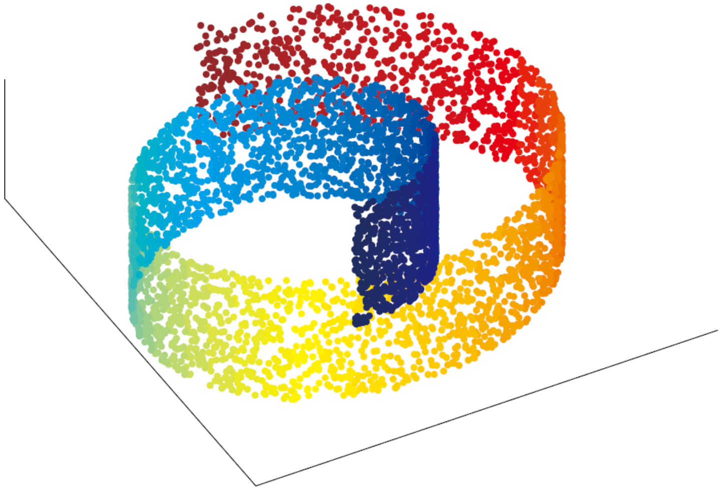
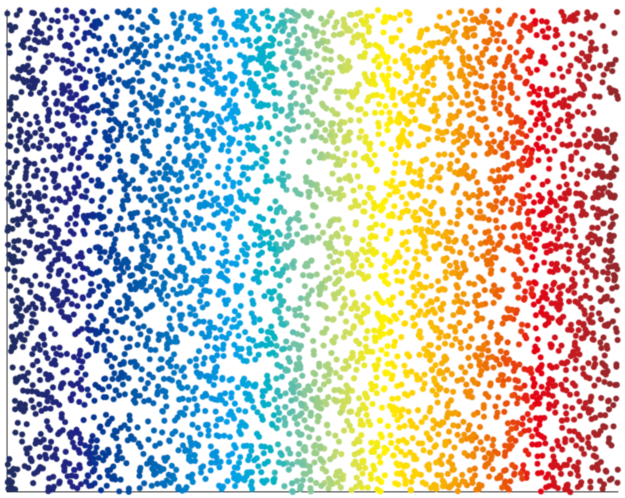
主成分分析

主成分分析 (Principal Component Analysis, PCA)

正交属性空间中的样本点，如何使用一个超平面对所有样本进行恰当的表达？





主成分分析 (Principal Component Analysis, PCA)

