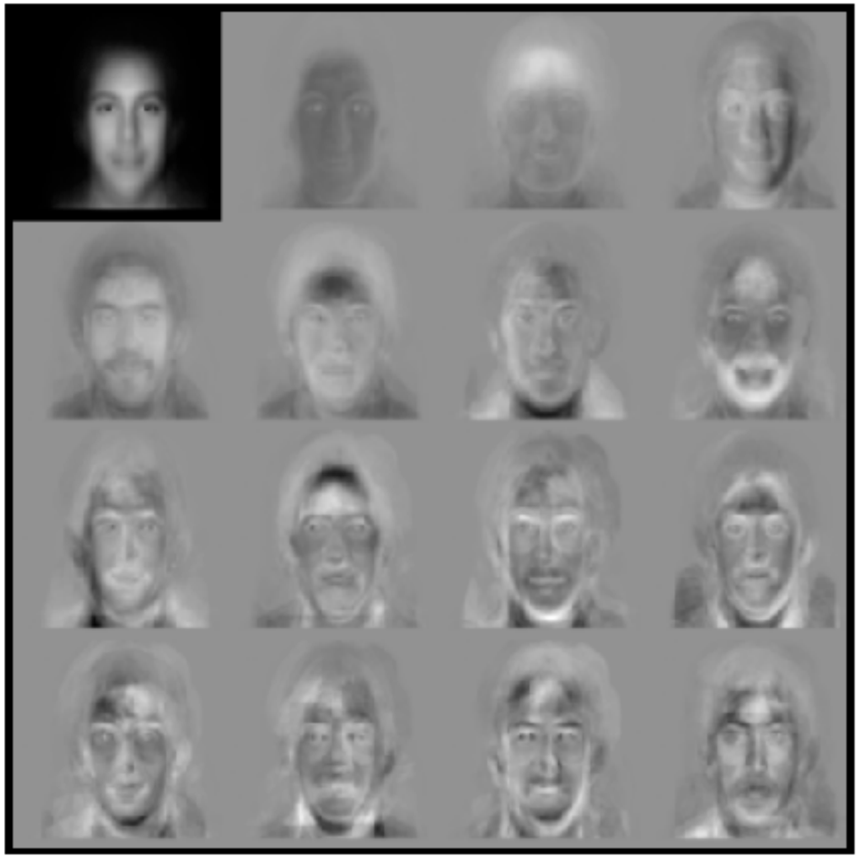


保留了比较大的方差都会起到降维的效果

但也一定会损失信息、(但这可能是噪声产生的)

Might lose some info, but if eigenvalues are small, do not lose much

Example: faces



Eigenfaces from 7562 images:

top left image is linear combination of rest.

Sirovich & Kirby (1987)
Turk & Pentland (1991)

Can represent a face image using just 15 numbers!

PCA Discussion

Strengths

Eigenvector method

No tuning of the parameters

No local optima

Weaknesses

会丢失高阶信息，

Limited to second order statistics

Limited to linear projections 可以用kernel解决

Kernel PCA (Kernel Principal Component Analysis)

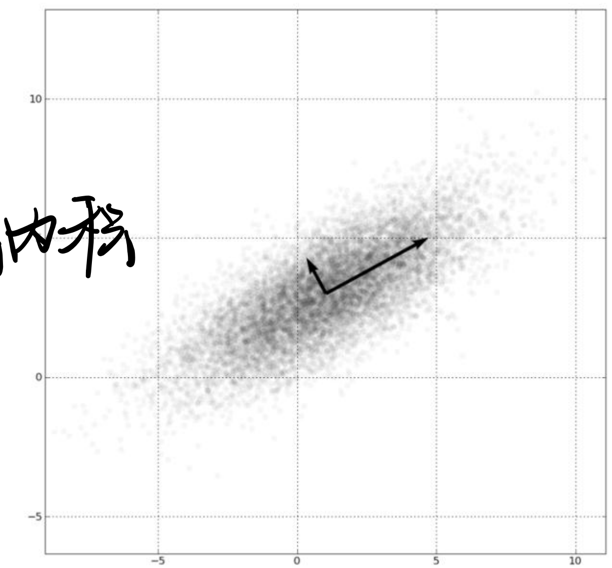
Useful when data lies on or near a low d-dimensional linear subspace of the ϕ-space associated with a kernel

