线性方法

刘家锋

哈尔滨工业大学

线性方法

FA: Factor Analysis

线性生成模型

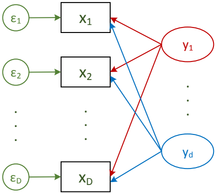
- **线性生成模型**
  - 主要介绍两种方法：
    - 因子分析：Factor Analysis
    - 概率主成分分析：Probabilistic PCA
  - 确切地说，FA和PPCA应该属于流形学习方法，但是同生成模型有着密切的联系
- **基本模型**
  - 假设观察到的高维随机矢量$\mathbf{x} \in R^D$，是由一个隐含的低维随机矢量$\mathbf{y} \in R^d \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$生成的
  - 生成的过程是由一个线性变换叠加噪声完成的：

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \boldsymbol{\epsilon}$$

其中，$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$为噪声矢量，$\mathbf{x}$和$\mathbf{y}$均为中心化的数据

Factor Analysis的基本思想

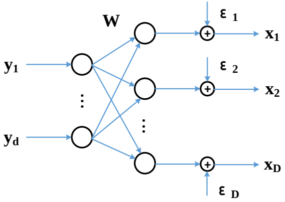
- **因子分析的基本假设**
  - 观察到的数据$\mathbf{x}$是由潜在数据$\mathbf{y}$的线性组合产生的，组合系数的绝对值为两者之间的相关性
  - 潜在数据$\mathbf{y}$服从标准正态分布
  - 噪声$\boldsymbol{\epsilon}$服从均值为$\mathbf{0}$，协方差矩阵$\boldsymbol{\Psi}$为对角阵的正态分布
  - $\mathbf{y}$与$\boldsymbol{\epsilon}$相互独立



$$\mathbf{x} = \mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}$$

Factor Analysis的基本思想

- 因子分析假设的含义
  - 数据$\mathbf{x}$是我们能够观察到的，例如一个人的长相、身高、肤色等等
  - 观察数据的产生是与一些潜在的“因子”$\mathbf{y}$有关的，如年龄、性别、种族等等
  - 因子之间是相互独立的，并且服从正态分布
  - 观察数据是由因子线性组合生成的，但会受到一定的噪声干扰，噪声服从正态分布



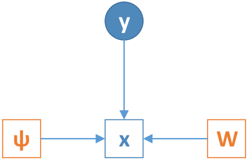
线性生成模型

- **模型的学习**
  - 根据训练数据集$\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$，推断变换矩阵$\mathbf{W}$和噪声协方差矩阵$\mathbf{\Psi}$
  - 应用于流形学习：推断生成观察数据$\mathbf{x}^*$的潜在低维因子

$$\mathbf{y}^* = \boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}^*} = \mathbb{E}_{p(\mathbf{y}|\mathbf{x}^*)}(\mathbf{y})$$

  - 应用于生成模型：随机抽样$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$，生成新的同分布数据

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \boldsymbol{\epsilon}$$



$\mathbf{x}$的分布

- 观察数据$\mathbf{x}$的分布
  - 观察数据$\mathbf{x}$由两个独立的正态分布随机变量$\mathbf{y}$和$\boldsymbol{\epsilon}$生成，服从什么样的分布？
  - 隐变量$\mathbf{y}$已知条件下，$\mathbf{x}$服从正态分布：

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{y}, \boldsymbol{\Psi})$$

  - 计算$\mathbf{x}$的边缘密度：

$$\begin{aligned}
p(\mathbf{x}) &= \int p(\mathbf{x}|\mathbf{y})p(\mathbf{y})d\mathbf{y} \\
&= \int \frac{1}{(2\pi)^{\frac{D+d}{2}}|\Psi|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{W}\mathbf{y})^t\boldsymbol{\Psi}^{-1}(\mathbf{x}-\mathbf{W}\mathbf{y}) - \frac{1}{2}\mathbf{y}^t\mathbf{y}\right]d\mathbf{y} \\
&= \cdots \\
&= \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{W}\mathbf{W}^t + \boldsymbol{\Psi})
\end{aligned}$$

y的分布

- **由观察x来推断隐变量y**
  - 可以证明，当已知观察数据x的条件下，y仍然服从正态分布，但不再是标准正态分布
  - 计算条件概率密度：

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} \\
&\propto \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{W}\mathbf{y})^t\mathbf{\Psi}^{-1}(\mathbf{x}-\mathbf{W}\mathbf{y})-\frac{1}{2}\mathbf{y}^t\mathbf{y}\right]}{\exp\left[-\frac{1}{2}\mathbf{x}^t(\mathbf{W}\mathbf{W}^t+\mathbf{\Psi})^{-1}\mathbf{x}\right]} \\
&\propto \exp\left\{-\frac{1}{2}\left[\mathbf{y}^t(\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W}+\mathbf{I})\mathbf{y}-2\mathbf{x}^t\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{y}+\mathbf{x}\mathbf{\Psi}^{-1}\mathbf{x}\right]\right\} \\
&= \mathcal{N}(\mathbf{y};\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}},\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}})
\end{aligned}
$$