正交属性空间中的样本点，如何使用一个超平面对所有样本进行恰当的表达？

若存在这样的超平面，那么它大概应具有这样的性质：

- 最大可分性：样本点在这个超平面上的投影能尽可能分开
- 最近重构性：样本点到这个超平面的距离都足够近

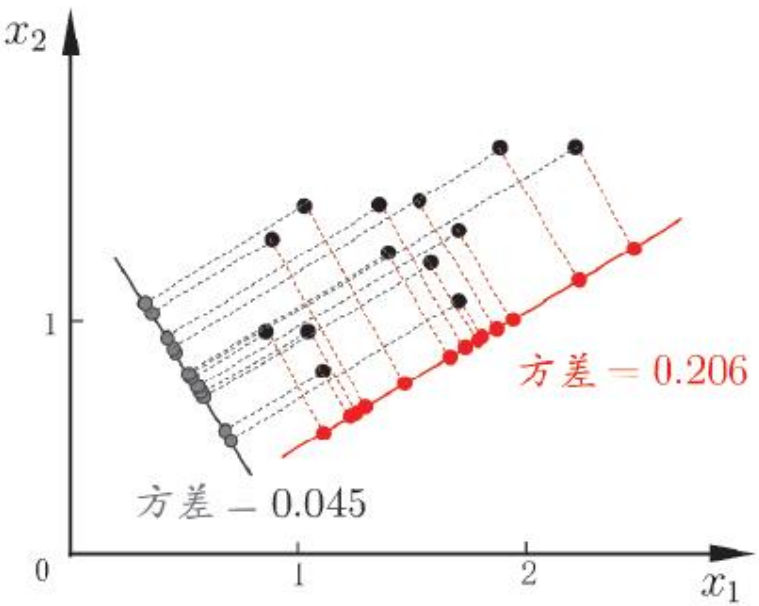
$\Rightarrow$ 主成分分析的两种等价推导

对样本进行中心化：$\sum_i \boldsymbol{x}_i = \mathbf{0}$

PCA - 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

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：

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

等价于：

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

PCA 求解

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

使用拉格朗日乘子法可得

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W} = \lambda\mathbf{W}.$$

只需对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 进行特征值分解，并将求得的特征值排序：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$，再取前 d' 个特征值对应的特征向量构成 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$，这就是主成分分析的解

关键变量：子空间方差

PCA - 最近重构性

对样本进行中心化：$\sum_i \boldsymbol{x}_i = \mathbf{0}$

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 $\boldsymbol{z}_i = (z_{i1}; z_{i2}; \ldots; z_{id'})$ $z_{ij} = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}_i$

若基于$\boldsymbol{z}_i$来重构$\boldsymbol{x}_i$，则会得到 $\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij} \boldsymbol{w}_j.$

PCA - 最近重构性（续）

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

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

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

$\boldsymbol{w}_j$ 是正交基，$\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$ 是协方差矩阵，于是由最近重构性，有：

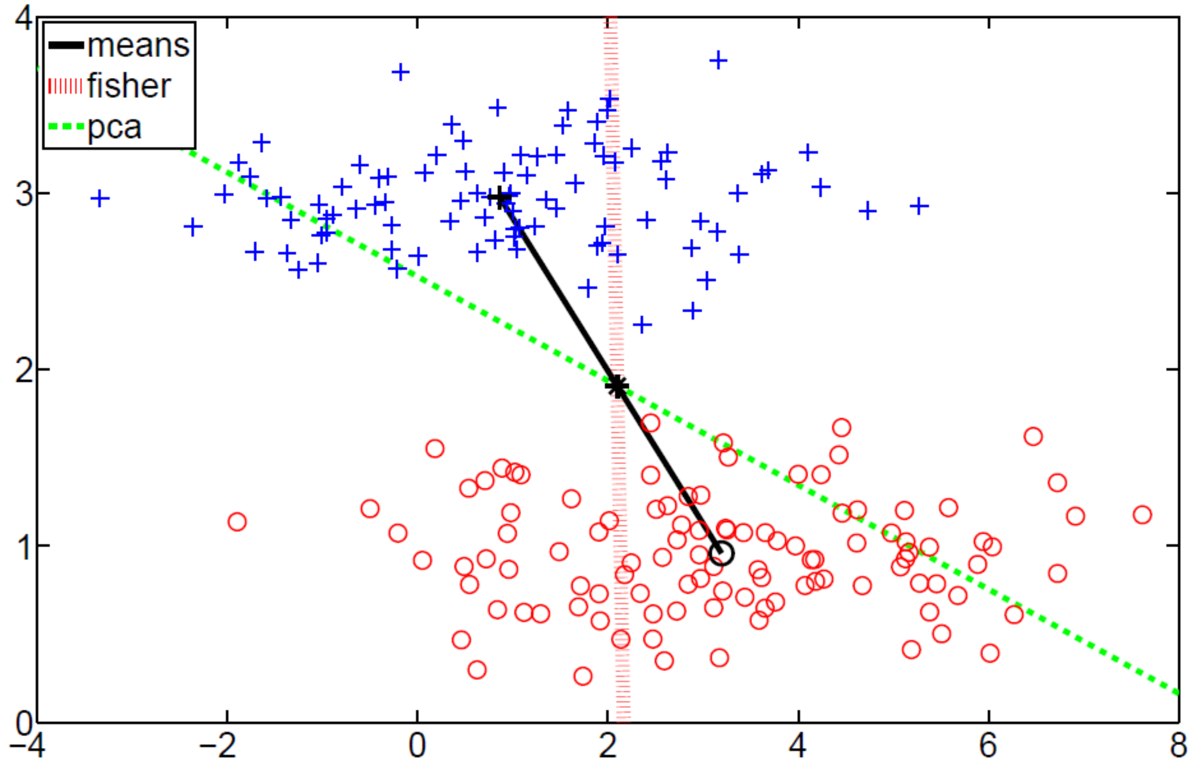
$$\min_{\mathbf{W}} \quad -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$$
$$\text{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}.$$

