《机器学习理论与应用》课件

4.4主成分分析（PCA）

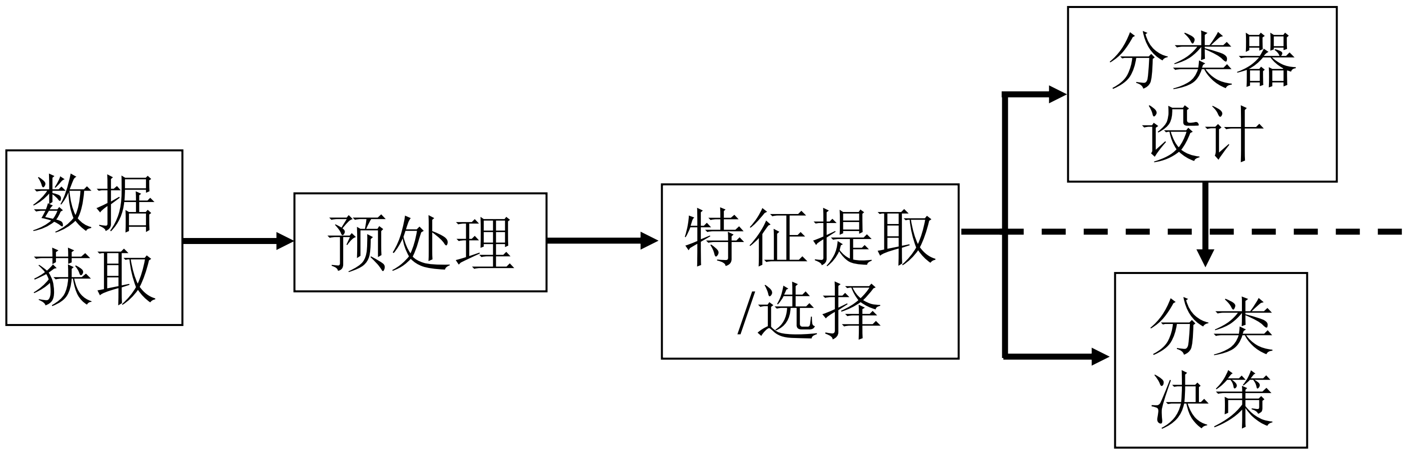
北京航空航天大學
BEIHANG UNIVERSITY

主成分分析

- 主成分分析方法是近几年来的研究热点，以它为代表的方法被称为子空间学习方法
- 在人脸识别领域获得成功的应用
- 主要用来进行特征提取

模式识别系统

- 模式识别系统的基本组成



特征提取：从输入信号中提取有效特征，其重要特点之一是：降维，简化了模式识别系统

特征提取

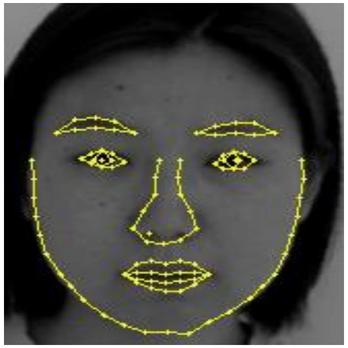
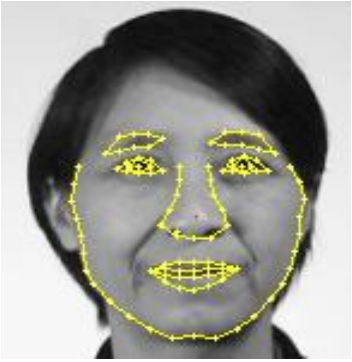
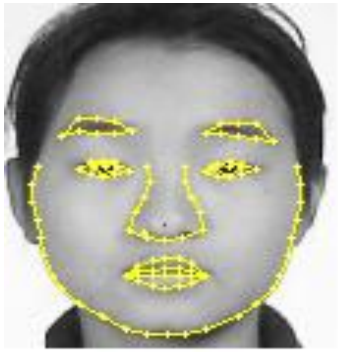
- 特征提取：从输入信号中提取有效特征，
- 其重要特点之一是：降维，简化了模式识别系统

人脸图像中的特征提取

- 从给定的一个输入人脸图像中提取有效信息。通常图片都比较大比如 **64x64**
- 通过特征提取为**83**个点，提取的点和点之间的几何位置信息进行识别。

参考**Active Shape Model, CVIU, 1995.**

不同的应用特征提取方法都不一致，是否存在一些通用的特征提取方法：基于主成份分析的特征提取方法


基于主成份分析的特征提取

- 主成份分析（Principal Component Analysis, PCA ）是一种利用线性映射来进行数据降维的方法，并去除数据的相关性；*且最大限度保持原始数据的方差信息。*

线性映射，去相关性，方差保持

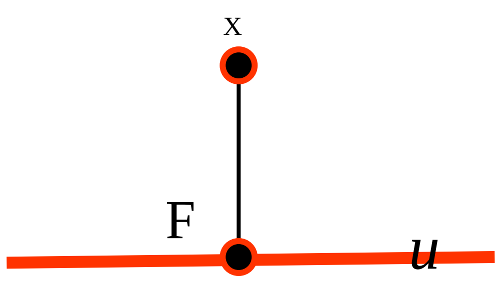
线性映射的意义

- **P**维向量**X**到一维向量**F**的一个线性映射表示为

$$\mathbf{F} = \sum_{i=1}^{p} u_i X_i = u_1 X_1 + u_2 X_2 + u_3 X_3 + \cdots + u_p X_p$$