其中：

$$
\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} = \mathbb{E}_{\mathbf{y}}(\mathbf{y}|\mathbf{x}) = (\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W}+\mathbf{I})^{-1}\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{x}
$$

$$
\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}} = \mathbb{E}_{\mathbf{y}}\left((\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}})(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}})^t|\mathbf{x}\right) = (\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W}+\mathbf{I})^{-1}
$$

EM for Factor Analysis

- **FA模型参数$\mathbf{W}$和$\mathbf{\Psi}$的学习**
  - 根据观察样本集$\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$学习模型参数$\mathbf{W}$和$\mathbf{\Psi}$，需要采用最大似然估计的方法
  - 观察数据生成过程对应的隐变量$\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$是未知的，因此需要使用EM算法迭代估计模型参数和隐变量
  - 首先计算$\mathbf{y}$的条件二阶矩：

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}(\mathbf{y}\mathbf{y}^t|\mathbf{x}) &= \mathbb{E}_{\mathbf{y}}\left((\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}})(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}})^t|\mathbf{x}\right) + \boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}}\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}}^t \\ &= (\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W}+\mathbf{I})^{-1} + \mathbf{M}\mathbf{x}\mathbf{x}^t\mathbf{M}^t\end{aligned}$$

    其中：

$$\mathbf{M} = (\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W}+\mathbf{I})^{-1}\mathbf{W}^t\mathbf{\Psi}^{-1}$$

EM for Factor Analysis

- 模型参数和隐变量的对数似然函数

$$
\begin{aligned}
l(\mathbf{W}, \mathbf{\Psi}, \{\mathbf{y}_i\}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\mathbf{y}_i) &= \sum_{i=1}^{n} \ln \mathcal{N}(\mathbf{x}_i; \mathbf{W}\mathbf{y}_i, \mathbf{\Psi}) \\
&= -\frac{nD}{2}\ln 2\pi - \frac{n}{2}|\mathbf{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}\{\mathbf{x}_i^t\mathbf{\Psi}^{-1}\mathbf{x}_i \\
&\quad -2\mathbf{x}_i^t\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{y}_i + \mathbf{y}_i^t\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{y}_i\}
\end{aligned}
$$

- **E步**：对数似然函数关于隐变量$\mathbf{y}$的期望

$$
\begin{aligned}
Q(\mathbf{W}, \mathbf{\Psi}) &= \mathbb{E}_{\mathbf{y}}(l(\mathbf{W}, \mathbf{\Psi}, \{\mathbf{y}_i\})) \\
&= -\frac{nD}{2}\ln 2\pi - \frac{n}{2}|\mathbf{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}\{\mathbf{x}_i^t\mathbf{\Psi}^{-1}\mathbf{x}_i \\
&\quad -2\mathbf{x}_i^t\mathbf{\Psi}^{-1}\mathbf{W}\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i) + tr\left[\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W}\mathbb{E}(\mathbf{y}_i\mathbf{y}_i^t|\mathbf{x}_i)\right]\}
\end{aligned}
$$

EM for Factor Analysis

- **M步：**计算期望似然函数的梯度和极值

对$\mathbf{W}$的梯度：

$$\frac{\partial Q(\mathbf{W},\boldsymbol{\Psi})}{\partial \mathbf{W}} = \sum_{i=1}^{n} \boldsymbol{\Psi}^{-1}\mathbf{x}_i \mathbb{E}(\mathbf{y}_i^t|\mathbf{x}_i) - \sum_{i=1}^{n} \boldsymbol{\Psi}^{-1}\mathbf{W}\mathbb{E}(\mathbf{y}_i\mathbf{y}_i^t|\mathbf{x}_i) = \mathbf{0}$$

得到：

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{x}_i \mathbb{E}(\mathbf{y}_i^t|\mathbf{x}_i) \left[\sum_{i=1}^{n} \mathbb{E}(\mathbf{y}_i\mathbf{y}_i^t|\mathbf{x}_i)\right]^{-1}$$

