# 第四章：线性模型

## -成分分析

# 成分分析

## Principal Component Analysis
## Linear Discriminant Analysis

# 问题的提出

- 在建立预测模型系统时，抽取的原始特征往往比较多，特征的维数比较大，这会给识别器的训练带来很大的困难，因此希望能够采用某种方法降低特征的维数。这些方法可以称作*成分分析*的方法。
  1. 主成分分析：寻找最小均方意义下，最能代表原始数据的投影方法
  2. 线性判别分析：寻找最小均方意义下，最能分开各类数据的投影方法

# 人脸识别举例

# 1 主成分分析

- PCA（Principal Component Analysis）是一种最常用的线性成分分析方法；
- PCA的主要思想是寻找到数据的主轴方向，由主轴构成一个新的坐标系（维数可以比原维数低），然后数据由原坐标系向新的坐标系投影；
- PCA的其它名称：离散K-L变换，Hotelling变换。

## 问题：有$n$个$\mathbf{d}$维样本，$\mathrm{x}_1,\mathrm{x}_2,..\mathrm{x}_n$,如何仅用一个样本$\mathrm{x}_0$代表这些样本，使误差准则函数最小？

$$J_0(\mathbf{x}_0)=\sum_{k=1}^{n}\left\|\mathbf{x}_0-\mathbf{x}_k\right\|^2 \longrightarrow \mathbf{x}_0=\mathbf{m}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$=\sum_{k=1}^{n}\left\|(\mathbf{x}_0-\mathbf{m})-(\mathbf{x}_k-\mathbf{m})\right\|^2$$

$$=\sum_{k=1}^{n}\left\|\mathbf{x}_0-\mathbf{m}\right\|^2-2\sum_{k=1}^{n}(\mathbf{x}_0-\mathbf{m})^t(\mathbf{x}_k-\mathbf{m})+\sum_{k=1}^{n}\left\|\mathbf{x}_k-\mathbf{m}\right\|^2$$

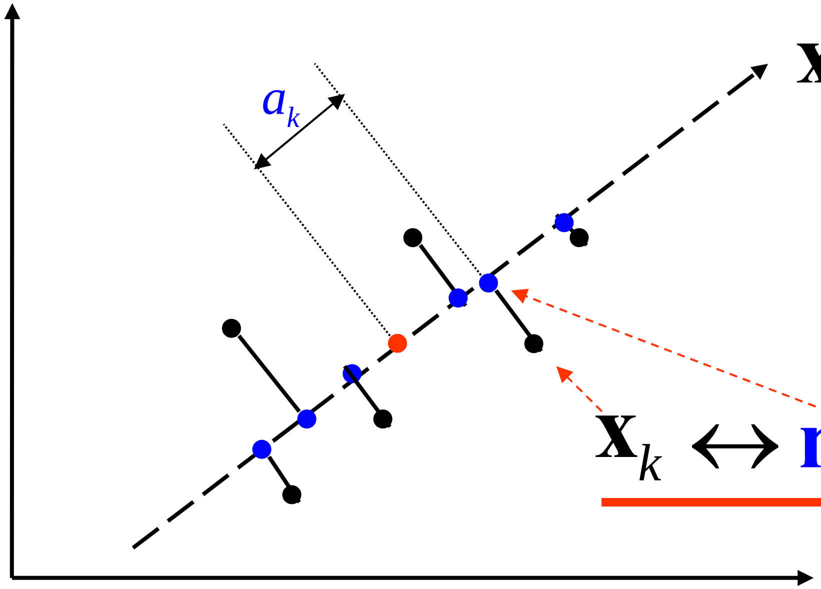
$$=\sum_{k=1}^{n}\underbrace{\left\|\mathbf{x}_0-\mathbf{m}\right\|^2}_{\mathbf{x}_0=\mathbf{m}\text{时取得最小值}}-2(\mathbf{x}_0-\mathbf{m})^t\underbrace{\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m})}_{=0}+\underbrace{\sum_{k=1}^{n}\left\|\mathbf{x}_k-\mathbf{m}\right\|^2}_{\text{不依赖于}\mathbf{x}_0}$$

样本均值是样本数据集的零维表达。
将样本数据集的空间分布，压缩为一个均值点。

简单，但不能反映样本间的差异

零维表达改为“一维”表达，将数据集空间，压缩为一条过均值点的线。

每个样本在直线上存在不同的投影，可以反映样本间的差异




$$\mathbf{x} = \mathbf{m} + a\mathbf{e}$$

$\mathbf{e}$为直线的单位向量

$a$为直线上的点到$\mathbf{m}$的距离

$$\mathbf{x}_k \leftrightarrow \mathbf{m} + a_k\mathbf{e}$$

$$J_0(\mathbf{m}) = \sum_{k=1}^{n} \left\| \mathbf{m} - \mathbf{x}_k \right\|^2 \longrightarrow J_1(a_1, \ldots a_n, \mathbf{e}) = \sum_{k=1}^{n} \left\| (\mathbf{m} + a_k\mathbf{e}) - \mathbf{x}_k \right\|^2$$

*0维平方误差* *1维平方误差*

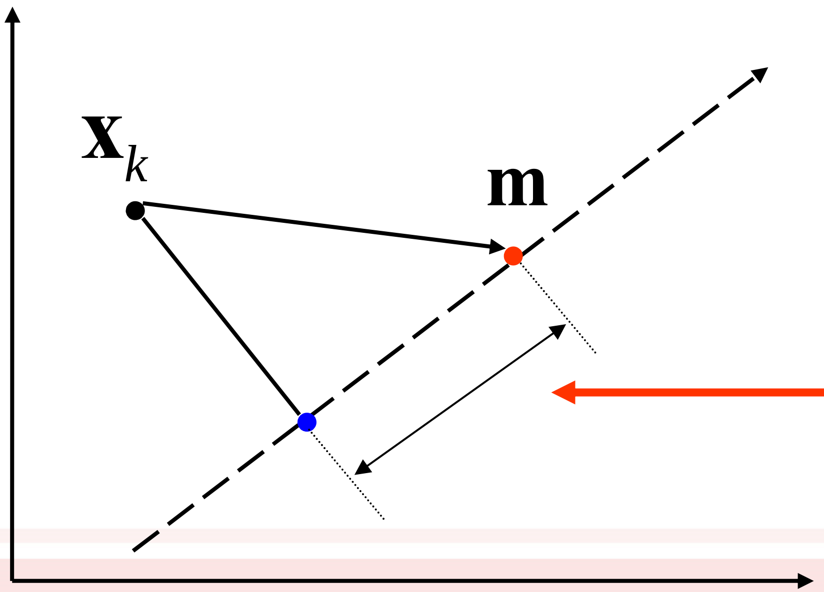
$$J_1(a_1,\ldots a_n,\mathbf{e})=\sum_{k=1}^{n}\left\|(\mathbf{m}+a_k\mathbf{e})-\mathbf{x}_k\right\|^2=\sum_{k=1}^{n}\left\|a_k\mathbf{e}-(\mathbf{x}_k-\mathbf{m})\right\|^2$$

