

7.2 特征工程

概括地说，特征工程就是把数据从一种形式变为另一种更适合利用的形式（图 7.17）。这一过程往往占数据分析总工作量的 70% 以上，是最重要但又缺乏系统方法的一个环节。它包括特征抽取 (feature extraction) 和特征选择 (feature selection) 两个子类。



图 7.17 特征工程

除了数据表示，以下所列的特征工程的方方面面也是非常重要的，但由于篇幅所限，它们在本书中并未被详尽地展开介绍，我们只能走马观花地介绍一些经典的方法。

- ❑ 数据整理：对数据进行汇总、分组、归类、编码、审核等，使之更加条理化，包括
 - 数据清洗（例如，独立成分分析，见 §7.2.3）。
 - 缺失数据分析（见 §8.3）。
 - 诱导特征：由已知的特征构造新的、有意义的特征（例如，由距离和时间定义平均速度）。
 - 数据压缩与重构：有损压缩和无损压缩（例如，主成分分析等）。
 - 数据合并：不同来源的同类数据（例如，不同实验室采集的某个癌症的质谱数据）如何合并成一个更大的数据集？
- ❑ 关系发现：如关联规则 (association rule)、知识图谱、因果分析，等等。
- ❑ 异常点检测：从多种角度（例如，分布、预测、模式识别等）判定异常点，然后找到引起异常的原因（根因分析更重要），特别应用于健康监测、故障诊断、欺诈检测、干扰分析等。

1883 年，高尔顿出版了心理学史上具有里程碑意义的著作《人类的才能及其发展研究》^[116]，提出“优生学” (eugenics) 这个术语。他试图从外在生物特征找出与智商、健康、犯罪的关系，并极力推广他的“智商测试”的理论（图 7.18）。高尔顿和他的学生卡尔·皮尔逊都热衷于采集人类颅骨的各种测量数据，他们试图通过颅骨测量术来评估智商，但最后均以失败告终。

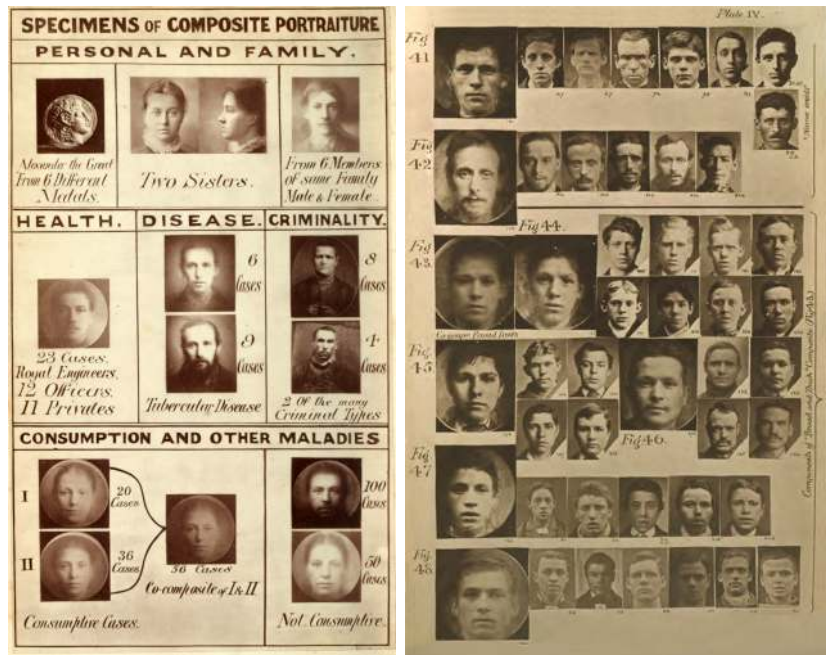


图 7.18 《人类的才能及其发展研究》中收集的数据

1893 年，高尔顿参观了巴黎的贝蒂隆刑事鉴定实验室，图 7.19 是当时的留影。不知 AI 识别系统能否单凭面相判定高尔顿是一位绅士，而不是一个种族主义者？

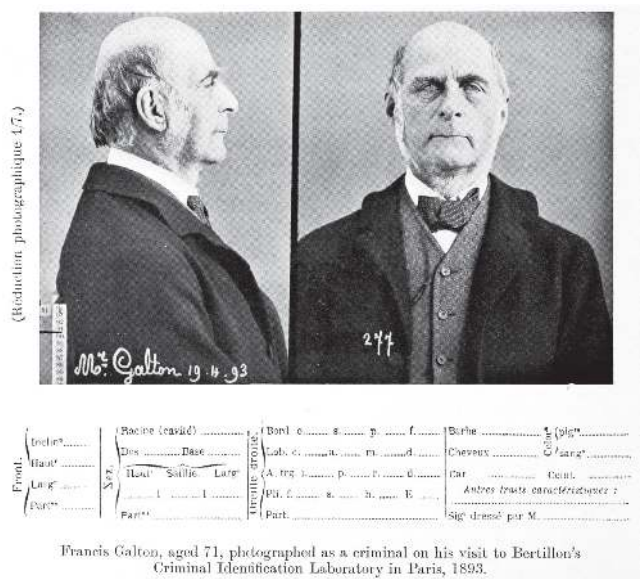


图 7.19 高尔顿

人体测量学 (anthropometry) 无法获取显著的特征（包括脑容量）来揭示智商的规律。可见，专业领域的可靠知识对于成功的特征工程、统计建模、数据分析是必不可少的，它指导我们应该采集怎样的数据，以及如何使用这些数据来佐证或反驳某个科学假设。毫无经验或知识的数据分析被称为“模型盲” (model-blind)，是应该尽量避免的（图 7.20）。



图 7.20 模型盲不足取

1. 特征选择

美籍匈牙利裔数学家、物理学家、计算机科学家约翰·冯·诺依曼 (John von Neumann, 1903—1957) 曾说“用四个参数，我就能拟合一头大象，用五个参数，我就能让它扭动鼻子。”冯·诺依曼（图 7.21）深知参数越多，表达的自由度越高。



图 7.21 冯·诺依曼

然而，我们并不知道大自然是怎样表达出数据的。作为常见的数据表示方法，特征选择只考虑总体 $\mathbf{X} = (X_1, \dots, X_d)^T$ 的部分分量 $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_k)^T$ ，其中 $1 \leq k \leq d$ ， \tilde{X}_i 是 \mathbf{X} 的某个分量，目标是为了使得应用（分类、预测等）效果更好或者效率更高。简而言之，特征选择扔掉与机器学习任务毫无帮助的分量。例如，对某个疾病的诊断来说，哪些医学指标或临床症状有助于精准地判别出有无该疾病？

在实际问题中，特征往往联合在一起产生“化学反应”，要找出很多特征的某个最优的子集并非易事，“组合爆炸”和应用效果都是拦路虎。目前，特征选择缺少系统的方法，仍是一个未解决的问题。本

书没有深入探讨这个话题，仅介绍一个简单的过滤法 (filter method)* 如下。

算法 7.1 (浅层特征选择) 给定一个二分类数据集 $\{(\mathbf{x}_j, t_j) : \mathbf{x}_j \in \mathbb{R}^m, t_j \in \{0, 1\}, \text{ 其中 } j = 1, 2, \dots, n\}$ 。此处，我们用 $\{0, 1\}$ 表示类标，每个样本 $\mathbf{x}_j \in \mathbb{R}^m$ 都是一个列向量。如表 7.1 所示，不妨设 $\mathbf{x}_1, \dots, \mathbf{x}_k$ 属于类 0，其余样本属于类 1。

表 7.1 二分类的浅层特征选择

特征	两类样本						重要性
	\mathbf{x}_1	\cdots	\mathbf{x}_k	\mathbf{x}_{k+1}	\cdots	\mathbf{x}_n	
f_1	$x_{1,1}$	\cdots	$x_{1,k}$	$x_{1,k+1}$	\cdots	$x_{1,n}$	$P(f_1 \alpha)$
f_2	$x_{2,1}$	\cdots	$x_{2,k}$	$x_{2,k+1}$	\cdots	$x_{2,n}$	$P(f_2 \alpha)$
\vdots	\vdots		\vdots	\vdots		\vdots	\vdots
f_m	$x_{m,1}$	\cdots	$x_{m,k}$	$x_{m,k+1}$	\cdots	$x_{m,n}$	$P(f_m \alpha)$
类	0	\cdots	0	1	\cdots	1	

浅层特征选择 (shallow feature selection) 的基本想法是：毋须假设总体分布，如果特征 $f_i, i = 1, 2, \dots, m$ 在统计意义上显著地区分了两个类，那么它对分类器来说是重要的。

- 利用两样本的 K-S 检验，在给定的显著水平 α 判定 $x_{i,1}, \dots, x_{i,k}$ 和 $x_{i,k+1}, \dots, x_{i,n}$ 的是否来自同一总体（零假设 $H_0^{(i)}$ 是它们来自同一总体），其中 $i = 1, 2, \dots, m$ 。
- 重抽样 $(N - 1)$ 次，对每次抽样，重复地做 K-S 检验。

重抽样弱化了极端值对 K-S 检验的影响。如果在多数试验中 $H_0^{(i)}$ 被拒绝，则特征 f_i 对分类来说是重要的。为刻画特征 f_i 的重要性，定义选择 f_i 的概率为

$$P(f_i|\alpha) = \frac{\#(\text{拒绝 } H_0^{(i)})}{N}$$

2. 特征抽取

在某个标准之下，将原始数据经过某个变换后使之达到最优的过程称作“特征抽取”。标准不同，所用的变换就不同，因此无法泛泛地评判孰优孰劣，一般要根据实际要求或应用效果来选择。譬如，针对信号处理里的盲源分离问题，独立成分分析是合适的方法。

相比特征选择，特征抽取所得到的数据表示可能缺乏“可解读性”，即经过变换使得数据分量的含义发生了改变，对人类而言，有的时候很难明确地解释它们。为了追求好的应用效果，人们常常会舍弃一些“可解读性”。

无论特征抽取，还是特征选择，都可以与某个具体的学习机“联手”，得到最适合该学习机的数据表示。但这种作法有待商榷，它可能导致数据表示对其他的学习机是不适合的——在此情况下，很难说数据表示抓住了数据的本质特征。

*过滤法逐一地考察特征，在统一标准下决定每个特征的去留。另外两种特征选择方法——封装法（考虑几个特征的共同作用）和嵌入法（同时考虑特征选择和学习机），本书不予涉猎。

面对一个机器学习任务，数据表示应不应该捆绑具体的学习机（如特征选择的嵌入法、深度学习等），还是要利用一些普适的准则有目的地挖掘数据本身的某些特征？这个问题学界尚无定论。如果要使得整理后的数据在多数学习机上都能取得改进的效果，后者或许更合理一些（本节所介绍的特征提取的方法，基本上都属于这一类），尤其对“可解释性”（即“可溯因性”）的要求较高时，有助于我们锁定问题出自模型还是数据。

有的时候，为了改善数据的可分性，我们需要把数据映射到高维空间（甚至无穷维的希尔伯特空间）。有的时候，我们需要对数据进行有损压缩，在保证压缩比的同时要使得数据在重构中的损失尽可能地少。有的时候，我们需要深入理解数据特征背后的规律，甚至探索一些不可直接观测的因素。有的时候，观测数据经过了某些未知的变换，我们需要对它们进行清洗，找出数据的原始形态……。

选择参数和数学建模一般需要特定的背景知识，有时解决一个难题可能需要“东市买骏马，西市买鞍鞮，南市买辔头，北市买长鞭。”（《木兰辞》）对统计分析和机器学习而言，一定数量的合适特征将使得对模式的描述更加丰富。如何得到这些特征？这是数据分析中最重要的环节——特征工程所关注的研究内容。有的时候，即便找到了特征，在特定环境下也难以用上。有诗为证，“雄兔脚扑朔，雌兔眼迷离；双兔傍地走，安能辨我是雄雌？”（图 7.22）



图 7.22 《木兰辞》

注：《木兰辞》是中国北朝的一首民歌，收于《乐府诗集》。花木兰隐藏了女性的特征替父从军，连与她朝夕相处的战友都没识别出来。

总而言之，“特征抽取”超越了费舍尔的简化数据的想法，有时为了抓取到本质特征，对数据进行必要的繁化也未尝不可。本节的后续内容主要聚焦于一些常见的特征抽取方法。

7.2.1 主成分分析

《随机之美》的第四章介绍了随机向量的主成分：随机向量 $\mathbf{X} = (X_1, \dots, X_d)^\top$ 经过某正交变换 $\mathbf{U}^\top \mathbf{X} = (Y_1, \dots, Y_d)^\top$ 后，各分量变得不相关（即相关系数为零），并且第一个分量 Y_1 的方差最大，第二分量 Y_2 的方差次之，……。这些分量被称为随机向量 \mathbf{X} 的主成分。该正交矩阵 \mathbf{U} 的列向量正是 \mathbf{X} 的方差-协方差矩阵的单位本征向量 $\mathbf{u}_1, \dots, \mathbf{u}_d$ ，对应的本征值满足

$$\lambda_1 \geq \dots \geq \lambda_d \geq 0$$

给定样本 $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ ，经过中心化 $\mathbf{Z} = \mathbf{Z}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ 后，式 (7.3) 定义的样本协方差矩阵 \mathbf{C} 是对总体 $\mathbf{X} = (X_1, \dots, X_d)^\top$ 协方差矩阵的估计，它的本征向量是否也有类似的性质？

随机向量主成分的概念是美国统计学家、经济学家哈罗德·霍特林 (Harold Hotelling, 1895—1973) 于 1933 年明确定义的，他还提出了多元统计的主成分分析 (principal component analysis, PCA) 的代数方法^[117]。霍特林 (图 7.23) 是美国大学统计学教育的主要推动者，他是费舍尔的追随者，深受其影响，曾全力推荐《研究者用的统计方法》一书。主成分分析常用于降维和最佳近似，是多元统计学里经典的数据表示方法之一。1901 年，统计学之父卡尔·皮尔逊曾给出该方法的几何推导。



图 7.23 霍特林

1. 经典主成分分析

主成分分析的主要想法是寻找一个单位向量 $\mathbf{u}_1 \in \mathbb{R}^d$ ，使得样本在这个方向上的投影坐标 $\mathbf{u}_1^\top \mathbf{x}_1, \dots, \mathbf{u}_1^\top \mathbf{x}_n$ 具有最大方差。单位向量 \mathbf{u}_1 被称为样本的第一主成分 (principal component)，它之所以特殊，是因为数据在 \mathbf{u}_1 这个方向上所含的信息量最大。即，

$$\begin{aligned} \mathbf{u}_1 &= \operatorname{argmax}_{\|\mathbf{u}\|=1} \|\mathbf{Z}^\top \mathbf{u}\|_2^2 \\ &= \operatorname{argmax}_{\|\mathbf{u}\|=1} \mathbf{u}^\top \mathbf{Z} \mathbf{Z}^\top \mathbf{u} \end{aligned} \quad (7.21)$$

□ 下面，利用拉格朗日乘子法求解这个最优化问题。该问题的拉格朗日函数是

$$\mathcal{L}(\mathbf{u}) = \mathbf{u}^\top \mathbf{Z} \mathbf{Z}^\top \mathbf{u} + \lambda(1 - \mathbf{u}^\top \mathbf{u})$$

为最大化该函数，令 $\nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}) = \mathbf{0}$ ，即

$$2\mathbf{Z} \mathbf{Z}^\top \mathbf{u} - 2\lambda \mathbf{u} = \mathbf{0}$$

我们得到

$$\mathbf{Z} \mathbf{Z}^\top \mathbf{u} = \lambda \mathbf{u} \quad (7.22)$$

因此，该最优化问题的解是散布矩阵 $\mathbf{S} = \mathbf{Z} \mathbf{Z}^\top$ 的本征向量。半正定矩阵 $\mathbf{Z} \mathbf{Z}^\top$ 的本征值都是非负的。

此时，为了使投影方差 $\|\mathbf{Z}^T \mathbf{u}\|_2^2 = \lambda$ 达到最大， λ 应是散布矩阵 $\mathbf{S} = \mathbf{Z}\mathbf{Z}^T$ 最大的本征值 λ_1 ，其本征向量 \mathbf{u}_1 就是第一主成分，见图 7.24。显然，样本在 \mathbf{u}_1 主成分上的方差要更大一些，所保留的原数据的信息量也更多一些。对于二维数据而言， \mathbf{u}_1 确定后， \mathbf{u}_2 也随之确定下来了。

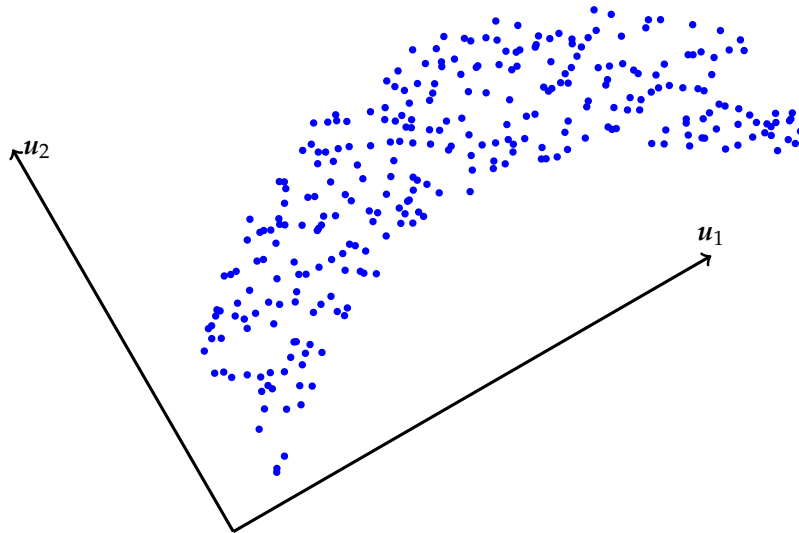


图 7.24 第一主成分的直观含义

□ 如果样本空间的维度大于 2，接下来，我们寻找第二主成分：一个与 \mathbf{u}_1 正交的单位向量，并且最大化目标函数

$$\mathcal{L}(\mathbf{u}) = \mathbf{u}^T \mathbf{Z}\mathbf{Z}^T \mathbf{u} + \lambda(1 - \mathbf{u}^T \mathbf{u}) + \alpha(\mathbf{u}^T \mathbf{u}_1)$$

令 $\nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}) = \mathbf{0}$ ，我们得到

$$2\mathbf{Z}\mathbf{Z}^T \mathbf{u} - 2\lambda \mathbf{u} + \alpha \mathbf{u}_1 = \mathbf{0}$$

上式两侧左乘 \mathbf{u}_1^T ，有

$$2\mathbf{u}_1^T \mathbf{Z}\mathbf{Z}^T \mathbf{u} - 2\lambda \mathbf{u}_1^T \mathbf{u} + \alpha \mathbf{u}_1^T \mathbf{u}_1 = 0$$