$$F = u^T X$$

相当于加权求和，每一组权重系数为一个主成份,它的维数跟输入数据**维数相同**



$$X = (1,1)^T$$

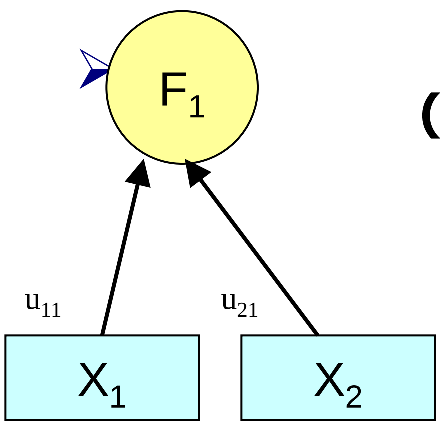
$$u = (1,0)^T$$

$$F = u^T X = 1*1+1*0 = 1$$

- 高等代数：**F**的几何意义表示为**x**在投影方向**u**上的**投影点**。上面例子在笛卡尔坐标系表示在横坐标上作一条垂线的交点。

基于线性映射主成份分析

- 主成份分析之计算方式：
- X是p维向量，主成份分析就是要把这p维原始向量通过***线性映射***变成K维新向量的过程.(k≤p）



(1,1)->(1)

$X=(1,1)^T$

$u_1=(1,0)^T$

$F_1=1*1+1*0=1$

$$F_1=u_{11}X_1+u_{12}X_2+u_{13}X_3+\cdots+u_{1p}X_p$$

$$F_2=u_{21}X_1+u_{22}X_2+u_{23}X_3+\cdots+u_{2p}X_p$$

$$\vdots$$

$$F_k=u_{k1}X_1+u_{k2}X_2+u_{k3}X_3+\cdots+u_{kp}X_p$$

去除数据的**相关性**，只需让各个主成份**正交**，**正交的基构成的空间称之为*子空间***

主成份分析的例子

- 在社会经济的研究中，为了全面系统的分析和研究问题，必须考虑许多经济指标，这些指标能从不同的侧面反映我们所研究的对象的特征，但在某种程度上存在信息的重叠，具有一定的相关性。
- 主成分分析是把各变量之间互相关联的复杂关系进行简化分析的方法。

主成份分析的例子

- 一项十分著名的工作是美国的统计学家斯通(stone)在1947年关于国民经济的研究。他曾利用美国1929一1938年各年的数据，得到了17个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。

- 在进行主成份分析后，竟以97.4％的精度，用三新变量就取代了原17个变量的**方差**信息。根据经济学知识，斯通给这三个新变量分别命名为总收入F1、总收入变化率F2和经济发展或衰退的趋势F3。

关于方差保持

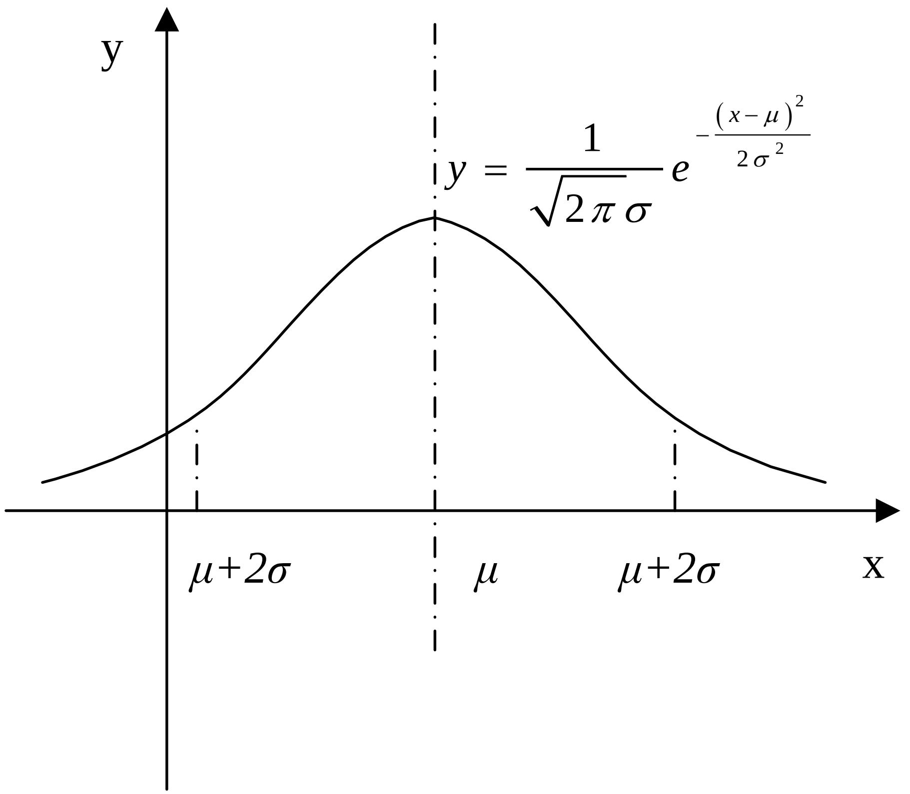
- 利用3维向量能够保持原始17维向量，97.4%的方差信息

- 核心提示是在低维空间能够尽可能多保持原始空间数据的方差

- 数据集合中各数据与平均样本的差的平方和的平均数叫做样本**方差**

- 在我们所讨论的问题中都有一个近似的假设，假定数据满足**高斯分布**或者**近似满足高斯分布**
- 问题：高斯分布需要几个参数刻画？
- 均值，方差（离散程度）

