Subject: Stanford CS229 Machine Learning, Lecture 15, PCA / ICA

Date: from December 21, 2025 to December 22, 2025

Contents

Stanford CS229 Machine Learning, PCA / ICA, 2022, Lecture 15

Link on YouTube: Stanford CS229 Machine Learning, PCA / ICA, 2022, Lecture 15

Introduction

Introduction

本节内容包括 PCA 和 ICA 两部分内容. 在 PCA 中通过 PCA 的两个等价目标以线代视角讲解 PCA 的算法原理. 在 ICA 中通过简单的案例讲解 ICA 的问题设定和目标.

Principle Component Analysis

Principle Component Analysis: Recall

Find the closet point to the line. 在给定了单位方向 u 之后，寻找点 x 在 u 方向的坐标就变成了寻找到 u 张成的直线的最近点，即：

$$\alpha^* = \arg\min_{\alpha} \|x - \alpha u\|^2. \tag{1}$$

由于 $\|x-\alpha u\|^2 = \langle x-\alpha u, x-\alpha u\rangle = \|x\|^2 + \alpha^2\|u\|^2 - 2\alpha\langle x,u\rangle$，因此

$$\|x-\alpha u\|^2 \xrightarrow{\|u\|^2=1} \|x\|^2+\alpha^2-2\alpha\langle x,u\rangle \Rightarrow \frac{d}{d\alpha}(\|x\|^2+\alpha^2-2\alpha\langle x,u\rangle) = 0 \Rightarrow \alpha^* = \langle x,u\rangle. \tag{2}$$

General to K components. 若已得到了 K 个单位方向 $u_1,\dots,u_k \in \mathbb{R}^d$，$x \in \mathbb{R}^d$，则式(1)拓展为：

$$(\alpha_1^*,\dots,\alpha_K^*) = \arg\min_{\alpha_1,\dots,\alpha_d} \left\| x - \sum_{k=1}^{d} \alpha_k u_k \right\|^2 \Rightarrow \alpha_k^* = \langle x, u_k\rangle, \tag{3}$$

其中 $\left\|x - \sum_{k=1}^{d} \alpha_k^* u_k\right\|^2$ 被称为残差(Residual)，其表示 x 中无法被子空间表示的部分.

Goal. PCA 的目标可以通过两种等价的方式来定义：

1. 最大化投影到子空间上的方差: $\max_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(u^\top x^{(i)})^2$ s.t. $\|u\|_2 = 1$.
2. 最小化重构残差（Residual）: $\min_{\alpha} \|x-\alpha u\|^2$.

Principle Component Analysis: Linear Algebra Review

Linear Algebra Review. ① 设 $A \in \mathbb{R}^{d\times d}$ 为对称矩阵(symmetric)，则有如下分解：

$$A = U\Lambda U^\top, \tag{4}$$

其中 U 为正交矩阵(orthogonal)，即 $UU^\top = I$；Λ 为对角阵(diagonal)，且 $\Lambda_{ii} = \lambda_i$.[a]

② 设 $x = \sum_{k=1}^{n} \alpha_k u_k$，其中 $U = [u_1,\dots,u_n]$，则：

$$Ax = U\Lambda U^\top x = U\Lambda \sum_{k=1}^{n} \alpha_k e_k = U\sum_{k=1}^{n} \lambda_k \alpha_k e_k = \sum_{k=1}^{n} \lambda_k \alpha_k u_k, \tag{5}$$

因此

$$x^\top Ax = (\sum_{k=1}^{n} \alpha_k u_k)^\top (\sum_{k=1}^{n} \lambda_k \alpha_k u_k) = \sum_{k=1}^{n} \lambda_k \alpha_k^2. \tag{6}$$

从而可得:

$$\max_{\|x\|^2=1} x^\top Ax \iff \max_{\|\alpha\|^2=\sum_k \alpha_k^2=1} \sum_{k=1}^{n} \lambda_k \alpha_k^2 \implies \alpha_1 = 1, \quad \alpha_{k\neq 1} = 0. \tag{7}$$

Note 1. 若 $\lambda_1 = \lambda_2$，则任意选择 λ_1, λ_2 使得 $\lambda_1^2 + \lambda_2^2 = 1$ 即可.