即，

$$2(\mathbf{Z}\mathbf{Z}^T \mathbf{u}_1)^T \mathbf{u} + \alpha = 0$$

进而，

$$\alpha = -2\lambda_1 \mathbf{u}_1^T \mathbf{u} = 0$$

于是, 最优问题的解依然满足 (7.22), 即 $\mathbf{S} = \mathbf{Z}\mathbf{Z}^\top$ 的本征向量。类似地, 为了使投影方差达到最大, 第二主成分应为第二大本征值 λ_2 所对应的本征向量 \mathbf{u}_2 。

算法 7.2 (主成分分析) 设中心化的解释矩阵 $\mathbf{Z}_{d \times n}$ 的秩为 r , 具有奇异值分解

$$\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

样本的前 $k \leq r$ 个主成分以及基于它们的数据表示可按下面的方法求得。

(1) 前 $k \leq r$ 个主成分即矩阵 \mathbf{U} 的前 k 列

$$\mathbf{u}_1, \dots, \mathbf{u}_k \in \mathbb{R}^d$$

(2) 样本 $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathbb{R}^d$ 在空间 $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$ 的投影 $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^k$ 即是 \mathbf{Z} 的 PCA 数据表示, 即

$$\begin{aligned} \mathbf{Y}_{k \times n} &= (\mathbf{y}_1, \dots, \mathbf{y}_n) \\ &= \mathbf{U}_k^\top \mathbf{Z} \\ &= \mathbf{\Sigma}_k \mathbf{V}_k^\top \end{aligned} \quad (7.23)$$

其中,

$$\begin{aligned} \mathbf{U}_k &= (\mathbf{u}_1, \dots, \mathbf{u}_k) \\ \mathbf{V}_k &= (\mathbf{v}_1, \dots, \mathbf{v}_k) \\ \mathbf{\Sigma}_k &= \text{diag}(\sigma_1, \dots, \sigma_k) \end{aligned}$$

证明: 结果 (7.23) 只需要代入奇异值分解验证即可。

$$\begin{aligned} \mathbf{U}_k^\top \mathbf{Z} &= \mathbf{U}_k^\top \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top \\ &= (\mathbf{I}_k, \mathbf{O}_{k \times (d-k)}) \text{diag}(\sigma_1, \dots, \sigma_r) (\mathbf{v}_1, \dots, \mathbf{v}_r)^\top \\ &= \mathbf{\Sigma}_k \mathbf{V}_k^\top \end{aligned}$$

□



算法 7.2 是一个通用的 PCA 算法。事实上, 对于 $d \ll n$ 的情形, 散布矩阵 $\mathbf{S}_{d \times d} = \mathbf{Z}\mathbf{Z}^\top$ 的谱分解比解释矩阵 \mathbf{Z} 的奇异值分解代价要小, 求前 k 个主成分莫不如直接求 \mathbf{S} 的前 k 个单位本征向量。

性质 7.6 算法 7.2 给出的样本的数据表示 (7.23) 是中心化的, 它的散布矩阵是

$$\mathbf{S}_Y = \text{diag}(\sigma_1^2, \dots, \sigma_k^2), \text{ 其中 } \sigma_1 \geq \dots \geq \sigma_k > 0 \text{ 是 } \mathbf{Z} \text{ 的前 } k \text{ 个奇异值}$$

证明: 样本 $\mathbf{y}_1 = \mathbf{U}_k^\top \mathbf{z}_1, \dots, \mathbf{y}_n = \mathbf{U}_k^\top \mathbf{z}_n$ 之和为 $\mathbf{0}$, 其散布矩阵为

$$\mathbf{S}_Y = \mathbf{Y}\mathbf{Y}^\top$$

$$\begin{aligned}
&= \Sigma_k V_k^T V_k \Sigma_k \\
&= \Sigma_k I_k \Sigma_k \\
&= \text{diag}(\sigma_1^2, \dots, \sigma_k^2)
\end{aligned}$$

□

例 7.8 考虑鸢尾花 (iris) 数据：样本维数 $d = 4$ ，样本容量 $n = 150$ ，分为 setosa、versicolor 和 virginica 三个类（每个类 50 个样本点）。我们将这三个类混在一起，利用 (7.23) 给出 iris 数据的二维 PCA 表示，如图 7.25 所示。

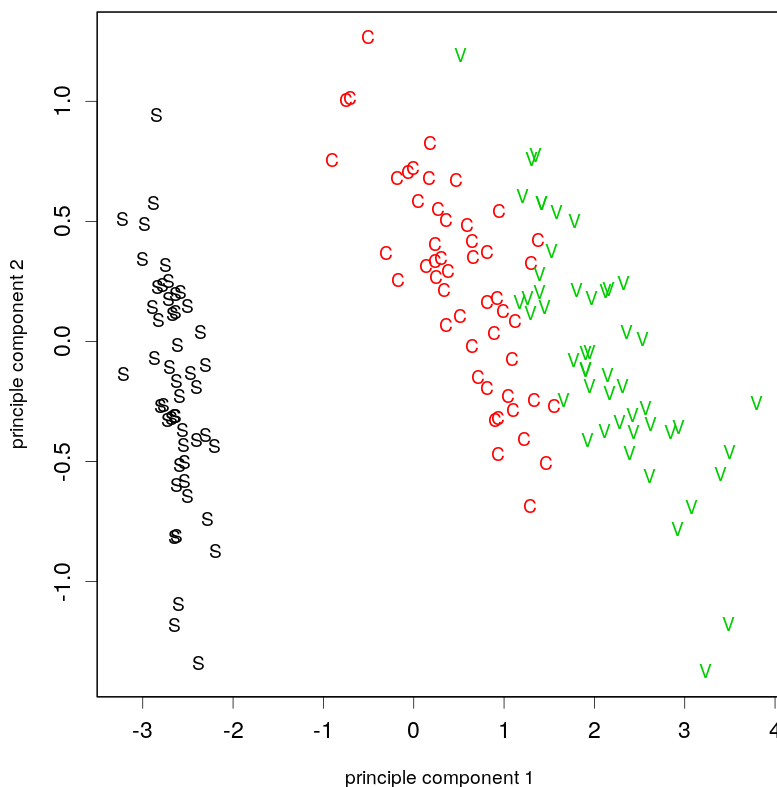


图 7.25 iris 数据的二维 PCA 表示

性质 7.7 在算法 7.2 中，按照下面的方式产生矩阵 $Z_{(k)}$ 作为 $Z_{d \times n}$ 的重构。

$$\begin{aligned}
Z_{(k)} &= U_k Y \\
&= U_k U_k^T Z \\
&= U_k \Sigma_k V_k^T
\end{aligned}$$

□ 由埃卡特-杨定理（见《随机之美》^[8] 的附录）可知，在所有秩为 k 的 $d \times n$ 矩阵中，按照 2-范数和 F -范数， $Z_{(k)}$ 都是 Z 的最佳近似，称为埃卡特-杨近似。即


$$Z_{(k)} = \underset{\text{rank}(B)=k}{\text{argmin}} \|B - Z\|_2$$

$$= \underset{\text{rank}(\mathbf{B})=k}{\text{argmin}} \|\mathbf{B} - \mathbf{Z}\|_F$$

其中, 误差分别为

$$\begin{aligned} \|\mathbf{Z}_{(k)} - \mathbf{Z}\|_2 &= \sigma_{k+1} \\ \|\mathbf{Z}_{(k)} - \mathbf{Z}\|_F &= \sqrt{\sum_{j=k+1}^r \sigma_j^2} \end{aligned}$$

□ 原解释矩阵 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ 可由 $\underbrace{(\bar{\mathbf{x}}, \dots, \bar{\mathbf{x}})}_{n \text{ 列}} + \mathbf{Z}_{(k)}$ 来近似。

 对于高维样本, 即 $d \gg n$, 散布矩阵 $\mathbf{S}_{d \times d}$ 的谱分解和解释矩阵 $\mathbf{Z}_{d \times n}$ 的奇异值分解代价都大, 精度也受影响。下面, 我们专门为高维数据的 PCA 设计一个算法。关键的想法来自性质 7.2, 不妨设 \mathbf{S} 的单位本征向量 \mathbf{u} 为

$$\mathbf{u} = \mathbf{Z}\mathbf{v} = \sum_{i=1}^n v_i \mathbf{z}_i, \text{ 其中 } \mathbf{v} = (v_1, \dots, v_n)^\top \text{ 使得 } \|\mathbf{u}\|_2 = 1 \quad (7.24)$$

用 \mathbf{Z}^\top 左乘 $\mathbf{S}\mathbf{u} = \lambda\mathbf{u}$ 两侧, 可得

$$\mathbf{Z}^\top \mathbf{Z} \mathbf{Z}^\top \mathbf{Z} \mathbf{v} = \lambda \mathbf{Z}^\top \mathbf{Z} \mathbf{v} \quad (7.25)$$

式 (7.25) 中, $\mathbf{G} = \mathbf{Z}^\top \mathbf{Z}$ 是样本的内积矩阵。经过整理, 式 (7.25) 变为

$$\mathbf{G}^2 \mathbf{v} = \lambda \mathbf{G} \mathbf{v}$$

解此方程组, 我们只需求下面的非零本征值问题。

$$\mathbf{G} \mathbf{v} = \lambda \mathbf{v}, \text{ 其中 } \lambda \text{ 是 } \mathbf{G} \text{ 的本征值} \quad (7.26)$$

式 (7.24) 要求 \mathbf{v} 满足

$$\mathbf{u}^\top \mathbf{u} = \mathbf{v}^\top \mathbf{Z}^\top \mathbf{Z} \mathbf{v} = \mathbf{v}^\top \mathbf{G} \mathbf{v} = 1$$

结合式 (7.26), 待定系数 \mathbf{v} 满足

$$\|\mathbf{v}\|_2 = \frac{1}{\sqrt{\lambda}} \quad (7.27)$$

显然, $\mathbf{Z}^\top \mathbf{u} = \mathbf{G} \mathbf{v}$ 是样本 $\mathbf{z}_1, \dots, \mathbf{z}_n$ 在单位向量 \mathbf{u} 上的投影坐标。有趣的是, 不必计算出主成分 $\mathbf{u}_1, \dots, \mathbf{u}_k$ 的具体结果, 样本 $\mathbf{z}_1, \dots, \mathbf{z}_n$ 在空间 $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$ 上的投影就变成求内积矩阵 \mathbf{G} 的满足条件 (7.27) 的本征向量。

算法 7.3 (高维数据的主成分分析) 给定样本 $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, 样本维数 $d \gg n$ 。设 $\mathbf{z}_1, \dots, \mathbf{z}_n$ 是中心化了的样本。

- (1) 计算内积矩阵 $\mathbf{G}_{n \times n} = \mathbf{Z}^\top \mathbf{Z}$ 的谱分解 $\mathbf{G} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top$, 其中 $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ 是 \mathbf{G} 的本征值按降序构成的对角阵, $\mathbf{V}_{n \times n} = (\mathbf{v}_1, \dots, \mathbf{v}_n)$ 是正交矩阵。
- (2) 样本 $\mathbf{z}_1, \dots, \mathbf{z}_n$ 由前 k 个主成分给出的数据表示是

$$\mathbf{Y}_{k \times n} = (\mathbf{y}_1, \dots, \mathbf{y}_n) = \mathbf{\Lambda}_k^{\frac{1}{2}} \mathbf{V}_k^\top \quad (7.28)$$

其中,

$$\begin{aligned} \mathbf{V}_k &= (\mathbf{v}_1, \dots, \mathbf{v}_k) \\ \mathbf{\Lambda}_k &= \text{diag}(\lambda_1, \dots, \lambda_k) \end{aligned}$$

证明: 样本 $\mathbf{z}_1, \dots, \mathbf{z}_n$ 在单位向量 $\mathbf{u}_j, j = 1, \dots, k$ 上的投影坐标是

$$\begin{aligned} \mathbf{G} \frac{\mathbf{v}_j}{\sqrt{\lambda_j}} &= \mathbf{V} \mathbf{\Lambda} \left(\mathbf{V}^\top \frac{\mathbf{v}_j}{\sqrt{\lambda_j}} \right) \\ &= \mathbf{V} \text{diag}(\lambda_1, \dots, \lambda_n) \left(0, \dots, 0, \frac{1}{\sqrt{\lambda_j}}, 0, \dots, 0 \right)^\top \\ &= (\mathbf{v}_1, \dots, \mathbf{v}_n) (0, \dots, 0, \sqrt{\lambda_j}, 0, \dots, 0)^\top \\ &= \sqrt{\lambda_j} \mathbf{v}_j \end{aligned} \quad \square$$

性质 7.8 在算法 7.3 中, 根据式 (7.28), $\mathbf{Z}_{(k)} = \mathbf{U}_k \mathbf{Y} = \mathbf{Z} \mathbf{V}_k \mathbf{V}_k^\top$ 是对 $\mathbf{Z}_{d \times d}$ 的近似。

证明: 根据 (7.24), 前 k 个主成分为 $\mathbf{u}_j = \frac{1}{\sqrt{\lambda_j}} \mathbf{Z} \mathbf{v}_j, j = 1, \dots, k$ 。于是,

$$\mathbf{U}_k \mathbf{Y} = \left(\mathbf{Z} \frac{\mathbf{v}_1}{\sqrt{\lambda_1}}, \dots, \mathbf{Z} \frac{\mathbf{v}_k}{\sqrt{\lambda_k}} \right) \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_k}) \mathbf{V}_k^\top = \mathbf{Z} \mathbf{V}_k \mathbf{V}_k^\top \quad \square$$

例 7.9 (潜在语义分析) 给定 m 个术语和 n 篇文档, 则第 j 篇文档可被抽象为一个 m 维向量 $\mathbf{d}_j \in \mathbb{R}^m$, 其中第 i 个分量就是第 i 个术语在该文档中出现的次数。于是, 数据可被整理为一个 $m \times n$ 矩阵, 称为术语-文档矩阵 (term-document matrix), 定义如下

$$\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_n)$$

不妨设 \mathbf{D} 的奇异值分解为 $\mathbf{D} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$, 则 $\mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top$ 是 \mathbf{D} 的近似, 即

$$(\mathbf{d}_1, \dots, \mathbf{d}_n) \approx (\mathbf{u}_1, \dots, \mathbf{u}_k) \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_k \end{pmatrix} (\mathbf{v}_1, \dots, \mathbf{v}_k)^\top$$

令 $\hat{\mathbf{d}}_j = \mathbf{U}_k^\top \mathbf{d}_j$, 它是 \mathbf{d}_j 在 $\mathbf{u}_1, \dots, \mathbf{u}_k$ 张成的 k 维空间 (称为主题空间) 里的投影。不难验证

$$(\hat{\mathbf{d}}_1, \dots, \hat{\mathbf{d}}_n) = \Sigma_k \mathbf{V}_k^\top = \begin{pmatrix} \sigma_1 \mathbf{v}_1^\top \\ \vdots \\ \sigma_k \mathbf{v}_k^\top \end{pmatrix}$$

如图 7.26 所示, 我们称 \mathbf{V}_k^\top 为文档空间。第 j 列向量的分量分别乘上权重 $\sigma_1, \dots, \sigma_k$ 所得向量就是 $\hat{\mathbf{d}}_j = \mathbf{U}_k^\top \mathbf{d}_j$, 也是矩阵 $\Sigma_k \mathbf{V}_k^\top$ 的第 j 列。术语-文档矩阵具有埃卡特-杨近似 (详见 [8] 附录) 如下。

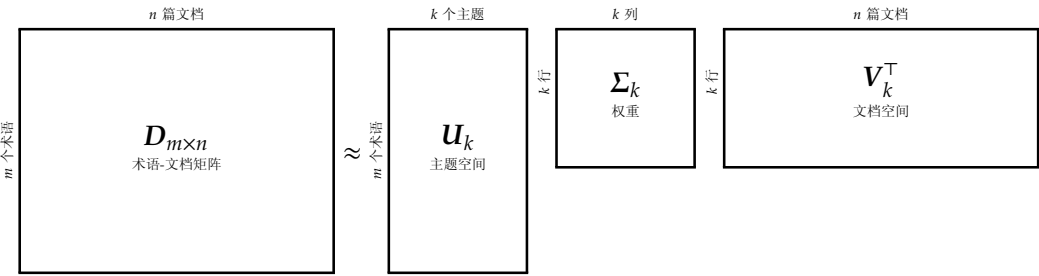


图 7.26 潜在语义分析的直观示意

潜在语义分析 (latent semantic analysis, LSA) 利用 $\hat{\mathbf{d}}_j$ 作为文档 \mathbf{d}_j 新的数据表示, 进而计算夹角余弦相似度等等。潜在语义分析利用奇异值分解降低样本的维数, 同时尽可能地保留了文档间的相似关系, 常用于文本聚类和信息检索。譬如, 考虑以下三篇中文文档, 经过切分 (segmentation) 后, 术语由下划线标出。

- (1) 整数 是自然数、自然数 的相反数 或者零。大多数数学 都包含整数。
- (2) 任何大于 1 的整数, 若只能被它本身和 1 整除, 则称之为素数 (见数论)。每一个整数 都有唯一一组素因子, 它们的乘积 等于该整数。例如, 整数 42 的素因子 是 2, 3 和 7。
- (3) 所有介子 的自旋 必须等于整数。自旋 等于整数 的粒子 称为玻色子。玻色子 有别于自旋 为非整数 的粒子——费米子, 它不服从 “泡利不相容原理” 这个物理规则。

在统计自然语言处理^[118] 的很多问题中, 高频无义词 (如助词 “着、了、过” 等) 和低频词一般不被考虑。不妨设词典 (lexicon) 由表 7.2 所示的一些术语构成。为简单起见, 文档中出现的某些词语未被收录。

设术语-文档矩阵 \mathbf{D} 的奇异值分解是 $\mathbf{D} = \mathbf{U} \Sigma \mathbf{V}^\top$ 。令 $k = 2$, 则文档在二维主题空间的数据表示为

$$(\hat{\mathbf{d}}_1, \hat{\mathbf{d}}_2, \hat{\mathbf{d}}_3) = \Sigma_2 \mathbf{V}_2^\top \approx \begin{pmatrix} -2.8945 & -4.0614 & -4.6177 \\ 1.4102 & 2.3422 & -2.9440 \end{pmatrix}$$

表 7.2 术语-文档矩阵

术语	d_1	d_2	d_3
玻色子	0	0	2
乘积	0	1	0
费米子	0	0	1
介子	0	0	1
粒子	0	0	2
泡利不相容原理	0	0	1
数论	0	1	0
数学	1	0	0
素数	0	1	0
素因子	0	2	0
物理规则	0	0	1
相反数	1	0	0
整除	0	1	0
整数	3	4	3
自然数	2	0	0
自旋	0	0	3