EM for Factor Analysis

- **M步：** 计算期望似然函数的梯度和极值

对$\mathbf{\Psi}^{-1}$的梯度：

$$\frac{\partial Q(\mathbf{W},\mathbf{\Psi})}{\partial \mathbf{\Psi}^{-1}} = \frac{n}{2}\mathbf{\Psi} - \frac{1}{2}\sum_{i=1}^{n}\left[\mathbf{x}_i\mathbf{x}_i^t - 2\mathbf{W}\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i)\mathbf{x}_i^t + \mathbf{W}\mathbb{E}(\mathbf{y}_i\mathbf{y}_i^t|\mathbf{x}_i)\mathbf{W}^t\right]$$
$$= \mathbf{0}$$

得到：

$$\begin{aligned}\mathbf{\Psi} &= \frac{1}{n}\left\{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t - 2\mathbf{W}\sum_{i=1}^{n}\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i)\mathbf{x}_i^t + \mathbf{W}\sum_{i=1}^{n}\mathbb{E}(\mathbf{y}_i\mathbf{y}_i^t|\mathbf{x}_i)\mathbf{W}^t\right\} \\ &= \frac{1}{n}\left\{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t - \mathbf{W}\sum_{i=1}^{n}\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i)\mathbf{x}_i^t\right\}\end{aligned}$$

EM for Factor Analysis

Algorithm 1 EM for Factor Analysis

E step: 给定$\mathbf{W}, \mathbf{\Psi}$

$$\mathbb{E}_{\mathbf{y}}(\mathbf{y}|\mathbf{x}) = (\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W} + \mathbf{I})^{-1}\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{x}$$
$$\mathbb{E}_{\mathbf{y}}(\mathbf{y}\mathbf{y}^t|\mathbf{x}) = (\mathbf{W}^t\mathbf{\Psi}^{-1}\mathbf{W} + \mathbf{I})^{-1} + \mathbf{M}\mathbf{x}\mathbf{x}^t\mathbf{M}^t$$

M step: 修正$\mathbf{W}, \mathbf{\Psi}$

$$\mathbf{W} = \sum_{i=1}^{n}\mathbf{x}_i\mathbb{E}(\mathbf{y}_i^t|\mathbf{x}_i)\left[\sum_{i=1}^{n}\mathbb{E}(\mathbf{y}_i\mathbf{y}_i^t|\mathbf{x}_i)\right]^{-1}$$
$$\mathbf{\Psi} = \frac{1}{n}diag\left\{\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t - \mathbf{W}\sum_{i=1}^{n}\mathbb{E}(\mathbf{y}_i|\mathbf{x}_i)\mathbf{x}_i^t\right\}$$

Factor Analysis的解

FA的解不是唯一的

如果$\mathbf{W}$和$\mathbf{y}$为Factor Analysis的一个解，令：

$$\mathbf{W}' = \mathbf{W}\mathbf{R} \quad \mathbf{y}' = \mathbf{R}^t\mathbf{y}$$

其中，$\mathbf{R}$为任意的单位正交矩阵，则：

$$\mathbb{E}(\mathbf{y}'\mathbf{y}'^t) = \mathbf{R}^t\mathbb{E}(\mathbf{y}\mathbf{y}^t)\mathbf{R} = \mathbf{R}^t\mathbf{R} = \mathbf{I}$$

因此，$\mathbf{W}'$和$\mathbf{y}'$也是Factor Analysis的一个解，两者相差一个旋转矩阵$\mathbf{R}$。

Factor Analysis vs. Principle Component Analysis

- 相同点
  - 高维数据由低维数据线性组合产生（低维线性流形嵌入）
  - PCA可以作为FA的一种简单求解方法
- 不同点
  - FA假设潜在数据是有具体某种意义的，PCA的主成分只具有几何意义
  - PCA的目的就是数据降维，FA的目的更多的是要理解数据的潜在结构
  - FA需要明确地模型化噪声项ϵ为正态分布，需要估计$\mathbf{\Psi}$
  - PCA的目标是要尽量保留观测数据$\mathbf{x}$的方差，而FA既要分析$\mathbf{x}$的方差，也要分析$\mathbf{x}$各维分量之间的相关性

PPCA：Probabilistic PCA

PPCA：Probabilistic PCA

- **基本模型**

  观察到的随机矢量$\mathbf{x} \in R^D$（已中心化），来自于一个隐含的随机矢量$\mathbf{y} \in R^d \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$：