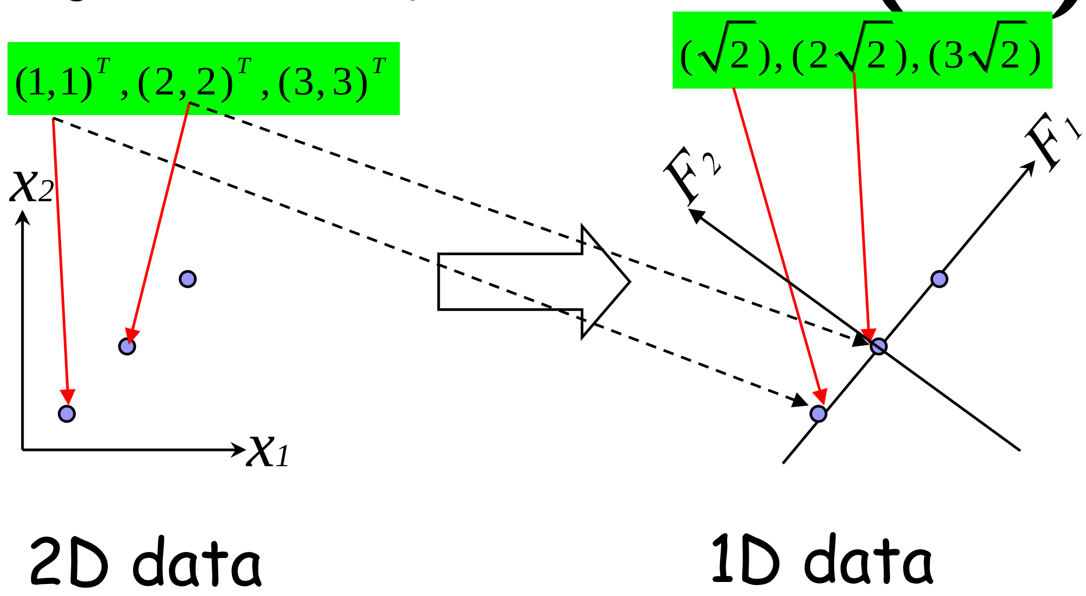
- **思考问题：为什么主成分分析基于协方差矩阵？**



➢ 基于主成份分析特征提取的基本思想

- 主成份分析试图在力保数据信息丢失最少的原则下，对高维空间的数据降维处理。
- 很显然，识别系统在一个低维空间要比在一个高维空间容易得多。
- 能够去除数据的相关性，从而进行有效的特征提取

➢ 主成份分析的例子（一）



方差越大，数据的分布越**分散**，从而越能保持原始空间中的距离信息 $\frac{1}{n}\sum_{l=1}^{n}(x_l-\bar{x})^T(x_l-\bar{x})$