例 7.9 中，潜在语义分析将文档和术语投射到二维主题空间里，如图 7.27 所示。一般地，语义上接近的对象，其夹角余弦相似度相对较高。

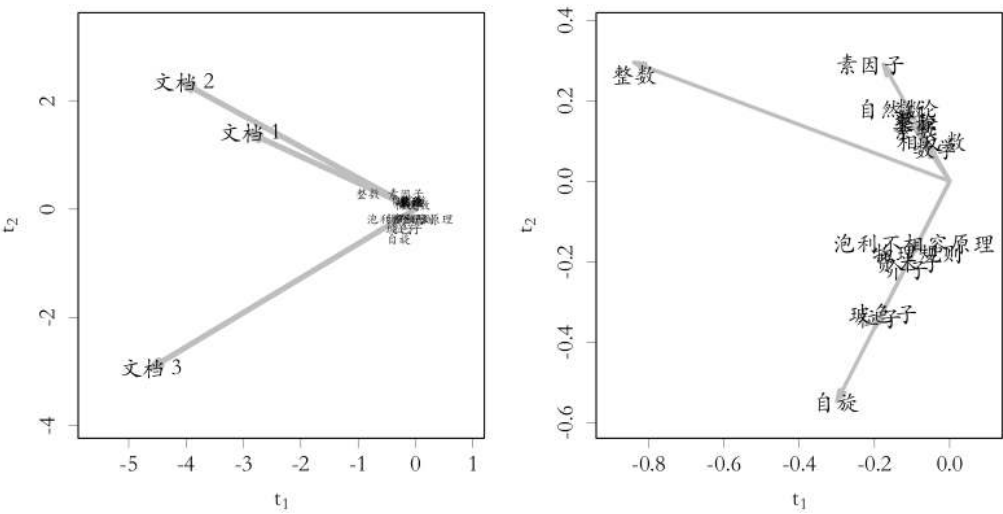


图 7.27 例 7.9 中文档和术语的二维表示

例 7.10 图像数据的维数一般都很高。例如，日本女性面部表情 (JAFPE) 数据库里每个图片的大小都是 256×256 像素，用一个 $d = 65536$ 维的向量来表示。每位女性的面部有七种表情，如图 7.28 所示，按列依次是愤怒、厌恶、恐惧、快乐、中性、悲伤和惊讶，每种表情 3 个样本，共有 $n = 21$ 个样本。

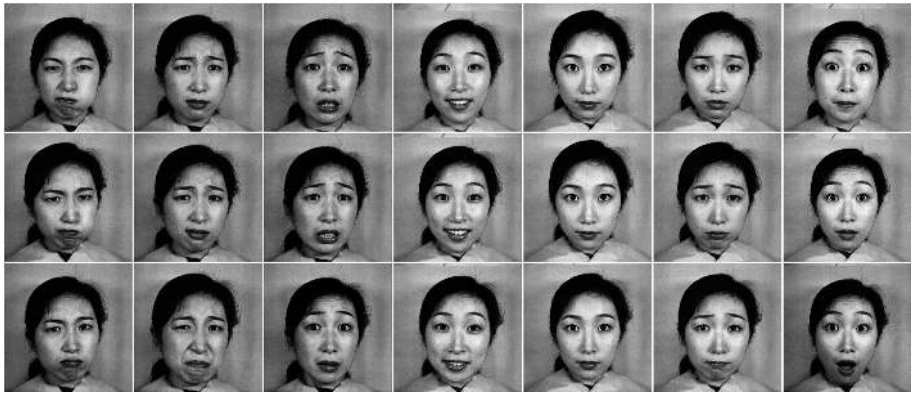
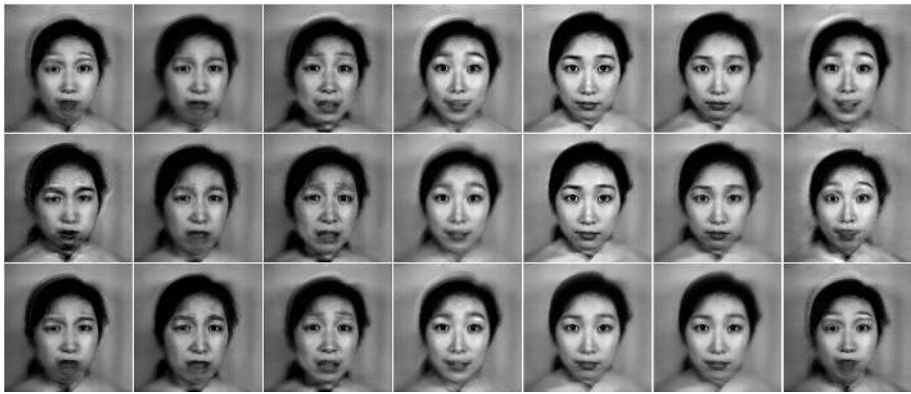


图 7.28 面部的七种表情



是样本均值，即平均表情。先利用算法 7.2（或算法 7.3）将每个图片压缩为一个 $k = 3$ 维向量，再利用性质 7.7（或性质 7.8）重构原始数据，效果见图 7.29。性质 7.7 和具体实验都显示： k 越大，近似效果越好。



(a) 利用样本均值和 3 维的 PCA 数据表示得到原始数据的近似



(b) PCA 数据表示的误差（为清楚地显示，采用反转片）

图 7.29 基于 PCA 的压缩与重构

2. 核主成分分析

对于给定的核函数 $\kappa(\mathbf{x}, \mathbf{y})$, 样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 经过特征映射变为 $\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_n)$, 简记作 $\varphi_1, \dots, \varphi_n$ 。先不妨设 $\Phi = (\varphi_1, \dots, \varphi_n)$ 是中心化了的, 即其均值为 $\mathbf{0}$, 稍后我们再来处理如何中心化的问题。

在特征空间里, 样本 $\varphi_1, \dots, \varphi_n$ 的维数一般很高 (甚至是无穷维)。仿照算法 7.3, 令 \mathbf{u} 是散布矩阵 $S_\varphi = \Phi\Phi^\top$ 的单位本征向量, 不难验证 \mathbf{u} 可由 $\varphi_1, \dots, \varphi_n$ 线性表出, 即 \mathbf{u} 落在 $\varphi_1, \dots, \varphi_n$ 张成的空间里。不妨设

$$\mathbf{u} = \Phi\mathbf{v} = \sum_{i=1}^n v_i \varphi_i, \text{ 其中 } \mathbf{v} = (v_1, \dots, v_n)^\top \text{ 使得 } \|\mathbf{u}\|_2 = 1 \quad (7.29)$$

用 Φ^\top 左乘 $S_\varphi \mathbf{u} = \lambda \mathbf{u}$ 两侧, 可得

$$\Phi^\top \Phi \Phi^\top \Phi \mathbf{v} = \lambda \Phi^\top \Phi \mathbf{v} \quad (7.30)$$

式 (7.30) 中, $\mathbf{K} = \Phi^\top \Phi$ 是样本的核矩阵。经过整理, 式 (7.30) 变为

$$\mathbf{K}^2 \mathbf{v} = \lambda \mathbf{K} \mathbf{v}$$

为解上述方程组, 我们求下面的非零本征值问题。

$$\mathbf{K} \mathbf{v} = \lambda \mathbf{v}, \text{ 其中 } \lambda \text{ 是 } \mathbf{K} \text{ 的本征值} \quad (7.31)$$

其中, 式 (7.29) 里的条件要求 \mathbf{v} 满足

$$\mathbf{u}^\top \mathbf{u} = \mathbf{v}^\top \Phi^\top \Phi \mathbf{v} = \mathbf{v}^\top \mathbf{K} \mathbf{v} = 1$$

结合式 (7.31), 待定系数 \mathbf{v} 满足

$$\|\mathbf{v}\|_2 = \frac{1}{\sqrt{\lambda}} \quad (7.32)$$

显然, $\Phi^\top \mathbf{u} = \mathbf{K} \mathbf{v}$ 是样本 $\varphi_1, \dots, \varphi_n$ 在单位向量 \mathbf{u} 上的投影坐标。于是, 样本 $\varphi_1, \dots, \varphi_n$ 的主成分分析就变成求核矩阵 \mathbf{K} 的满足条件 (7.32) 本征向量 \mathbf{v} 这一线性代数问题, 简称核主成分分析 (kernel PCA)。

算法 7.4 (核主成分分析) 给定样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 和核函数 $\kappa(\cdot, \cdot)$, 不妨设 φ 是特征映射。在特征空间里, 假设样本 $\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_n)$ 是中心化了的。

- (1) 定义核矩阵 $\mathbf{K}_{n \times n} = (k_{ij})$, 其中 $k_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ 。
- (2) 给出核矩阵 \mathbf{K} 的谱分解 $\mathbf{K} = \mathbf{V} \Lambda \mathbf{V}^\top$, 其中 $\Lambda = \text{diag}(\lambda_1^*, \dots, \lambda_n^*)$ 是 \mathbf{K} 的本征值按降序构成的对角阵, $\mathbf{V}_{n \times n} = (\mathbf{v}_1, \dots, \mathbf{v}_n)$ 是正交矩阵。

(3) 在特征空间里, 样本 $\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_n)$ 由前 k 个主成分给出的数据表示是

$$(\mathbf{y}_1, \dots, \mathbf{y}_n)_{k \times n} = \begin{pmatrix} \mathbf{v}_1^\top / \sqrt{\lambda_1^*} \\ \vdots \\ \mathbf{v}_k^\top / \sqrt{\lambda_k^*} \end{pmatrix} \mathbf{K} = \begin{pmatrix} \sqrt{\lambda_1^*} \mathbf{v}_1^\top \\ \vdots \\ \sqrt{\lambda_k^*} \mathbf{v}_k^\top \end{pmatrix}$$

上述算法简直就是算法 7.3 的翻版。值得关注的是, 上述解法虽然借助了样本散布矩阵 \mathbf{S} 的本征向量, 但并未求其显式表达, 而是巧妙地绕过了 \mathbf{S} 的谱分解, 直奔主题——样本在主成分上的投影坐标。之所以能这样做, 就是利用“核技巧”抹掉了所有显式的 φ , 用核函数 κ 取而代之。其中, 式 (7.29) 和式 (7.30) 是关键。

要比较主成分分析和核主成分分析的效果, 常用 (中心化的) 解释矩阵与它的秩为 k 的埃卡特-杨近似之间的误差的 F -范数之比来衡量, 具体为

$$\frac{\|\Phi - \hat{\Phi}\|_F}{\|\mathbf{Z} - \hat{\mathbf{Z}}\|_F} = \frac{\sqrt{\sum_{i=k+1}^n \lambda_i^*}}{\sqrt{\sum_{i=k+1}^n \lambda_i}}$$

最后, 我们来解决数据在特征空间中心化的问题。一般情况下, 我们无法得到数据的显式表示, 但我们只需要利用“核技巧”搞清楚中心化后的数据之间的内积, 就可以得到核矩阵。

性质 7.9 (核矩阵) 令 $\mathbf{K}_{n \times n} = (k_{ij}) = (\kappa(\mathbf{x}_i, \mathbf{x}_j))$ 是未中心化的样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 的核矩阵。

□ 为在特征空间里将样本中心化, 定义新的特征映射

$$\tilde{\varphi}(\mathbf{x}) = \varphi(\mathbf{x}) - \frac{1}{n} \sum_{k=1}^n \varphi(\mathbf{x}_k)$$

显然, $\tilde{\varphi}(\mathbf{x}_1), \dots, \tilde{\varphi}(\mathbf{x}_n)$ 是中心化的。该特征映射定义的核函数为

$$\begin{aligned} \tilde{\kappa}(\mathbf{x}, \mathbf{y}) &= \langle \tilde{\varphi}(\mathbf{x}), \tilde{\varphi}(\mathbf{y}) \rangle \\ &= \kappa(\mathbf{x}, \mathbf{y}) - \frac{1}{n} \sum_{k=1}^n \kappa(\mathbf{x}_k, \mathbf{x}) - \frac{1}{n} \sum_{k=1}^n \kappa(\mathbf{x}_k, \mathbf{y}) + \frac{1}{n^2} \sum_{s,t=1}^n \kappa(\mathbf{x}_s, \mathbf{x}_t) \end{aligned}$$

□ 设核函数 $\tilde{\kappa}$ 在样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 上所对应的核矩阵为 $\tilde{\mathbf{K}}_{n \times n} = (\tilde{k}_{ij})$, 其中

$$\begin{aligned} \tilde{k}_{ij} &= \tilde{\kappa}(\mathbf{x}_i, \mathbf{x}_j) \\ &= k_{ij} - \frac{1}{n} \sum_{k=1}^n k_{ki} - \frac{1}{n} \sum_{k=1}^n k_{kj} + \frac{1}{n^2} \sum_{s,t=1}^n k_{st} \end{aligned} \quad (7.33)$$

不难看出, 式 (7.33) 中的 $\frac{1}{n} \sum_{k=1}^n k_{ki}$, $\frac{1}{n} \sum_{k=1}^n k_{kj}$ 和 $\frac{1}{n^2} \sum_{s,t=1}^n k_{st}$ 分别是核矩阵 \mathbf{K} 的第 i 列的均值, 第 j 列的均值和整个核矩阵 \mathbf{K} 的均值。

7.2.2 因子分析

随机向量 $\mathbf{X} = (X_1, \dots, X_d)^\top$ 各分量之间的关系早就写在联合分布之中了，推导这些关系靠的是纯粹数学，答案总是唯一的。但当对总体 \mathbf{X} 的分布一无所知的时候，通过样本来研究 X_1, \dots, X_d 的关系就是统计学的任务了。另外，测量并非易事，世间充满了不可测量的事物。因子分析利用可测量的变量来解释潜在因子的语义，如智力、消费者态度等，有助于探索事物的深层规律。

20 世纪初，英国心理学家、统计学家查尔斯·斯皮尔曼 (Charles Spearman, 1863—1945) 在分析学生各科成绩的时候发现了统计相关性，即某科成绩好，其他科也不赖。为此，斯皮尔曼 (图 7.30) 假想存在一个潜在的 (即不可直接测量的) 一般智力 (general intelligence) 因子影响着学生的成绩。1904 年，斯皮尔曼发表因子分析 (factor analysis) 的首篇论文。如今，因子分析已发展成多元统计的经典方法，它常用来分析 d 个随机变量之间的相关关系，目的是为了将这些变量分为 $k < d$ 组 (每组称为一个因子)，使得组内的变量之间是高度相关的^[106]。我们可以把因子视为变量的聚类，每个因子具有潜在的语义，可以由背景知识给出合理的解释。



图 7.30 斯皮尔曼



图 7.31 因子分析起源于儿童心理学的研究

因子分析的方法不唯一，本节只介绍其中比较常用的三种方法：① 基于主成分、② 基于最大似然估计，③ 基于最小二乘估计的因子分析。其他因子分析的方法可参阅 [106, 119–122]。

1. 基于主成分的因子分析

定义 7.9 因子分析的一般模型是考虑均值为 $\mathbf{0}$ 的随机向量 $\mathbf{X} = (X_1, \dots, X_d)^\top$ 是否可能被少数几个无关的随机变量线性表出，即

$$\mathbf{X} = \mathbf{L}_{d \times k} \mathbf{F} + \boldsymbol{\epsilon}, \text{ 其中 } k < d \quad (7.34)$$

在式 (7.34) 中，随机向量 $\mathbf{F} = (F_1, \dots, F_k)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_d)^\top$ 的各分量之间不相关，并且满足

$$\begin{aligned} \mathbf{E}(\mathbf{F}) &= \mathbf{0} & \text{Cov}(\mathbf{F}) &= \mathbf{I}_k \\ \mathbf{E}(\boldsymbol{\epsilon}) &= \mathbf{0} & \text{Cov}(\boldsymbol{\epsilon}) &= \boldsymbol{\Psi}, \text{ 其中 } \boldsymbol{\Psi} = \text{diag}(\psi_1, \dots, \psi_d) \end{aligned}$$

矩阵 $L_{d \times k}$ 是未知的, 称为因子载荷矩阵 (factor loading matrix)。 k 维随机向量 F 被称为公共因子 (common factor), 它和特殊因子 (unique factor) ϵ 都是不可直接测量的。特殊方差矩阵 Ψ 也是未知的。

公共因子 F 的直观含义是它的分量是独立的随机变量, 使得观测数据 X 按照 (7.34) 用 F 线性表示出来 (图 7.32), 样本协方差矩阵 (或样本相关系数矩阵) 与原有的样本协方差矩阵 (或样本相关系数矩阵) 尽可能地接近。

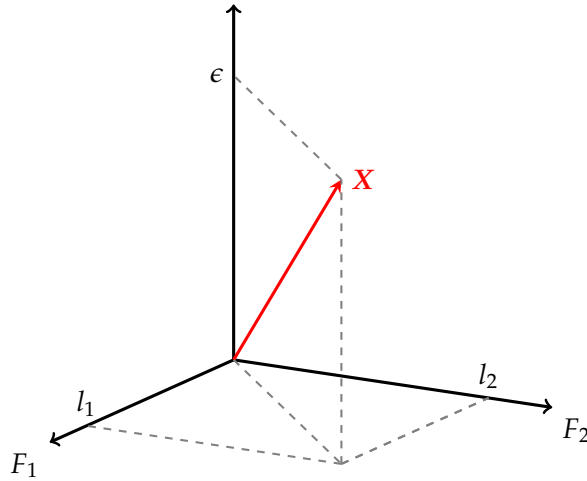


图 7.32 数据 X 由公共因子 F 线性表示

令 Σ 是 X 的协方差矩阵, 则

$$\begin{aligned}
 \Sigma &= E(XX^T) \\
 &= E[(LF + \epsilon)(LF + \epsilon)^T] \\
 &= LE(FF^T)L + LE(F\epsilon^T) + E(\epsilon\epsilon^T) \\
 &= LL^T + \Psi
 \end{aligned} \tag{7.35}$$

另外,

$$\begin{aligned}
 \text{Cov}(\epsilon, F) &= E(\epsilon F^T) = 0 \\
 \text{Cov}(X, F) &= E(XF^T) = LE(FF^T) + E(\epsilon F^T) = L
 \end{aligned}$$

显然, 满足条件 (7.35) 的 L, Ψ 不唯一。譬如, 若 L 满足 (7.35), 则 $\tilde{L} = LU$ 也满足 (7.35), 其中 U 是任意正交矩阵。并且, $\tilde{F} = U^T F$ 满足

$$\begin{aligned}
 E(\tilde{F}) &= U^T E(F) = 0 \\
 \text{Cov}(\tilde{F}) &= U^T \text{Cov}(F)U = U^T U = I
 \end{aligned}$$

从 X 的观测结果, 无法区分载荷矩阵 L 和 \tilde{L} , 并且因子 F 和 \tilde{F} 具有相同的统计性质。所以, 载荷矩阵只需精确到相差一个正交变换即可。

算法 7.5 样本 $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$ 的协方差矩阵 $\mathbf{C}_{d \times d} = \frac{1}{n-1} \mathbf{Z} \mathbf{Z}^\top$ 是 $\mathbf{\Sigma}$ 的无偏估计。

(1) 利用对称矩阵 \mathbf{C} 的谱分解 (奇异值分解的特例)

$$\mathbf{C} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top, \text{ 其中 } r = \text{rank}(\mathbf{\Lambda})$$

由埃卡特-杨定理可知, 在所有秩为 $k \leq r$ 的 $d \times k$ 矩阵之中,

$$\hat{\mathbf{L}} = (\sqrt{\lambda_1} \mathbf{u}_1, \dots, \sqrt{\lambda_k} \mathbf{u}_k) \text{ 使得 } \hat{\mathbf{L}} \hat{\mathbf{L}}^\top \text{ 是 } \mathbf{C} \text{ 的最佳逼近}$$