关键变量：重构误差

这就是主成分分析的优化目标

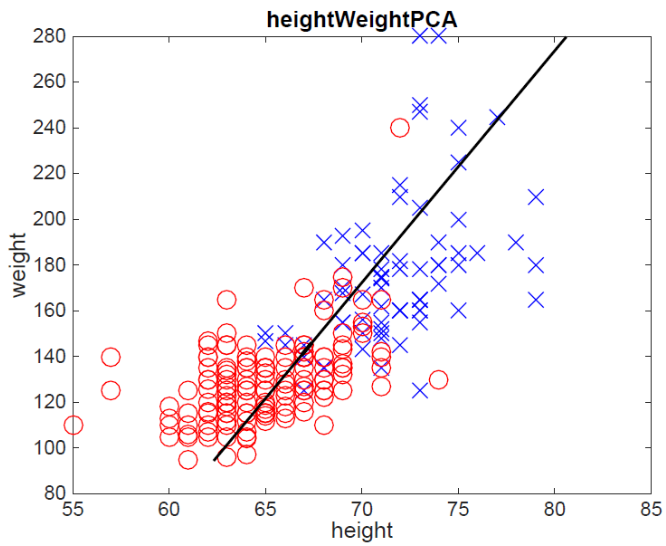
PCA - FDA

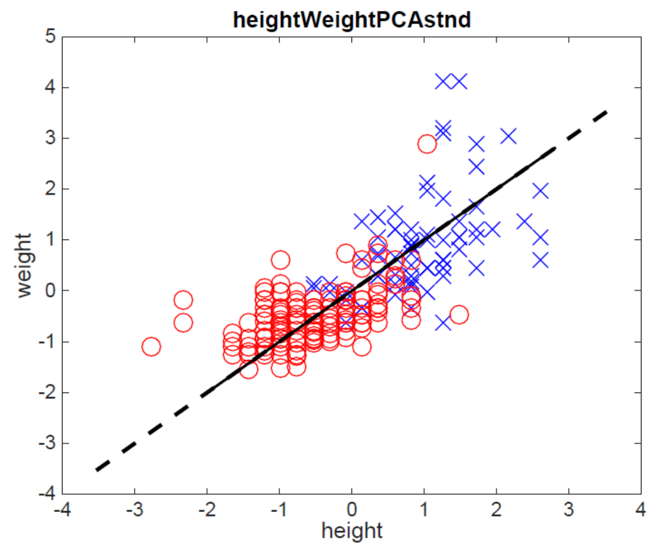
PCA是无监督学习方法，而FDA是监督学习方法，考虑了标记的作用。



PCA 应用

协方差矩阵易受到特征尺度影响





通过对数据进行标准化，使所有特征在同一尺度上

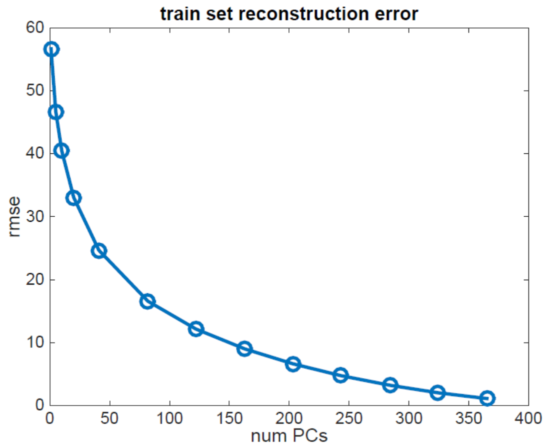
PCA 应用