原始数据空间中,类别信息没有丢失

但是维度减少50%

➢ 主成份分析的例子（一）

- ➢ 在几何上投影方向总是沿着数据的分布最分散方向

- ➢ 为了去掉相关性，投影方向之间应该保持正交

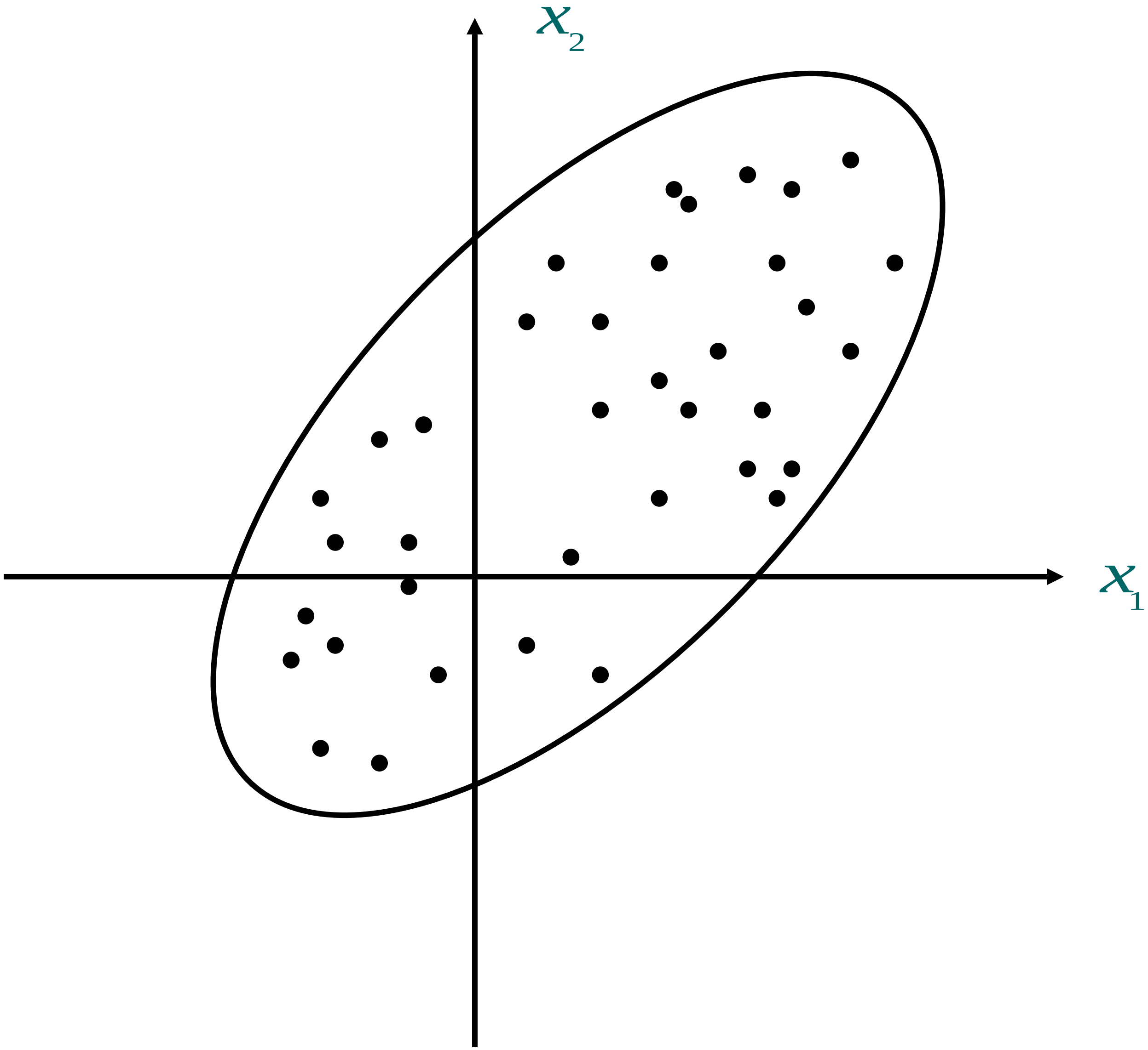
➢ 主成份分析的例子（二）

- 为了加深理解，我们在二维空间中讨论主成份的几何意义。

- 设有n个样本，每个样本有二维即x_l和x_2，在由x_l和x_2 所确定的二维平面中，n个样本点所散布的情况如椭圆状。

平移、旋转坐标轴

主成份分析的几何解释



- 由图可以看出这n个样本点无论是沿着x_l 轴方向或x_2轴方向都具有较大的离散性，其离散程度可以分别用观测变量x_l 的方差和x_2的方差定量地表示。

- 如果只考虑x_l和x_2 中的任何一个，那么包含在原始数据中的信息将会有损失。

➢ 将x_l 轴和x_2轴先平移，再同时按逆时针方向旋转θ角度，得到新坐标轴F_l和F_2，则

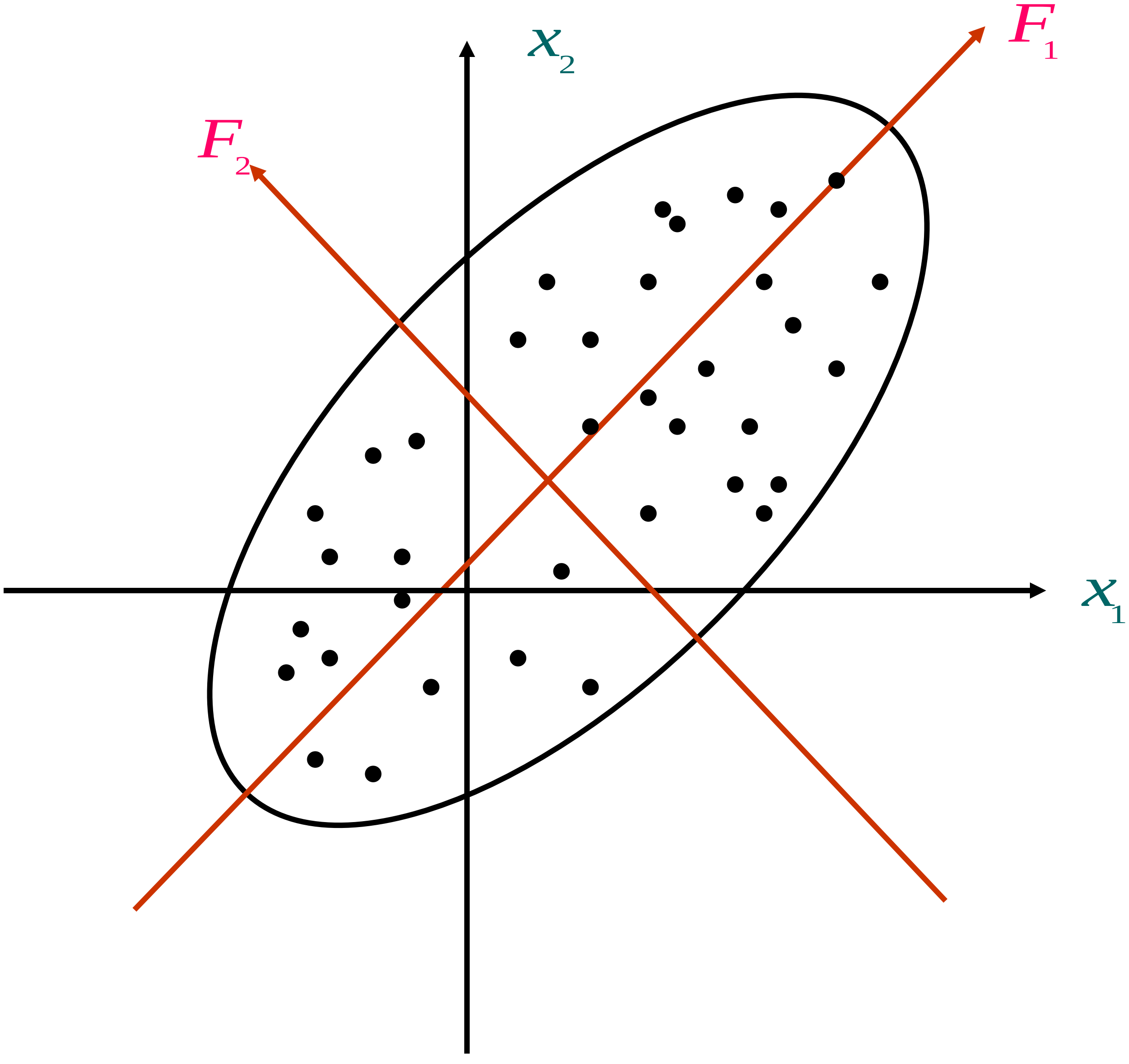
$$\begin{cases} y_1 = x_1 \cos\theta + x_2 \sin\theta \\ y_2 = -x_1 \sin\theta + x_2 \cos\theta \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = U'x$$