(2) 矩阵 $\mathbf{C} - \hat{\mathbf{L}} \hat{\mathbf{L}}^\top$ 对角线上的元素构成的对角阵 $\hat{\mathbf{\Psi}} = \text{diag}(\hat{\psi}_1, \dots, \hat{\psi}_d)$ 是对 $\mathbf{\Psi}$ 的估计。残差矩阵 $\mathbf{C} - (\hat{\mathbf{L}} \hat{\mathbf{L}}^\top + \hat{\mathbf{\Psi}})$ 中所有元素的平方和具有如下上界,

$$\|\mathbf{C} - (\hat{\mathbf{L}} \hat{\mathbf{L}}^\top + \hat{\mathbf{\Psi}})\|_F^2 \leq \sum_{j=k+1}^r \lambda_j$$

(3) 因子个数 k 的选择, 一般遵循如下规则: 选择 k 使得以下比例刚好不超过给定的阈值 t 。一般讲, k 越小, $\hat{\mathbf{L}} \hat{\mathbf{L}}^\top + \hat{\mathbf{\Psi}}$ 拟合 \mathbf{C} 的效果就越差, 此时的因子分析是没有意义的。

$$R(k) = \frac{\lambda_{k+1} + \dots + \lambda_r}{\lambda_1 + \dots + \lambda_r} \leq t \quad (7.36)$$

然而在实践中, 为了消除不同变量的均值和量纲的影响, 人们经常使用标准化的样本 (定义 7.3)。根据性质 7.1, 其协方差矩阵就是相关系数矩阵。所以, 在上述解法中, 将样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 的协方差矩阵 \mathbf{C} 替换为样本相关系数矩阵 (定义 7.4) 即可。

例 7.11 问卷调查是伦敦统计学会 (Statistical Society of London), 即皇家统计学会 (图 7.33) 于 1838 年发明的一种抽样调查 (sampling survey) 方法, 它通过一系列设计好的问题采集被访者的意见、反应、感受等, 进而推断整体人群的想法。例如, 民意测验、舆情分析等。调查问卷 (图 7.33) 的设计需要在预设模型的基础上, 尽可能全面地收集信息以达成统计分析之目的。同时, 在伦理标准的指导下, 避免触碰和泄露个人隐私, 不能引起被访者的反感和抵触。譬如, 姓名、年龄如果不在模型考虑之内就无需询问, 收入按设定好的区间调查等。



图 7.33 皇家统计学会与调查问卷 (questionnaire)

例如，影响是否购买某电子产品的几个潜在因素是价格 (price)、配套软件 (software)、外观 (looks)、品牌 (brand)、朋友建议 (friend)、家庭适用 (home)。针对这几个变量，我们设计调查问卷如下（用从 1 到 10 的打分来刻画程度）。

- ☐ 价格 (P) 吸引你的程度？
- ☐ 对品牌 (B) 的满意度？
- ☐ 对配套软件 (S) 的满意度？
- ☐ 朋友 (F) 建议影响你的程度？
- ☐ 对外观 (L) 的满意度？
- ☐ 适用于整个家庭 (H) 的程度？

对随机遇到的 17 个消费者进行问卷调查，打分的结果见表 7.3。

表 7.3 例 7.11 的问卷调查结果

变量	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P	7	2	8	7	10	1	2	9	4	8	10	2	1	7	8	3	4
S	10	4	5	10	4	4	5	6	7	5	10	8	9	10	7	7	2
L	10	9	5	2	2	10	7	5	8	2	1	8	7	1	4	8	1
B	8	10	7	1	2	6	8	5	10	3	4	10	7	4	5	7	9
F	9	7	6	1	2	2	8	10	4	5	8	8	5	9	10	2	7
H	10	8	5	5	3	5	2	7	5	7	10	7	7	10	8	4	10

这些变量之间有怎样的关系？我们可以通过样本相关系数来粗略理解。在表 7.4 所示的样本相关系数矩阵 ρ 中，不难发现：价格 (P) 和外观 (L)、品牌 (B) 都是负相关，意味着对外观和品牌越满意，对价格越不满意（可能是因为价格太贵）。朋友建议 (F) 和家庭适用 (H) 是正相关的，意味着朋友建议可能很多都是考虑到家庭适用的经验之谈。

表 7.4 例 7.11 中样本相关系数矩阵

	P	S	L	B	F	H
P	1.0000	0.1965	-0.6636	-0.6735	0.1981	0.2138
S	0.1965	1.0000	-0.0388	-0.2401	0.1588	0.3196
L	-0.6636	-0.0388	1.0000	0.6617	-0.0587	-0.2485
B	-0.6735	-0.2401	0.6617	1.0000	0.2802	0.0915
F	0.1981	0.1588	-0.0587	0.2802	1.0000	0.5996
H	0.2138	0.3196	-0.2485	0.0915	0.5996	1.0000

对矩阵 ρ 进行谱分解，得到 $\rho = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ ，其中 $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_r)$ ，满足 $\lambda_1 \geq \lambda_2 \leq \dots \geq \lambda_r > 0$ 。具体为

$$\mathbf{\Lambda} = \text{diag}(2.4597, 1.7609, 0.9330, 0.4508, 0.2417, 0.1538)$$

给定阈值 $t = 0.15$ ，由算法 7.5 求得 $R(k), k = 1, 2, \dots, 5$ ，不难发现首个使得不等式 (7.36) 成立的是 $k = 3$ ，于是因子个数选为 3。进而，得到 $\hat{\mathbf{L}}_{6 \times 3}$ 如表 7.5 所示。

因子 1 与产品本身的价格 (P)、外观 (L)、品牌 (B) 有关。因子 2 与朋友建议 (F) 和家庭适用 (H) 有关，是一些外在因素。因子 3 是配套软件 (S)。这三个因子的重要性依次从高至低。

表 7.5 利用主成分得到例 7.11 的因子分析结果

	因子 1	因子 2	因子 3
P	-0.8848	0.0701	-0.1323
S	-0.3793	-0.3236	0.8553
L	0.8452	-0.1657	0.3056
B	0.7901	-0.5136	-0.1555
F	-0.2042	-0.8526	-0.2526
H	-0.3909	-0.7957	-0.0521

由 \hat{L} 和 $\hat{\Psi} = \text{diag}(0.1948, 0.0199, 0.1648, 0.0878, 0.1676, 0.2114)$, 我们得到例 7.11 中残差矩阵的 F -范数为

$$\|\rho - (\hat{L}\hat{L}^\top + \hat{\Psi})\|_F \approx 0.3733$$

接下来, 分别介绍基于最大似然估计和基于最小二乘估计的因子分析, 将它们应用于表 7.3 的问卷调查数据上, 结果分别见表 7.6 和表 7.7。

2. 基于最大似然估计的因子分析

假设 $F \sim N_k(\mathbf{0}, I_k)$ 与 $\epsilon \sim N_d(\mathbf{0}, \Psi)$ 相互独立, 其中 $\Psi = \text{diag}(\psi_1, \dots, \psi_d)$ 。由 $X = LF + \epsilon$ 得到

$$X \sim N_d(\mathbf{0}, LL^\top + \Psi)$$

将 F 视为隐藏变量, 则它与 X 的联合分布是正态分布

$$\begin{pmatrix} F \\ X \end{pmatrix} \sim N_{d+k} \left(\mathbf{0}, \begin{pmatrix} I_k & L^\top \\ L & LL^\top + \Psi \end{pmatrix} \right)$$

根据定理 7.1, 给定 $X = x$, 随机向量 F 的条件分布是

$$F|X = x \sim N_k(L^\top(LL^\top + \Psi)^{-1}x, I_k - L^\top(LL^\top + \Psi)^{-1}L)$$

显然,

$$\begin{aligned} E(F|X = x) &= L^\top(LL^\top + \Psi)^{-1}x \\ &= Ax, \text{ 其中 } A_{k \times d} = L^\top(LL^\top + \Psi)^{-1} \\ E(FF^\top|X = x) &= \text{Cov}(F|X = x) + E(F|X = x)[E(F|X = x)]^\top \\ &= I_k - AL + Axx^\top A^\top, \text{ 显然 } Axx^\top A^\top \text{ 是对称矩阵} \end{aligned}$$

给定观测样本 $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, 对数似然函数 (忽略掉常数) 是

$$\begin{aligned}\ell(\mathbf{L}, \boldsymbol{\Psi}) &= -\frac{n}{2} \ln |\boldsymbol{\Psi}| - \frac{1}{2} \sum_{j=1}^n (\mathbf{x}_j - \mathbf{L} \mathbf{f}_j)^\top \boldsymbol{\Psi}^{-1} (\mathbf{x}_j - \mathbf{L} \mathbf{f}_j) \\ &= -\frac{n}{2} \ln |\boldsymbol{\Psi}| - \frac{1}{2} \sum_{j=1}^n (\mathbf{x}_j^\top \boldsymbol{\Psi}^{-1} \mathbf{x}_j - 2 \mathbf{x}_j^\top \boldsymbol{\Psi}^{-1} \mathbf{L} \mathbf{f}_j + \mathbf{f}_j^\top \mathbf{L}^\top \boldsymbol{\Psi}^{-1} \mathbf{L} \mathbf{f}_j) \\ &= -\frac{n}{2} \ln |\boldsymbol{\Psi}| - \frac{1}{2} \sum_{j=1}^n [\mathbf{x}_j^\top \boldsymbol{\Psi}^{-1} \mathbf{x}_j - 2 \mathbf{x}_j^\top \boldsymbol{\Psi}^{-1} \mathbf{L} \mathbf{f}_j + \text{tr}(\mathbf{L}^\top \boldsymbol{\Psi}^{-1} \mathbf{L} \mathbf{f}_j \mathbf{f}_j^\top)]\end{aligned}$$

上式中 \mathbf{f}_j 和 $\mathbf{f}_j \mathbf{f}_j^\top$ 是未知的, 我们分别用 $\mathbf{E}(\mathbf{F}|\mathbf{x}_j)$ 和 $\mathbf{E}(\mathbf{F}\mathbf{F}^\top|\mathbf{x}_j)$ 来替换它们, 则上式变为

$$Q = -\frac{n}{2} \ln |\boldsymbol{\Psi}| - \frac{1}{2} \sum_{j=1}^n [\mathbf{x}_j^\top \boldsymbol{\Psi}^{-1} \mathbf{x}_j - 2 \mathbf{x}_j^\top \boldsymbol{\Psi}^{-1} \mathbf{L} \mathbf{E}(\mathbf{F}|\mathbf{x}_j) + \text{tr}(\mathbf{L}^\top \boldsymbol{\Psi}^{-1} \mathbf{L} \mathbf{E}(\mathbf{F}\mathbf{F}^\top|\mathbf{x}_j))]$$

于是,

$$\frac{\partial Q}{\partial \mathbf{L}} = - \sum_{j=1}^n \boldsymbol{\Psi}^{-1} \mathbf{x}_j [\mathbf{E}(\mathbf{F}|\mathbf{x}_j)]^\top + \sum_{j=1}^n \boldsymbol{\Psi}^{-1} \mathbf{L} \mathbf{E}(\mathbf{F}\mathbf{F}^\top|\mathbf{x}_j)$$

令上面的梯度矩阵为零矩阵, 我们得到:

$$\mathbf{L} \sum_{j=1}^n \mathbf{E}(\mathbf{F}\mathbf{F}^\top|\mathbf{x}_j) = \sum_{j=1}^n \mathbf{x}_j [\mathbf{E}(\mathbf{F}|\mathbf{x}_j)]^\top \quad (7.37)$$

类似地,

$$\frac{\partial Q}{\partial \boldsymbol{\Psi}^{-1}} = \frac{n}{2} \boldsymbol{\Psi} - \frac{1}{2} \sum_{j=1}^n [\mathbf{x}_j \mathbf{x}_j^\top - 2 \mathbf{L} \mathbf{E}(\mathbf{F}|\mathbf{x}_j) \mathbf{x}_j^\top + \mathbf{L} \mathbf{E}(\mathbf{F}\mathbf{F}^\top|\mathbf{x}_j) \mathbf{L}^\top]$$

将式 (7.37) 代入上式, 得到

$$\frac{\partial Q}{\partial \boldsymbol{\Psi}^{-1}} = \frac{n}{2} \boldsymbol{\Psi} - \frac{1}{2} \sum_{j=1}^n [\mathbf{x}_j \mathbf{x}_j^\top - \mathbf{L} \mathbf{E}(\mathbf{F}|\mathbf{x}_j) \mathbf{x}_j^\top]$$

令上面的梯度矩阵为零矩阵, 我们得到:

$$\boldsymbol{\Psi} = \frac{1}{n} \sum_{j=1}^n [\mathbf{x}_j \mathbf{x}_j^\top - \mathbf{L} \mathbf{E}(\mathbf{F}|\mathbf{x}_j) \mathbf{x}_j^\top] \quad (7.38)$$

直接求 $\mathbf{L}, \boldsymbol{\Psi}$ 的最大似然估计的显式解很困难, 然而用数值逼近的方法估计 $\mathbf{L}, \boldsymbol{\Psi}$ 却是可行的。1982 年, 美国统计学家唐纳德·鲁宾 (Donald Rubin, 1943—) 给出了下述算法, 其理论基础是期望最大化算法^[123], 详见第 8 章。

算法 7.6 设样本是经过标准化的。初始化 $\Psi_{(0)}$ 和 $L_{(0)}$, 使得矩阵 $L_{(0)}L_{(0)}^\top + \Psi_{(0)}$ 可逆。

□ 按照当前的参数 L 和 Ψ , 令

$$A_{(t)} = L_{(t)}^\top (L_{(t)}L_{(t)}^\top + \Psi_{(t)})^{-1}$$

计算 F 和 FF^\top 的条件期望如下,

$$\begin{aligned} E(F|x_j) &= A_{(t)}x_j \\ E(FF^\top|X=x_j) &= I_k - A_{(t)}L_{(t)} + A_{(t)}x_jx_j^\top A_{(t)}^\top \end{aligned}$$

□ 令 S 是散布矩阵。由结果 (7.37) 和 (7.38), 更新 L, Ψ 如下,

$$\begin{aligned} L_{(t+1)} &= \left[\sum_{j=1}^n x_j [E(F|x_j)]^\top \right] \left[\sum_{j=1}^n E(FF^\top|x_j) \right]^{-1} \\ &= SA_{(t)}^\top [n(I_k - A_{(t)}L_{(t)}) + A_{(t)}SA_{(t)}^\top]^{-1} \\ \Psi_{(t+1)} &= \frac{1}{n} \text{diag} \left[\sum_{j=1}^n x_jx_j^\top - L_{(t)}E(F|x_j)x_j^\top \right] \\ &= \frac{1}{n} \text{diag}[(I_d - L_{(t)}A_{(t)})S] \end{aligned}$$

因为对角阵 $\hat{\Psi}$ 中的元素必须是非负的, 所以需要将其对角线元素更新为各自的绝对值。

□ 重复上述两个步骤, 直至达到预定的收敛标准。

例 7.12 利用算法 7.6 求得例 7.11 的因子分析结果 (表 7.6), 其结论与表 7.5 的相似。

表 7.6 利用最大似然估计得到例 7.11 的因子分析结果

	因子 1	因子 2	因子 3
P	-0.7967	0.02072	-0.1303
S	-0.3757	-0.35379	0.8215
L	0.7416	-0.11674	0.2445
B	0.8024	-0.50428	-0.1253
F	-0.1742	-0.74628	-0.2191
H	-0.3199	-0.66265	-0.0655

3. 基于最小二乘估计的因子分析

对问题 (7.34) 来说, 估计 L, Ψ 的方法并不唯一。除了上述基于主成分和最大似然估计的方法, 还有最小化以下目标函数的最小二乘估计^[106]。

$$f_{\text{LSE}}(L, \Psi) = \frac{1}{2} \text{tr}[C - (LL^\top + \Psi)]^2, \text{ 其中 } C \text{ 是样本协方差矩阵}$$

将上式右侧展开, 得到

$$\begin{aligned}
 f_{\text{LSE}}(\mathbf{L}, \mathbf{\Psi}) &= \frac{1}{2} \text{tr}(\mathbf{C}^2 - \mathbf{C}\mathbf{L}\mathbf{L}^\top - \mathbf{C}\mathbf{\Psi} \\
 &\quad - \mathbf{L}\mathbf{L}^\top \mathbf{C} + \mathbf{L}\mathbf{L}^\top \mathbf{L}\mathbf{L}^\top + \mathbf{L}\mathbf{L}^\top \mathbf{\Psi} \\
 &\quad - \mathbf{\Psi}\mathbf{C} + \mathbf{\Psi}\mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}^2) \\
 &= \frac{1}{2} \text{tr}(\mathbf{C}^2) + \frac{1}{2} \text{tr}(\mathbf{L}\mathbf{L}^\top \mathbf{L}\mathbf{L}^\top) + \frac{1}{2} \text{tr}(\mathbf{\Psi}^2) \\
 &\quad - \text{tr}(\mathbf{C}\mathbf{L}\mathbf{L}^\top) - \text{tr}(\mathbf{C}\mathbf{\Psi}) + \text{tr}(\mathbf{L}\mathbf{L}^\top \mathbf{\Psi})
 \end{aligned}$$

利用梯度矩阵 (见《随机之美》^[8] 的附录), 不难得到

$$\begin{aligned}
 \frac{\partial f}{\partial \mathbf{L}} &= \frac{1}{2} \frac{\partial \text{tr}(\mathbf{C}^2)}{\partial \mathbf{L}} + \frac{1}{2} \frac{\partial \text{tr}(\mathbf{L}\mathbf{L}^\top \mathbf{L}\mathbf{L}^\top)}{\partial \mathbf{L}} + \frac{1}{2} \frac{\partial \text{tr}(\mathbf{\Psi}^2)}{\partial \mathbf{L}} \\
 &\quad - \frac{\partial \text{tr}(\mathbf{C}\mathbf{L}\mathbf{L}^\top)}{\partial \mathbf{L}} - \frac{\partial \text{tr}(\mathbf{C}\mathbf{\Psi})}{\partial \mathbf{L}} + \frac{\partial \text{tr}(\mathbf{L}\mathbf{L}^\top \mathbf{\Psi})}{\partial \mathbf{L}} \\
 &= \mathbf{O} + 2\mathbf{L}\mathbf{L}^\top \mathbf{L} + \mathbf{O} - 2\mathbf{C}\mathbf{L} + \mathbf{O} + 2\mathbf{\Psi}\mathbf{L} \\
 &= 2\mathbf{L}\mathbf{L}^\top \mathbf{L} - 2\mathbf{C}\mathbf{L} + 2\mathbf{\Psi}\mathbf{L}
 \end{aligned}$$

令上述梯度矩阵为零矩阵, 我们得到

$$\mathbf{L}\mathbf{L}^\top \mathbf{L} = (\mathbf{C} - \mathbf{\Psi})\mathbf{L}$$

将上式两边同时右乘 \mathbf{L}^\top , 因为 $\mathbf{L}\mathbf{L}^\top$ 是正定的, 所以总有

$$\mathbf{L}\mathbf{L}^\top = \mathbf{C} - \mathbf{\Psi}$$

类似地,

$$\frac{\partial f}{\partial \mathbf{\Psi}} = \mathbf{\Psi} - \mathbf{C} + \mathbf{L}\mathbf{L}^\top$$