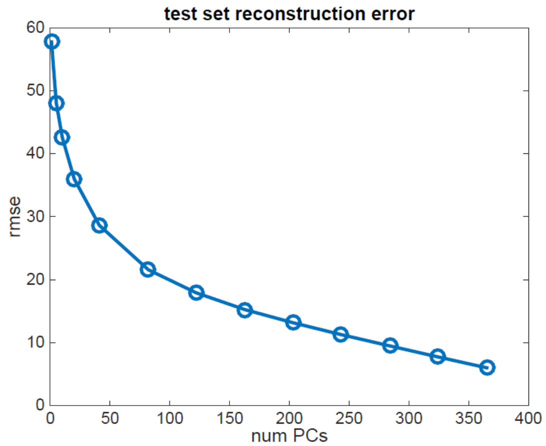
d' 的设置：

- 用户指定

- 通过重构误差判断？

$$\sum_{i=1}^{m} \|x_i - \hat{x}_i\|_2^2$$





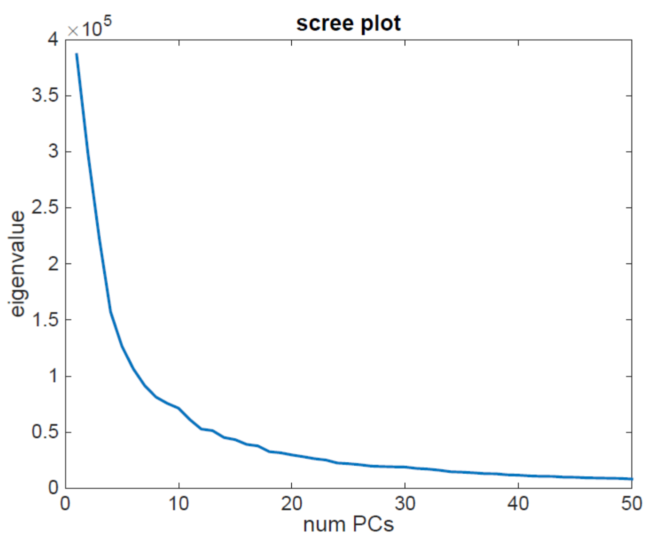
模型越复杂，重构误差越低

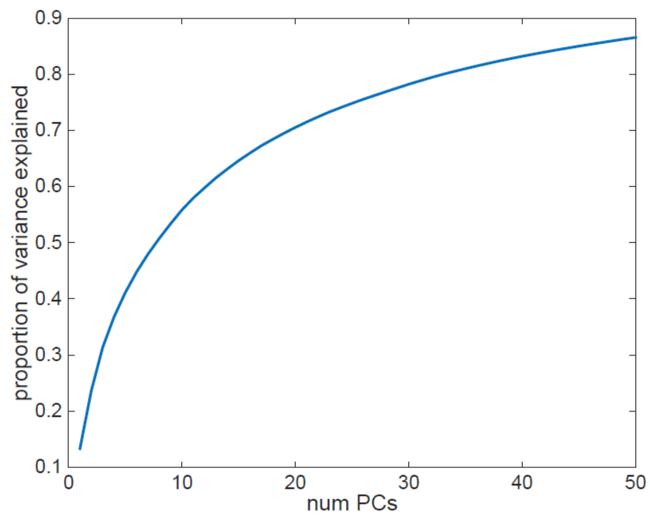
PCA 应用

d' 的设置：

- 用户指定
- 在低维空间中对k近邻或其他分类器进行交叉验证