Properties of PCA

- Given a set of n centered observations $x_i \in R^D$, 1st PC is the direction that maximizes the variance
  - $X = (x_1, x_2, \ldots, x_n)$
  - $v_1 = argmax_{\|v\|=1} \frac{1}{n} \sum_i (v^\top x_i)^2$

$$= argmax_{\|v\|=1} \frac{1}{n} v^\top X X^\top v$$

需要 $X^\top X \in M_{n \times n}$ 才有内积



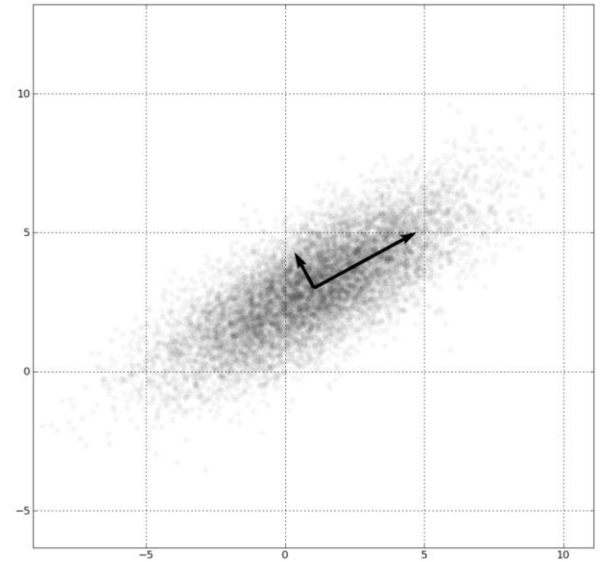
- Covariance matrix $C = \frac{1}{n} X X^\top$

- v_1 can be found by solving the eigenvalue problem:
  - $C v_1 = \lambda v_1$ (of maximum λ)

Properties of PCA

- Given a set of n centered observations $x_i \in R^D$, 1st PC is the direction that maximizes the variance
  - $X = (x_1, x_2, \ldots, x_n)$
  - $v_1 = argmax_{\|v\|=1} \frac{1}{n} \sum_i (v^\top x_i)^2$

$$= argmax_{\|v\|=1} \frac{1}{n} v^\top X X^\top v$$



- Covariance matrix $C = \frac{1}{n} X X^\top$ is a DxD matrix

the (i,j) entry of $XX^\top$ is the correlation of the i-th coordinate ofexamples with jth coordinate of examples

- To use kernels, need to use the inner-product matrix $X^T X$.

Alternative expression for PCA

- The principal component lies in the span of the data

$$v_1 = \sum_i \alpha_k x_i = X\alpha$$

Why? 1st PC is direction of largest variance, and for any direction outside of the span of the data, only get more variance if we project that direction into the span.

- Plug this in we have

$$C v_1 = \frac{1}{n} X X^\top X\alpha = \lambda\, X\alpha$$

- Now, left-multiply the LHS and RHS by X^T.

$$\frac{1}{n} X^\top X X^\top X\alpha = \lambda X^\top X\alpha$$

Only depends on the inner product matrix

Kernel PCA

- **Key Idea:** Replace inner product matrix by kernel matrix
  - PCA: $\frac{1}{n} X^\top X X^\top X \alpha = \lambda X^\top X \alpha$
  - Let $K = \left[K\left(x^i, x^j\right)\right]_{ij}$ be the matrix of all dot-products in the ϕ-space.
  - Kernel PCA: replace "$X^T X$" with K.

    $\frac{1}{n} K K \alpha = \lambda K \alpha$, or equivalently, $\boxed{\frac{1}{n} K \alpha = \lambda\, \alpha}$
- **Key computation:** form an n by n kernel matrix K, and then perform eigen-decomposition on K.

What You Should Know

- Principal Component Analysis (PCA)
  - What PCA is, what is useful for.
  - Both the maximum variance subspace and the minimum reconstruction error viewpoint.
- Kernel PCA

Additional material on computing the principal components and ICA

Power method for computing PCs

Given matrix $X \in R^{D \times n}$, compute the top eigenvector of $X\,X^T$

Initialize with random $\hat{v} \in R^D$

Repeat

$\hat{\mathrm{v}} \leftarrow \mathrm{X}\,\mathrm{X}^{\mathrm{T}}\hat{\mathrm{v}}$

$\hat{\mathrm{v}} \leftarrow \hat{\mathrm{v}}/||\hat{\mathrm{v}}||$

Claim

For any $\epsilon > 0$, whp over choice of initial vector, after $O\left(\frac{1}{\epsilon}\log\frac{d}{\epsilon}\right)$ iterations, we have $\hat{v}^T X X^T \hat{v} \geq (1-\epsilon)\lambda_1$.