U'为正交旋转变换矩阵

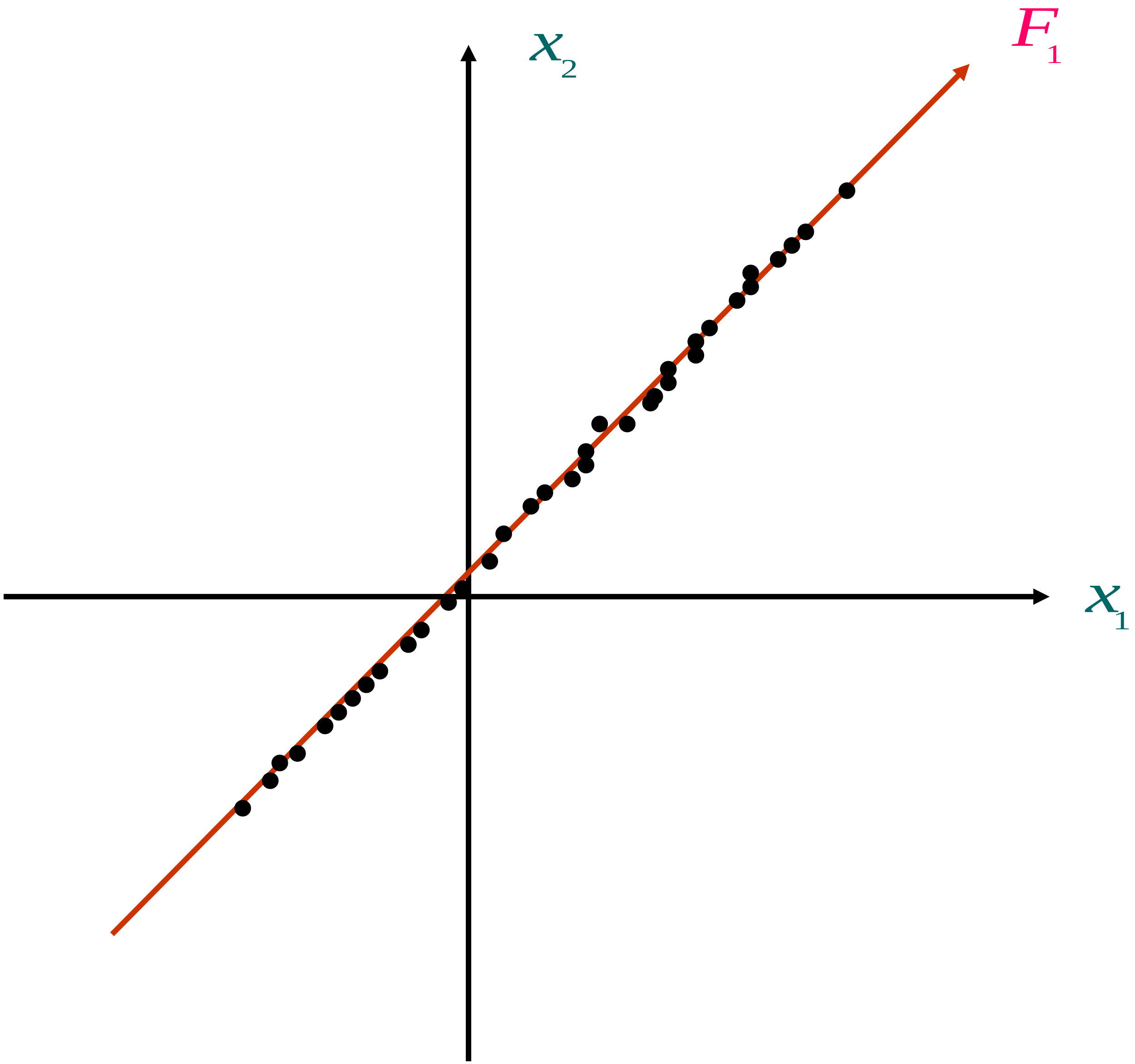
平移、旋转坐标轴

主成份分析的几何解释



- 由图可以看出这n个样本点沿着F_l 轴方向有最大的离散性，这是第一个主成份
- 为了去掉相关性，第二个主成份应该正交于第一个主成份
- 如果只考虑F_l和F_2 中的任何一个，那么包含在原始数据中的信息将会有损失。
- 根据系统精度的要求，可以只选择F_l

主成份分析的几何解释



表示方法

- 实际问题总是变成数学问题，然后才是用机器去解决
- X 表示变量
- 如果X表示向量，X_i表示向量的第i个分量
- 如果X表示矩阵，X_i表示矩阵的第i个向量
- X_{ij}表示第j个样本的第i个分量

数学模型

- 假设我们所讨论的实际问题中，X是p维变量，记为X_1，X_2，…，X_p，主成分分析就是要把这p个变量的问题，转变为讨论p个变量的线性组合的问题，而这些新的分量F_1，F_2，…，F_k($k \leqslant p$)，按照保留主要信息量的原则充分反映原变量的信息，并且相互独立。

➢ 这种由讨论多维变量降为维数较低的变量的过程在数学上就叫做降维。主成份分析通常的做法是，寻求向量的线性组合F_i。

$$F_1 = u_{11}X_1 + u_{12}X_2 + \quad + u_{1p}X_p$$

$$F_2 = u_{21}X_1 + u_{22}X_2 + \quad + u_{2p}X_p$$

$$F_k = u_{k1}X_1 + u_{k2}X_2 + \quad + u_{kp}X_p$$

满足如下的条件：

- **每个主成份的系数平方和为1, 即**

$$u_{i1}^{2} + u_{i2}^{2} + \quad + u_{ip}^{2} = 1$$

- **主成份之间相互独立，即无重叠的信息, 即**

$$Cov（F_i，F_j）= 0，i \neq j，i，j = 1，2，\Lambda，p$$

- **主成份的方差依次递减，重要性依次递减，即**

$$Var（F_1）\geq Var(F_2) \geq \Lambda \geq Var(F_p)$$

➢主成份的数学上的计算

一、两个线性代数的结论

➢ 1、若A是p阶正定或者半正定实阵，则一定可以找到正交阵U，使

$$\mathbf{U}^T\mathbf{A}\mathbf{U}=\begin{bmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & 0 \\ & & & \\ 0 & 0 & & \lambda_p \end{bmatrix}_{p \times p}$$