于是,

$$\mathbf{\Psi} = \text{diag}(\mathbf{C} - \mathbf{L}\mathbf{L}^\top)$$

算法 7.7 假设样本是经过标准化的。初始化 $\hat{\mathbf{\Psi}}$, 譬如, $\hat{\mathbf{\Psi}} = [\text{diag}(\mathbf{C}^{-1})]^{-1}$, 或者 $\hat{\mathbf{\Psi}}$ 为均匀分布 $U(0, 1)$ 的 d 个随机数构成的对角阵。

□ 令 $\mathbf{A} = \mathbf{C} - \hat{\mathbf{\Psi}}$ 。显然, \mathbf{A} 是对称矩阵。设 \mathbf{A} 的本征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ 所对应的本征向量是 $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_d$ 。记

$$\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_d)$$

$$V = (v_1, v_2, \dots, v_d)$$

对称矩阵 A 具有谱分解 $A = V\Lambda V^T$ 。令

$$\hat{L} = V_k \Lambda_k^{\frac{1}{2}}$$

其中,

$$\Lambda_k = \text{diag}(\lambda_1, \dots, \lambda_k)$$

$$V_k = (v_1, v_2, \dots, v_k)$$

这一步骤与算法 7.5 的第一步是类似的。显然, $\hat{L}\hat{L}^T$ 是对 A 的近似。

□ 设当前对 L 的估计是 \hat{L} 。令

$$\hat{\Psi} = \text{diag}(C - \hat{L}\hat{L}^T)$$

与算法 7.6 类似, 需要将 $\hat{\Psi}$ 的对角线元素更新为其绝对值。

□ 重复上述两个步骤, 直至达到预定的收敛标准。例如, $\hat{\Psi}$ 中元素的最大变化不超过 0.01。

例 7.13 利用算法 7.7 求得例 7.11 的因子分析结果 (表 7.7), 其结论也与表 7.5 的相似。

表 7.7 利用最小二乘法得到例 7.11 的因子分析结果

	因子 1	因子 2	因子 3
P	-0.8028	-0.006056	0.1728
S	-0.4110	0.456305	-0.7898
L	0.7664	0.107173	-0.2798
B	0.8547	0.473126	0.1332
F	-0.1335	0.755712	0.2962
H	-0.2980	0.680853	0.1536

7.2.3 独立成分分析

在演唱会现场（图 7.34），麦克风的数量可以是一个，也可以是两个、三个……，每个麦克风都采集了乐器和歌者的声音（即观测数据）。音源和麦克风的个数是已知的，如何将这几个独立的音源分离出来？



图 7.34 演唱会现场

上述问题在信号处理里属于盲源分离 (blind source separation, BSS)，即从观测信号中将各个源信号分离出来（图 7.35）。这类问题有两个难点：一是源信号不能直接观测，二是信号模型是未知的，即我们并不了解源信号是如何混合的^[124–126]。譬如，如何分离在鸡尾酒会上同时说话的每个人的声音信号。



图 7.35 盲源分离

盲源分离是信号处理的难题之一，它可以一般化为：对于任意的随机向量 $\mathbf{X} = (X_1, \dots, X_d)^\top$ ，是否存在一个可逆变换 $\mathbf{A} = (a_{ij})_{d \times d}$ 使得随机向量 \mathbf{X} 可分解为 $\mathbf{X} = \mathbf{AS}$ ，其中，矩阵 \mathbf{A} 和随机向量 $\mathbf{S} = (S_1, \dots, S_d)^\top$ 都是未知的，然而 \mathbf{S} 的各个分量是相互独立的？

对于多元正态分布，它的主成分是相互独立的^[8]，毫无疑问上述分解是存在的。下面，介绍一种经典的盲源分离方法——独立成分分析 (independent component analysis, ICA)，利用最大似然估计寻找最优可能的分解

$$\mathbf{X} = \mathbf{AS}$$

设 $\mathbf{W} = (w_{ij})_{d \times d}$ 是 \mathbf{A} 的逆变换, 则 $\mathbf{S} = \mathbf{W}\mathbf{X}$, 满足分量独立的条件, 于是随机向量 \mathbf{S} 的协方差矩阵应为单位矩阵。给定观测数据 $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, 潜在的源数据是

$$\mathbf{s}_j = \mathbf{W}\mathbf{x}_j, \text{ 其中 } j = 1, \dots, n$$

令 \mathbf{C}' 是源数据 $\mathbf{s}_1, \dots, \mathbf{s}_n$ 的协方差矩阵, 它应为单位矩阵。令 \mathbf{C} 是样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 的协方差矩阵, 它与 \mathbf{C}' 的具体关系是

$$\begin{aligned} \mathbf{C} &= \frac{1}{n-1} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^\top \\ &= \frac{1}{n-1} \sum_{j=1}^n (\mathbf{A}\mathbf{s}_j)(\mathbf{A}\mathbf{s}_j)^\top \\ &= \mathbf{A} \left(\frac{1}{n-1} \sum_{j=1}^n \mathbf{s}_j \mathbf{s}_j^\top \right) \mathbf{A}^\top \\ &= \mathbf{A}\mathbf{C}'\mathbf{A}^\top \\ &= \mathbf{A}\mathbf{A}^\top \end{aligned}$$

设 \mathbf{A} 的奇异值分解是 $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (其几何含义见图 7.36), 代入上式, 得到

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$$

上式是半正定的对称矩阵 \mathbf{C} 的谱分解, 其中 $\mathbf{\Sigma}^2 = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ 是 \mathbf{C} 的本征值构成的对角阵, $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_d)$ 是相应本征向量构成的正交阵。我们只需要搞清楚 \mathbf{V} , 即可得到

$$\begin{aligned} \mathbf{W} &= \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top \\ &= \mathbf{V}\tilde{\mathbf{W}} \end{aligned}$$

独立成分分析试图重新发现数据的原始形态, 也就是求潜在变换 \mathbf{A} 的逆变换 \mathbf{W} 。虽然 \mathbf{A} 是未知的, 但它的奇异值分解的部分信息是可以从样本协方差矩阵中获取的, 即旋转 \mathbf{U}^\top 和拉伸 $\mathbf{\Sigma}^{-1}$ 。再经过一个旋转 \mathbf{V} , 源数据就出现了。

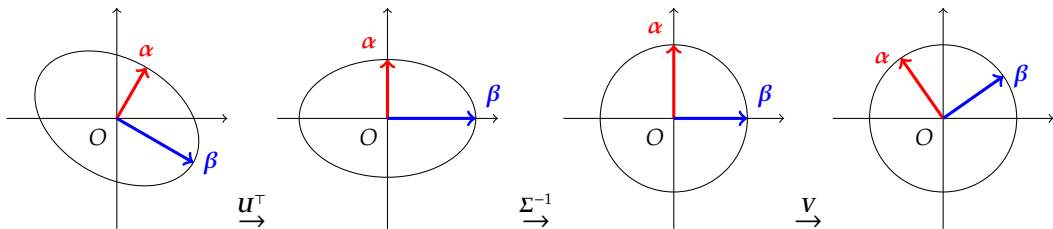


图 7.36 奇异值分解 $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ 的几何含义

我们称 $\tilde{\mathbf{W}} = \mathbf{\Sigma}^{-1}\mathbf{U}^\top$ 为白化矩阵 (whitening matrix), 称 $\mathbf{z}_j = \tilde{\mathbf{W}}\mathbf{x}_j, j = 1, \dots, n$ 为白化数据 (whitened

data)。从白化数据到源数据就差一个旋转变换 V 。从白化数据可以得到各个变量的边缘分布，如果这些变量是独立的，其联合分布的抽样可以通过边缘分布的抽样得到，其形态应该与白化数据相差无几。遗憾的是，白化数据中的变量一般不满足独立性，见下面的例 7.14。

例 7.14 令源数据是 $(-1,1) \times (-1,1)$ 上均匀分布的随机数，经过变换 $A = \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix}$ 得到观测数据，见图 7.37(a)。从图 7.37(b) 不难看出，白化数据离源数据只有一步之遥。图 7.37(c) 是从白化数据的经验边缘分布抽样所得到的重构数据，与白化数据相距甚远，说明两个边缘分布并不独立（图 7.37）。

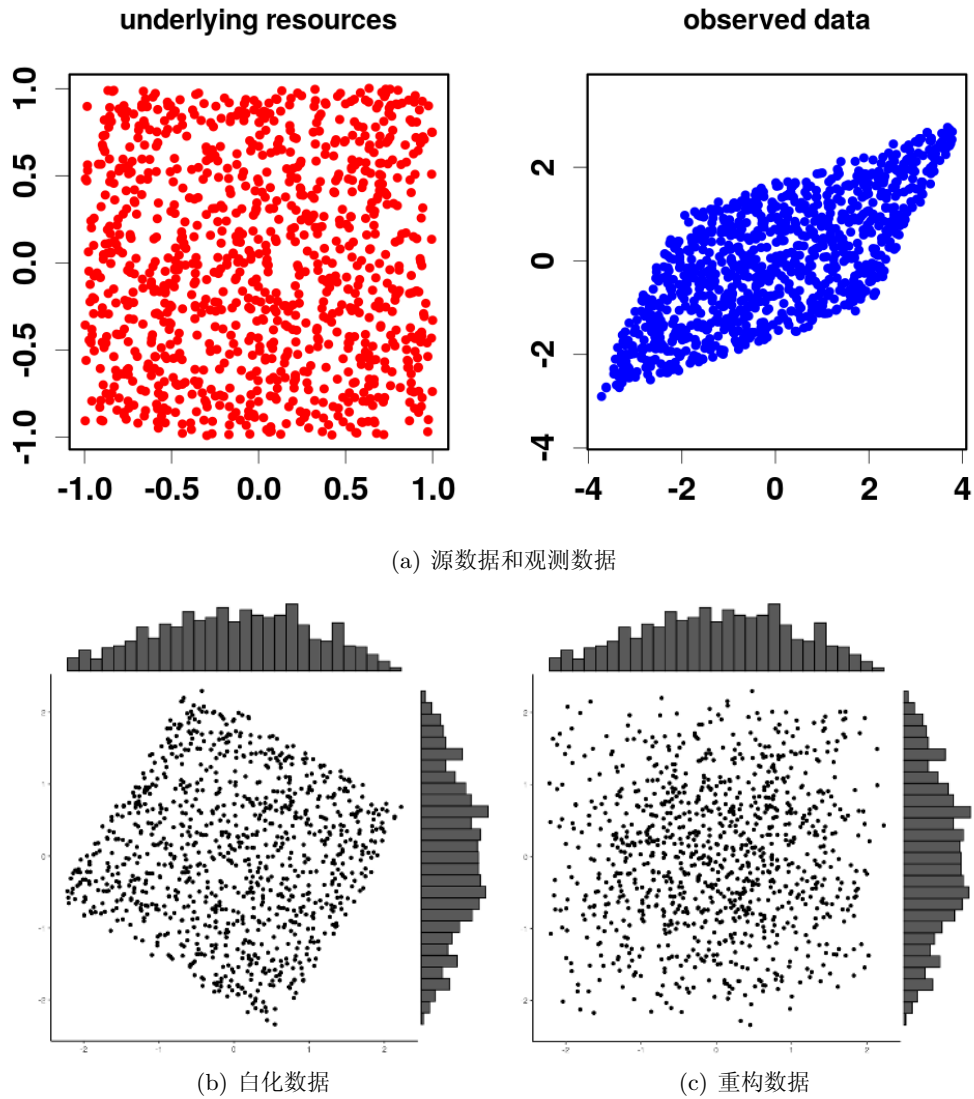


图 7.37 一个不成功的重构

假设源信号 S 不服从正态分布，否则白化数据做任何旋转都可以得到独立成分，因此解不唯一，独立成分分析没有任何意义。为了得到 V ，下面我们分两个步骤考虑它的最大似然估计问题，其中源数据的独立成分不是正态分布。

□ 白化数据的总体分布：若已知 \mathbf{S} 的密度函数是

$$f_{\mathbf{S}}(\mathbf{s}) = \prod_{i=1}^d f_{S_i}(s_i)$$

随机向量 \mathbf{S} 与 $\mathbf{Z} = \boldsymbol{\Sigma}^{-1} \mathbf{U}^{\top} \mathbf{X}$ 的关系是 $\mathbf{S} = \mathbf{VZ}$ ，或者等价地

$$\mathbf{Z} = \mathbf{V}^{\top} \mathbf{S}$$

因此， \mathbf{Z} 的密度函数是

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) &= \left| \frac{\partial \mathbf{s}}{\partial \mathbf{z}} \right| f_{\mathbf{S}}(\mathbf{s}) \\ &= |\mathbf{V}| f_{\mathbf{S}}(\mathbf{Vz}) \\ &= |\mathbf{V}| \prod_{i=1}^d f_{S_i}(\mathbf{v}_i^{\top} \mathbf{z}) \end{aligned}$$

其中， $\mathbf{v}_i = (v_{i1}, \dots, v_{id})^{\top}$ ，即 \mathbf{V} 的第 i 行向量转置而成的列向量。显然，

$$\mathbf{V}^{\top} = (\mathbf{v}_{1\cdot}, \dots, \mathbf{v}_{d\cdot})_{d \times d}$$

□ 由白化数据的对数似然函数求旋转变换 \mathbf{V} ：设白化数据 $\mathbf{z}_1, \dots, \mathbf{z}_n$ 是来自总体 \mathbf{Z} 的样本，则对数似然函数为

$$\ell(\mathbf{V}) = n \ln |\mathbf{V}| + \sum_{j=1}^n \sum_{i=1}^d \ln f_{S_i}(\mathbf{v}_i^{\top} \mathbf{x}_j)$$

定义逐量向量函数 (component-wise vector function) $\mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 如下，

$$\mathbf{g}(\mathbf{y}) = \begin{pmatrix} g_1(y_1) \\ \vdots \\ g_d(y_d) \end{pmatrix}, \text{ 其中 } \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix}$$

$$g_i = \frac{f'_{S_i}}{f_{S_i}}, \text{ 其中 } i = 1, \dots, d$$

特别地，当所有 g_i 都等于某个函数 g 时， $\mathbf{g}(\mathbf{y})$ 也简写作 $g(\mathbf{y})$ 。在 R 语言里，向量值函数 $\mathbf{g}(\mathbf{y})$ 的定义也是

$$\mathbf{g}(\mathbf{y}) = \begin{pmatrix} g(y_1) \\ \vdots \\ g(y_d) \end{pmatrix}$$

对数似然函数对 \mathbf{V} 的梯度矩阵是

$$\begin{aligned}
 \frac{\partial \ell(\mathbf{V})}{\partial \mathbf{V}} &= \frac{n}{|\mathbf{V}|} \frac{\partial |\mathbf{V}|}{\partial \mathbf{V}} + \sum_{j=1}^n \sum_{i=1}^d \frac{f'_{S_i}(\mathbf{v}_i^\top \mathbf{z}_j)}{f_{S_i}(\mathbf{v}_i^\top \mathbf{z}_j)} \cdot \frac{\partial (\mathbf{v}_i^\top \mathbf{z}_j)}{\partial \mathbf{V}} \\
 &= \frac{n}{|\mathbf{V}|} |\mathbf{V}| (\mathbf{V}^\top)^{-1} + \sum_{j=1}^n \sum_{i=1}^d g_i(\mathbf{v}_i^\top \mathbf{z}_j) \begin{pmatrix} 0 \\ \mathbf{z}_j^\top \\ 0 \end{pmatrix} \leftarrow \text{第 } i \text{ 行} \\
 &= n\mathbf{V} + \sum_{j=1}^n \begin{pmatrix} g_1(\mathbf{v}_1^\top \mathbf{z}_j) \mathbf{z}_j^\top \\ \vdots \\ g_d(\mathbf{v}_d^\top \mathbf{z}_j) \mathbf{z}_j^\top \end{pmatrix} \\
 &= n\mathbf{V} + \sum_{j=1}^n \mathbf{g}(\mathbf{V} \mathbf{z}_j) \mathbf{z}_j^\top
 \end{aligned}$$

算法 7.8 令上述梯度矩阵 $\frac{\partial \ell(\mathbf{V})}{\partial \mathbf{V}}$ 为零矩阵，我们得到

$$\mathbf{V} \leftarrow -\frac{1}{n} \sum_{j=1}^n \mathbf{g}(\mathbf{V} \mathbf{z}_j) \mathbf{z}_j^\top \quad (7.39)$$

将 $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_d)$ 的列向量都变为单位向量，即

$$\mathbf{V} \leftarrow \left(\frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}, \dots, \frac{\mathbf{v}_d}{\|\mathbf{v}_d\|} \right)$$

然后，再次重复 (7.39)，直至再无很多更新。

例 7.15 算法 7.8 的一个具体实现：假设 \mathbf{S} 的各个分量 $S_1, \dots, S_d \stackrel{\text{iid}}{\sim} \text{Logistic}(0, 1)$ ，其密度函数是

$$f_{S_i}(s) = \frac{\exp\{-s\}}{(1 + \exp\{-s\})^2}$$

于是，

$$\begin{aligned}
 g_i(s) &= \frac{f'_{S_i}(s)}{f_{S_i}(s)} \\
 &= 1 - \frac{2}{1 + \exp\{-s\}} \\
 &= 1 - 2S(s)
 \end{aligned}$$

上式中， $S(x)$ 是机器学习和模式识别里大名鼎鼎的 S 形函数。由更新方法 (7.39) 可得，

$$\mathbf{V} \leftarrow -\frac{1}{n} \sum_{j=1}^n \left(\mathbf{1}_d - \frac{2}{1 + \exp\{-\mathbf{V} \mathbf{z}_j\}} \right) \mathbf{z}_j^\top$$

$$= \frac{2}{n} \sum_{j=1}^n \frac{1}{1 + \exp\{-Vz_j\}} z_j^T - \begin{pmatrix} \bar{z}^T \\ \vdots \\ \bar{z}^T \end{pmatrix}_{d \times d}$$

其中, $\bar{z} = \frac{1}{n} \sum_{j=1}^n z_j$ 是白化数据的均值。

例 7.16 接着例 7.14, 从图 7.37(a) 所示的观测数据, 我们得到 PCA 数据表示和 ICA 数据表示如图 7.38 所示。其中, PCA 数据表示 (7.23) 无法得到源数据, 它只关心数据散布最大的正交方向。ICA 数据表示可能与源数据相差一个常尺度的拉伸, 在某种意义上算是恢复了源数据。