- 设置重构阈值，例如 t =95%，然后选取最小的 d' 使得

$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq t.$$



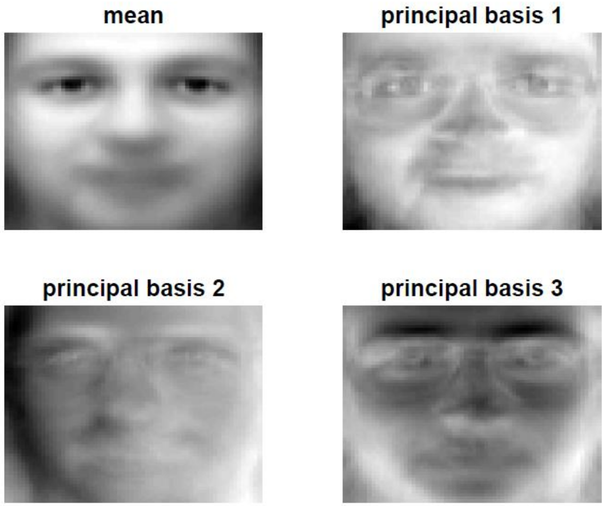


PCA 应用（续）

PCA 是最常用的降维方法，在不同领域有不同的称谓

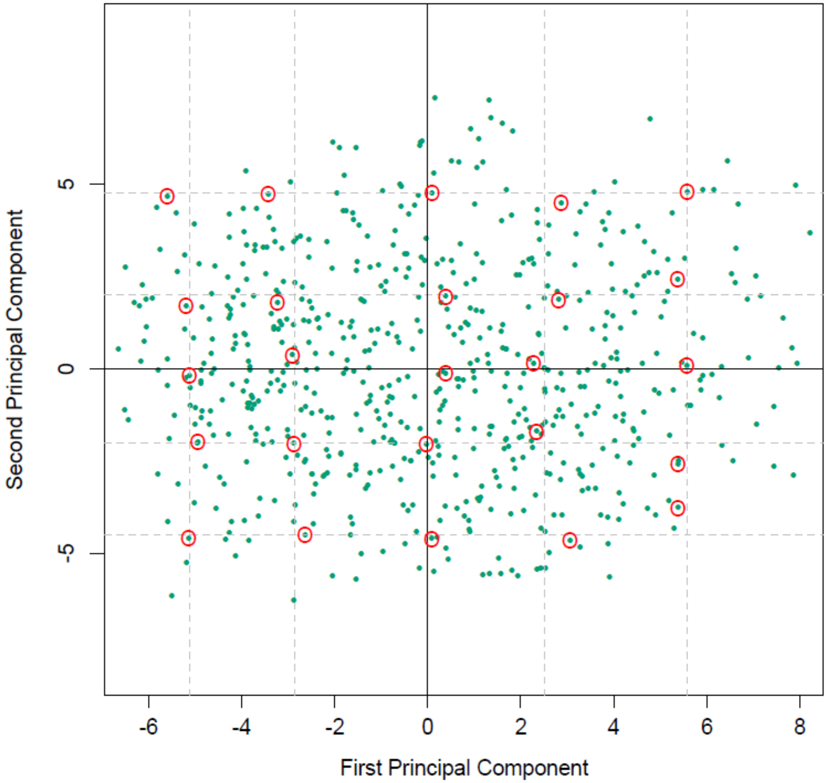
例如在人脸识别中该技术被称为“特征脸”(eigenface)