其中$\lambda_i, i=1.2.\Lambda\ p$ 是A的特征根。

- 2、若上述矩阵的特征根所对应的单位特征向量为 $\mathbf{u}_1,\cdots,\mathbf{u}_p$

令 $$\mathbf{U}=(\mathbf{u}_1,\cdots,\mathbf{u}_p)=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

则实对称阵 $\mathbf{A}$ 属于不同特征根所对应的特征向量是正交的，即有 $\mathbf{U}^T\mathbf{U}=\mathbf{U}\mathbf{U}^T=\mathbf{I}$

(二) 基于协方差矩阵的特征值分解

$$F = u^T X, \bar{F} = \frac{1}{n}\sum_F F$$

$$\text{Max}: \frac{1}{n-1}\sum_F (F-\bar{F})(F-\bar{F})^T = \frac{1}{n-1}\sum_x (u^T(X-\bar{X}))(u^T(X-\bar{X}))^T$$

$$\text{Max}: \frac{1}{n-1}\sum_x u^T(X-\bar{X})(X-\bar{X})^T u = u^T\left(\underline{\frac{1}{n-1}\sum_x (X-\bar{X})(X-\bar{X})^T}\right)u$$

$$\text{Constraint}: u^T u = 1$$

$$\sum\nolimits_x$$

引入拉格朗日乘子

$$J(u) = u^T \sum\nolimits_x u - \lambda(u^T u - 1)$$

$$\frac{\partial J(u)}{\partial u} = 2\sum\nolimits_x u - 2\lambda u = 0$$

$$\sum\nolimits_x u = \lambda u -> u^T \sum\nolimits_x u = \lambda$$

$$\sum\nolimits_x = \frac{1}{n-1}\sum_x (X-\bar{X})(X-\bar{X})^T \quad \begin{pmatrix} x_{1j}-\bar{x}_1 \\ x_{2j}-\bar{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ x_{pj}-\bar{x}_p \end{pmatrix}(x_{1j}-\bar{x}_1, x_{2j}-\bar{x}_2, \ldots, x_{pj}-\bar{x}_p)$$

- 设X的协方差阵为
$$\boldsymbol{\Sigma}_{\mathbf{x}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

- 由于Σ_{x}为对称阵，则利用线性代数的知识可得，存在正交阵U，使得

$$\mathbf{U}^T \boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{U} = \begin{bmatrix} \lambda_1 & & 0 \\ & & \\ 0 & & \lambda_p \end{bmatrix}$$

主成份分析的步骤

- 基于协方差矩阵

$$\Sigma_x = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \right)_{p\times p}$$

$$\mathbf{X}_i = \left(x_{1i}, \quad x_{2i}, \quad \cdots, \quad x_{pi} \right)' \quad (i = 1,2, \quad \cdots, \quad n)$$

第一步：由X的协方差阵Σ$_x$，求出其特征根，即解方程 $|\boldsymbol{\Sigma} - \lambda \mathbf{I}| = 0$，可得特征根 $\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_p \geq 0$

第二步：求出分别所对应的特征向量U_1，U_2，…，U_p，

$$\mathbf{U}_{\mathbf{i}} = \left(u_{1i}, \quad u_{2i}, \quad , \quad u_{pi}\right)^T$$

第三步：给出恰当的主成分个数。

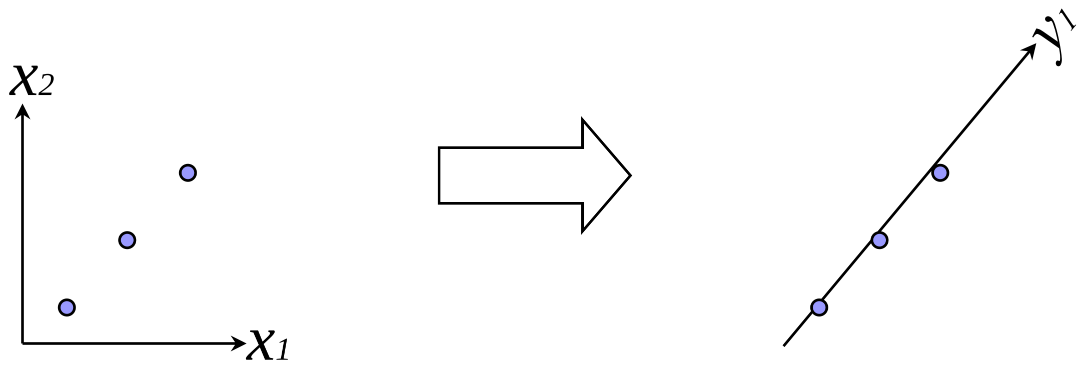
$$F_{\mathbf{i}} = \mathbf{U}_i^T \mathbf{X}, \quad i = 1,2, \quad , \quad k\,(k \le p)$$

第四步：计算所选出的k个主成份的得分。将原始数据的中心化值：

$$\mathbf{X}_{\mathbf{i}}^* = \mathbf{X}_{\mathbf{i}} - \overline{\mathbf{X}} = \left(x_{1i} - \overline{x}_1, \quad x_{2i} - \overline{x}_2, \Lambda \ , \quad x_{pi} - \overline{x}_p\right)'$$