$$=\sum_{k=1}^{n}a_k^2\underbrace{\|\mathbf{e}\|^2}_{=1}-2\sum_{k=1}^{n}a_k\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

$$\frac{\partial J_1(a_1,\ldots a_n,\mathbf{e})}{\partial a_k}=2a_k-2\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})=0$$

$$\Downarrow$$

$$a_k=\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})$$




只需把向量 $\mathbf{x}_k$ 向过 $\mathbf{m}$ 的直线垂直投影就能得到最小方差

$$\Downarrow$$

如何找到直线的最优方向？

$$J_1(a_1,\ldots a_n,\mathbf{e})=\sum_{k=1}^{n}\left\|(\mathbf{m}+a_k\mathbf{e})-\mathbf{x}_k\right\|^2=\sum_{k=1}^{n}\left\|a_k\mathbf{e}-(\mathbf{x}_k-\mathbf{m})\right\|^2$$

$$=\sum_{k=1}^{n}a_k^2\underbrace{\|\mathbf{e}\|^2}_{=1}-2\sum_{k=1}^{n}a_k\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

$$a_k=\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})$$

$$J_1(\mathbf{e})=\sum_{k=1}^{n}a_k^2-2\sum_{k=1}^{n}a_k^2+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

$$=-\sum_{k=1}^{n}\left(\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})\right)^2+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

$$=-\sum_{k=1}^{n}\mathbf{e}^t(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})^t\mathbf{e}+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2$$

协方差矩阵的**n-1**倍：散度矩阵

$$=-\mathbf{e}^t\mathbf{S}\mathbf{e}+\sum_{k=1}^{n}\|\mathbf{x}_k-\mathbf{m}\|^2 \qquad \mathbf{S}=\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})^t$$

最小化$J_1(\mathbf{e})$ $\longrightarrow$ 最大化 $\mathbf{e}^t\mathbf{S}\mathbf{e}$ ，约束条件为: $\|\mathbf{e}\|=1$

最大化 $\mathbf{e}^t\mathbf{S}\mathbf{e}$，约束条件为：$\|\mathbf{e}\|=1$ $\longrightarrow$ **Lagrange乘子法**

$$u = \mathbf{e}^t\mathbf{S}\mathbf{e} - \lambda\mathbf{e}^t\mathbf{e}$$

散度矩阵

$$\frac{\partial u}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} = 0 \quad \Rightarrow \quad \underline{\mathbf{S}}\mathbf{e} = \underline{\lambda}\mathbf{e}$$

散度矩阵的特征值

$$\mathbf{e}^t\mathbf{S}\mathbf{e} = \mathbf{e}^t\lambda\mathbf{e} = \lambda$$

为了最大化 $\mathbf{e}^t\mathbf{S}\mathbf{e}$

选取散度矩阵最大特征值 $\lambda_{\max}$

选取$\lambda_{\max}$对应的特征向量作为投影直线 $\mathbf{e}$ 的方向

# PCA算法——从0维，1维到*d'* 维

有$n$个$d$维样本，$x_1,x_2,..x_n$,

零维表达：仅用一个样本$\mathbf{x_0}$代表这些样本，使误差最小？

$$\mathbf{x}_0 = \mathbf{m} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

简单，但不能反映样本间的差异

一维表达：将这些样本，映射到过**m**的一条直线上使误差最小？

**1,选取散度矩阵** $\mathbf{S} = \sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^t$ **最大特征值** $\lambda_{\max}$

**2,选取** $\lambda_{\max}$ **对应的特征向量作为直线方向** $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

**3,将样本向直线做垂直投影**

*d'* 维表达：将这些样本，映射到以**m**为原点的*d'*维空间中*,* 使误差准则函数最小？

# PCA算法$d'$维表达：

有样本集合$\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$，其中$\mathbf{x}=(x_1,\cdots,x_d)^t$，以样本均值$\mathbf{m}$为坐标原点建立新的坐标系，则有：$\mathbf{x}=\mathbf{m}+\sum_{i=1}^{d}a_i\mathbf{e}_i$，其中$\{\mathbf{e}_i\}$为标准正交向量基：

$$\mathbf{e}_i^t\mathbf{e}_j=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$$

因此有：

$$a_i=\mathbf{e}_i^t(\mathbf{x}-\mathbf{m})$$

将特征维数降低到$d'<d$，则有对$\mathbf{x}$的近似：$\hat{\mathbf{x}}=\mathbf{m}+\sum_{i=1}^{d'}a_i\mathbf{e}_i$

误差平方和准则函数：

$$J(\mathbf{e})=\sum_{k=1}^{n}\left\|\mathbf{x}_k-\hat{\mathbf{x}}_k\right\|^2=\sum_{k=1}^{n}\left\|\sum_{i=1}^{d}a_{ik}\mathbf{e}_i-\sum_{i=1}^{d'}a_{ik}\mathbf{e}_i\right\|^2=\sum_{k=1}^{n}\left\|\sum_{i=d'+1}^{d}a_{ik}\mathbf{e}_i\right\|^2$$

$$=\sum_{k=1}^{n}\sum_{i=d'+1}^{d}a_{ik}^2=\sum_{i=d'+1}^{d}\sum_{k=1}^{n}\left[\mathbf{e}_i^t(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})^t\mathbf{e}_i\right]$$