Then can subtract the $\hat{v}$ component off of each example and repeat to get the next.

Eigendecomposition

Any symmetric matrix $A = XX^T$ is guaranteed to have an eigendecomposition with real eigenvalues: $A = V\,\Lambda\,V^T$.

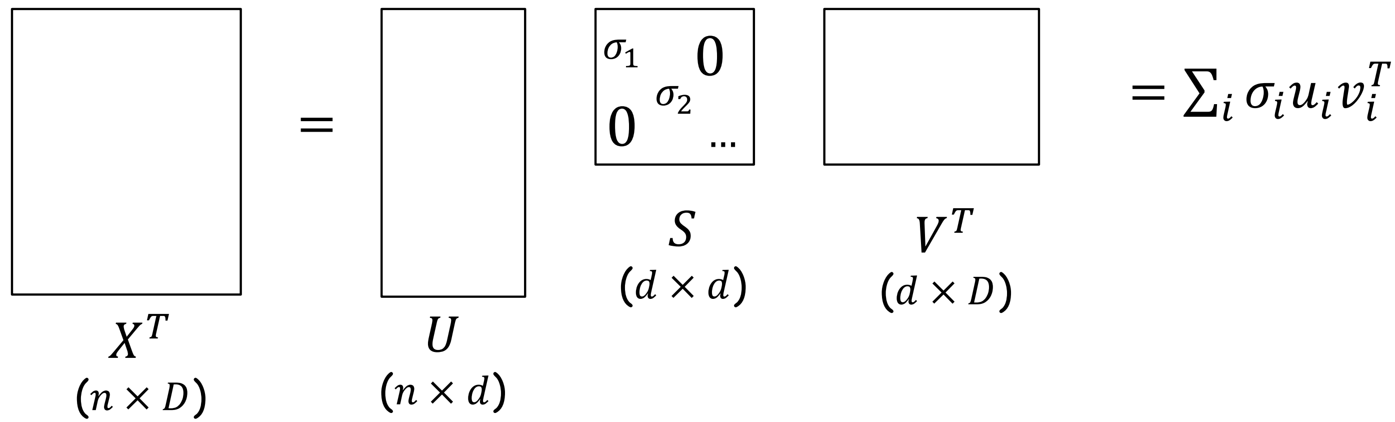
$$\underset{\substack{\text{A}\\(\text{DxD})}}{\boxed{}} = \underset{\substack{\text{V}\\(\text{DxD})}}{\boxed{}}\ \underset{\substack{\Lambda\\(\text{DxD})}}{\boxed{\begin{matrix}\lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_3 \\ & & & \dots\end{matrix}}}\ \underset{\substack{V^T\\(\text{DxD})}}{\boxed{}} = \sum_i \lambda_i v_i v_i^T$$

Matrix Λ is diagonal with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ on the diagonal. Matrix V has the eigenvectors as the columns.

Singular Value Decomposition (SVD)

Eigendecomp of XX^T is closely related to SVD of X.

Given a matrix $X \in R^{D \times n}$, the SVD is a decomposition: $X^T = USV^T$



- S is a diagonal matrix with the singular values $\sigma_1, \dots, \sigma_d$ of X.
- Columns of U, V are orthogonal, unit length.
- So, $XX^T = VSU^TUSV^T = VS^2V^T =$ eigendecomposition of XX^T.

So, $\lambda_i = \sigma_i^2$ and can read off the solution from the SVD.

Singular Value Decomposition (SVD)

Eigendecomp of XX^T is closely related to SVD of X.

Given a matrix $X \in R^{D\times n}$, the SVD is a decomposition: $X^T = USV^T$

$$\underset{(n \times D)}{X^T} = \underset{(n \times d)}{U} \; \underset{(d \times d)}{S} \; \underset{(d \times D)}{V^T} = \sum_i \sigma_i u_i v_i^T$$

$$S = \begin{bmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ 0 & & \dots \end{bmatrix}$$

- In fact, can view the rows of US as the coordinates of each example along the axes given by the d eigenvectors.

So, $\lambda_i = \sigma_i^2$ and can read off the solution from the SVD.