[a]即为特征值，其中 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$. 能够排列是因为对角阵的特征值均为实数因此可以比较大小.

Principle Component Analysis: Eigenvalue

在上述线性代数基础上，对于 PCA 的目标可以进行如下改写:

$$\max_{u\in\mathbb{R}^d, \|u\|_2=1} \quad \frac{1}{n}\sum_{i=1}^{n}(u^\top x^{(i)})^2 \xlongequal{(u^\top x^{(i)})^2 = u^\top x^{(i)} x^{(i)\top} u} u^\top \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)} x^{(i)\top}\right) u, \tag{8}$$

其中由于数据已经过中心化，因此 $\Sigma = \frac{1}{n}\sum\limits_{i=1}^{n} x^{(i)} x^{(i)\top}$ 即为协方差矩阵，故问题转化为:

$$\max_{\|u\|^2=1} \quad u^\top \Sigma u. \tag{9}$$

根据式(7)结论可知每一次只需要选取最大的特征值即可，因此原始数据 $x^{(i)}$ 表示为:

$$x^{(i)} \quad \longmapsto \quad \sum_{j=1}^{k} (x^{(i)\top} u_j)\, u_j, \quad \mathbb{R}^d \quad \longrightarrow \quad \mathbb{R}^k. \tag{10}$$

k 的选取一般是使得累计解释方差$\geqslant 90\%$，即选取累计占比 $\geqslant 90\%$ 的特征值:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geqslant 0.9. \tag{11}$$

Note 2. *PCA* 仍然是使用与原先坐标一样的 $\mathbb{R}^d$ 维基向量，但不同的是原来 $\mathbb{R}^d$ 维空间有 d 个基向量、需要 d 个坐标，例如三维空间中 *(1, 2, 1)* 才能表示一个点，但 *PCA* 重新选取了 $k \ll d$ 个基向量，只需要 k 个系数就可以将该坐标表示，因此数据的表示 $\mathbb{R}^d \to \mathbb{R}^k$.

Note 3. 式*(11)*中分母虽然包含了所有的特征值，但是并不需要计算，因为只需要计算 *trace*(Σ) 即可，因此还是只需要计算 k 个特征值. 如果有相同的特征值，那么可能每次进行 *PCA* 得到的主成分不相同.

Independent Component Analysis (ICA)

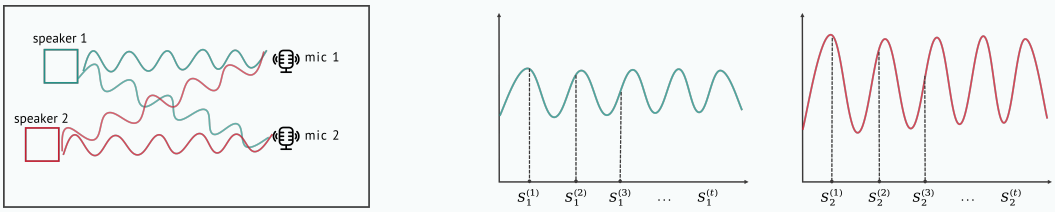
Independent Component Analysis: Motivation

假设在一个房间中有两个说话者(speaker) S_1, S_2 和两个麦克风(microphone)，他们的位置均固定(如图1a). 说话者可以发出 2 个独立信号 $s_1^{(t)}, s_2^{(t)}$ (如图1b)，麦克风可以观测到

2 个混合信号 $x_1^{(t)}, x_2^{(t)}$. 假设仅观测数据已知 $x_1^{(t)}, x_2^{(t)}$，且其为 $s_1^{(t)}, s_2^{(t)}$ 的线性组合：

$$x_j(t) = a_{j1}s_1^{(t)} + a_{j2}s_2^{(t)},\ j = \{1,2\} \Rightarrow x(t) = As(t) \tag{12}$$

其中 $A, s(t)$ 均视为 latent (未知). 我们的目的是想通过观测恢复背后"相互独立的源信号".



(a) (b) Signal from speakers

Figure 1: Example for ICA

下面将上述例子拓展至一般形式.