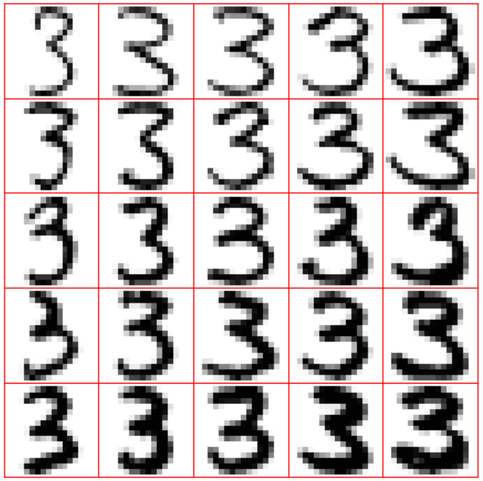
因为若将前 d' 个特征值对应的特征向量还原为图像，则得到

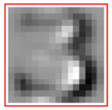


降维体现在哪里？

PCA 应用（续）

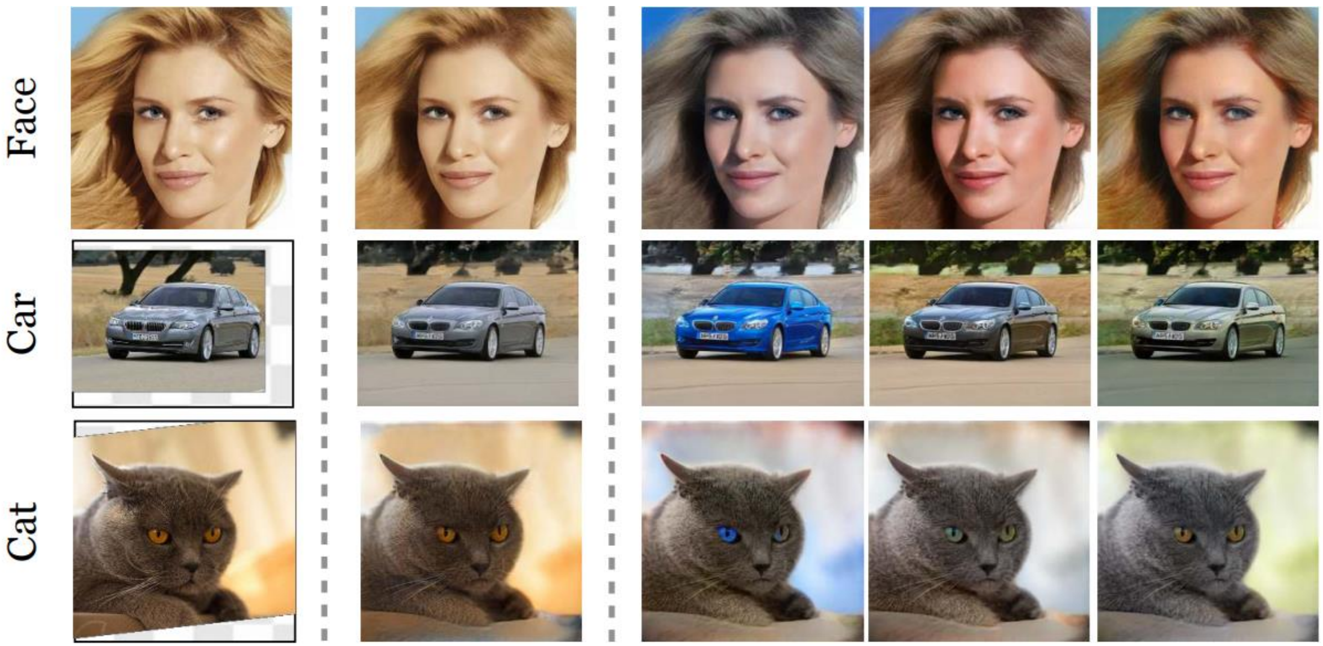




$$\hat{f}(\lambda) =$$  $$+ \lambda_1 \cdot$$  $$+ \lambda_2 \cdot$$  .

PCA 应用（续）

PCA 可有效降低模型的自由度，可用于实现高维空间中的数据分析和可视化。



在前 d' 主成分方向进行随机扰动，从而实现通过较少的参数控制高维空间中图像风格的变化

PCA 的拓展1 – 自编码

从减小重构误差角度理解PCA

$$\min_{\mathrm{W}^{\top}\mathrm{W}=\mathrm{I}} \sum_{i=1}^{m} \|x_i - \mathrm{W}z_n\|_2^2 = \sum_{i=1}^{m} \|x_i - \mathrm{W}\mathrm{W}^{\top}x_i\|_2^2$$

对W的形式进行推广

$$\min_{f,g} \sum_{i=1}^{m} \|x_i - g \circ f(x_i)\|_2^2 = \sum_{i=1}^{m} \|x_i - decoder \circ encoder(x_i)\|_2^2$$

PCA 的拓展2 – Robust PCA

- PCA的解是基于协方差的特征向量，记为$\Sigma \propto X^\top X = U_\Sigma \Lambda_\Sigma U_\Sigma^\top$
- X的奇异值分解记为$X \approx U_X S_X V_X^\top$