Independent Component Analysis (ICA)

Find a linear transformation

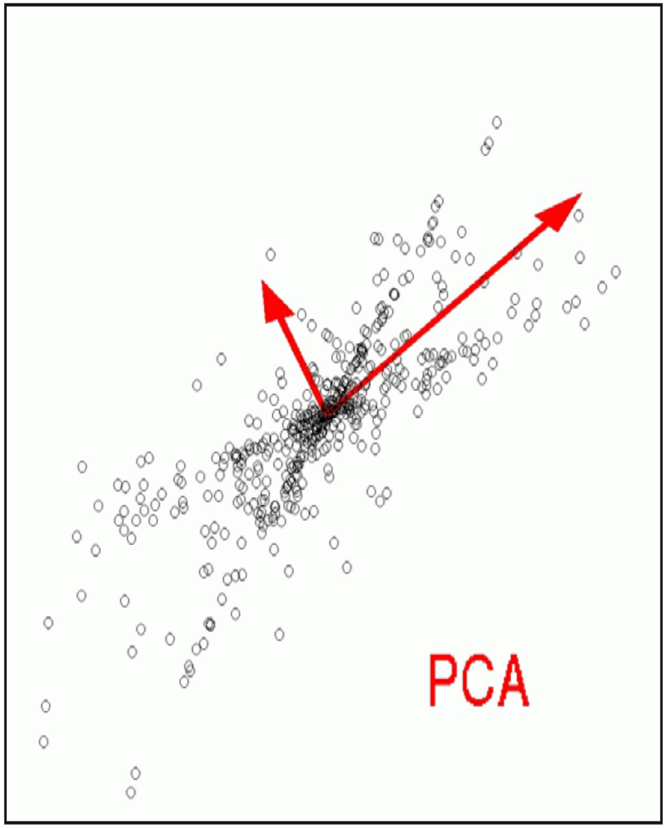
$$\boldsymbol{x} = V \cdot \boldsymbol{s}$$

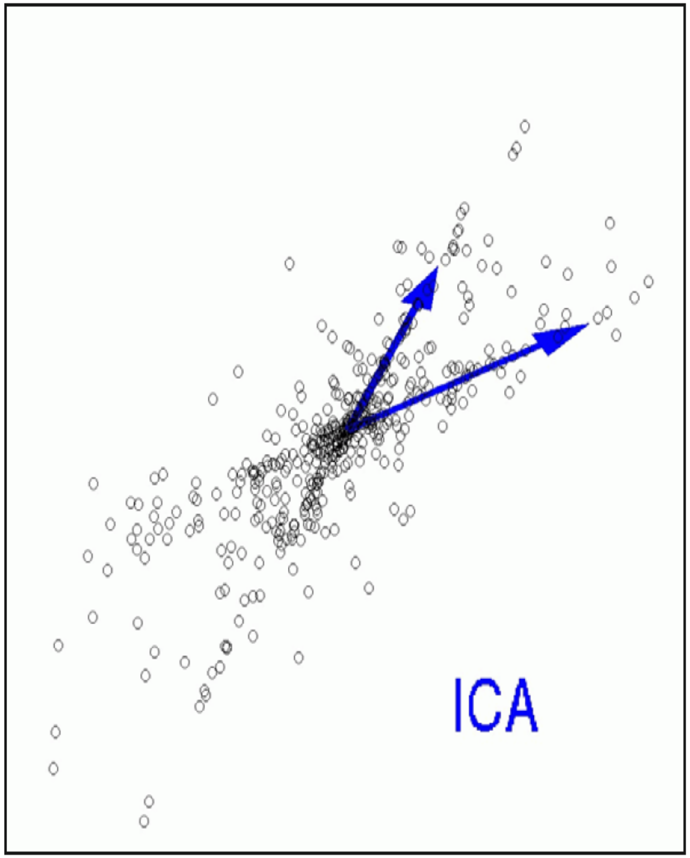
for which coefficients $\boldsymbol{s} = (s_1, s_2, \dots, s_D)^T$ are **statistically independent**

$$p(s_1, s_2, \dots, s_D) = p_1(s_1)p_2(s_2) \dots p_n(s_D)$$

Algorithmically, we need to identify matrix V and coefficients s, s.t. under the condition $\boldsymbol{x} = V^T \cdot \boldsymbol{s}$ the **mutual information** between $s_1, s_2, \dots, s_D$ is minimized:

$$I(s_1, s_2, \dots, s_D) = \sum_{i=1}^{D} H(s_i) - H(s_1, s_2, \dots, s_D)$$





PCA finds directions of maximum variation,
ICA would find directions most “aligned” with data.