Subject: Stanford CS229 Machine Learning, Lecture 14, Factor Analysis / PCA

Date: from August 26, 2025 to August 28, 2025

Contents

Stanford CS229 Machine Learning, Factor Analysis / PCA, 2022, Lecture 14

Link on YouTube: Stanford CS229 Machine Learning, Factor Analysis / PCA, 2022, Lecture 14

Introduction

Introduction

本节内容主要介绍高维小样本情形下的子空间型无监督学习方法，包括基于概率模型的因子分析（Factor Analysis）与非概率方法的主成分分析（PCA）.

Factor Analysis

Factor Analysis: Idea

Challenge. 在现实情况下，有时数据量(n)很少但数据的维度(d)很大，即 $n \ll d$，此时使用模型例如最小二乘拟合这些数据得到的模型有很多可能的解. (未知数数量> 方程数)

Idea. 类似于 *GMM*，此时可以假设存在 latent variable，其本身结构简单、便于估计，但能够刻画数据的生成机制（从而解释观测数据的结构）.

Example 1: Fit Gaussian. 对于高斯分布 $x \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$, $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$，此时已有数据 $x^{(1)}, \dots, x^{(n)}$，则参数估计为：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \quad (\in \mathbb{R}^d), \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \hat{\mu}\right)\left(x^{(i)} - \hat{\mu}\right)^\top \quad (\in \mathbb{R}^{d\times d}). \tag{1}$$

但是由于 $\text{rank}(\Sigma) < \{n, d\}, n \ll d$，因此 $\text{rank}(\Sigma) < n$，因此 Σ 并非满秩，这就导致 ① $|\Sigma|^{1/2} = 0$; ② Σ^{-1} 不存在. (*MLE* 细节详见附录A)

Factor Analysis: Examples

Example 1. 假设此时的高斯分布为独立(independent) 且各向同性(identical)，即：

$$\begin{aligned} &\text{(independent): } x^{(1)}, \dots, x^{(n)} \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \Sigma) \\ &\text{(identical): } \Sigma = \sigma^2 I, \quad \sigma^2 \in \mathbb{R}_+ \text{ (scalar)}, \end{aligned} \tag{2}$$

那么此时由于各向同性，此高斯分布的等高线为圆形(如图1a)并且此时行列式 $|\Sigma| = |\sigma^2 I| = \sigma^{2d}$. 此时 *MLE* 如下：

$$\begin{aligned} &\max_{\mu,\sigma^2} \ell(\mu, \sigma^2) = \max_{\mu,\sigma^2} \sum_{i=1}^{n}\left[-\frac{1}{2\sigma^2}(x^{(i)} - \mu)^\top (x^{(i)} - \mu) - \frac{d}{2}\log\sigma^2\right] \\ &\Rightarrow \min_{\mu,\sigma^2} \sigma^{-2}\sum_{i=1}^{n}\|x^{(i)} - \mu\|^2 + nd\log\sigma^2 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}, \hat{\sigma}^2 = \frac{1}{nd}\sum_{i=1}^{n}\|x^{(i)} - \hat{\mu}\|^2 \end{aligned} \tag{3}$$

Example 2. 假设此时的高斯分布独立且各维独立(协方差矩阵为对角阵)，即：

$$\begin{aligned}&\text{(independent): } x^{(1)},\dots,x^{(n)} \overset{i.i.d.}{\sim} \mathcal{N}(\mu,\Sigma)\\&\text{(diagonal): } \Sigma=\operatorname{diag}(\sigma_1^2,\sigma_2^2,\dots,\sigma_d^2),\end{aligned} \tag{4}$$

由于各维独立，此高斯分布的等高线为轴对齐的椭圆(如图1b). 记 $z_j=\sigma_j^2$，此时 *MLE*：

$$\begin{aligned}&\max_{\mu,\{z_j\}} \ell(\mu,\{z_j\}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}\left[z_j^{-1}(x_j^{(i)}-\mu_j)^2+\log z_j\right]\\&\Rightarrow \min_{z_1,\dots,z_d}\sum_{i=1}^{n}\sum_{j=1}^{d}\left(z_j^{-1}(x_j^{(i)}-\mu_j)^2+\log z_j\right).\end{aligned} \tag{5}$$