# PCA算法d' 维表达：

$$J(\mathbf{e})=\sum_{k=1}^{n}\left\|\mathbf{x}_k-\hat{\mathbf{x}}_k\right\|^2=\sum_{k=1}^{n}\left\|\sum_{i=1}^{d}a_{ik}\mathbf{e}_i-\sum_{i=1}^{d'}a_{ik}\mathbf{e}_i\right\|^2=\sum_{k=1}^{n}\left\|\sum_{i=d'+1}^{d}a_{ik}\mathbf{e}_i\right\|^2$$

$$=\sum_{k=1}^{n}\sum_{i=d'+1}^{d}a_{ik}^2=\sum_{i=d'+1}^{d}\sum_{k=1}^{n}\left[\mathbf{e}_i^t(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})\mathbf{e}_i^t\right]$$

$$=\sum_{i=d'+1}^{d}\mathbf{e}_i^t\left[\underbrace{\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m})(\mathbf{x}_k-\mathbf{m})}_{\text{散度矩阵}}\right]\mathbf{e}_i^t=\sum_{i=d'+1}^{d}\mathbf{e}_i^t\mathbf{S}\mathbf{e}_i^t$$

最小化$J(\mathbf{e})$，约束条件为：$\|\mathbf{e}\|=1$ 使用拉格朗日乘数法：

$$J'(\mathbf{e})=\sum_{i=d'+1}^{d}\left[\mathbf{e}_i^T\mathbf{S}\mathbf{e}_i-\lambda_i\left(\mathbf{e}_i^T\mathbf{e}_i-1\right)\right]$$

$$J'(\mathbf{e}) = \sum_{i=d'+1}^{d} \left[ \mathbf{e}_i^T \mathbf{S} \mathbf{e}_i - \lambda_i \left( \mathbf{e}_i^T \mathbf{e}_i - 1 \right) \right]$$

$$\frac{\partial J'(\mathbf{e})}{\partial \mathbf{e}_i} = 2\mathbf{S}\mathbf{e}_i - 2\lambda_i \mathbf{e}_i = 0 \longrightarrow \mathbf{S}\mathbf{e}_i = \lambda_i \mathbf{e}_i$$

$\lambda_i$ 为 $\mathbf{S}$ 的特征值，$\mathbf{e}_i$ 为 $\mathbf{S}$ 的特征矢量。

$$J(\mathbf{e}) = \sum_{i=d'+1}^{d} \mathbf{e}_i^T \mathbf{S} \mathbf{e}_i = \sum_{i=d'+1}^{d} \lambda_i \mathbf{e}_i^T \mathbf{e}_i = \sum_{i=d'+1}^{d} \lambda_i$$

要使 $J(\mathbf{e})$ 最小，只需将s的特征值由大到小排序，选择最大的前 $d'$ 个特征值对应的特征向量构成一个新的 $d'$ 维坐标系，将样本向新的坐标系的各个轴上投影，计算出新的特征矢量

$$(x_1, \cdots, x_d)^T \to (a_1, \cdots, a_{d'})^T \quad 其中 \quad a_i = \mathbf{e}_i^T (\mathbf{x} - \mathbf{m})$$

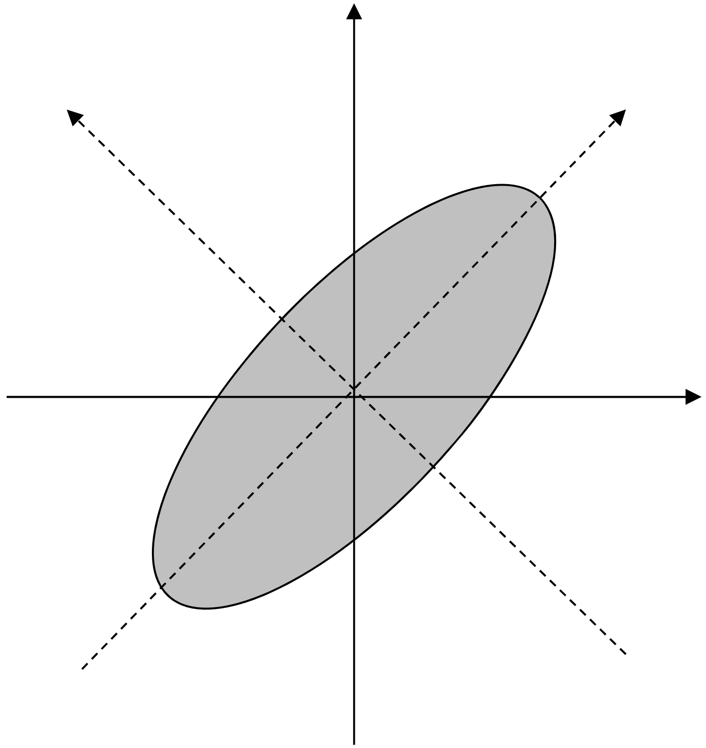
# PCA算法

1. 利用训练样本集合计算样本的均值m和散度矩阵S；
2. 计算S的特征值，并由大到小排序；
3. 选择前d’个特征值对应的特征向量作成一个变换矩阵$E=[e_1, e_2, \ldots, e_{d'}]$；
4. 训练和识别时，每一个输入的d维特征矢量x可以转换为d’维的新特征矢量y：