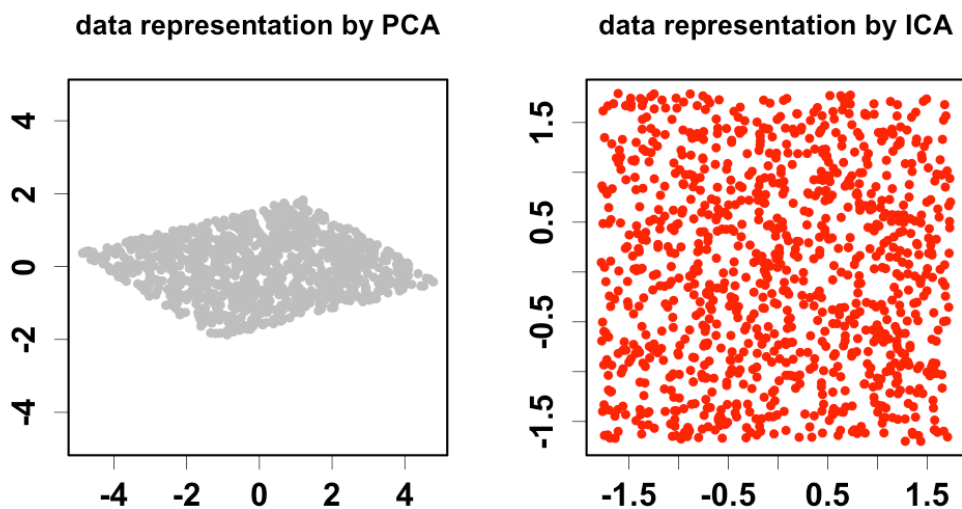


图 7.38 例 7.14 的 PCA 和 ICA 数据表示

7.2.4 多维缩放与等距映射

在几何上,如何把样本的维数降下来,同时又尽量不破坏它们之间的距离关系?多维缩放 (multidimensional scaling, MDS) 就是实现这类数据表示的方法。

已知样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 的距离矩阵 $\Delta = (d_{ij})_{n \times n}$, 譬如, 欧氏距离矩阵。我们要寻找的数据表示 $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^k$ 要尽可能地保证

$$\|\mathbf{y}_i - \mathbf{y}_j\|_2 = d_{ij}, \forall i, j \in \{1, \dots, n\}$$

换句话说, 就是求解下面的全局最优化问题。

$$\operatorname{argmin}_{\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^k} \sum_{i < j} (\|\mathbf{y}_i - \mathbf{y}_j\|_2 - d_{ij})^2 \quad (7.40)$$

实际上, 问题 (7.40) 并不需要真实样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$, 有也可无也可。它探讨的是如何仅仅通过给定的距离矩阵 Δ 重构样本 $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^k$, 使其距离矩阵就是 Δ 。重构样本的维数 k 是由 Δ 决定的。1938 年, 美国数学家盖尔·杨 (Gail Young, 1915—1999) 和阿尔斯通·斯科特·豪斯霍尔德 (Alston Scott Householder, 1904—1993) (图 7.39) 利用由欧氏距离矩阵 Δ 构造的内积矩阵 (见附录 D 的性质 D.8) 的谱分解给出了一个答案^[127]。



图 7.39 杨和豪斯霍尔德

算法 7.9 给定欧氏距离矩阵 Δ , 下述算法给出问题 (7.40) 的一个解。

- (1) 根据性质 D.8, 由 Δ 构造内积矩阵 \mathbf{G} 。
- (2) 计算 \mathbf{G} 的谱分解

$$\begin{aligned} \mathbf{G} &= \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \\ &= \mathbf{V} \operatorname{diag}(\lambda_1, \dots, \lambda_r) \mathbf{V}^T \end{aligned}$$

- (3) 我们得到问题 (7.40) 的一个解

$$\begin{aligned} \mathbf{Y} &= \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{V}^T \\ &= \operatorname{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_r}) \mathbf{V}^T \end{aligned}$$

另外, $\mathbf{Y}_{n \times n} = \mathbf{V} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{V}^T$ 也是问题 (7.40) 的一个解。

证明: 矩阵 $\mathbf{Y}_{r \times n} = \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{V}^T$ 满足 $\mathbf{Y}^T \mathbf{Y} = \mathbf{G}$, 即 \mathbf{A} 的列向量的内积矩阵就是 \mathbf{G} , 进而根据性质 D.8, 这些列向量的欧氏距离矩阵为 Δ 。□

例 7.17 从普林斯顿大学 WordNet 词汇语义知识库或者类似的知识图谱中, 可以抽取词汇距离 (例如, 两个概念节点之间按照某种关系的最短路径的长度)。多维缩放算法 7.9 使得这些词汇或概念在欧氏空间里

得以向量化，为自然语言处理提供了数据表示。下面，通过一个虚构的例子讲解算法 7.9：表 7.8 是一个词汇距离矩阵，只用于演示多维缩放。

表 7.8 虚构的词汇距离数据

Δ	狗	猫	人类	机器人	车
狗	0	3	8	12	16
猫	3	0	9	13	16
人类	8	9	0	6	15
机器人	12	13	6	0	4
车	16	16	15	4	0

根据表 7.8，由算法 7.9 计算出词汇的二维向量表示，见表 7.9。

表 7.9 二维词汇向量

	狗	猫	人类	机器人	车
y_1	6.26	6.49	2.49	-5.50	-9.74
y_2	1.70	3.13	-5.52	-2.48	3.17

表 7.9 中的二维词汇向量近似地保留了表 7.8 所描述的词汇距离信息。在算法 7.9 中，投射空间的维度不超过内积矩阵 G 的秩。词汇向量的维度越高，距离信息越精确。

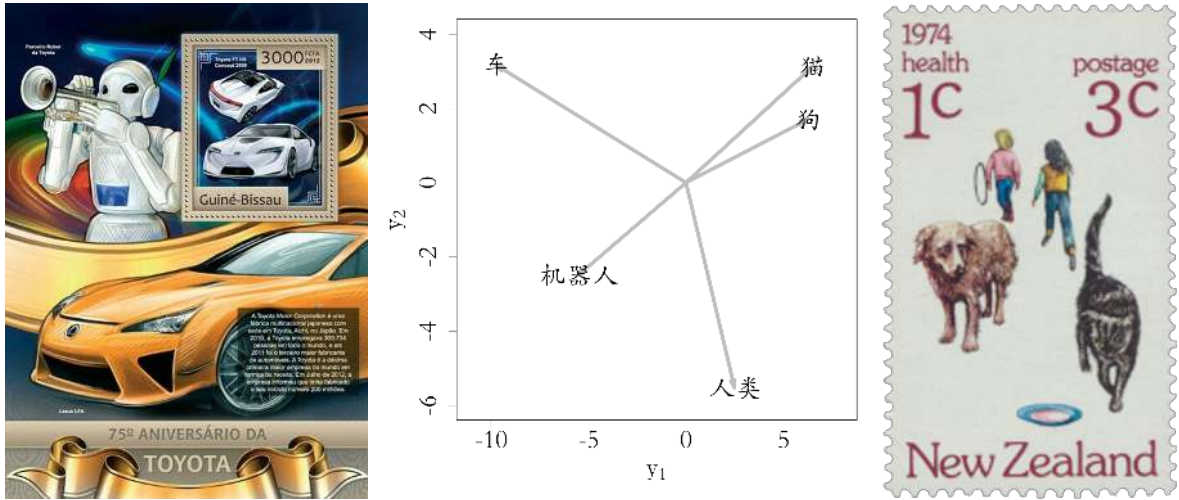


图 7.40 在平面内的二维词汇向量

以上的讨论都是基于 Δ 是欧氏距离矩阵。在多维缩放中，采用欧氏距离的作法是值得商榷的。譬如，数据若分布在一个球面上，用欧氏距离来实现多维缩放就是错误的。一般地，数据在空间的分布是未知的，甚至连数据所在的空间我们也是知之甚少。若要考虑比球面更一般的几何对象，就需要流形 (manifold) 的概念。

流形不严格地定义为每个点的局部都可近似地看作欧氏空间的几何对象（如球面、环面等）。例如，足球是一些正五边形“拼接”而成的（图 7.41(a)）。再如，美国建筑师、发明家巴克敏斯特·富勒 (Buckminster

Fuller, 1895—1983) (图 7.41(b)) 的球型屋顶, 也是用欧氏碎片“拼接”得到的。大量的几何对象都可以通过这种“拼接”方式构造出来。



图 7.41 流形

按照流形的定义, 我们可按照欧氏几何来研究流形的局部性质, 见图 7.42(a)。在一个流形上, 局部两点之间的最短路径被称为测地线 (geodesic)。例如, 在球面上, 两点之间的最短路径是过此两点和球心的平面截出来的大圆上的一段弧线, 见图 7.42(b)。其中, 球面三角形 ABC 是由三条测地线构成的, 三个内角之和大于 π 。

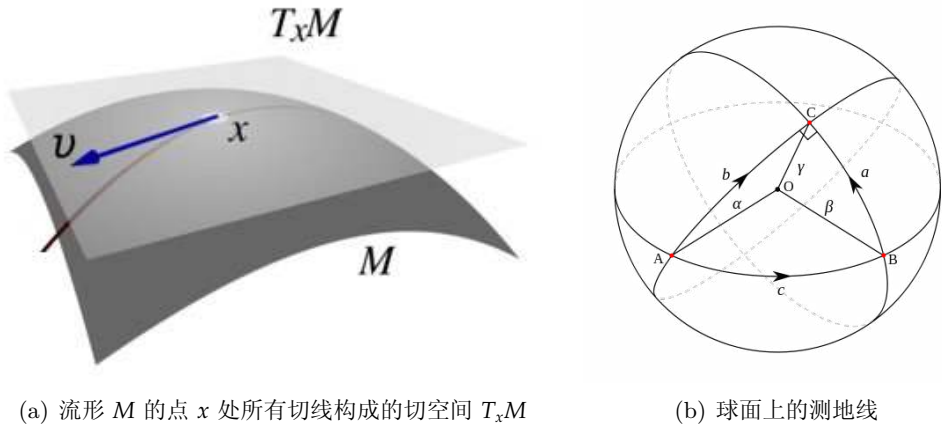


图 7.42 微分流形上的切平面与测地线

在不清楚流形具体长啥样的情况下, 如何测得任意两样本点之间的测地距离 (geodesic distance) 呢? 可以利用流形的定义, 对每个样本点, 先局部测好与近邻 (nearest neighbor, NN) 的距离, 然后再一段一段地拼接起来。1816—1855 年, 俄国天文学家、地理学家瓦西里·雅可夫列维奇·斯特鲁维 (Vasily Yakovlevich Struve, 1793—1864)* (图 7.43) 从挪威到黑海构建了一组三角测量点, 所得到的三角测量链被称为“斯特鲁维测地弧” (Struve geodetic arc), 全长 2820 公里。

*斯特鲁维的德文名字是弗里德里希·格奥尔格·威廉·冯·斯特鲁维 (Friedrich Georg Wilhelm von Struve), 他的家族出了好几位天文学家, 月球上的斯特鲁维环形山正是以他和他的儿孙天文学家命名的。



图 7.43 斯特鲁维测地弧

流形学习 (manifold learning) 中常见的等距映射 (isometric map, Isomap) 算法^[128] 借鉴了斯特鲁维测地弧的基本想法, 该算法是麻省理工学院认知科学家乔舒亚·特南鲍姆 (Joshua Tenenbaum, 1972—) 于 2000 年提出的。

算法 7.10 (等距映射) 先按照欧氏距离算得每个点 p 的 k 个近邻, 这些近邻与 p 点的测地距离可近似为欧氏距离。

- (1) 构造一个近邻图, 边的权重就是欧氏距离。
- (2) 利用戴克斯特拉算法*或者弗洛伊德-沃舍尔算法[†]求任意两点间的最短路径, 所得的距离之和即为测地距离的近似, 称为“近似测地距离”。
- (3) 最后利用 MDS 进行非线性降维。

假设数据分布在一个弯曲空间里, 则两点之间的距离应采用测地距离, 它的近似计算参见算法 7.10 的前两个步骤。如图 7.44 所示, 数据分布在“卷筒”流形 (也称“瑞士卷”) 上图 7.44(a), 不能直接用欧

*荷兰计算机科学家、1972 年图灵奖得主艾兹赫尔·戴克斯特拉 (Edsger Dijkstra, 1930—2002) 于 1956 年发现的广度优先搜索算法, 用于寻找赋权有向图的单源最短路径。

[†]美国计算机科学家、1978 年图灵奖得主罗伯特·弗洛伊德 (Robert Floyd, 1936—2001) 和美国计算机科学家斯蒂芬·沃舍尔 (Stephen Warshall, 1935—2006) 于 1962 年发现的求最短路径的一种算法。

氏距离来刻画两点间的测地距离。而是应该利用欧氏距离为每个点找出近邻后，再用最短路径的欧氏距离之和来逼近测地距离，见图 7.44(b) 的近似测地线。如果将瑞士卷平铺开来，如图 7.44(c) 所示，测地线与近似测地线的关系就一目了然了。

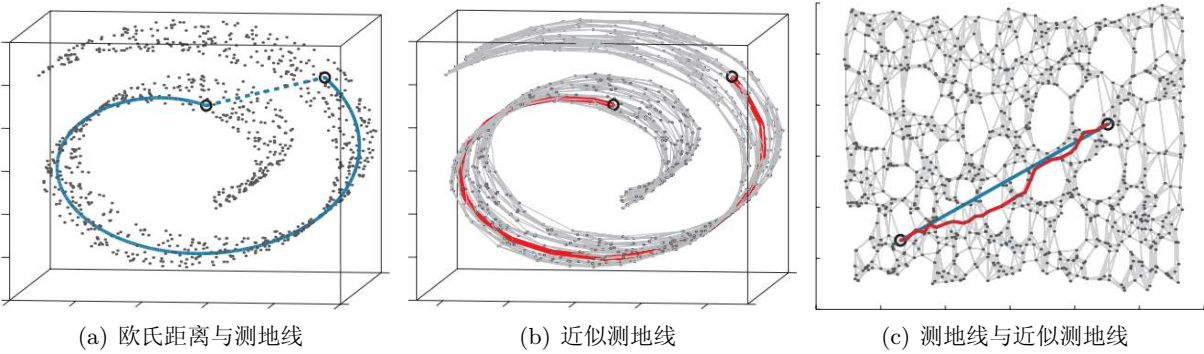


图 7.44 等距映射算法 7.10 中的近似测地距离

因为要算任意两点之间的欧氏或者测地距离，基于全局寻优的多维缩放算法的计算复杂度较高。如果我们只关注近邻之间的性质，就可以把全局寻优降格为局部寻优。下一小节的“局部线性嵌入”和“拉普拉斯本征映射”就是一类只考虑保留局部性质的降维方法，也属于流形学习的范畴。

7.2.5 局部嵌入的降维

与多维缩放保证全局最优的降维方法不同，如果只想让某个局部性质在降维过程中保持不变，例如，在每个局部的样本点之间的线性相关性（即每个样本点可由它的几个近邻线性表出），或者局部的近邻关系等等，我们无需顾忌数据在空间分布的整体性质，搜索空间大大缩小。该如何把这类保持局部性质不变的降维抽象为一个最优化问题呢？

1. 局部线性嵌入

2000 年，美国机器学习专家山姆·罗维斯 (Sam Roweis, 1972—2010) 提出局部线性嵌入 (locally-linear embedding, LLE) 的降维方法，它的基本想法如图 7.45 所示，即每个样本点在局部被其 k 个近邻线性表出，在降维时保留近邻间的线性关系^[129]。该方法潜在的假设是，样本间的局部线性关系是数据的本质特征，在降维后仍该予以保留。

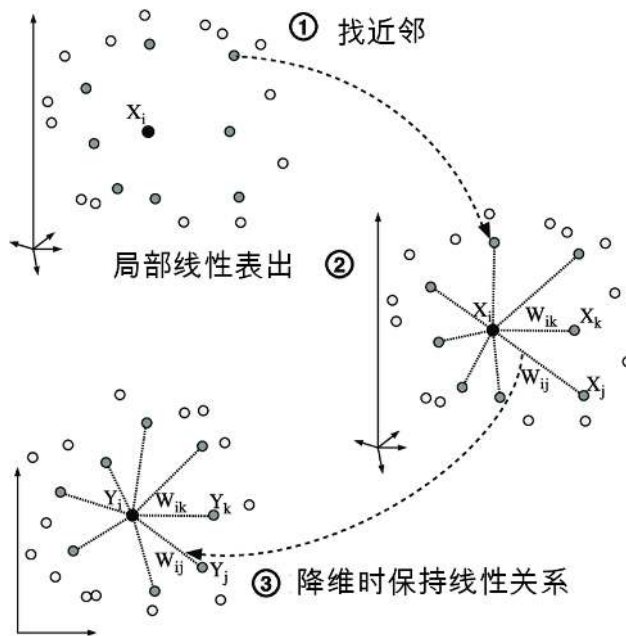


图 7.45 局部线性嵌入的基本想法^[129]

(1) 局部线性表示：对每个样本点 $x_i \in \mathbb{R}^D, i = 1, 2, \dots, n$ 考虑其近邻 $x_j, j \in N(i)$ ，其中 $N(i)$ 表示 x_i 的 k 个近邻的指标集合。寻找一个稀疏矩阵 $W_{n \times n} = (w_{ij})$ ，使之最小化下面的误差函数

$$E(W) = \sum_{i=1}^n \left\| x_i - \sum_{j=1}^n w_{ij} x_j \right\|^2$$

其中， w_{ij} 满足以下条件

$$w_{ij} = 0, \text{ 如果 } j \notin N(i)$$

$$\sum_{j=1}^n w_{ij} = 1$$

在此条件之下，每个点 \mathbf{z} 各自寻求其 k 个近邻 $\mathbf{z}_1, \dots, \mathbf{z}_k$ 的局部线性表出。令 $\mathbf{w} = (w_1, \dots, w_k)^\top$ ，误差函数 $E(\mathbf{w})$ 可以定义为

$$\begin{aligned} E(\mathbf{w}) &= \left\| \mathbf{z} - \sum_{j=1}^k w_j \mathbf{z}_j \right\|^2 \\ &= \left\| \sum_{j=1}^k w_j (\mathbf{z} - \mathbf{z}_j) \right\|^2 \\ &= \mathbf{w}^\top (\mathbf{z} - \mathbf{z}_j)(\mathbf{z} - \mathbf{z}_j)^\top \mathbf{w} \\ &= \mathbf{w}^\top \mathbf{Z} \mathbf{w}, \text{ 其中 } \mathbf{Z} = (\mathbf{z} - \mathbf{z}_j)(\mathbf{z} - \mathbf{z}_j)^\top \end{aligned}$$

定义拉格朗日函数（附录 B）如下，

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^\top \mathbf{Z} \mathbf{w} + \lambda(1 - \mathbf{w}^\top \mathbf{1}_k)$$

令 $\partial \mathcal{L} / \partial \mathbf{w} = \mathbf{0}$ ，得到

$$2\mathbf{Z} \mathbf{w} - \lambda \mathbf{1}_k = \mathbf{0}$$

按照条件，对 \mathbf{w} 进行归一化（便消掉了 λ ），于是

$$\mathbf{w} = \frac{\mathbf{Z}^{-1} \mathbf{1}_k}{\mathbf{1}_k^\top \mathbf{Z}^{-1} \mathbf{1}_k}$$