Given: 观测数据 $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$，$d$ 为麦克风数 = 说话源数.

Do: 找到 $s^{(1)}, \ldots, s^{(n)} \in \mathbb{R}^d$ 及 $A \in \mathbb{R}^{d\times d}$，使得 $x^{(t)} = As^{(t)}$，其中称 A 为混合矩阵(mixing matrix)，称 $W = A^{-1}$ 为 unmixing matrix.

如果已知 A 或 W，那么 $s^{(t)} = Wx^{(t)}$，写成行向量为

$$W = \begin{bmatrix} w_1^\top \\ \vdots \\ w_d^\top \end{bmatrix} \Rightarrow s_j^{(t)} = w_j^\top x^{(t)}. \tag{13}$$

Caveats. 需要注意的是：

1. 假设源信号相互独立，且混合矩阵 A 不随时间变化.
2. ICA 存在内在的不确定性(intrinsic ambiguity). 源信号在如下方面不可区分：
   - 顺序(permutation): 例如 speaker 1 与 speaker 2 可交换，不能确定顺序
   - 尺度/ 强度(scaling): 由于 $(CA)(C^{-1}s^{(t)}) = As^{(t)}$，因此对于已观测到的 $As^{(t)}$，任意改变 A 都可以间接改变源信息 $s^{(t)}$，因此只能相对区分强度.
3. 源信号不能为高斯分布：若源信号为高斯分布 $s \sim \mathcal{N}(\mu_s, I)$，则

$$x^{(i)} \sim \mathcal{N}(\mu_x, AA^\top) \Rightarrow \text{若 } U^\top U = I,\ AU \text{ 生成与之相同的观测分布} \tag{14}$$

Note 4. 若源信号为高斯分布 $s \sim \mathcal{N}(\mu_s, I)$，则 $x = As \sim \mathcal{N}(\mu_x, AA^\top)$. 由于 $\mathcal{N}(\mu_x, AA^\top)$ 的 *covariance matrix* $AA^\top$ 是对称的，因此其具有旋转不变性*(rotation invariance)*. 取正交矩阵 $UU^\top = I$，定义新源 $s' = Us$，由于高斯分布在正交变换下不变因此 $s' \sim \mathcal{N}(\mu_s, I)$，故 $x = As = A(U^\top U)s = (AU^\top)\tilde{s}$，即新源可以生成一样的观测分布，从而根本无法区分.

Independent Component Analysis: Algorithm

Review: Density under linear transform. 对于均匀分布 $s \sim U([0,1]), p_s(x) = \begin{cases} 1, & [0,1] \\ 0, & \text{o.w.} \end{cases}$

若 $u = 2s$，则归一化常数必须改变: $p_u(x) = p_s\left(\frac{x}{2}\right) \cdot \frac{1}{2}$. 对于一般情形 $x = As$:

$$p_x(x) = p_s(A^{-1}x) \cdot \left|\det(A^{-1})\right| \xrightarrow{W=A^{-1}} p_x(x) = p_s(Wx) \cdot |\det(W)| \tag{15}$$

ICA is *MLE.* 由于源信号相互独立，因此 $p(s) = \prod_{j=1}^{d} p_s(s_j)$，其中 d 为源信号数，故

$$x = As, W = A^{-1}, s = Wx \Rightarrow p(x) = \prod_{j=1}^{d} p_s(w_j^\top x) \cdot |\det(W)|. \tag{16}$$

Key technical trick. 在计算概率时使用 likelihood function g 进行替代：$p_s(s_j) \propto g'(s_j)$，其中 $g(x) = (1 + e^{-x})^{-1}$，此时 Log-likelihood:

$$\ell(W) = \sum_{t=1}^{n} \left[\sum_{j=1}^{d} \log g'(w_j^\top x^{(t)}) + \log|\det(W)| \right]. \tag{17}$$

Note 5. 事实上此处的 *likelihood function* g 只需要是非 *relationally invariant* 即可. 此处 $g(x) = (1 + e^{-x})^{-1}$ 为 *sigmoid*，使用的是其导数 g'，即 *logistic density.*

Last update: December 22, 2025