$$y = E^t(x-m)。$$

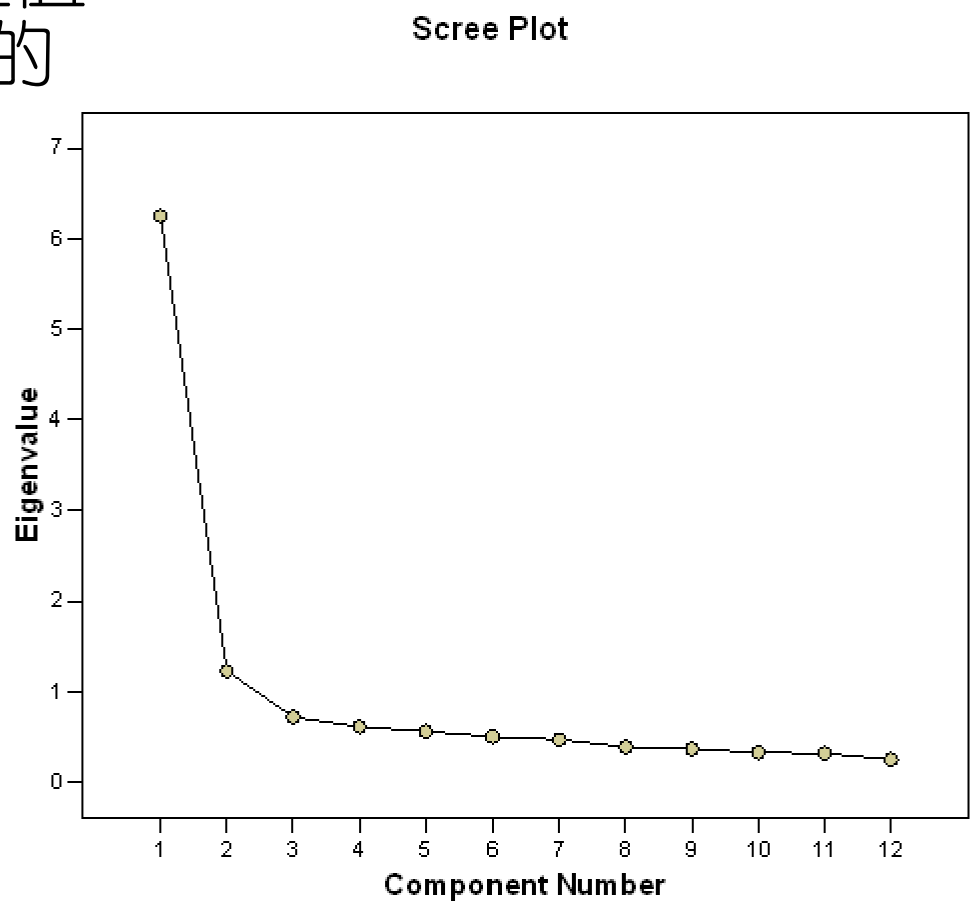
# PCA的讨论

- 由于S是实对称阵，因此特征向量是正交的；
- 将数据向新的坐标轴投影之后，特征之间是不相关的；
- PCA仅依赖于样本数据的均值和协方差矩阵，有些分布可以由这两个量刻画，有些不行。



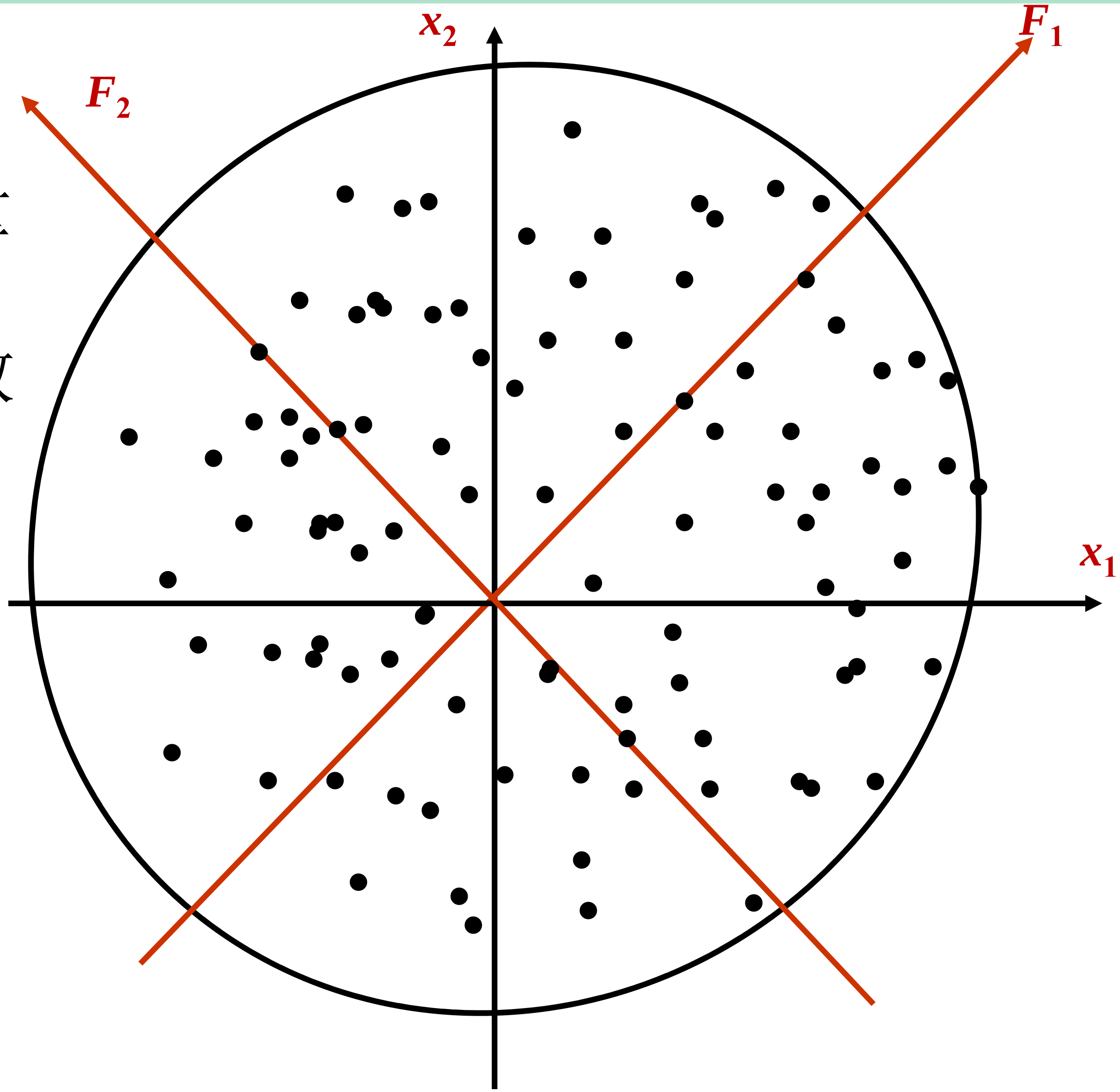
# PCA的讨论

- 通常，最大的几个特征值占据了所有特征值之和的绝大部分；
- 少数几个最大特征值对应的特征向量即可表示原数据中的绝大部分信息，而剩下的小部分，通常可以认为是数据噪声而丢掉。