对每个点（并行地）进行局部线性表出，便得到矩阵 \mathbf{W} 。

- (2) 保证局部线性关系的降维：令 $\mathbf{y}_i \in \mathbb{R}^d$ 是 \mathbf{x}_i 降维后的结果，其中局部线性关系 $\mathbf{x}_i \approx \sum_{j=1}^n w_{ij} \mathbf{x}_j$ 保留至 $\mathbf{y}_i \approx \sum_{j=1}^n w_{ij} \mathbf{y}_j$ （即降维尽可能不破坏每个样本点的局部线性表示）。我们的任务是寻找解释矩阵 $\mathbf{Y}_{d \times n} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ ，使得下面的损失函数达到最小。

$$\begin{aligned} L(\mathbf{Y}) &= \sum_{i=1}^n \left\| \mathbf{y}_i - \sum_{j=1}^n w_{ij} \mathbf{y}_j \right\|^2 \\ &= \|\mathbf{Y} - \mathbf{Y} \mathbf{W}^\top\|_F^2, \text{ 由 } \|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A} \mathbf{A}^\top) \\ &= \text{tr}[(\mathbf{Y} - \mathbf{Y} \mathbf{W}^\top)(\mathbf{Y} - \mathbf{Y} \mathbf{W}^\top)^\top] \\ &= \text{tr}(\mathbf{Y} \mathbf{M} \mathbf{Y}^\top), \text{ 其中 } \mathbf{M}_{n \times n} = (\mathbf{I}_n - \mathbf{W})^\top (\mathbf{I}_n - \mathbf{W}) \end{aligned}$$

$$= \sum_{j=1}^d \tilde{\mathbf{y}}_j^\top \mathbf{M} \tilde{\mathbf{y}}_j, \text{ 其中 } \mathbf{Y}^\top = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_d)$$

为了解的唯一性，我们要求 $\mathbf{y}_1, \dots, \mathbf{y}_n$ 满足以下标准化（即均值为零，样本协方差矩阵为单位矩阵）的条件。

$$\begin{aligned} \mathbf{Y} \mathbf{1}_n &= \sum_{i=1}^n \mathbf{y}_i = \mathbf{0} \\ \frac{1}{n-1} \mathbf{Y} \mathbf{Y}^\top &= \frac{1}{n-1} \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^\top = \mathbf{I}_d \end{aligned}$$

不难看出，第一个条件通过中心化就可轻易实现。第二个条件要求 $\mathbf{Y}^\top = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_d)$ 的列向量是正交的且欧氏长度都是 \sqrt{n} 。于是，向量 $\tilde{\mathbf{y}}_j / \sqrt{n}$ 都落在单位球面上，根据第 570 页的定理 B.1，当 $\mathbf{x} = \tilde{\mathbf{y}}_j / \sqrt{n}$ 是 \mathbf{M} 的单位本征向量时，使得二次型 $\mathbf{x}^\top \mathbf{M} \mathbf{x}$ 取得极值，即本征值 λ_j 。

对称矩阵 \mathbf{M} 是半正定的，其所有本征值都是非负的，本征向量都是正交的。其中， $\lambda = 0$ 一定是 \mathbf{M} 的本征值，其本征向量为 $\mathbf{1}_n$ ，因为

$$\begin{aligned} \mathbf{M} \mathbf{1}_n &= (\mathbf{I}_n - \mathbf{W})^\top (\mathbf{I}_n - \mathbf{W}) \mathbf{1}_n \\ &= (\mathbf{I}_n - \mathbf{W})^\top \left[\mathbf{1}_n - \begin{bmatrix} \sum_{j=1}^n w_{1j} \\ \vdots \\ \sum_{j=1}^n w_{nj} \end{bmatrix} \right] \\ &= \mathbf{0} \end{aligned}$$

要使得损失函数 $L(\mathbf{Y})$ 达到最小，我们只需考虑从小到大前 d 个正的本征值 $\lambda_1, \dots, \lambda_d$ 所对应的长度为 \sqrt{n} 的本征向量 $\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_d \in \mathbb{R}^n$ ，并令 $\mathbf{Y}_{d \times n} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_d)^\top$ ，此时的损失是

$$L(\mathbf{Y}) = n(\lambda_1 + \dots + \lambda_d)$$

如果不知道定理 B.1 这个结果也无所谓，我们下面老老实实地用拉格朗日乘子法（附录 B）来求解：令 $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$ ，定义拉格朗日函数 $\mathcal{L}(\mathbf{Y}, \mathbf{\Lambda})$ 如下，

$$\mathcal{L}(\mathbf{Y}, \mathbf{\Lambda}) = \text{tr}(\mathbf{Y} \mathbf{M} \mathbf{Y}^\top) - \text{tr}\{\mathbf{\Lambda}[\mathbf{Y} \mathbf{Y}^\top - (n-1)\mathbf{I}_d]\}$$

利用以下结果^[8]，

$$\frac{\partial \text{tr}(\mathbf{A} \mathbf{X} \mathbf{B})}{\partial \mathbf{X}} = \mathbf{A}^\top \mathbf{B}^\top \quad (7.41)$$

$$\frac{\partial \text{tr}(\mathbf{A} \mathbf{X} \mathbf{B} \mathbf{X}^\top \mathbf{C})}{\partial \mathbf{X}} = \mathbf{C} \mathbf{A} \mathbf{X} \mathbf{B} + \mathbf{A}^\top \mathbf{C}^\top \mathbf{X} \mathbf{B}^\top \quad (7.42)$$

□ 令 $\partial \mathcal{L} / \partial \Lambda = \mathbf{O}$ ，即得到第二个条件。

□ 令 $\partial \mathcal{L} / \partial \mathbf{Y} = \mathbf{O}$ ，得到

$$2\mathbf{Y}\mathbf{M} - 2\Lambda\mathbf{Y} = \mathbf{O}$$

上式转置后得到

$$\mathbf{M}\mathbf{Y}^\top = \mathbf{Y}^\top \Lambda$$

或者，

$$\mathbf{M}\tilde{\mathbf{y}}_j = \lambda_j \tilde{\mathbf{y}}_j, \text{ 其中 } j = 1, \dots, d$$

即， $\lambda_1, \dots, \lambda_d$ 是对称半正定阵 \mathbf{M} 的本征值。

例 7.18 图 7.46 展示了局部线性嵌入将分布在“卷筒”流形上的数据投射到二维平面，保持局部的线性关系一致^[129]。

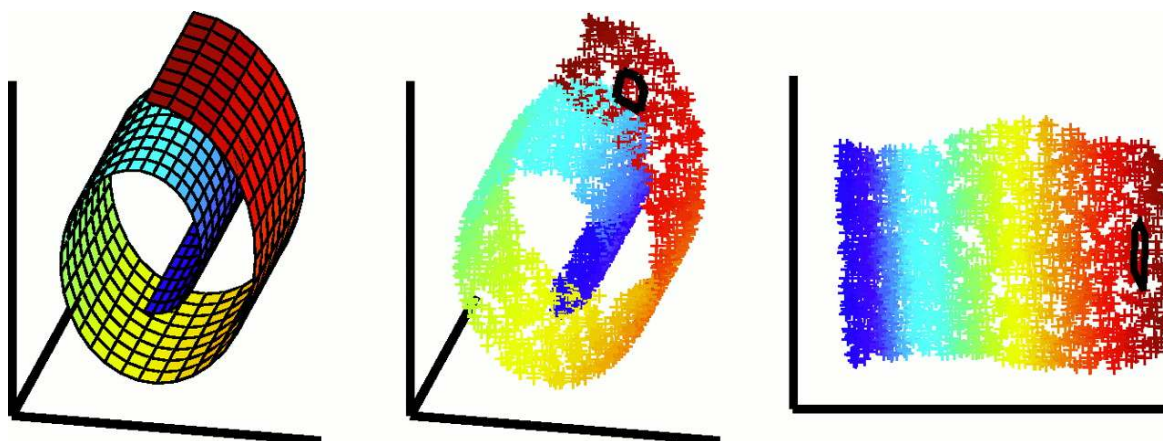


图 7.46 “卷筒”流形上数据的局部线性嵌入

利用局部线性嵌入，将四维空间里的 iris 数据投射到二维平面。如图 7.47 所示，LLE 对近邻个数 k 是敏感的，导致降维结果迥然不同。然而，类 S 都被映为原点。

2. 拉普拉斯本征映射

与局部线性嵌入方法类似，2003 年美国机器学习专家米哈伊尔·贝尔金 (Mikhail Belkin) 和帕萨·尼约吉 (Partha Niyogi) 提出了一种新的局部嵌入的降维方法——拉普拉斯本征映射 (Laplacian eigenmap, LE)^[130]。这种降维方法只是粗略地要求相近的样本点 $\mathbf{x}_i, \mathbf{x}_j$ 降维后变为 $\mathbf{y}_i, \mathbf{y}_j \in \mathbb{R}^d$ 依然很接近。

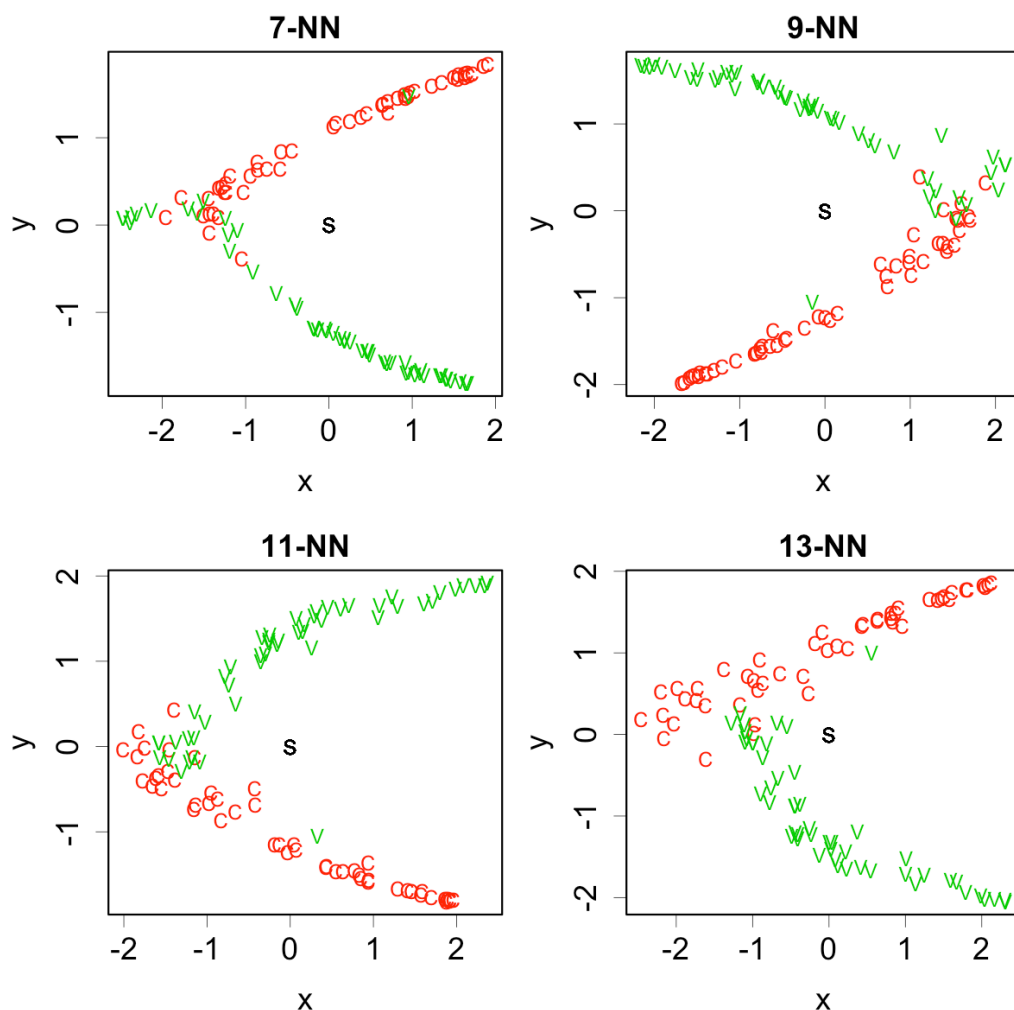


图 7.47 iris 数据的局部线性嵌入

该方法也是从 k 近邻开始考虑:构建一个无向图 $G = (V, E)$, 每个节点代表一个样本点, $V = \{1, 2, \dots, n\}$ 。如果点 i 是点 j 的 k 近邻之一, 或者 j 是 i 的 k 近邻之一, 则 i, j 之间有一条边相连, 记作 $(i, j) \in E$ 。拉普拉斯本征映射化归为一个最优化问题

$$\min_Y \sum_{i,j=1}^n w_{ij} \|y_i - y_j\|^2, \text{ 其中 } Y_{d \times n} = (y_1, \dots, y_n)$$

上式中, w_{ij} 是一个惩罚因子, 定义为

$$w_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{t}\right) & , \text{ 如果 } x_i, x_j \text{ 连通} \\ 0 & , \text{ 否则} \end{cases}$$

或者, 简单地将其定义

$$w_{ij} = \begin{cases} 1 & \text{, 如果 } \mathbf{x}_i, \mathbf{x}_j \text{ 之间有一条边} \\ 0 & \text{, 否则} \end{cases} \quad (7.43)$$

显然, $\mathbf{W}_{n \times n} = (w_{ij})$ 是一个对称矩阵。 w_{ij} 越接近 0, 表示 $\mathbf{x}_i, \mathbf{x}_j$ 的距离越远, 我们对 $\|\mathbf{y}_i - \mathbf{y}_j\|^2$ 的大小越不在意。相反, w_{ij} 越接近 1, 则 $\mathbf{x}_i, \mathbf{x}_j$ 的距离越近, $\|\mathbf{y}_i - \mathbf{y}_j\|^2$ 的大小直接影响目标函数。

$$\begin{aligned} \sum_{i,j=1}^n w_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 &= \sum_{i,j=1}^n w_{ij} (\|\mathbf{y}_i\|^2 + \|\mathbf{y}_j\|^2 - 2\mathbf{y}_i^\top \mathbf{y}_j) \\ &= \sum_{i=1}^n \|\mathbf{y}_i\|^2 \sum_{j=1}^n w_{ij} + \sum_{j=1}^n \|\mathbf{y}_j\|^2 \sum_{i=1}^n w_{ij} - 2 \sum_{i,j=1}^n \mathbf{y}_i^\top w_{ij} \mathbf{y}_j \\ &= 2 \sum_{i=1}^n \mathbf{y}_i^\top d_i \mathbf{y}_i - 2 \sum_{i,j=1}^n \mathbf{y}_i^\top w_{ij} \mathbf{y}_j, \text{ 其中 } d_i = \sum_{j=1}^n w_{ij} \\ &= 2\text{tr}[\mathbf{Y}(\mathbf{D} - \mathbf{W})\mathbf{Y}^\top], \text{ 其中 } \mathbf{D} = \text{diag}(d_1, \dots, d_n) \\ &= 2\text{tr}[\mathbf{Y}\mathbf{L}\mathbf{Y}^\top], \text{ 其中 } \mathbf{L} = \mathbf{D} - \mathbf{W} \end{aligned}$$

d_i 越大, 意味着 \mathbf{x}_i 和它的 k 近邻越凝聚。将 \mathbf{D} 类比作无向图中一个节点的度 (即该节点所连的边数), 将 \mathbf{W} 类比作邻接矩阵 (adjacency matrix), 对称矩阵 $\mathbf{L} = \mathbf{D} - \mathbf{W}$ 被称为拉普拉斯矩阵 (Laplacian matrix), 这个概念来自于图论 (见《随机之美》^[8] 的附录 “矩阵计算的一些结果”)。

性质 7.10 拉普拉斯矩阵是一个对称、半正定矩阵。事实上, 这是因为

$$\begin{aligned} \mathbf{x}^\top \mathbf{L} \mathbf{x} &= \mathbf{x}^\top (\mathbf{D} - \mathbf{W}) \mathbf{x} \\ &= \sum_{i=1}^n d_i x_i^2 - 2 \sum_{(i,j) \in E} x_i x_j \\ &= \sum_{(i,j) \in E} (x_i - x_j)^2 \geq 0 \end{aligned}$$

为了移除局部嵌入中的任意比例因子, 加上限制条件 $\mathbf{Y}\mathbf{D}\mathbf{Y}^\top = \mathbf{I}_d$ 。拉普拉斯本征映射将原始数据变为满足以下条件的 $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$,

$$\begin{aligned} &\text{argmin} \text{tr}(\mathbf{Y}\mathbf{L}\mathbf{Y}^\top) \\ &\mathbf{Y}\mathbf{D}\mathbf{Y}^\top = \mathbf{I}_d \end{aligned} \quad (7.44)$$

显然, 样本点之间的近邻关系被抽象为 (\mathbf{W}, \mathbf{D}) , 它们甚至可以是像 (7.43) 这样没有度量的拓扑关系。换句话说, 只要 (\mathbf{W}, \mathbf{D}) 一样, 不管原数据怎样, 拉普拉斯本征映射的结果都一样。

□ 最优化问题 (7.44) 与局部线性嵌入在形式上如此之像。令 $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$, 定义拉格朗日函数

$\mathcal{L}(\mathbf{Y}, \mathbf{A})$ 如下,

$$\mathcal{L}(\mathbf{Y}, \mathbf{A}) = \text{tr}(\mathbf{Y}\mathbf{L}\mathbf{Y}^\top) - \text{tr}[\mathbf{A}(\mathbf{Y}\mathbf{D}\mathbf{Y}^\top - \mathbf{I}_d)]$$

利用结果 (7.41) 和 (7.42), 令 $\partial\mathcal{L}/\partial\mathbf{Y} = \mathbf{O}$, 我们得到

$$\mathbf{Y}\mathbf{L} - \mathbf{A}\mathbf{Y}\mathbf{D} = \mathbf{O}$$

上式转置后得到

$$\mathbf{L}\mathbf{Y}^\top = \mathbf{D}\mathbf{Y}^\top\mathbf{A}$$

或者,

$$\mathbf{L}\tilde{\mathbf{y}}_j = \lambda_j \mathbf{D}\tilde{\mathbf{y}}_j, \text{ 其中 } \mathbf{Y}^\top = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_d) \quad (7.45)$$

我们称满足式 (7.45) 的 $\lambda_j, \tilde{\mathbf{y}}_j$ 为拉普拉斯本征值和拉普拉斯本征向量。特别地, 如果 $d_i > 0, i = 1, \dots, n$, 则 $\tilde{\mathbf{y}}_j, j = 1, \dots, n$ 是半正定阵 $\mathbf{D}^{-1}\mathbf{L}$ 的本征值。忽略掉那些 $\lambda_j = 0$, 因为 $\mathbf{L}\tilde{\mathbf{y}}_j = \mathbf{0}$ 。其中, $\lambda = 0$ 一定是拉普拉斯本征值, 其拉普拉斯本征向量为 $\mathbf{1}_n$, 因为

$$\begin{aligned} (\mathbf{D} - \mathbf{W})\mathbf{1}_n &= \mathbf{D}\mathbf{1}_n - \mathbf{W}\mathbf{1}_n \\ &= \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} - \begin{pmatrix} \sum_{j=1}^n w_{1j} \\ \vdots \\ \sum_{j=1}^n w_{nj} \end{pmatrix} \\ &= \mathbf{0} \end{aligned}$$