代入前k个主成分的表达式，分别计算出各单位k个主成分的得分，并按得分值的大小排队。

➢ 主成分分析的例子（一）



2D data 1D data

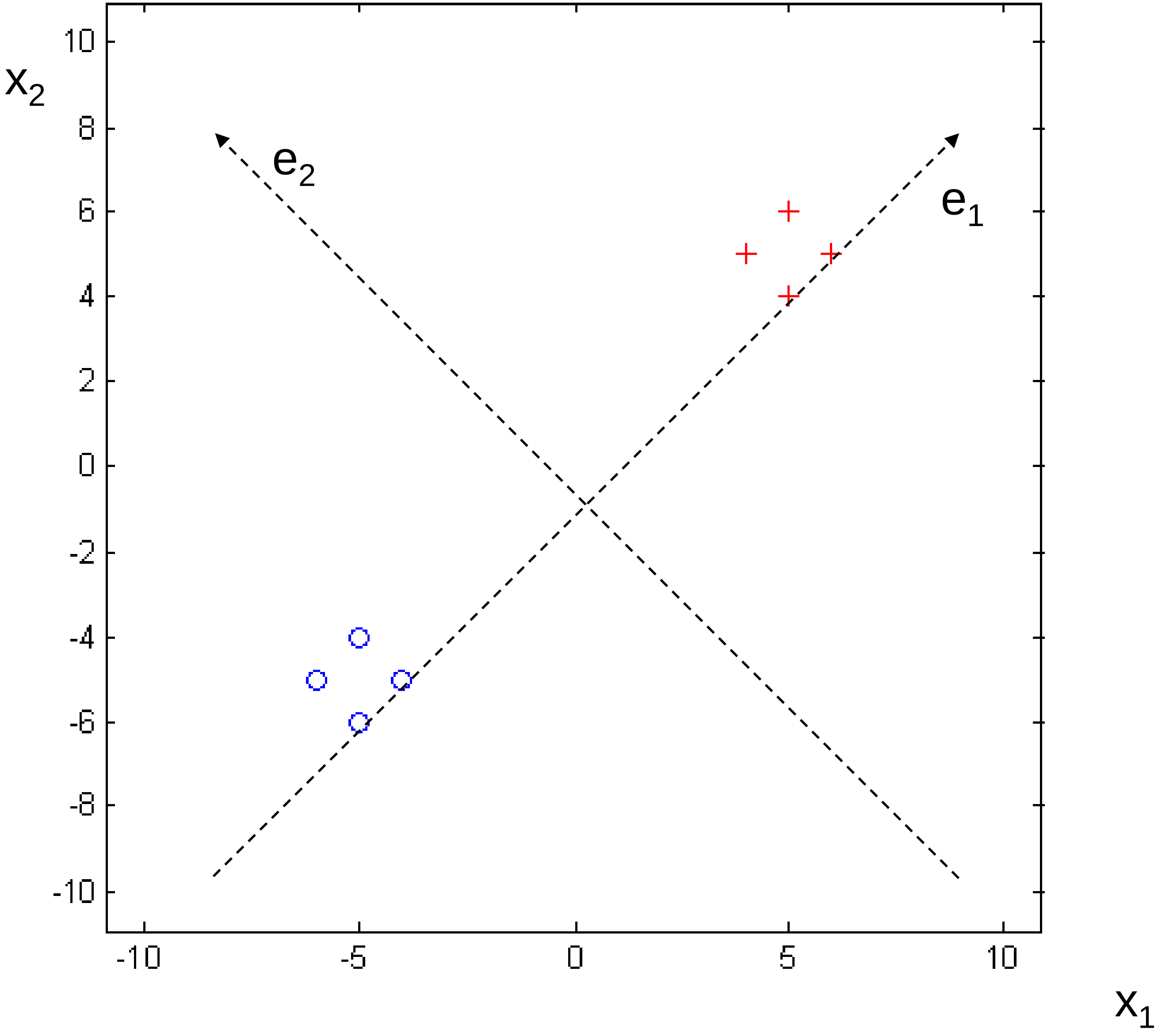
- 3个点(1,1)(2,2)(3,3)
  - 特征向量？特征值？

- 已知数据集合：

$$\Omega_1:(-5,-5)^T,(-5,-4)^T,(-4,-5)^T,(-5,-6)^T,(-6,-5)^T$$

$$\Omega_2:(5,5)^T,(5,4)^T,(4,5)^T,(5,6)^T,(6,5)^T$$

将特征由2维压缩为1维。



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

➢主成份分析的性质

➢一、均值 $E(\mathbf{U}^T x) = \mathbf{U}^T \bar{x}$

➢二、方差为所有特征根之和

$$\sum_{i=1}^{p} Var(F_i) = \lambda_1 + \lambda_2 + \Lambda + \lambda_p = \sigma_1^2 + \sigma_2^2 + \Lambda + \sigma_p^2$$

说明主成分分析把P维随机变量的总方差分解成为P个不相关的随机变量的方差之和。

协方差矩阵Σ的对角线上的元素之和等于特征根之和。

➢ 三、如何选择主成份个数

1）贡献率：第i个主成份的方差在全部方差中所占比重 $\lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ ，称为贡献率，反映了原来i个特征向量的信息，有多大的提取信息能力 。

2）累积贡献率：前k个主成份共有多大的综合能力，用这k个主成份的方差和在全部方差中所占比重

$$\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$$

来描述，称为累积贡献率。

➢ 三、如何选择主成份个数

累计贡献率大小反映m个主成分提取了$x_1, x_2, \ldots, x_p$的多少信息，但没有表达某个变量被提取了多少信息，为此引入下述概念。

3）将前m个主成分$y_1, y_2, \ldots, y_m$对原始变量x_i的贡献率定义为x_i与$y_1, y_2, \ldots, y_m$之间的相关系数的平方，

即$v_i^{(m)} = \sum_{i=1}^{m} \frac{\lambda_k u_{ik}^2}{\sigma_{ii}}$

- 我们进行主成份分析的目的之一是希望用尽可能少的主成分F_1，F_2，…，F_k（$k \leqslant p$）代替原来的P维向量。
- 到底应该选择多少个主成份，在实际工作中，主成分个数的多少取决于能够反映原来变量95%以上的信息量为依据，即当累积贡献率≥95%时的主成分的个数就足够了。

例 设 x_1, x_2, x_3 的协方差矩阵为

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

解得特征根为 $\lambda_1 = 5.83$，$\lambda_2 = 2.00$，$\lambda_3 = 0.17$

, ,

$$U_1 = \begin{bmatrix} 0.383 \\ -0.924 \\ 0.000 \end{bmatrix} \quad U_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad U_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}$$