# PCA的讨论

- 当原始变量不相关时，PCA没有效果。



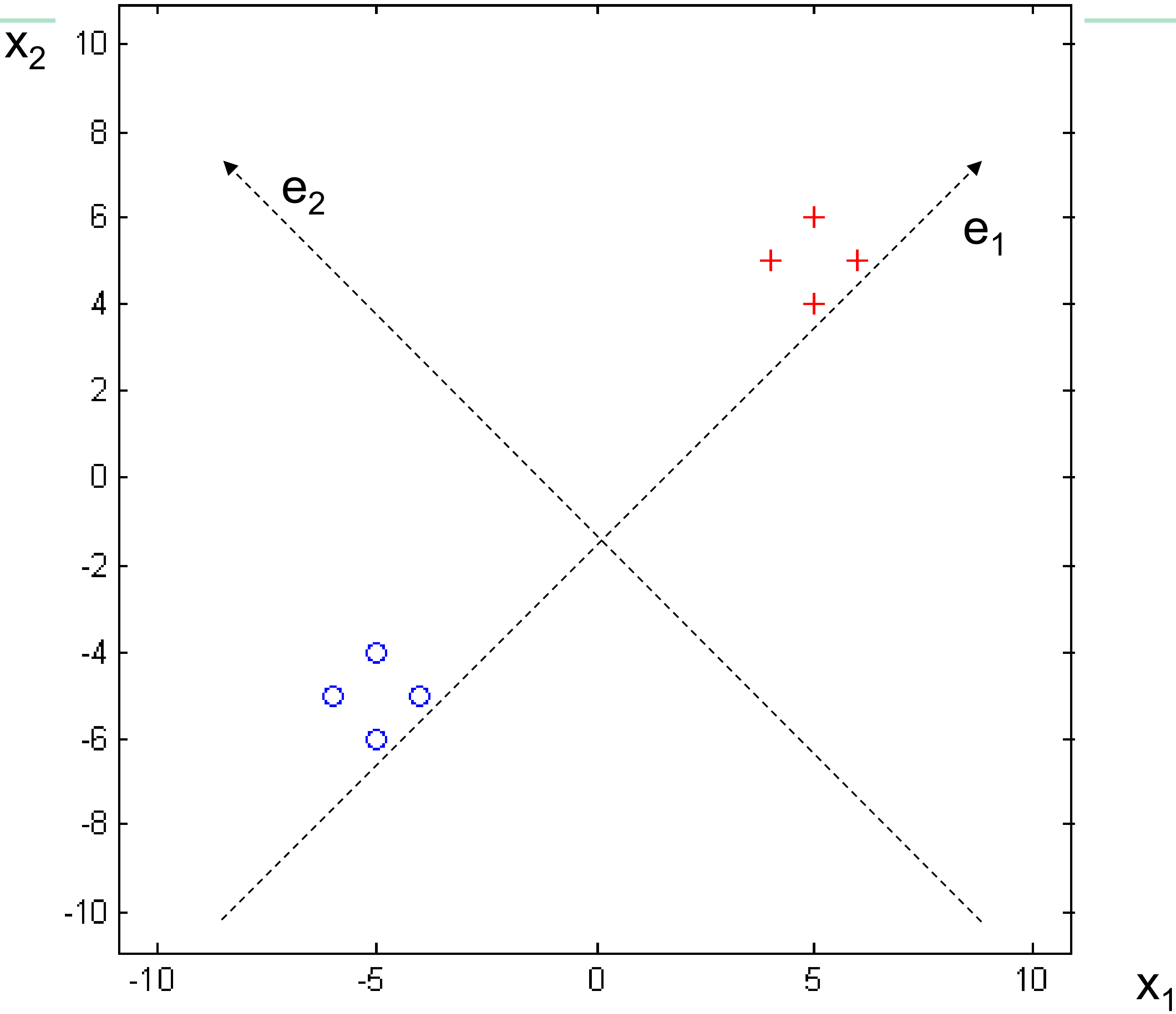
# 例

- 有两类问题的训练样本：

$$\omega_1:(-5,-4)^t,(-4,-5)^t,(-5,-6)^t,(-6,-5)^t$$

$$\omega_2:(5,4)^t,(4,5)^t,(5,6)^t,(6,5)^t$$

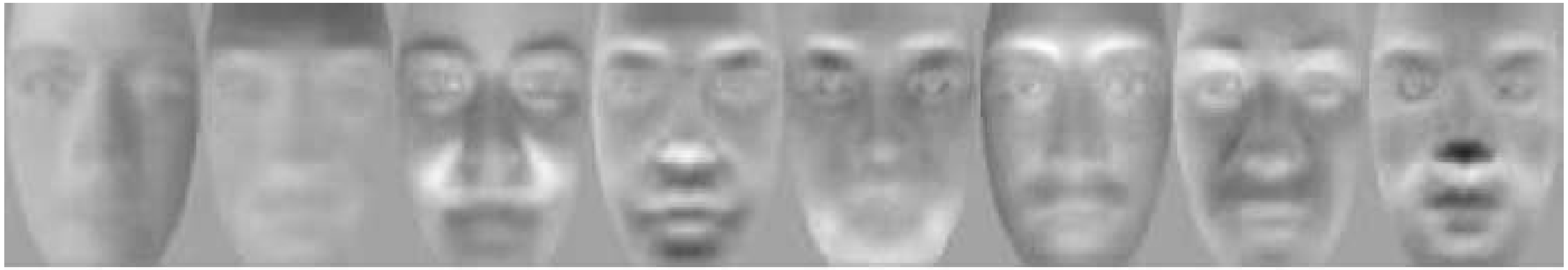
将特征由2维压缩为1维。



x2
10
8
6
4
2
0
-2
-4
-6
-8
-10
-10
-5
0
5
10
x1
e2
e1

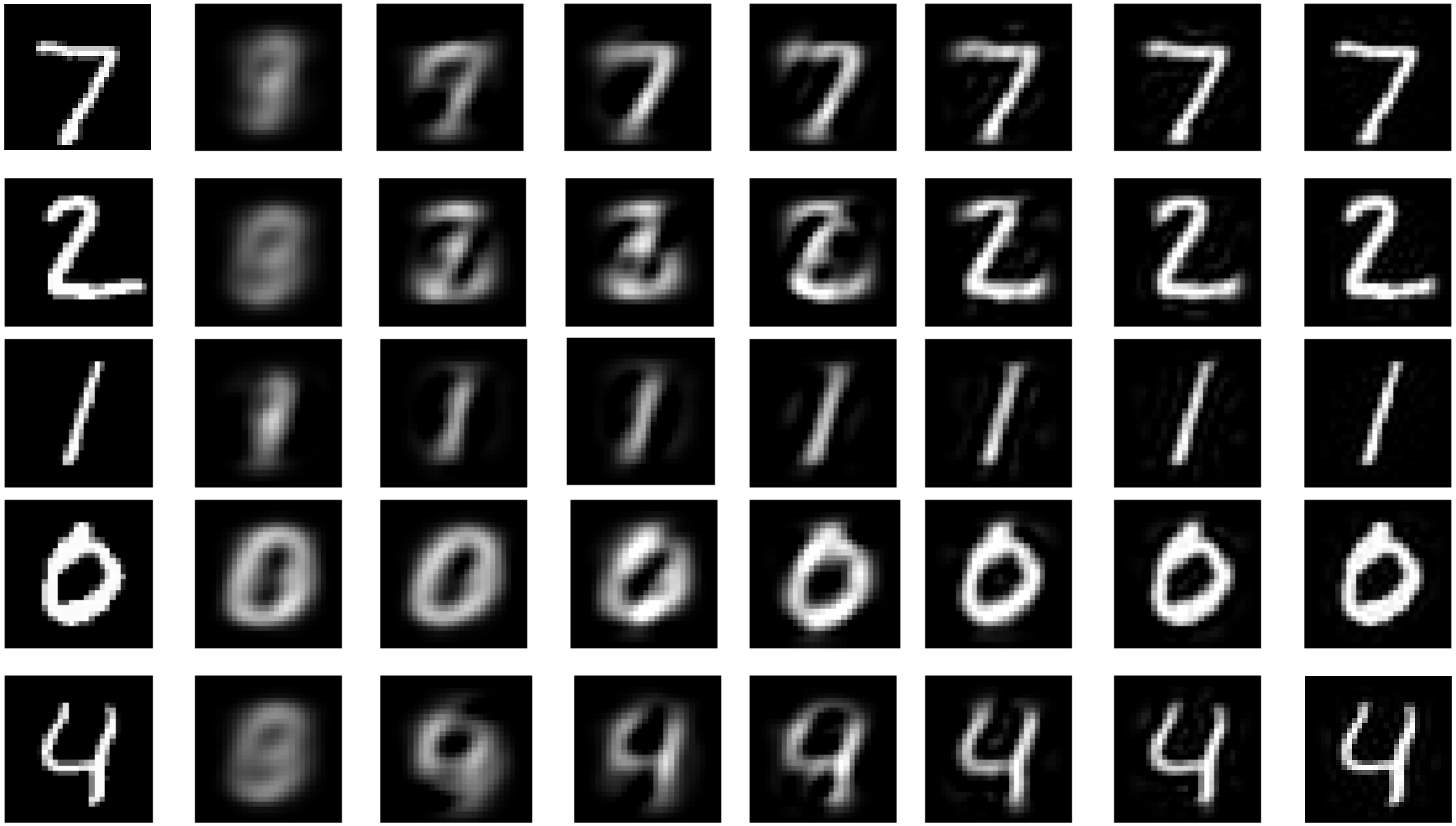
# 特征人脸

$e_1$ $e_2$ $e_3$ $e_4$ $e_5$ $e_6$ $e_7$ $e_8$



# PCA重构

原图像 d’=1 5 10 20 50 100 200



# 主成分分析

## 最近重构性

- 对样本进行中心化，$\sum_i \boldsymbol{x}_i = \mathbf{0}$，再假定投影变换后得到的新坐标系为$\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d\}$，其中$\boldsymbol{w}_i$是标准正交基向量，

$$||\boldsymbol{w}_i||_2 = 1, \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{w}_j = 0 (i \neq j).$$

- 若丢弃新坐标系中的部分坐标，即将维度降低到 $d' < d$，则样本点在低维坐标系中的投影是$\boldsymbol{z}_i = (z_{i1}; z_{i2}; \ldots; z_{id'})$，$z_{ij} = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}_i$ 是 $\boldsymbol{x}_i$ 在低维坐标下第$j$维的坐标，若基于$\boldsymbol{z}_i$来重构$\boldsymbol{x}_i$，则会得到

$$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij} \boldsymbol{w}_j .$$