由于目标函数在各维度独立，因此此时只需要对于各个维度 j 作 *MLE*:

$$\begin{aligned}&\min_{z_j}\sum_{i=1}^{n}\left(z_j^{-1}(x_j^{(i)}-\mu_j)^2+\log z_j\right) \Rightarrow -z_j^{-2}\sum_{i=1}^{n}(x_j^{(i)}-\mu_j)^2+\frac{n}{z_j}=0\\&\Rightarrow \hat{\sigma}_j^2=z_j=\frac{1}{n}\sum_{i=1}^{n}(x_j^{(i)}-\mu_j)^2.\end{aligned} \tag{6}$$

(a) Example 1　　(b) Example 2

Figure 1: Examples

Factor Analysis: Factor Model

Parameter. 在因子模型(factor model) 中，共有 $\mu\in\mathbb{R}^d,\Lambda\in\mathbb{R}^{d\times s},\Phi\in\mathbb{R}^{d\times d}$ 三个参数，其中 μ 为均值向量，Λ 为因子载荷矩阵，Φ 为噪声协方差矩阵.

Model. 与之前的建模相同，此时的生成模型仍然包含 latent variable z，即为

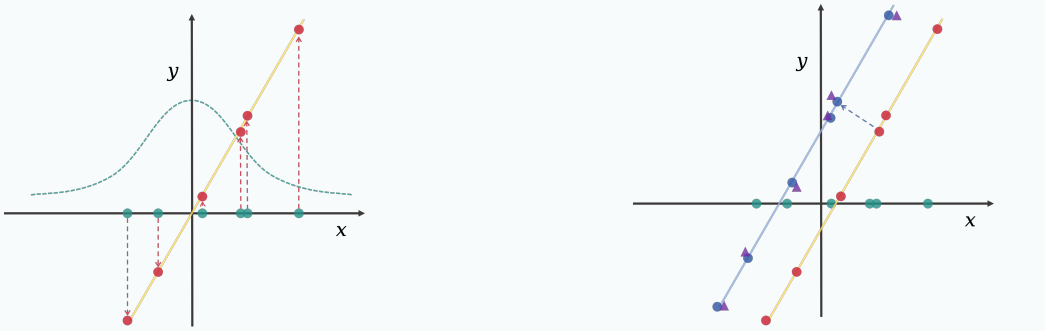
$$p(x,z)=p(x\mid z)\,p(z), \tag{7}$$

其中 z 为 latent variable，且 $z\sim\mathcal{N}(0,I)\in\mathbb{R}^s,\quad s<d.$ 此时条件分布 $x\mid z$ 被建模为 $x\mid z\sim\mathcal{N}(\mu+\Lambda z,\ \Phi)$，实际计算方式即：

$$x=\mu+\Lambda z+\epsilon,\epsilon\sim\mathcal{N}(0,\Phi). \tag{8}$$

Example. 以 $d=2,s=1,n=5$ 为例，最终的模型输出为 $x=\mu+\Lambda z+\epsilon$:

1. Step 1. 隐变量生成： $z^{(1)}, \ldots, z^{(n)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \to$ 对应图2a 中绿色点.
2. Step 2. 载荷矩阵（假设）： $\Lambda = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \to$ 对应图2a 中将点映射至直线 $y = 2x$ 上.
3. Step 3. 加入偏移量 μ： $\tilde{x} = \mu + \Lambda z \to$ 对应图2b 中将点移动至蓝色直线上.
4. Step 4. 加入噪声 ϵ： $x = \mu + \Lambda z + \epsilon \to$ 对应图2b 中将点进行随机扰动至紫色点.