第一个主成分的贡献率为5.83/（5.83+2.00+0.17）=72.875%，尽管第一个主成分的贡献率并不小，但在本题中第一主成分不含第三个原始变量的信息，所以应该取两个主成分。

➢ 四、原始变量与主成份之间的相关系数

$$F_j = u_{1j}x_1 + u_{2j}x_2 + \Lambda \ + u_{pj}x_p \qquad j = 1,2,\Lambda \ , m, m \le p$$

$$\mathbf{F} = \mathbf{U}^T\mathbf{X} \qquad \mathbf{U}\mathbf{F} = \mathbf{X}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \\ x_p \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & & u_{1p} \\ u_{21} & u_{22} & & u_{2p} \\ & & & \\ u_{p1} & u_{p2} & & u_{pp} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \\ F_p \end{bmatrix}$$

➢ 四、原始变量与主成份之间的相关系数

- 主成分y_k与原始变量x_i之间的相关系数$\rho(y_k, x_i)$称为因子负荷量（或因子载荷量），并且

$$\rho(y_k, x_i) = \frac{\sqrt{\lambda_k} u_{ik}}{\sqrt{\sigma_{ii}}} (k, i = 1,2, \dots, p)$$

➢ 四、原始变量与主成份之间的相关系数

- 证明，$\rho(y_k, x_i) = \frac{Cov(y_k, x_i)}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}} = \frac{Cov(u_k' x, e_i' x)}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}}$，其中$e_i' = (0, \dots, 0,1,0, \dots, 0)$。
- 于是，$Cov(u_k' x, e_i' x) = u_k' Cov(x, x) e_i = u_k' \sum e_i = e_i' \sum u_k = \lambda_k e_i' u_k = \lambda_k u_{ik}$

基于主成分分析的人脸识别方法

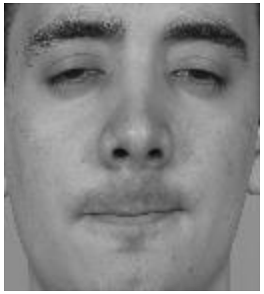
- 人世间找不到两张完全一样的脸！
  - 人脸是人类赖以区分不同人的基本途径
- 谁决定了你的长相？
  - 基因
  - 成长环境
  - 夫妻相

人脸识别的定义

- 生物特征识别的一种
- 计算机以人的脸部图像或者视频作为研究对象，从而进行人的身份确认

➢ 人脸识别的核心问题是提取特征

- 人脸的相关性很大，冗余信息多
- 如何去掉？利用主成份分析(Eigenface)
- 如何从一个矩阵变成一个向量？



➢ Eigenface->计算方法

- 计算过程为

计算样本均值m → 中心平移每个训练样本 x_i → 计算训练集合的样本协方差矩阵 → 对协方差矩阵进行特征值分解 → 取协方差矩阵的特征向量形成变换矩阵 W

参考文献：Face recognition using eigenfaces, in proc. of CVPR, 1991,citation 2315

➢ Eigenface->协方差矩阵计算

- 输入训练样本集合的协方差矩阵定义为：

$$\sum_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

其中 $\bar{x}$ 是人脸样本均值。

➢ PCA：用于降维

- 按照其所相应的特征值的大小对特征向量排序
- 选择头k个对应最大特征值的特征向量构成变换矩阵W_{pxk}

从p维空间到k维空间的投影 ($k < p$)！

原始数据(p维)

压缩（d维)

$$y = W^T x$$

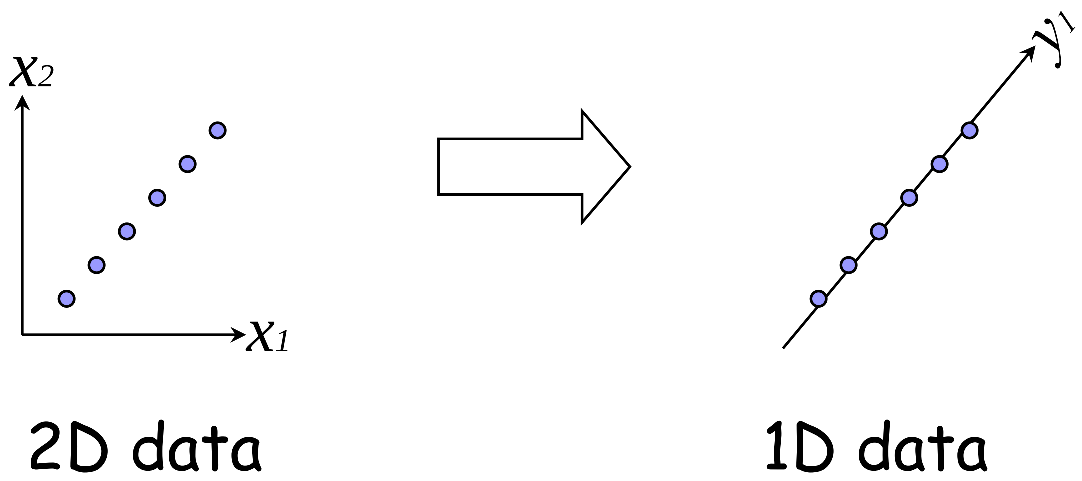
➢ 数据集合

➢ 特征人脸

8个主成份的可视化表示，它可以用来提取8维向量作为特征，原始数据的维数为64x64！

后续的课程还会讲解如何在此基础上进行识别

➢ 数据降维：理想情况图示



原始数据空间中，其中一维数据的方差为0，没有信息，可以完全去掉，而没有任何损失！

总结和作业

- 深刻理解特征值分解与特征提取之间的关系
- 如何计算协方差矩阵
- 推导协方差矩阵得到主成份和特征值
- 计算各个特征向量或者主成份的提取率