# 主成分分析

## 最近重构性

- 考虑整个训练集，原样本点$\boldsymbol{x}_i$与基于投影重构的样本点$\hat{\boldsymbol{x}}_i$之间的距离为

$$\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i+\mathrm{const}$$

$$\propto-\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\left(\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right)\mathbf{W}\right).$$

- 根据最近重构性应最小化上式。考虑到$\boldsymbol{w}_j$是标准正交基，$\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$是协方差矩阵，有

$$\begin{aligned}&\min_{\mathbf{W}}\quad -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})\\&\ \mathrm{s.t.}\quad \mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}.\end{aligned}$$

这就是主成分分析的优化目标。

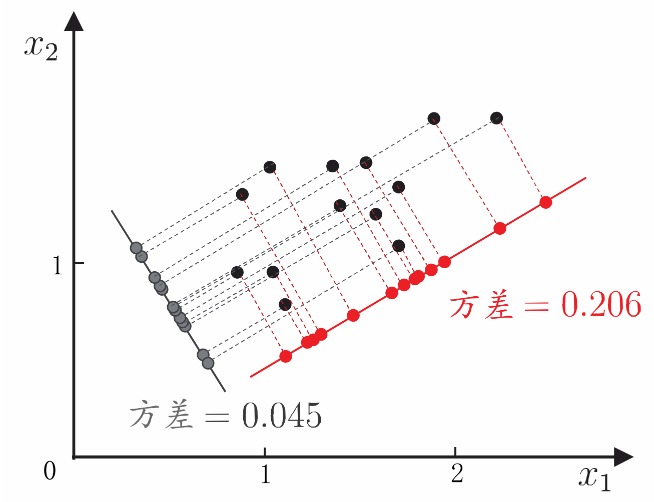
# 主成分分析

## 最大可分性

- 样本点$\boldsymbol{x}_i$在新空间中超平面上的投影是$\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i$，若所有样本点的投影能尽可能分开，则应该使得投影后样本点的方差最大化。若投影后样本点的方差是$\sum_i \mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\mathbf{W}$，于是优化目标可写为

$$\max_{\mathbf{W}} \quad \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$$
$$\mathrm{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}.$$