(a) Example for Factor Model: Step 1-2 (b) Example for Factor Model: Step 3-4

Figure 2: Example for Factor Model

Fact for Gaussian. 记 $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^d$，其中 $x_1 \in \mathbb{R}^{d_1}, x_2 \in \mathbb{R}^{d_2}, d = d_1 + d_2$，协方差矩阵写为分块矩阵 $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Sigma_{ij} \in \mathbb{R}^{d_i \times d_j}, i, j \in \{1, 2\}$. 回顾 Gaussian 分布的两个基本事实:

1. **Marginalization.** 边缘分布 $p(x_1) = \int p(x_1, x_2)\, dx_2$，Gaussian有 $p(x_1) \sim \mathcal{N}(\mu_1, \Sigma_{11})$
2. **Conditioning.** $p(x_1 \mid x_2) \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$，其中 $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$，$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Note 1. $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ 推导可由矩阵求逆引理（*Matrix Inversion Lemma*）得到.

根据上述事实，对于 factor analysis $x = \mu + \Lambda z + \epsilon, \epsilon \sim \mathcal{N}(0, \Phi)$，此时由于 $\mathbb{E}[z] = 0, \mathbb{E}[x] = \mu$，因此联合分布 $\begin{pmatrix} z \\ x \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma\right)$. 对于协方差矩阵 $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$:

$$
\begin{aligned}
&\Sigma_{11} = \mathrm{Cov}(z) = \mathbb{E}[zz^\top] = I \\
&\Sigma_{12} = \mathrm{Cov}(z, x) = \mathbb{E}\left[z(x-\mu)^\top\right] = \mathbb{E}\left[z(\Lambda z + \varepsilon)^\top\right] = \mathbb{E}[zz^\top]\Lambda^\top + \mathbb{E}[z\varepsilon^\top] = I\Lambda^\top + 0 = \Lambda^\top \\
&\Sigma_{21} = \Sigma_{12}^\top = \Lambda \\
&\Sigma_{22} = \mathbb{E}\left[(x-\mu)(x-\mu)^\top\right] = \mathbb{E}\left[(\Lambda z + \varepsilon)(\Lambda z + \varepsilon)^\top\right] = \Lambda\mathbb{E}[zz^\top]\Lambda^\top + \mathbb{E}[\varepsilon\varepsilon^\top] = \Lambda\Lambda^\top + \Phi.
\end{aligned} \tag{9}
$$

因此 $\Sigma = \begin{pmatrix} I & \Lambda^\top \\ \Lambda & \Lambda\Lambda^\top + \Phi \end{pmatrix}$.

Factor Analysis: Conclusion

对于 factor analysis 模型，已经假设存在 latent variable $z \sim \mathcal{N}(0, I)$，并有 μ, Λ, Φ 三个参数需要估计，此时 $x = \mu + \Lambda z + \epsilon, \epsilon \sim \mathcal{N}(0, \Phi)$，将其写为联合分布 $\begin{pmatrix} z \\ x \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \Sigma\right)$，使用 *EM* 算法：

1. ***E*-step.** $Q_i(z) = p(z^{(i)} \mid x^{(i)}, \theta)$
2. ***M*-step.** 使用闭式解 $\Sigma = \begin{pmatrix} I & \Lambda^\top \\ \Lambda & \Lambda\Lambda^\top + \Phi \end{pmatrix}$ 更新参数.

Principle Component Analysis (PCA)

PCA: Introduction

无监督学习的分类可以根据是否是概率模型，也可以根据是"聚类"型还是"子空间"型进行区分，直至目前所接触到的无监督学习方法区分如下：

Structure	Probabilistic Model	Non-probabilistic Method
Clustering	GMM	K-means
Subspace	Factor Analysis	PCA

PCA example. 给定一批水果，记录其体积(x-轴)、质量(y-轴)两种数据，我们的目的是区分其品种，但是仅根据体积大小或质量并不能显著区分. 此时可以使用 $y = x$ 方向作为区分标准，并将其解释为"密度"，这样将所有的点投影至 $y = x$ 上可以将不同品种很好地区分开. 在这个例子中，$y = x$ 实际上就是对数据变化(variation) 解释性更强的方向.

Idea. *PCA* 的基本思想是寻找主成分方向(principle component)，使得数据在这些方向上能够更容易区分开（例如使数据点投影后更分散开），即更能够刻画数据的变化(variation).