不妨设 $0 < \lambda_1 \leq \dots \leq \lambda_d$, 则拉普拉斯本征映射后的结果是

$$\mathbf{Y} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_d)^\top$$

□ 令 $\mathbf{Z} = \mathbf{Y}\mathbf{D}^{\frac{1}{2}}$, 其中 $\mathbf{D}^{\frac{1}{2}} = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})$ 。如果 \mathbf{D} 非奇异, 则

$$\min_{\mathbf{Y}\mathbf{D}\mathbf{Y}^\top = \mathbf{I}_d} \text{tr}(\mathbf{Y}\mathbf{L}\mathbf{Y}^\top) = \min_{\mathbf{Z}\mathbf{Z}^\top = \mathbf{I}_d} \text{tr}[\mathbf{Z}(\mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}})\mathbf{Z}^\top]$$

即, 拉普拉斯本征映射模型和局部线性嵌入模型在形式上是一致的。

7.2.6 塔克分解

实践中,我们经常会遇到多维数组 (multidimensional array)。例如,商品为行、消费者为列、时间为层来记录一个时间段内消费者的购物情况,数据表示就是如图 7.48 所示的三维数组,记作 $\mathbf{X} \in \mathbb{R}^I \times \mathbb{R}^J \times \mathbb{R}^K$ (有时,简记作 $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$)。

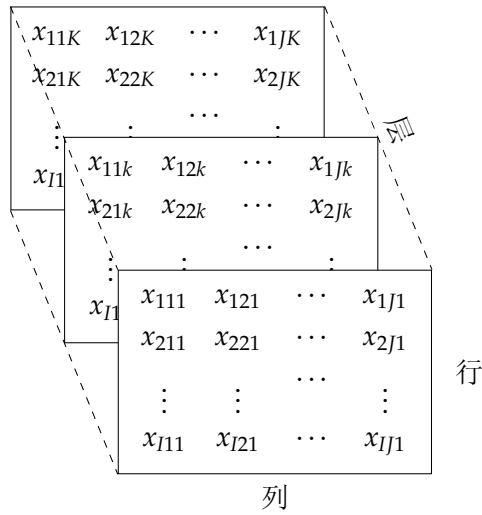


图 7.48 三维数组

可以从不同的角度来理解图 7.48 所示的三阶张量,每一切片都是一个矩阵^[131]: 有的是层相同,有的是列相同,有的是行相同 (图 7.49)。

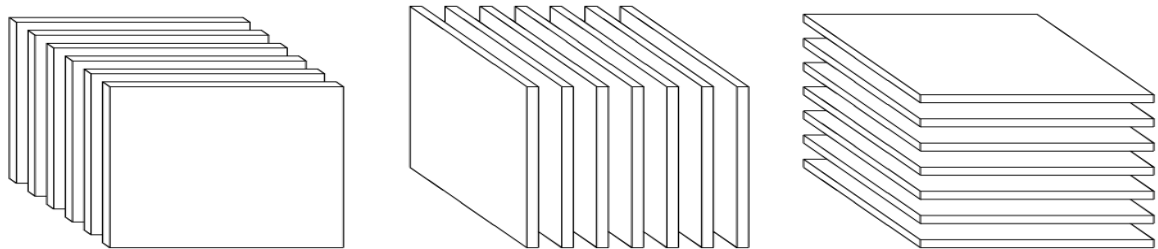


图 7.49 三维数组的切片

张量可由多维数组表示,它是向量、矩阵的一般化 (附录 E): 向量是一阶张量,矩阵是二阶张量。主成分分析的理论基础是矩阵的奇异值分解,对于高阶张量,是否也有类似的分解? 人们之所以对它感兴趣,主要是想通过多维数组的压缩与重构,找出数据中的规律性。1966 年,美国统计学家、心理学家莱德亚德·塔克 (Ledyard Tucker, 1910—2004) (图 7.50) 提出张量的“塔克分解”,最初被描述为因子分析和主成分分析的扩展。塔克在《心理测量学》期刊发表论文,将三维数组近似为



图 7.50 塔克
(7.46)

$$\hat{x}_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R \sigma_{pqr} u_i^p v_j^q w_k^r$$

其中, $P \leq I, Q \leq J, R \leq K$ 是给定的。我们的目标是寻找秩为 (P, Q, R) 的张量 $\Sigma_{P \times Q \times R} = (\sigma_{pqr}) \in \mathbb{R}^{P \times Q \times R}$ 和正交矩阵 $\mathbf{U}_{I \times P}, \mathbf{V}_{J \times Q}, \mathbf{W}_{K \times R}$, 使得式 (7.46) 是对张量 $\mathbf{X}_{I \times J \times K} = (x_{ijk})$ 的最佳逼近, 即误差平方和 $\sum_{i,j,k} (x_{ijk} - \hat{x}_{ijk})^2$ 最小。秩为 (P, Q, R) 的塔克分解简记作

$$\mathbf{X}_{(P,Q,R)} = [\Sigma_{P \times Q \times R}; \mathbf{U}_{I \times P}, \mathbf{V}_{J \times Q}, \mathbf{W}_{K \times R}] \quad (7.47)$$

按照爱因斯坦记法, $a^j b_j$ 表示 $a^1 b_1 + a^2 b_2 + \dots$, 即对重复的上下标 (这里, a^j 不是 a 的 j 次方! 而是第 j 个分量) 求和。于是, 式 (7.46) 可简记为

$$\hat{x}_{ijk} = \sigma_{pqr} u_i^p v_j^q w_k^r \quad (7.48)$$

从数据压缩与重构的角度, 塔克分解是对张量 \mathbf{X} 的有损压缩, 式 (7.48) 是对 \mathbf{X} 的重构。塔克分解 (7.47) 是从 $\mathbf{X}_{(I,J,K)} = [\Sigma_{I \times J \times K}; \mathbf{U}_{I \times I}, \mathbf{V}_{J \times J}, \mathbf{W}_{K \times K}]$ 裁剪而得的, 其直观解释见图 7.51, 如同埃卡特-杨近似是奇异值分解的一个“简化版本”一样。因此, 塔克分解 (7.46) 也被称为高阶奇异值分解 (higher order singular value decomposition, HOSVD), 也有类似埃卡特-杨近似 (图 7.26) 的最佳逼近。其中, $\mathbf{U}_{I \times P}$ 是正交矩阵 $\mathbf{U}_{I \times I}$ 的前 P 列, 简记作 \mathbf{U}_P ; $\mathbf{V}_{J \times Q}$ 是正交矩阵 $\mathbf{V}_{J \times J}$ 的前 Q 列, 简记作 \mathbf{V}_Q ; $\mathbf{W}_{K \times R}$ 是正交矩阵 $\mathbf{W}_{K \times K}$ 的前 R 列, 简记作 \mathbf{W}_R 。

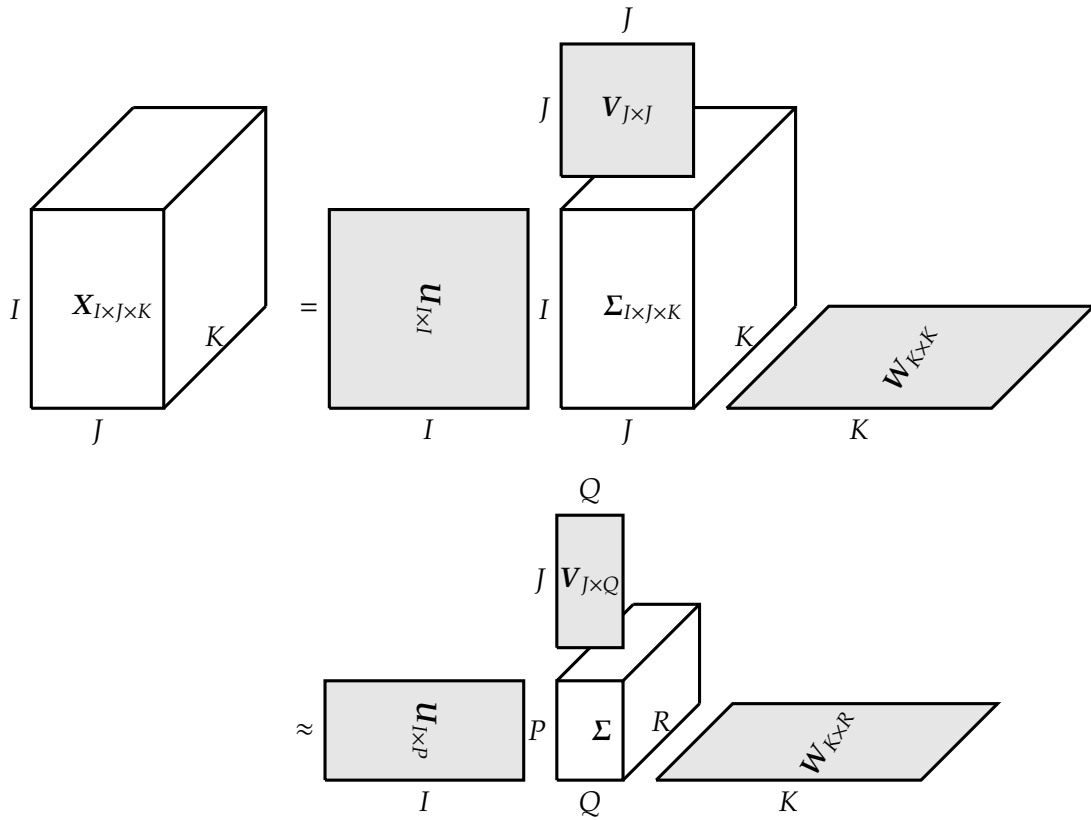


图 7.51 塔克分解与塔克近似

塔克分解并不惟一，基于 (7.48)，构造新的塔克分解如下：

$$\begin{aligned}\hat{x}_{ijk} &= (\sigma_{pqr} a_\alpha^p b_\beta^q c_\gamma^r) (u_i^p a_p^\alpha) (v_j^q b_q^\beta) (w_k^r c_r^\gamma) \\ &= \tilde{\sigma}_{\alpha\beta\gamma} \tilde{u}_i^\alpha \tilde{v}_j^\beta \tilde{w}_k^\gamma\end{aligned}$$

其中， $a_\alpha^p, b_\beta^q, c_\gamma^r$ 分别是非奇异矩阵 $a_p^\alpha, b_q^\beta, c_r^\gamma$ 的逆矩阵。

定义 7.10 (张量的矩阵展开) 对于三阶张量 $\mathbf{X}_{I \times J \times K}$ ，有三种常见的矩阵展开方式，分别记作 $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}$ 。以图 7.52 所示的张量为例，

13	16	19	22
14	17	20	23
15	18	21	24
1	4	7	10
2	5	8	11
3	6	9	12

图 7.52 按列将 $1, 2, \dots, 24$ 存为 $I = 3, J = 4, K = 2$ 的三阶张量

矩阵 $\mathbf{X}_1 = \begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix}$ 和矩阵 $\mathbf{X}_2 = \begin{pmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{pmatrix}$ “按列合并” 定义为

$$\mathbf{X}_1 \triangleleft \mathbf{X}_2 = \begin{pmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{pmatrix}$$

定义张量的三种矩阵展开，

$$\begin{aligned}\mathbf{X}^{(1)} &= \mathbf{X}_1 \triangleleft \mathbf{X}_2 \\ \mathbf{X}^{(2)} &= \mathbf{X}_1^\top \triangleleft \mathbf{X}_2^\top \\ &= \begin{pmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{pmatrix} \\ \mathbf{X}^{(3)} &= [\text{vec}(\mathbf{X}_1) \triangleleft \text{vec}(\mathbf{X}_2)]^\top \\ &= \begin{pmatrix} 1 & 2 & 3 & \cdots & 10 & 11 & 12 \\ 13 & 14 & 15 & \cdots & 22 & 23 & 24 \end{pmatrix}\end{aligned}$$

其中， $\text{vec}(\mathbf{A}_{m \times n})$ 是将 \mathbf{A} 的列向量首尾相接“拉直”为一个 mn 维列向量。

一般地, 对于张量 $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, 其第 n 种矩阵展开定义为

$$\mathbf{X}^{(n)} \in \mathbb{R}^{I_n} \times \mathbb{R}^{D/I_n}, \text{ 其中 } D = \prod_{j=1}^N I_j$$

定义 7.11 (张量积) 向量 $\mathbf{a} \in \mathbb{R}^I, \mathbf{b} \in \mathbb{R}^J, \dots, \mathbf{c} \in \mathbb{R}^K$ 的张量积 (tensor product) 是外积 (7.2) 的一般化, 定义为

$$\mathbf{X}_{I \times J \times \cdots \times K} = \mathbf{a} \circ \mathbf{b} \circ \cdots \circ \mathbf{c}, \text{ 其中 } x_{ij \cdots k} = a_i b_j \cdots c_k$$

塔克分解 (7.48) 可由张量积简单地表示为

$$\mathbf{X}_{(P,Q,R)} = \sigma_{pqr} \mathbf{u}^p \circ \mathbf{v}^q \circ \mathbf{w}^r$$

其中, $\mathbf{u}^p, \mathbf{v}^q, \mathbf{w}^r$ 分别是矩阵 $\mathbf{U}, \mathbf{V}, \mathbf{W}$ 的第 p, q, r 列。

算法 7.11 (高阶奇异值分解, HOSVD) 对于三阶张量 $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$, 其塔克分解 $\mathbf{X}_{(P,Q,R)}$ 可由下述方法求得。

- (1) 依次求得 \mathbf{X} 的矩阵展开 $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}$ 的前 P, Q, R 个左奇异向量*, 将这些矩阵分别记作 $\mathbf{U}, \mathbf{V}, \mathbf{W}$ 。显然,

$$\begin{aligned} \mathbf{U}^\top \mathbf{U} &= \mathbf{I}_P \\ \mathbf{V}^\top \mathbf{V} &= \mathbf{I}_Q \\ \mathbf{W}^\top \mathbf{W} &= \mathbf{I}_R \end{aligned}$$

- (2) 张量 $\Sigma_{P \times Q \times R}$ 按如下方法计算:

$$\begin{aligned} \Sigma_{P \times Q \times R} &= [\mathbf{X}_{I \times J \times K}; \mathbf{U}^\top, \mathbf{V}^\top, \mathbf{W}^\top] \\ &= x_{ijk} \tilde{\mathbf{u}}^i \circ \tilde{\mathbf{v}}^j \circ \tilde{\mathbf{w}}^k \end{aligned} \quad (7.49)$$

其中, $\tilde{\mathbf{u}}^i, \tilde{\mathbf{v}}^j, \tilde{\mathbf{w}}^k$ 分别是矩阵 $\mathbf{U}^\top, \mathbf{V}^\top, \mathbf{W}^\top$ 的第 i, j, k 列。

这个算法可以自然地推广到 n 阶张量, 只是要考虑更多的矩阵展开 $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}$ 。其他做法与算法 7.11 都是类似的。

*矩阵 $\mathbf{A}_{m \times n}$ 的左奇异向量即 $\mathbf{A}\mathbf{A}^\top$ 的单位本征向量。按照本征值的降序, 对应可得左奇异向量的序。这里, $\mathbf{X}^{(1)}$ 是一个 $I \times JK$ 矩阵, 它的前 P 个左奇异向量按列构成了矩阵 $\mathbf{U}_{I \times P}$ 。

定义 7.12 矩阵 $A_{m \times n} = (a_{ij})$ 与 $B_{p \times q}$ 的克罗内克积 (Kronecker product) 定义为

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}_{(mp) \times (nq)}$$

显然, 两个向量 x, y 的外积是克罗内克积的一个特例, 即

$$x \circ y = x \otimes y^T = y^T \otimes x$$

性质 7.11 克罗内克积不满足交换律, 但满足结合律。并且有

$$(A \otimes B)^T = A^T \otimes B^T$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}, \text{ 其中 } A, B \text{ 是可逆方阵}$$

$$|A_{m \times m} \otimes B_{p \times p}| = |A|^p |B|^m$$

$$\text{tr}(A \otimes B) = \text{tr}(A)\text{tr}(B)$$

$$\text{rank}(A \otimes B) = \text{rank}(A)\text{rank}(B)$$

性质 7.12 由算法 7.11 所得的塔克分解, 可以重构张量 X 的三个矩阵展开。

$$\hat{X}^{(1)} = U_{I \times P} \Sigma_{P \times QR}^{(1)} (W_{K \times R} \otimes V_{J \times Q})^T$$

$$\hat{X}^{(2)} = V_{J \times Q} \Sigma_{Q \times PR}^{(2)} (W_{K \times R} \otimes U_{I \times P})^T$$

$$\hat{X}^{(3)} = W_{K \times R} \Sigma_{R \times PQ}^{(3)} (V_{J \times Q} \otimes U_{I \times P})^T$$

HOSVD 算法并未使得 $\|X^{(1)} - \hat{X}^{(1)}\|_F$ 达到最小, 我们需要继续利用下述高阶正交迭代 (higher order orthogonal iteration, HOOI) 算法求得最优塔克分解。

算法 7.12 (高阶正交迭代, HOOI) 在 HOSVD 结果的基础上, 令 $\tilde{u}^i, \tilde{v}^j, \tilde{w}^k$ 分别是矩阵 U^T, V^T, W^T 的第 i, j, k 列。依次更新张量 U, V, W, Σ 如下,

□ 依次更新 U, V, W , 直至满足预定的收敛条件:

$$T_{I \times Q \times R} \leftarrow x_{ijk} \tilde{v}^j \circ \tilde{w}^k$$

$T_{I \times Q \times R}$ 是一个张量。更新 $U_{I \times P}$ 为 $T^{(1)}$ 的前 P 个左奇异向量。

$$T_{P \times J \times R} \leftarrow x_{ijk} \tilde{u}^i \circ \tilde{w}^k$$

更新 $V_{J \times Q}$ 为 $T^{(2)}$ 的前 Q 个左奇异向量。

$$T_{P \times Q \times K} \leftarrow x_{ijk} \tilde{u}^i \circ \tilde{v}^j$$

更新 $W_{K \times R}$ 为 $T^{(3)}$ 的前 R 个左奇异向量。

□ 当 U, V, W 不再更新时, 张量 $\Sigma_{P \times Q \times R}$ 按 (7.49) 更新。

例 7.19 彩色图像在计算机里是由红、绿、蓝三个基色的图像叠加而成。莱娜的彩色图像是 $512 \times 512 \times 3$ 的张量 X , 图 7.53 的第一行是这三种基色的原始图像 (即 $X^{(1)}$, 按灰度图像显示)。令 $P = Q = 100, R = 2$, 经过塔克分解, HOSVD 和 HOOI 的效果见第二、三行。



图 7.53 莱娜灰度图像的重构

图 7.54 的左图是莱娜彩色图像的原图，中图和右图是经过塔克分解 ($P = Q = 100, R = 2$), HOSVD 和 HOOI 的重构。HOOI 的效果似乎比 HOSVD 的要好一些。



图 7.54 莱娜彩色图像的重构

固定 $R = 3$ 或 1 ，分别设置 $P = Q = 50, 100, 150$ ，图 7.55 列出了更多莱娜彩色图像的塔克分解 HOOI 重构。不难看出， R 越大，对颜色保留的信息越多。 $P = Q$ 越大，图像的解析度越高。



(a) $R = 3, P = Q = 50, 100, 150$



(b) $R = 1, P = Q = 50, 100, 150$

图 7.55 莱娜彩色图像的 HOOI 重构