显然与
$$\min_{\mathbf{W}} \quad -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$$
$$\mathrm{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}.$$
等价。

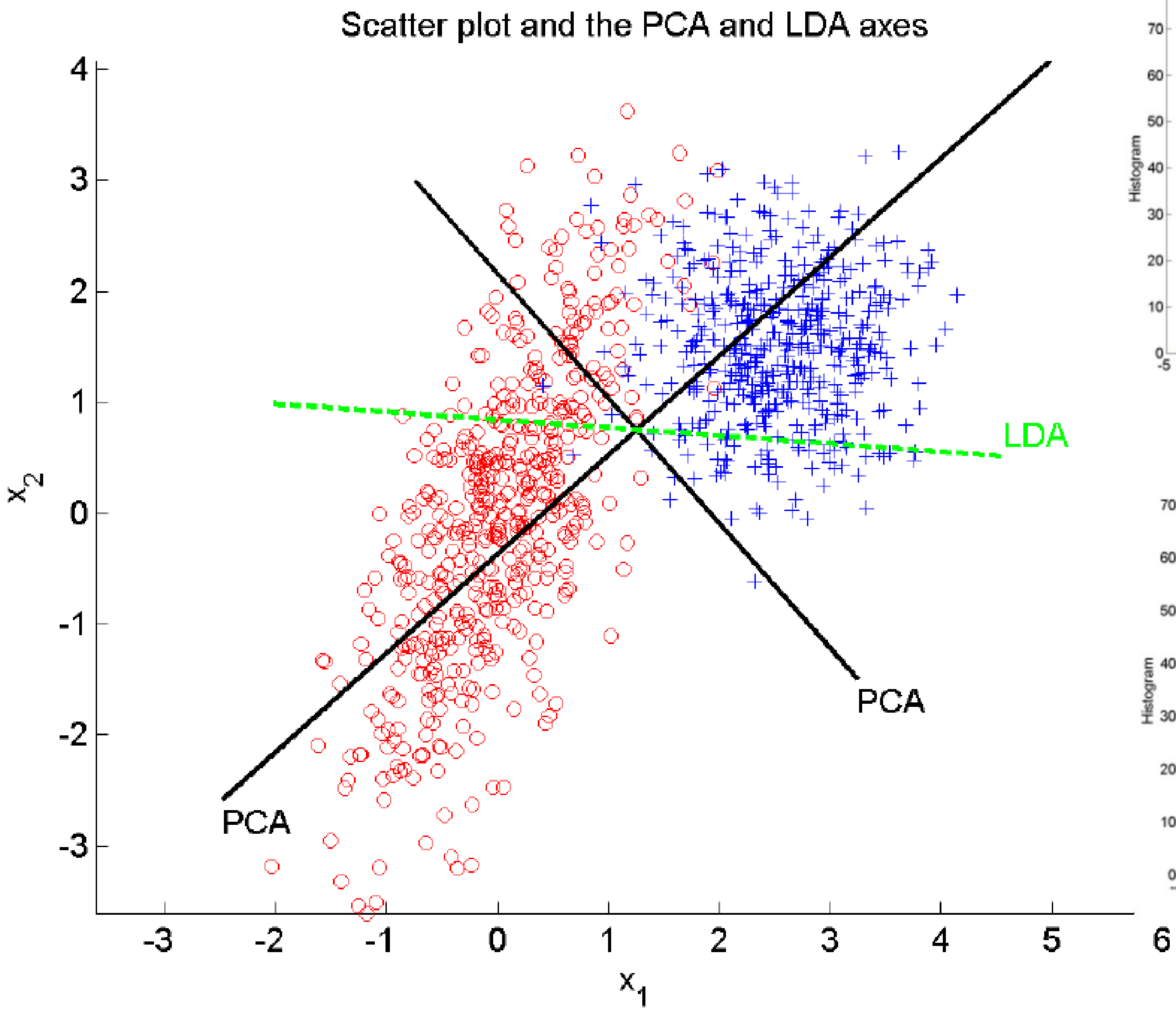


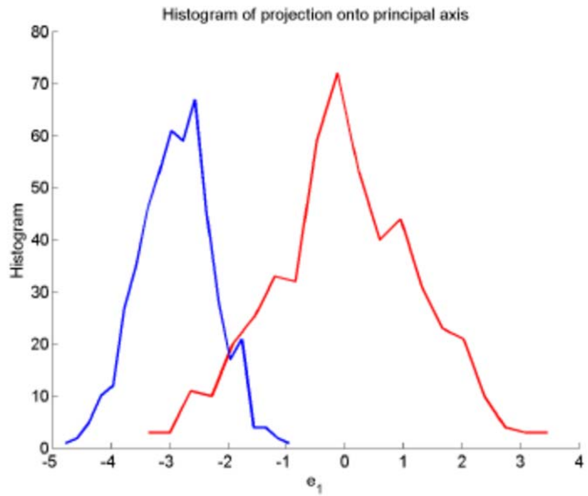
# LDA与PCA

- PCA将所有的样本作为一个整体对待，寻找一个均方误差最小意义下的最优线性映射，也就是寻找用来有效表示数据（从最小均方误差的意义上讲）的主轴方向，而没有考虑样本的类别属性，它所忽略的投影方向有可能恰恰包含了重要的可分性信息；