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \boldsymbol{\epsilon}$$

  其中，$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$为噪声矢量

- **PPCA是对FA的简化**
  - 噪声矢量$\boldsymbol{\epsilon}$的协方差矩阵：$\boldsymbol{\Psi} = \sigma^2\mathbf{I}$
  - 这样的假设可以简化学习过程，模型参数的估计不再需要EM迭代计算

Probabilistic PCA

- **PPCA的基本问题**：给定样本集$\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$
  - 参数估计问题：估计模型参数$\mathbf{W}, \sigma^2$
  - 降维问题：计算隐含的矢量集$\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$

$$\mathbf{y}_i = \mathbb{E}_{p(\mathbf{y}|\mathbf{x}_i)}(\mathbf{y})$$

- $\mathbf{x}$**的分布密度**
  - 条件密度：

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{y}, \sigma^2\mathbf{I})$$

  - 边缘密度：

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{W}\mathbf{W}^t + \sigma^2\mathbf{I})$$

Probabilistic PCA：参数估计问题

- **对数似然函数**

$$
\begin{aligned}
l(\mathbf{W},\sigma^2) &= \sum_{i=1}^{n} \ln p(\mathbf{x}_i) \\
&= -\frac{nD}{2}\ln 2\pi - \frac{n}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i^t\mathbf{C}^{-1}\mathbf{x}_i \\
&= -\frac{n}{2}\left[D\ln 2\pi + \ln|\mathbf{C}| + tr(\mathbf{C}^{-1}\mathbf{\Sigma}_{\mathbf{x}})\right]
\end{aligned}
$$

其中，$\mathbf{C} = \mathbf{W}\mathbf{W}^t + \sigma^2\mathbf{I}$，$\mathbf{\Sigma}_{\mathbf{x}}$为样本集的协方差矩阵

$$
\mathbf{\Sigma}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t
$$

Probabilistic PCA：参数估计问题

- **计算梯度**
  - 对$\mathbf{W}$的梯度

$$\frac{\partial l}{\partial \mathbf{W}} = -n\mathbf{C}^{-1}\mathbf{W} + n\mathbf{C}^{-1}\mathbf{\Sigma_x}\mathbf{C}^{-1}\mathbf{W} = \mathbf{0}$$

得到：

$$\mathbf{W} = \mathbf{\Sigma_x}\mathbf{C}^{-1}\mathbf{W}$$

  - 无意义解：
    - ✓ 解1：$\mathbf{W} = \mathbf{0}$
    - ✓ 解2：$\mathbf{C} = \mathbf{W}\mathbf{W}^t + \sigma^2\mathbf{I} = \mathbf{\Sigma_x}$

矩阵$W \in R^{D\times d}$，因此$rank(\mathbf{W}\mathbf{W}^t) = d$，对应的后$D-d$个特征值$\tau_{d+1}, \cdots, \tau_D = 0$。由于

$$(\mathbf{W}\mathbf{W}^t + \sigma^2\mathbf{I})\mathbf{u} = \mathbf{W}\mathbf{W}^t\mathbf{u} + \sigma^2\mathbf{u} = (\tau + \sigma^2)\mathbf{u}$$

因此，$\mathbf{W}\mathbf{W}^t + \sigma^2\mathbf{I}$的特征值为$\tau_1 + \sigma^2, \cdots, \tau_d + \sigma^2, \sigma^2, \cdots, \sigma^2$，要求协方差矩阵$\Sigma_{\mathbf{x}}$的后$D-d$个特征值刚好都是$\sigma^2$显然是不合理的。

Probabilistic PCA：参数估计问题

- **$\mathbf{W}$的解**
  - 考虑$\mathbf{C} \neq \mathbf{\Sigma_x}$的情况
  - $\mathbf{W}$的奇异值分解：$\mathbf{W} = \mathbf{ULV}^t$
    - $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_D)$和$\mathbf{V}$分别为单位正交的左、右奇异矩阵
    - $\mathbf{L} = diag\{l_1, \cdots, l_d\}$为奇异值矩阵
  - 代入等式：

$$\begin{aligned}
\mathbf{W} = \mathbf{ULV}^t &= \mathbf{\Sigma_x}\mathbf{C}^{-1}\mathbf{W} \\
&= \mathbf{\Sigma_x}(\mathbf{WW}^t + \sigma^2\mathbf{I})^{-1}\mathbf{W} \\
&= \mathbf{\Sigma_x}(\mathbf{ULL}^t\mathbf{U}^t + \sigma^2\mathbf{I})^{-1}\mathbf{ULV}^t \\
&= \mathbf{\Sigma_x}\mathbf{U}(\mathbf{LL}^t + \sigma^2\mathbf{I})^{-1}\mathbf{LV}^t
\end{aligned}$$