PCA的优化目标可改写为

$$\min_{rank(\hat{X})=d'} \left\| X - \hat{X} \right\|_2^2$$

Robust PCA

$$\min_{\hat{X}} \left\| X - \hat{X} \right\|_0 + rank(\hat{X})$$

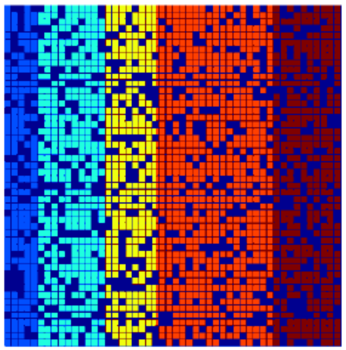
$$\Downarrow$$

$$\min_{\hat{X}} \left\| X - \hat{X} \right\|_1 + \left\| \hat{X} \right\|_*$$

PCA 的拓展2 – Robust PCA

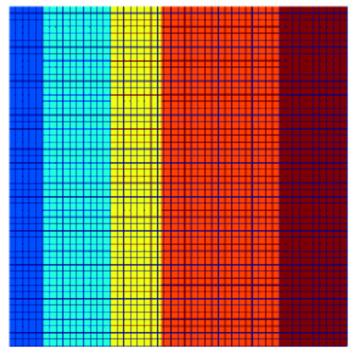
- Robust PCA

$$\min_{\hat{X}} \|X - \hat{X}\|_1 + \|\hat{X}\|_*$$



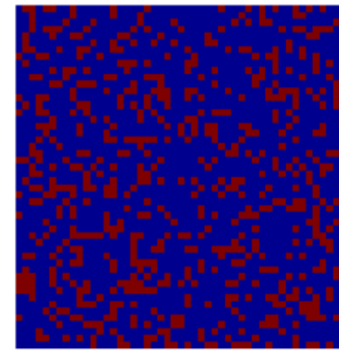
输入数据

=



低秩

+

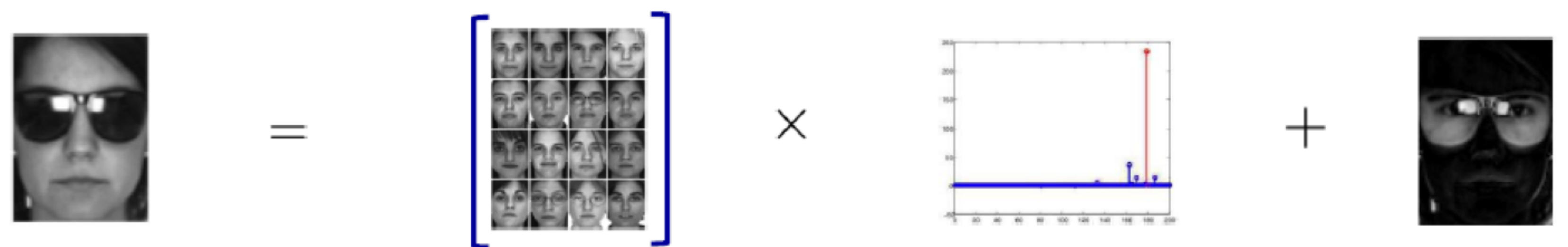


稀疏

PCA 的拓展2 – Robust PCA

- Robust PCA

$$\min_{\hat{X}} \left\| X - \hat{X} \right\|_1 + \left\| \hat{X} \right\|_*$$



$\boldsymbol{y} \in \mathbb{R}^m$ Test image $=$ $\boldsymbol{A} = [\boldsymbol{A}_1 \mid \boldsymbol{A}_2 \mid \cdots \mid \boldsymbol{A}_k]$ Combined training dictionary $\times$ $\boldsymbol{x} \in \mathbb{R}^n$ coefficients $+$ $\boldsymbol{e} \in \mathbb{R}^m$ corruption, occlusion

PCA 的拓展3 – 函数推广

PCA构建一组正交基，对数据进行重构

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

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

推广到**函数空间**，对任意（周期）函数$x(t)$进行重构

- 构建函数正交基
  - $\{\cos n\omega t\,, \sin n\omega t\}, n = 0,1,2, \ldots, \infty$
- 优化重构系数

$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t$$

$$a_n = \frac{\langle x, \cos n\omega t\rangle}{\langle \cos n\omega t\,, \cos n\omega t\rangle} = \frac{2}{T}\int_{t_0}^{t_0+T} x(t)\cos n\omega t\,\mathrm{d}t$$