PCA

Pre-processing. 给定数据 $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$，首先需要对数据进行预处理：

1. **Center the data.** 首先中心化数据：

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}, x^{(i)} \longleftarrow x^{(i)} - \mu. \tag{10}$$

2. **Rescaling.** 再进行特征缩放以归一化消除量纲：

$$x_j^{(i)} \longleftarrow x_j^{(i)}/\sigma_j. \tag{11}$$

Note 2. 中心化使 *PCA* 关注方差而非偏移，如果不中心化 *PCA* 会浪费一个主成分去解释均值. 由于 $Var(kx) = k^2 Var(x)$，因此不归一化数值大的特征会主导主成分，因此为更关注方向上的方差而非绝对大小需要进行归一化.

***PCA* as Optimization Problem.** 当只有一个主成分方向(principle component) $u_1 \in \mathbb{R}^d$ 后，需要在子空间 $\mathcal{S}_1 = \{tu_1 : t \in \mathbb{R}\}$ 中寻找距离数据 x 最近的点，即投影点：

$$\begin{aligned} \alpha_1 &= \arg\min_{\alpha} \|x - \alpha u_1\|^2 = \|x\|^2 + \alpha^2\|u_1\|^2 - 2\alpha\langle u_1, x\rangle \\ &\Rightarrow \frac{\partial}{\partial\alpha}\left(\|x\|^2 + \alpha^2 - 2\alpha\langle u_1, x\rangle\right) = 2\alpha - 2\langle u_1, x\rangle = 0 \Rightarrow \alpha = \langle u_1, x\rangle. \end{aligned} \tag{12}$$

因此数据可以被重新表示为 $\alpha u_1 = \langle u_1, x\rangle u_1$，$\alpha$ 就是数据在新坐标轴 u_1 上的坐标.

类似地，已有 k 个方向 $u_1, \ldots, u_k \in \mathbb{R}^d$，且满足 $\|u_i\| = 1, u_i^\top u_j = \delta_{ij}$，此时：

$$\arg\min_{\alpha_1,\ldots,\alpha_k} \|x - \sum_{i=1}^{k} \alpha_i u_i\|^2 \Rightarrow \alpha_i = \langle u_i, x\rangle. \tag{13}$$

PCA 可以从两个完全等价的角度来理解：

1. **Maximizing Projected Subspace.** 最大化投影：寻找一个单位向量（或子空间），使数据投影后的方差最大.
2. **Minimizing Residual.** 最小化残差：寻找一个低维子空间，使数据点到该子空间的重构误差最小.

Last update: December 19, 2025

A Proof of *MLE* of Gaussian

Proof of *MLE* of Gaussian

Proof. 对于高斯分布 $x \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \quad \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d}$，其 *MLE* 为：

$$\max_{\mu,\Sigma} \ell(\mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)} \mid \mu, \Sigma) = \min_{\mu,\Sigma} \sum_{i=1}^{n} \left[(x^{(i)}-\mu)^\top \Sigma^{-1}(x^{(i)}-\mu) + \log|\Sigma|\right]. \quad (14)$$

再对 μ 求梯度（假设Σ 满秩）寻找一阶最优性条件：

$$\nabla_\mu f(\mu) = \sum_{i=1}^{n} \Sigma^{-1}(\mu - x^{(i)}) = 0 \Rightarrow \sum_{i=1}^{n}(\mu - x^{(i)}) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \quad (15)$$

代入 $\mu = \hat{\mu}$，定义

$$S \triangleq \sum_{i=1}^{n}(x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^\top \Rightarrow \ell(\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}(\Sigma^{-1}S). \quad (16)$$

对矩阵 Σ 进行求导，得到：

$$\frac{\partial}{\partial\Sigma}\log|\Sigma| = \Sigma^{-1}, \quad \frac{\partial}{\partial\Sigma}\operatorname{tr}(\Sigma^{-1}S) = -\Sigma^{-1}S\Sigma^{-1}, \quad (17)$$

得到一阶最优性条件：

$$-\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}S\Sigma^{-1} = 0 \Rightarrow S = n\Sigma \Rightarrow \hat{\Sigma} = \frac{1}{n}S. \quad (18)$$

□

References