Probabilistic PCA：参数估计问题

- **W的解**

可以推导得到：

$$\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{u}_i = (l_i^2 + \sigma^2)\mathbf{u}_i$$

显然，协方差矩阵$\boldsymbol{\Sigma}_{\mathbf{x}}$的特征矢量是矩阵$\mathbf{W}$的左奇异矢量$\{\mathbf{u}_1, \cdots, \mathbf{u}_D\}$，对应的特征值$\lambda_i$为$l_i^2 + \sigma^2$。因此，矩阵$\mathbf{W}$的奇异值：

$$l_i = (\lambda_i - \sigma^2)^{\frac{1}{2}}$$

重构$\mathbf{W}$：

$$\mathbf{W} = \mathbf{U}_d(\boldsymbol{\Lambda}_d - \sigma^2\mathbf{I}_d)^{\frac{1}{2}}\mathbf{R}$$

其中，$\mathbf{U}_d$为$\boldsymbol{\Sigma}_{\mathbf{x}}$最大$d$个特征矢量构成的矩阵，$\boldsymbol{\Lambda}_d$为最大$d$个特征值构成的对角阵，$\mathbf{R}$为任意的单位正交矩阵

Probabilistic PCA：参数估计问题

- **计算梯度**

将$\mathbf{W}$代入对数似然函数，可以得到：

$$l(\mathbf{W}, \sigma^2) = -\frac{n}{2}\left[\ln 2\pi + \ln|\mathbf{C}| + tr(\mathbf{C}^{-1}\mathbf{\Sigma_x})\right]$$

$$= -\frac{n}{2}\left[\ln 2\pi + \sum_{i=1}^{d}\ln\lambda_i + (D-d)\ln\sigma^2 + d + \frac{1}{\sigma^2}\sum_{i=d+1}^{D}\lambda_i\right]$$

对$\sigma^{\mathbf{2}}$的梯度：

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\left\{\frac{D-d}{\sigma^2} - \frac{1}{\sigma^4}\sum_{i=d+1}^{D}\lambda_i\right\} = 0 \quad \Rightarrow \quad \sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D}\lambda_i$$

Probabilistic PCA：参数估计问题

Algorithm 2 Probabilistic PCA

1: 输入：样本集$\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$
2: 计算协方差矩阵

$$\mathbf{\Sigma}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t$$

3: 计算$\mathbf{\Sigma}_{\mathbf{x}}$的特征值分解：

$$\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t$$

4: 计算：

$$\sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D}\lambda_i, \quad \mathbf{W} = \mathbf{U}_d(\mathbf{\Lambda}_d - \sigma^2\mathbf{I}_d)^{\frac{1}{2}}$$

Probabilistic PCA：降维

- $\mathbf{y}$的分布

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})} \\
&= \frac{\mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{y}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{y}; \mathbf{0}, \mathbf{I})}{\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{W}\mathbf{W}^t + \sigma^2\mathbf{I})} \\
&= \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}})
\end{aligned}
$$

其中：

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} &= (\mathbf{W}^t\mathbf{W} + \sigma^2\mathbf{I})^{-1} \\
\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} &= (\mathbf{W}^t\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^t\mathbf{x}
\end{aligned}
$$

Probabilistic PCA：降维

- **降维**：矢量$\mathbf{x}^*$在潜在空间的投影

$$\mathbf{y}^* = \mathbb{E}_{p(\mathbf{y}|\mathbf{x}^*)}(\mathbf{y}) = \boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}^*} = (\mathbf{W}^t\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^t\mathbf{x}^*$$

- **与PCA的关系**：当$\sigma^2 = 0$时

$$\mathbf{W} = \mathbf{U}_d\boldsymbol{\Lambda}^{\frac{1}{2}}$$

$$\mathbf{y}^* = (\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}_d^t\mathbf{U}_d\boldsymbol{\Lambda}^{\frac{1}{2}})^{-1}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}_d^t\mathbf{x}^* = \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{U}_d^t\mathbf{x}^*$$

其中，矩阵$\mathbf{U}_d = (\mathbf{u}_1, \cdots, \mathbf{u}_d)$是由协方差矩阵$\boldsymbol{\Sigma}_\mathbf{x}$的特征矢量构成的，矩阵$\boldsymbol{\Lambda}$为对应的特征值。此时，PPCA与PCA的降维结果只是相差一个尺度因子$1/\sqrt{\lambda_i}$。