PCA 的拓展3 – 函数推广

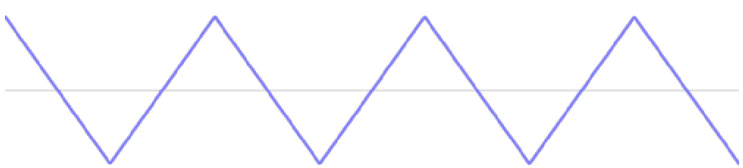
傅里叶级数：推广到**函数空间**，对任意（周期）函数$x(t)$进行重构

- 构建函数正交基
  - $\{\cos n\omega t, \sin n\omega t\}, n = 0,1,2,\dots,\infty$

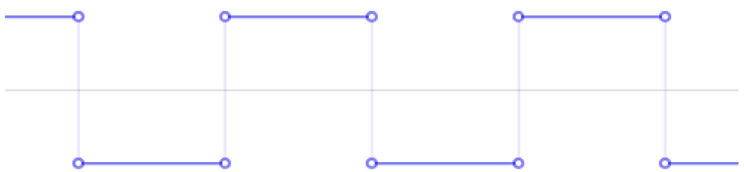
$$a_n = \frac{\langle x, \cos n\omega t\rangle}{\langle \cos n\omega t, \cos n\omega t\rangle} = \frac{2}{T}\int_{t_0}^{t_0+T} x(t)\cos n\omega t \,\mathrm{d}t$$

- 优化重构系数

$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t$$



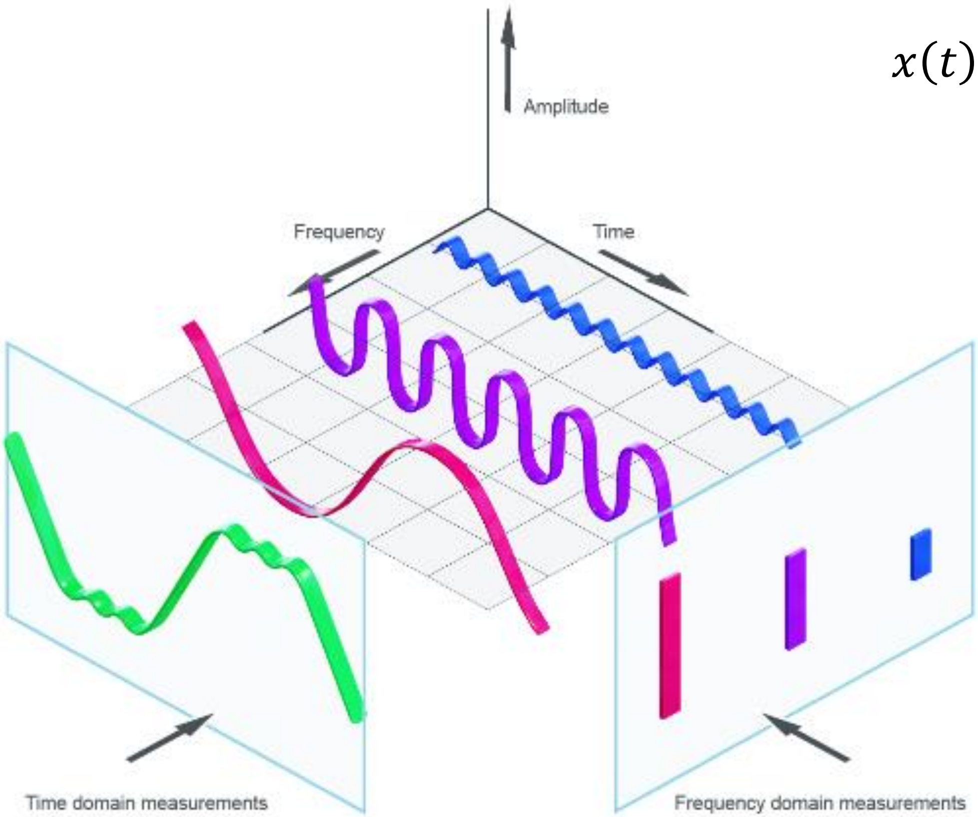
N = 0



N = 0

PCA 的拓展3 – 函数推广

傅里叶级数：推广到**函数空间**，对任意（周期）函数$x(t)$进行重构



$$x(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t$$