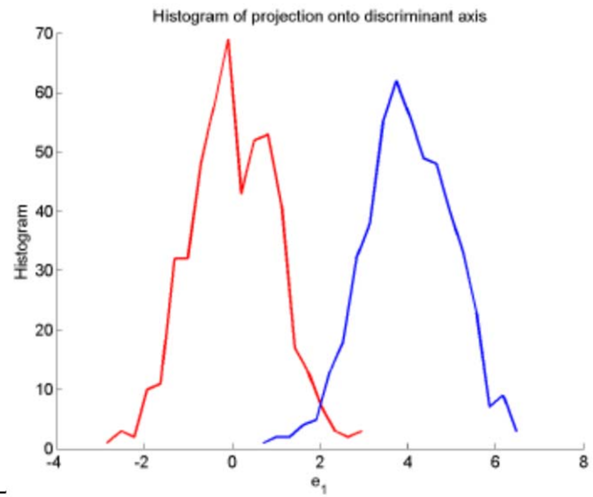
- LDA则是在可分性最大意义下的最优线性映射，充分保留了样本的类别可分性信息，是寻找用来有效分类的方向。

# LDA与PCA




投影到主成分方向




投影到LDA方向

# 成分分析的其它问题

- 独立成分分析( ICA, Independent Component Analysis )：PCA去除掉的是特征之间的相关性，但不相关不等于相互独立，独立是更强的要求。ICA试图使特征之间相互独立
- 多维尺度变换(MDS, Multidimensional Scaling)
- 典型相关分析(CCA, Canonical Correlation Analysis)
- 偏最小二乘(PLS, Partial Least Square)

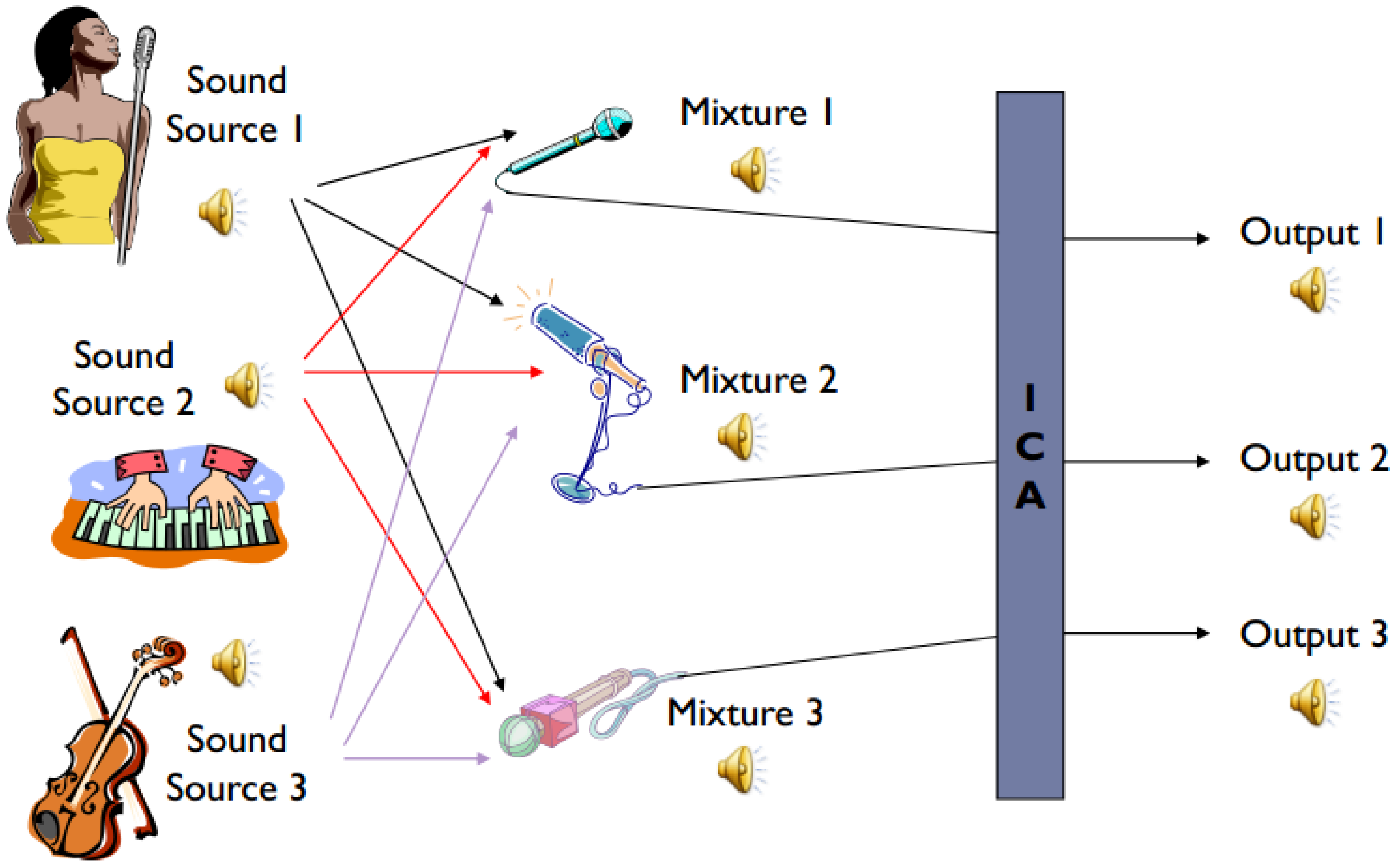
# 成分分析的其它问题

- Unsupervised
  - Latent Semantic Indexing (LSI): truncated SVD
  - Independent Component Analysis (ICA)
  - Principal Component Analysis (PCA)
  - Manifold learning algorithms
- Supervised
  - Linear Discriminant Analysis (LDA)
  - Canonical Correlation Analysis (CCA)
  - Partial Least Squares (PLS)
- Semi-supervised

# 成分分析的其它问题

- Linear
  - Latent Semantic Indexing (LSI): truncated SVD
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Canonical Correlation Analysis (CCA)
  - Partial Least Squares (PLS)
- Nonlinear
  - Nonlinear feature reduction using kernels
  - Manifold learning

# Independent Component Analysis



# Manifold Learning

- Discover low dimensional representations (smooth manifold) for data in high dimension.
- A manifold is a topological space which is locally Euclidean
- An example of nonlinear manifold: