第二章　主成分分析

2.1　简 介

2.1 简 介

- 在实际问题的研究中, 往往涉及多变量, 且不同变量之间有一定的相关性. 当变量较多时, 将增加分析问题的复杂性. 由于变量较多且变量之间存在相关性, 使得观测到的数据在一定程度上有所重叠. 因此利用变量间信息的重叠, 通过变量的降维, 可以使得复杂问题得到简化.
- 主成分分析 (principal component analysis, PCA) 是利用降维的思想, 在尽量减少信息损失的前提下, 将多变量转化为少数几个综合变量的一种机器学习方法. 例如, 在商品经济中, 用主成分分析可以将复杂的经济数据综合成几个商业指数, 如物价指数、生活费用指数以及商业活动指数等. 主成分分析是由皮尔逊 (Pearson)[11] 提出, 后来被霍特林 (Hotelling)[12] 发展起来的. 通常将生成的综合变量称为主成分, 这些主成分保留原始变量的绝大部分信息,都是原始变量的线性组合, 而且各个主成分之间互不相关. 在研究复杂问题时, 通过主成分分析, 可以从事物错综复杂的关系中找出一些主要成分, 揭示事物内部变量之间的规律, 简化问题, 提高分析效率.

2.2　总体的主成分

2.2.1 总体主成分的定义

- 主成分分析的目标是找到原始变量的一个能够按照“重要性”排序并且信息不重复的线性组合. 具体地, 假设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ 为 p 维随机向量, 其均值为 $\boldsymbol{\mu}$, 协方差矩阵为 $\boldsymbol{\Sigma}$. 对 $\mathbf{X}$ 进行线性变换, 可以形成新的综合变量, 记为 $\boldsymbol{Y}=(Y_1, Y_2, \cdots, Y_p)^{\mathrm{T}}$, 即

$$\begin{cases} Y_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p = \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{X}, \\ Y_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p = \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{X}, \\ \vdots \\ Y_p = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p = \boldsymbol{a}_p^{\mathrm{T}}\boldsymbol{X}. \end{cases}$$

  - 首先用一个综合变量 Y_1 来替代原始的 p 个变量, 为了使得 Y_1 在 $X_1, X_2, \cdots, Xp$ 的所有线性组合中最具代表性, 应使其方差最大化, 以最大限度地保留原始变量的方差和协方差结构信息. 由于

$$\mathrm{Var}\left(Y_1\right) = \mathrm{Var}\left(\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{X}\right) = \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_1,$$

2.2.1 总体主成分的定义

- 若对 $\boldsymbol{a}_1$ 不加以约束, 可使得 Y_1 的方差任意增大, 那么方差最大化就变得没有意义. 因此主成分分析限制 $\boldsymbol{a}_1$ 为单位向量, 即 $\boldsymbol{a}^{\mathrm{T}}_1\,\boldsymbol{a}_1 = 1$, 寻求向量 $\boldsymbol{a}_1$, 使得 $\mathrm{Var}(Y_1) = \mathrm{Var}(\boldsymbol{a}^{\mathrm{T}}_1\,\boldsymbol{X})$ 达到最大, Y_1 就称为第一主成分. 若第一主成分所含信息不够多, 不足以代表原始数据的 p 个变量, 则需要考虑 Y_2, 为了使得Y_2 中所含信息与 Y_1 不重叠, 则应该要求

$$\mathrm{Cov}\left(Y_1, Y_2\right) = 0,$$

- 即 Y_1 和 Y_2 不相关. 因此主成分分析在上述约束和 $\boldsymbol{a}^{\mathrm{T}}_2\,\boldsymbol{a}_2 = 1$ 的条件下寻求$\boldsymbol{a}_2$, 使得 $\mathrm{Var}(Y_2) = \mathrm{Var}(\boldsymbol{a}^{\mathrm{T}}_2\,\boldsymbol{X})$ 达到最大, 所求的 Y_2 称为第二主成分. 类似地, 可以定义第三主成分 · · · · · 第 p 主成分. 各主成分在总方差中所占比重依次递减. 实际应用中, 通常只需挑选前几个主成分, 达到简化问题, 抓住问题本质的目的.

2.2.2 总体主成分的求法

- 本节将阐述求解 $\boldsymbol{X}$ 主成分的计算过程. 假设 $\boldsymbol{\Sigma}$ 是 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ 的协方差矩阵, $\boldsymbol{\Sigma}$ 的特征值及其相应的正交单位化特征向量分别为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > \lambda_{r+1} = \cdots = \lambda_p = 0$ 及 $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_p$, 其中 $r = \operatorname{rank}(\boldsymbol{\Sigma})$.
  - 首先求出 $\boldsymbol{X}$ 的第一个主成分 $Y_1 = \boldsymbol{a}^{\mathrm{T}}_1 \boldsymbol{X}$. 由于第一主成分的系数 $\boldsymbol{a}_1$ 应在条件 $\boldsymbol{a}^{\mathrm{T}}_1 \boldsymbol{a}_1 = 1$ 下, 使得 $\boldsymbol{X}$ 的所有线性变换中方差

$$\operatorname{Var}\left(\boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{X}\right) = \boldsymbol{a}_1^{T} \boldsymbol{\Sigma} \boldsymbol{a}_1$$

  - 最大化. 因此, 求第一主成分就转换为求解以下约束最优化问题:

$$\max_{\boldsymbol{a}_1} \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{a}_1, \ \text{s.t.} \quad \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{a}_1 = 1.$$

  - 根据拉格朗日乘子法, 定义拉格朗日函数

$$L(\boldsymbol{a}_1, \lambda) = \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{a}_1 - \lambda\left(\boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{a}_1 - 1\right),$$

2.2.2 总体主成分的求法

- 其中 λ 为拉格朗日乘子. 将拉格朗日函数 $L(\boldsymbol{a}_1, \lambda)$ 分别对参数 $\boldsymbol{a}_1, \lambda$ 求导, 令其为 0, 即得

$$\begin{cases} \boldsymbol{\Sigma a}_1 - \lambda \boldsymbol{a}_1 = 0, \\ \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{a}_1 = 1. \end{cases}$$

- 因此, λ 是协方差矩阵 Σ 的特征值, $\boldsymbol{a}_1$ 是其对应的单位特征向量. 可得目标函数

$$\boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{\Sigma a}_1 = \boldsymbol{a}_1^{\mathrm{T}} \lambda \boldsymbol{a}_1 = \lambda \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{a}_1 = \lambda.$$

- 因此若 $\boldsymbol{a}_1$ 是 $\boldsymbol{\Sigma}$ 的最大特征值 λ_1 对应的单位特征向量, 则 $\boldsymbol{a}_1$ 与 λ_1 是上述最优化问题的解. 即可得第一个主成分 $Y_1 = \boldsymbol{a}^{\mathrm{T}}{}_1 \boldsymbol{X}$, 其方差为协方差矩阵 $\boldsymbol{\Sigma}$的最大特征值 λ_1, 其系数 $\boldsymbol{a}_1$ 是 λ_1 对应的单位特征向量.

- 下面求解 $\boldsymbol{X}$ 的第二个主成分 $Y_2 = \boldsymbol{a}^{\mathrm{T}}{}_2 \boldsymbol{X}$. 由于第二个主成分的系数 $\boldsymbol{a}_2$应满足以下条件：单位向量 $\boldsymbol{a}^{\mathrm{T}}{}_2 \boldsymbol{a}_2 = 1$, 且 $Y_2 = \boldsymbol{a}^{\mathrm{T}}{}_2 \boldsymbol{X}$ 与 $Y_1 = \boldsymbol{a}^{\mathrm{T}}{}_1 \boldsymbol{X}$ 不相关,并使得 $\boldsymbol{X}$ 的所有线性变换中方差

$$\mathrm{Var}\left(\boldsymbol{a}_2^{\mathrm{T}} \boldsymbol{X}\right) = \boldsymbol{a}_2^{\mathrm{T}} \boldsymbol{\Sigma a}_2$$

- 达到最大.

2.2.2 总体主成分的求法

- 因此, 求第二主成分就转换为求解以下约束最优化问题:

$$\max_{a_2} \ \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_2, \ \text{s.t.} \ \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_2 = 1, \quad \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_1 = 0.$$

- 由于

$$\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_2 = \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_1 = \boldsymbol{a}_2^{\mathrm{T}}\lambda_1\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_2,$$

- 则有

$$\boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_1 = 0, \quad \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_2 = 0.$$

- 定义拉格朗日函数

$$L(\boldsymbol{a}_2, \lambda, \phi) = \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_2 - \lambda(\boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_2 - 1) - \phi\boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_1,$$

- 其中 λ, ϕ 为拉格朗日乘子. 将拉格朗日函数 $L(\boldsymbol{a}_2, \lambda, \phi)$ 分别对参数 $\boldsymbol{a}_2, \lambda, \phi$求导, 令其为 0, 即得

$$\begin{cases} 2\boldsymbol{\Sigma}\boldsymbol{a}_2 - 2\lambda\boldsymbol{a}_2 - \phi\boldsymbol{a}_1 = 0, \\ \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_2 = 1, \\ \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_1 = 0. \end{cases}$$

2.2.2 总体主成分的求法

- 由于

$$2\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_2 - 2\lambda\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_2 - \phi\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_1 = 0,$$

- 则 $\phi = 0$, 进而可以得到

$$\boldsymbol{\Sigma}\boldsymbol{a}_2 - \lambda\boldsymbol{a}_2 = 0.$$

- 显然, λ 是协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值, $\boldsymbol{a}_2$ 是其对应的单位特征向量. 此时, 目标函数可表示为

$$\boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{a}_2 = \boldsymbol{a}_2^{\mathrm{T}}\lambda\boldsymbol{a}_2 = \lambda\boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_2 = \lambda,$$

- 因此若$\boldsymbol{a}_2$ 是$\boldsymbol{\Sigma}$ 的第二大特征值 λ_2 对应的单位特征向量, 则$\boldsymbol{a}_2$ 与 λ_2 是上述最优化问题的解. 即可得第二个主成分 $Y_2=\boldsymbol{a}^{\mathrm{T}}{}_2\,\boldsymbol{X}$, 其方差为协方差矩阵Σ 的第二大特征值 λ_2, 系数向量 $\boldsymbol{a}_2$ 是 λ_2 对应的单位特征向量.

- 以此类推, 可知第 k 个主成分 $Y_k = \boldsymbol{a}^{\mathrm{T}}{}_k\,\boldsymbol{X}$, 其方差为协方差矩阵 $\boldsymbol{\Sigma}$ 的第k 大特征值 λ_k, 系数向量 $\boldsymbol{a}_{\mathrm{k}}$ 是 λ_k 对应的单位特征向量.

2.2.2 总体主成分的求法

- 因此, 假设 $\boldsymbol{X}$ 的第 k 个主成分为

$$Y_k = \boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{X} = a_{k1} X_1 + a_{k2} X_2 + \cdots + a_{kp} X_p,$$

  - 其中 $\boldsymbol{a}_k = (\boldsymbol{a}_{k1}, \cdots, \boldsymbol{a}_{kp})^{\mathrm{T}}$. 显然有:

$$\begin{cases} \mathrm{Var}(Y_k) = \boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{a}_k = \lambda_k \boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{a}_k = \lambda_k, k = 1,2,\cdots,p, \\ \mathrm{Cov}(Y_k, Y_j) = \boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{a}_j = \lambda_k \boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{a}_j = 0, k \neq j. \end{cases}$$

  - 即令 $\boldsymbol{A} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_p)$, 则 $\boldsymbol{A}$ 是一个正交矩阵, 且 $\boldsymbol{A}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{A} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$, 其中 $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$ 表示对角矩阵. 因此, 求主成分问题就转化成了求协方差矩阵的特征值和特征向量

2.2.3 总体主成分的性质

- 1. 主成分的协方差矩阵 $\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{\Lambda}$. 即 $\mathrm{Var}(Y_j) = \lambda_j, j = 1, 2, \cdots, p$ 且$Y_1, Y_2, \cdots, Y_p$ 互不相关.
- 2. 假设 $\boldsymbol{\Sigma} = (\sigma_{jk})_{p\times p}$ 表示 $\boldsymbol{X}$ 的协方差矩阵, 则总体主成分的方差之和可表示为

$$\sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \sigma_{jj}.$$

  - 事实上, 由于 $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\mathrm{T}}$, 则

$$\sum_{j=1}^{p} \lambda_j = \mathrm{tr}\left(\boldsymbol{\Lambda}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right) = \mathrm{tr}\left(\boldsymbol{\Lambda}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right) = \mathrm{tr}\left(\boldsymbol{\Sigma}\right) = \sum_{j=1}^{p} \sigma_{jj}.$$

  - 由此可知, 主成分分析是把 p 个随机变量 $X_1, X_2, \cdots, X_p$ 的总方差分解为 p个不相关的随机变量 $Y_1, Y_2, \cdots, Y_p$ 的方差之和.

2.2.3 总体主成分的性质

- 在主成分分析中, 令 η_k 表示第 k 个主成分的方差贡献率, 定义为

$$\eta_k = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}, k = 1,2,\cdots,p,$$

- 其含义是第 k 个主成分 Y_k 所提取的信息占总信息的比例. 根据主成分分析的算法原理, 第一主成分的贡献率最大, 意味着 Y_1 综合原始变量 $X_1, X_2, \cdots, X_p$ 所含的信息能力最强, 而 $Y_2, Y_3, \cdots, Y_p$ 的综合能力依次减弱.

- 前 m 个主成分 $Y_1, Y_2, \cdots, Y_m$ 的方差贡献率之和定义为

$$\sum_{j=1}^{m} \eta_j = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j},$$

- 表示主成分 $Y_1, Y_2, \cdots, Y_m$ 的累积贡献率, 其含义是前 m 个主成分综合提供原始变量信息的能力. 在实际应用中, 通常选取 $m < p$, 使得前 m 个主成分的累积贡献率达到较高的比例 (例如, 大于 85%). 此时, 用 $Y_1, Y_2, \cdots, Y_m$ 代替原始随机变量 $X_1, X_2, \cdots, X_p$ 不但使得变量的维数降低, 而且也不损失太多的信息.

2.2.3 总体主成分的性质

- 3. 设矩阵 $\boldsymbol{A}$ 的第 k 行第 j 列元素为 $\boldsymbol{A}_{kj}$. 由于 $\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}$, 则有$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{Y}$, 故而有 $X_j=A_{j1}Y_1+A_{j2}Y_2+\cdots+A_{jp}Y_p$, $\mathrm{Cov}(Y_k, X_j)=\lambda_k A_{jk}$,则可得主成分 Y_k 与原始变量 X_{j} 的相关系数为

$$\rho Y_k, X_j=\frac{\mathrm{Cov}\left(Y_k, X_j\right)}{\sqrt{\mathrm{Var}\left(Y_k\right)}\sqrt{\mathrm{Var}\left(X_j\right)}}=\frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}}A_{jk}.$$

  - 它给出了主成分 Y_{k} 与原始变量 X_{j} 的线性关联性的度量, 也称为因子负荷量或因子载荷量.

- 4. 之前所提到的累积贡献率度量了前 m 个主成分 $Y_1, Y_2, \cdots, Y_m$ 综合提供原始变量 $X_1, X_2, \cdots, X_p$ 所含的信息能力, 那么前 m 个主成分中包含原始变量 Xj 有多少信息应该如何度量呢？这个指标为前 m 个主成分 $Y_1, Y_2, \cdots, Y_m$ 与原始变量 X_j 的相关系数的平方和, 我们称之为 $Y_1, Y_2, \cdots, Y_m$ 对原始变量 X_j 的贡献率.

2.2.3 总体主成分的性质

下面我们通过一个例子阐述总体主成分的计算方法.

例 2.1 设随机变量 $\boldsymbol{X}=(X_1, X_2, X_3)^{\mathrm{T}}$ 的协方差矩阵为

$$\boldsymbol{\Sigma}=\begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

则知 Σ 的特征值为 $\lambda_1=3+\sqrt{8}, \lambda_2=2, \lambda_3=3-\sqrt{8}$, 相应的单位正交特征向量为

$$\boldsymbol{a}_1=\begin{pmatrix} 0.383 \\ -0.924 \\ 0.000 \end{pmatrix}, \quad \boldsymbol{a}_2=\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \boldsymbol{a}_3=\begin{pmatrix} 0.924 \\ 0.383 \\ 0.000 \end{pmatrix}.$$

2.2.3 总体主成分的性质

- 因此, 主成分为

$$Y_1 = 0.383X_1 - 0.924X_2,$$
$$Y_2 = X_3,$$
$$Y_3 = 0.924X_1 + 0.383X_2.$$

- 取 $m = 1$ 时, Y_1 的累积贡献率为 $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{3+\sqrt{8}}{8} = 72.8\%$; 取 $m = 2$时, Y_1, Y_2 的累积贡献率为 97.85%. 下表列出 m 个主成分对变量 X_i 的贡献率.

j	ρ_{Y_1,X_j}	$\rho^2_{Y_1,X_j}$	ρ_{Y_2,X_j}	$\rho^2_{Y_2,X_j}$
1	0.925	0.855	0.000	0.000
2	−0.998	0.996	0.000	0.000
3	0.000	0.000	1.000	1.000

- 由此可见, 当 $m = 1$ 时, Y_1 的累积贡献率已达 72.8%, 但是 Y_1 对 X_3 的贡献率为 0, 这是因为在 Y_1 中没有包含 X_3 的任何信息, 因此仅取 $m = 1$ 不够,故而取 $m = 2$, 这时 Y_1, Y_2 的累积贡献率为 97.85%, 且 Y_1, Y_2 对 X_j ($j = 1, 2, 3$)的贡献率也比较高.

2.2.4 标准化变量的主成分

在实际问题中, 通常有两种情形不适合直接从协方差矩阵 $\boldsymbol{\Sigma}$ 出发求主成分. 一种是各变量的单位不全相同, 对同样的变量使用不同的单位进行主成分分析, 其结果一般是不一样的, 甚至差异较大, 这样做出来的分析也没有意义. 另一种是各变量的单位虽相同, 但是其变量方差的差异较大, 以至于主成分分析的结果往往倾向于方差大的变量, 而方差小的变量几乎被忽略. 因此,对这两种情形, 通常先将各原始变量做标准化处理, 然后从标准化变量的协方差矩阵出发求主成分. 常用的标准化变换为

$$X_j^* = \frac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}, i = 1,2,\cdots,p,$$

- 其中 $\mu_j = E(X_j)$, $\sigma_{jj} = \mathrm{Var}(X_j)$. 此时 $\boldsymbol{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)^{\mathrm{T}}$ 的协方差矩阵为原始变量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ 的相关系数矩阵 $\boldsymbol{\rho} = (\rho_{kj})_{p\times p}$, 其中

$$\rho_{kj} = \frac{\mathrm{Cov}(X_k, X_j)}{\sqrt{\mathrm{Var}(X_k)}\sqrt{\mathrm{Var}(X_j)}},$$

2.2.4 标准化变量的主成分

- 因此只需直接从相关系数矩阵 $\boldsymbol{\rho}$ 出发求主成分, 此时的主成分分析将均等地对待每一个原始变量. 从相关系数矩阵出发求的主成分与从协方差矩阵出发是完全类似的, 并且主成分的一些性质具有更简单的数学形式.

设 $\boldsymbol{\rho}$ 的特征值 $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_r^* \geq \lambda_{r+1}^* = \cdots = \lambda_p^* = 0$ 其中 $r = \text{rank}(\boldsymbol{\rho})$, $\boldsymbol{\rho}$ 的 p 个单位特征向量为 $a_1^*, \cdots, a_p^*$, 且相互正交, 则 p 个主成分为: $Y_1^* = a_1^{*\mathrm{T}} \boldsymbol{X}^*, Y_2^* = a_2^{*\mathrm{T}} \boldsymbol{X}^*, \cdots, Y_p^* = a_p^{*\mathrm{T}} \boldsymbol{X}^*$. 记 $\boldsymbol{Y}^* = \left(Y_1^*, Y_2^*, \cdots, Y_p^*\right)^{\mathrm{T}}, \boldsymbol{A}^* = \left(a_1^*, a_2^*, \cdots, a_p^*\right)$, , 则有 $\boldsymbol{Y}^* = \boldsymbol{A}^{*\mathrm{T}} \boldsymbol{X}^*$.上述主成分具有的性质可以概括如下:

- 1. $E\left(\boldsymbol{Y}^*\right) = 0$, $\text{Var}\left(\boldsymbol{Y}^*\right) = \boldsymbol{\Lambda}^* = \text{diag}\left(\lambda_1^*, \cdots \lambda_p^*\right)$.
- 2. $\sum_{j=1}^{p} \text{Var}\left(Y_j^*\right) = \sum_{j=1}^{p} \lambda_j^* = \sum_{j=1}^{p} \text{Var}\left(X_j^*\right) = p$.
- 3.第 k 个主成分 Y_k^* 的贡献率为 $\frac{\lambda_k^*}{p}$ 前 m 个主成分 Y_1^* , Y_2^* , $\cdots$, Y_m^* 的累积贡献率为 $\frac{\sum_{j=1}^{m} \lambda_j^*}{p}$.
- 4. 主成分 Y_k^* 与 X_j^* 的相关系数为 $\rho_{Y_k^*, X_j^*} = \sqrt{\lambda_k^*} A_{jk}^*$.

2.2.4 标准化变量的主成分

- 下面通过一个例子说明分别从协方差矩阵和相关系数矩阵出发求主成分的差异.

- **例 2.2** 随机变量 $\boldsymbol{X}=(X_1, X_2, X_3)^{\mathrm{T}}$ 的协方差矩阵为

$$\boldsymbol{\Sigma}=\begin{pmatrix} 16 & 2 & 30 \\ 2 & 1 & 4 \\ 30 & 4 & 100 \end{pmatrix},$$

  - 其相关系数矩阵为

$$\boldsymbol{\rho}=\begin{pmatrix} 1 & 0.5 & 0.75 \\ 0.5 & 1 & 0.4 \\ 0.75 & 0.4 & 1 \end{pmatrix}.$$

2.2.3 总体主成分的性质

- 经计算可知 Σ 的特征值为 $\lambda_1 = 109.793, \lambda_2 = 6.469, \lambda_3 = 0.738$, 相应的单位正交特征向量为

$$\boldsymbol{a}_1 = \begin{pmatrix} 0.305 \\ 0.041 \\ 0.951 \end{pmatrix}, \quad \boldsymbol{a}_2 = \begin{pmatrix} 0.944 \\ 0.120 \\ -0.308 \end{pmatrix}, \quad \boldsymbol{a}_3 = \begin{pmatrix} -0.127 \\ 0.992 \\ -0.002 \end{pmatrix}.$$

- 因此, 主成分为

$$\begin{aligned} Y_1 &= 0.305X_1 + 0.041X_2 + 0.951X_3, \\ Y_2 &= 0.944X_1 + 0.120X_2 - 0.308X_3, \\ Y_3 &= -0.127X_1 + 0.992X_2 - 0.002X_3. \end{aligned}$$

- Y_1 的贡献率为 $\dfrac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0.938$. 如此高的贡献率归因于 X_1 的大方差, 以及 X_1, X_2, X_3 之间存在一定的相关性.

2.2.3 总体主成分的性质

- 相关系数矩阵 $\boldsymbol{\rho}$ 的特征值为 $\lambda_1^* = 2.114, \lambda_2^* = 0.646, \lambda_3^* = 0.240$ 相应的单位正交特征向量为

$$\boldsymbol{a}_1^* = \begin{pmatrix} 0.627 \\ 0.497 \\ 0.600 \end{pmatrix}, \quad \boldsymbol{a}_2^* = \begin{pmatrix} -0.241 \\ 0.856 \\ -0.457 \end{pmatrix}, \quad \boldsymbol{a}_3^* = \begin{pmatrix} -0.741 \\ 0.142 \\ -0.656 \end{pmatrix}.$$

- 因此, 主成分为

$$\begin{aligned} Y_1^* &= 0.627X_1^* + 0.497X_2^* + 0.600X_3^*, \\ Y_2^* &= -0.241X_1^* + 0.856X_2^* - 0.457X_3^*, \\ Y_3^* &= -0.741X_1^* + 0.142X_2^* + 0.656X_3^*. \end{aligned}$$

- Y_1^* 的贡献率为 $\frac{\lambda_1^*}{3} = 0.705$, Y_1^* 和 Y_2^* 的累积贡献率为 $\frac{\lambda_1^* + \lambda_2^*}{3} = 0.920$. 比较从 $\boldsymbol{\Sigma}$ 出发和从 $\boldsymbol{\rho}$ 出发的主成分分析结果. 可知从 ρ 出发的 Y_1^* 的贡献率0.705 明显小于从 $\boldsymbol{\Sigma}$ 出发的 Y_1 的贡献率 0.938, 事实上, 原始变量方差之间的差异越大, 这一点往往越明显. 此例也说明标准化后的结论可能会发生很大的变化, 因此标准化并不是无关紧要的.

2.3　样本主成分

2.3 样本主成分

- 在上一节, 我们可以从协方差矩阵 $\boldsymbol{\Sigma}$ 或相关系数矩阵 $\boldsymbol{\rho}$ 出发求主成分.但是在实际问题中, $\boldsymbol{\Sigma}$ 和 $\boldsymbol{\rho}$ 一般都是未知的, 需要通过样本来进行估计得到.设

$$\boldsymbol{X}_i = \left(X_{i1}, \cdots, X_{ip}\right)^{\mathrm{T}}, i = 1,2, \cdots, n$$

- 为取自样本数据矩阵 $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^{\mathrm{T}}$ 的一个简单随机样本, 其中 p 表示特征维数, n 表示样本数. 样本的协方差矩阵和相关系数矩阵分别为

$$\boldsymbol{S} = \left(s_{\boldsymbol{kj}}\right)_{\boldsymbol{p}\times\boldsymbol{p}} = \frac{1}{n-1}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}\left(\boldsymbol{X}_{\boldsymbol{i}} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_{\boldsymbol{i}} - \overline{\boldsymbol{X}}\right)^{T},$$

$$\boldsymbol{R} = \left(r_{kj}\right)_{p\times p} = \frac{8_{kj}}{\sqrt{8_{kk}8_{jj}}},$$

2.3 样本主成分

- 其中

$$\overline{\boldsymbol{X}}=\left(\overline{X}_1,\cdots,\overline{X}_{\boldsymbol{p}}\right)^{\mathrm{T}},\qquad \overline{X}_{\boldsymbol{j}}=\sum_{i=1}^{n}X_{ij},\qquad j=1,2,\cdots,p,$$

$$s_{kj}=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ik}-\overline{X}_k\right)\left(X_{ij}-\overline{X}_j\right)^{\mathrm{T}}\qquad k,j=1,2,\cdots,p$$

- 分别以 $\boldsymbol{S},\boldsymbol{R}$ 作为 $\boldsymbol{\Sigma},\boldsymbol{\rho}$ 的估计, 再按照总体主成分的方法求得的主成分称之为样本主成分.

- 设 $\hat{\lambda}_1\geq\hat{\lambda}_2\geq\cdots\geq\hat{\lambda}_r>\hat{\lambda}_{r+1}=\cdots=\hat{\lambda}_p=0$ 为样本协方差矩阵 $\boldsymbol{S}$ 的特征值, $\hat{\boldsymbol{a}}_1,\cdots,\hat{\boldsymbol{a}}_p$ 为对应的正交单位化特征向量, 其中 $\hat{\boldsymbol{a}}_1=\left(\hat{a}_{1i},\cdots,\hat{a}_{pi}\right)^{\mathrm{T}}$, 则第 m 个样本的第 j 个主成分可表示为 $Y_{mj}=\hat{a}_j^{\mathrm{T}}\boldsymbol{X}_m$. 此时, 可以得到

- 1. Y_j 的样本方差为

$$\mathrm{Var}\left(Y_j\right)=\hat{\boldsymbol{a}}_j^{\mathrm{T}}\boldsymbol{S}\hat{\boldsymbol{a}}_j=\hat{\lambda}_j,\qquad j=1,2,\cdots,p.$$

2.3 样本主成分

- 2. Y_k 与 Y_j 的样本协方差为

$$\mathrm{Cov}\left(Y_k, Y_j\right) = \hat{\boldsymbol{a}}_k^{\mathrm{T}} \boldsymbol{S} \hat{\boldsymbol{a}}_j = 0.$$

- 3. 样本总方差为

$$\sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \hat{\lambda}_j.$$

- 第 j 个主成分的贡献率为

$$\hat{\eta}_j = \frac{\hat{\lambda}_j}{\sum_{j=1}^{p} \hat{\lambda}_j}, \qquad j = 1,2,\cdots,p,$$

- 前 k 个主成分累积贡献率为

$$\sum_{j=1}^{p} \hat{\eta}_j = \frac{\sum_{j=1}^{k} \hat{\lambda}_j}{\sum_{j=1}^{p} \hat{\lambda}_j}.$$

2.3 样本主成分

- 类似地, 为了避免单位不统一或者变量之间差异性较大产生的影响, 可以对样本进行标准化处理, 即令

$$\boldsymbol{X}_i^* = \left(\frac{X_{i1} - \overline{X}_1}{\sqrt{s_{11}}}, \frac{X_{i2} - \overline{X}_2}{\sqrt{s_{22}}}, \cdots, \frac{X_{ip} - \overline{X}_p}{\sqrt{s_{pp}}} \right)^{\mathrm{T}}, \quad i = 1,2,\cdots,n,$$

  - 标准化后数据的样本协方差矩阵即为原始数据的样本相关系数矩阵 $\boldsymbol{R}$. 由 $\boldsymbol{R}$出发所求的样本主成分称为标准化样本主成分. 计算 $\boldsymbol{R}$ 的特征值及相应的正交单位化特征向量即可求得标准化样本主成分. 此时, 标准化的样本总方差为 p.

- 选取前 m 个样本主成分, 使其累积贡献率达到一定的要求 (例如, 大于85%), 用这 m 个样本主成分代替原始数据进行分析, 可以达到在保留大部分信息的前提下, 降低原始数据维数的目的.

2.4　非线性主成分分析

2.4 非线性主成分分析

- 传统主成分分析一般适应于线性降维, 其对于非线性数据往往不能达到较好的效果, 例如, 不同人之间的人脸图像存在非线性关系, 用传统的线性主成分分析结果不尽人意. 下面介绍几种非线性主成分分析算法

2.4.1 核主成分分析

- 核主成分分析 (kernel principal component analysis, KPCA) 是对传统主成分分析 (PCA) 算法的非线性拓展. 简单地说, 通过将非线性不可分问题映射到维度更高的特征空间, 使其在新的特征空间上线性可分. 设样本数据矩阵 $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^{\mathrm{T}}$, 其中 p 表示特征维数, n 表示样本数. $\boldsymbol{X} \in \mathbf{R}^p$ 为取自 $\mathbf{X}$ 的一个简单随机样本, 为了将其映射到维度更高的 k 维子空间, 定义如下非线性映射函数 ϕ:

$$\phi : \mathbf{R}^{p} \rightarrow \mathbf{R}^{k} \left(p \ll k \right)$$

- 换句话说, 利用 KPCA, 可以通过非线性变换将数据映射到一个高维空间, 然后在此高维空间中使用标准 PCA 将其映射到另外一个低维空间中, 并通过线性分类器进行划分. 但是, 由于协方差矩阵中每个元素都是向量的内积, 因此映射到高维度空间后, 向量维度增加导致计算量大幅度增大. 故而,可以利用核函数忽略映射函数的具体形式, 直接得到低维数据映射到高维后的内积.

2.4.1 核主成分分析

- 假设 $\phi(\mathbf{X})$ 是一个映射后的中心化的矩阵, 维数是 $n \times k$, 可以计算得到协方差矩阵为

$$\Sigma = \frac{1}{n}\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X}).$$

  - 然而, 由于我们没有显式的定义映射 ϕ, 无法计算 $\boldsymbol{\Sigma}$, 传统的 PCA 算法失效.但是, 由 $\frac{1}{n}\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})v = \lambda v$,两边同除以 λ, 得到$\boldsymbol{V} = \frac{1}{n\lambda}\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})v = \phi(\mathbf{X})^{\mathrm{T}}\boldsymbol{a}$, 因此, 特征向量可表示为 $\boldsymbol{v} = \phi(\mathbf{X})^{\mathrm{T}}\boldsymbol{a}$, 代入 $\boldsymbol{\Sigma v} = \lambda \boldsymbol{v}$, 则可得

$$\frac{1}{n}\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X})\phi(\mathbf{X})^{\mathrm{T}}\boldsymbol{a} = \lambda\phi(\mathbf{X})^{\mathrm{T}}\boldsymbol{a},$$

  - 两边左乘 $\phi(\boldsymbol{X})$, 则可得

$$\frac{1}{n}\phi(\mathbf{X})\phi(\mathbf{X})^{\mathrm{T}}\boldsymbol{a} = \lambda\boldsymbol{a},$$

  - 即

$$\frac{1}{n}\boldsymbol{K}\boldsymbol{a} = \lambda\boldsymbol{a}, \qquad (2.4.1)$$

2.4.1 核主成分分析

- 其中 $\boldsymbol{K}$ 为核矩阵：$\boldsymbol{K}=\phi(\mathbf{X})\phi(\mathbf{X})^{\mathrm{T}}$. 显然, $\boldsymbol{K}$ 由高维空间中的内积决定, 为了避免由此带来的复杂计算, 我们不需要显式定义映射 $\phi(\mathbf{X})$, 可通过定义核函数表示样本点在高维空间中的内积, 这就是核技巧. 令 $k(\boldsymbol{X},\boldsymbol{Y})$ 表示核函数, $\boldsymbol{X}=(X_1,\cdots,X_p)^{\mathrm{T}}$、$\boldsymbol{Y}=(Y_1,\cdots,Y_p)^{\mathrm{T}}$ 表示样本. 此时矩阵 $\boldsymbol{K}$ 的第 i行, 第 j 列元素 $K_{ij}=k(\boldsymbol{X}_i,\boldsymbol{Y}_j)$. 由单位特征向量的假定知 $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=1$, 推出$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{a}=1$, 因此得到条件

$$n\lambda\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}=1. \tag{2.4.2}$$

- 利用式 (2.4.1) 和条件 (2.4.2) 可以求解出未知向量 $\boldsymbol{a}$, 以及对应的特征值和特征向量. 接下来, 对于一个新的样本 $\boldsymbol{X}$, 我们可以得到它的第一主成分是

$$\phi(\boldsymbol{X})\boldsymbol{v}_1=\sum_{i=1}^{n}\boldsymbol{a}_i k(\boldsymbol{X},\boldsymbol{X}_i)$$

- 其中 $\boldsymbol{v}_1$ 是最大特征值对应的特征向量.

2.4.1 核主成分分析

- 常用的核函数有：
  - 1. 线性核函数：

$$k(\boldsymbol{X},\boldsymbol{Y})=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+c,$$ 其中 c 为参数.

  - 2. 多项式核函数：

$$k(\boldsymbol{X},\boldsymbol{Y})=\left(a\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+c\right)^{d},$$ 其中 a, b, c 为参数.

  - 3. 高斯核函数：

$$k(\boldsymbol{X},\boldsymbol{Y})=\exp\left(-\frac{\|\boldsymbol{X}-\boldsymbol{Y}\|^{2}}{2\sigma^{2}}\right),$$

其中 σ 为参数, 高斯核函数是径向基函数核的一个典型代表.

  - 4. 指数核函数：

$$k(\boldsymbol{X},\boldsymbol{Y})=\exp\left(-\frac{\|\boldsymbol{X}-\boldsymbol{Y}\|}{2\sigma^{2}}\right),$$

其中 $\boldsymbol{\sigma}$ 为参数. 指数核函数也是径向基函数核代表, 与高斯核函数很像, 只是将 L_2 范数变成 L_1 范数.

2.4.1 核主成分分析

- 5. 拉普拉斯核函数：

$$k(\boldsymbol{X},\boldsymbol{Y})=\exp\left(-\frac{\|\boldsymbol{X}-\boldsymbol{Y}\|}{\sigma}\right).$$

拉普拉斯核函数完全等价于指数核, 区别在于前者对参数的敏感性降低, 也是一种径向基函数核.

- 值得注意的是, 上述理论均是在 $\phi(\mathbf{X})$ 已经中心化的前提下完成的. 在实际应用中, 应首先将矩阵 $\phi(\mathbf{X})$ 中心化, 即

$$\tilde{\phi}(\boldsymbol{X})=\phi(\boldsymbol{X})-1_n\cdot\phi(\boldsymbol{X})$$

- 其中 $1_n=\frac{1}{n}1_{n\times1}1_{n\times1}^{\mathrm{T}}$, 为 $n\times n$ 矩阵, 其每个元素为 $\frac{1}{n}$. 由上述可知, 不需要显示计算 $\tilde{\phi}(\mathbf{X})$, 只需得到中心化后的核矩阵即可：

$$\tilde{\boldsymbol{K}}=\tilde{\phi}(\mathbf{X})\tilde{\phi}(\mathbf{X})^{\mathrm{T}}=\boldsymbol{K}-\boldsymbol{K}\cdot1_n-1_n\cdot\boldsymbol{K}+1_n\cdot\boldsymbol{K}\cdot1_n.$$

- 因此上面介绍 KPCA 使用的 $\phi(\mathbf{X})$ 和 $\boldsymbol{K}$ 本质上就是这里的 $\tilde{\phi}(\mathbf{X})$和 $\tilde{\boldsymbol{K}}$.

2.4.2 *t*-SNE 非线性降维算法

- *t* 分布随机邻域嵌入 (t-distributed stochastic neighbor embedding,t-SNE) 是一种针对高维数据的非线性降维算法, 在 2008 年由 Laurens vander Maaten 和 Geoffrey Hinton [13] 提出. 传统主成分分析是一种线性算法,不能解释特征之间的复杂多项式关系. 而 *t*-SNE 是基于邻域图上随机游走的概率分布寻找数据内的结构, 将数据点之间的相似度转化为条件概率, 原始空间中数据点的相似度由正态分布表示, 嵌入空间中数据点的相似度由 *t* 分布表示. 通过原始空间和嵌入空间的联合概率分布的 Kullback Leibler(库尔贝克–莱布勒, KL) 散度 (用于评估两个分布的相似度的指标) 来评估嵌入效果的好坏.

1. SNE 算法

- *t*-SNE 算法是从 SNE 改进而来, 所以先介绍 SNE . 给定一组高维数据$\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^{\mathrm{T}}$, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^{\mathrm{T}}$. 目标是将这组数据降维到二维,SNE 的基本思想是如果两个数据在高维空间中是相似的, 那么降维到二维空间时距离应当较近.

2.4.2 *t*-SNE 非线性降维算法

1. SNE 算法

- 随机邻域嵌入 (SNE) 首先通过将数据点之间的高维欧几里得距离转换为相似性的条件概率来描述两个数据之间的相似性.

- 假设高维空间中的两个点 X_i, X_j , 以点 X_i 为中心构建方差为 σ_i 的高斯分布. 用 Pj|i 表示 X_j 在 X_i 邻域的概率, 若 X_j 与 X_i 相距很近, 则 $P_{j|i}$很大; 反之, $P_{j|i}$ 很小, $P_{j|i}$ 可以表示为

$$P_{j|i}=\frac{\exp\left(-\left\|X_j-X_i\right\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\left\|X_j-X_i\right\|^2/2\sigma_i^2\right)},\ P_{j|i}=0,$$

  - 其中 $||X_i-X_j||$表示点 X_i, X_j 的欧氏距离, 这里只关注不同数据点间的距离所以我们设置 $P_{i|i}=0$. 高斯核的带宽 σ_i 是条件概率中所涉及的范围. 有些特征点是稀疏的, 有些比较紧密, 因此带宽大小也是不同的. 一般来说, 数据密度高的区域带宽要小于数据密度低的区域.每个数据点的高斯核带宽的最优值可以通过简单的二进制搜索[14] 等得到

2.4.2 *t*-SNE 非线性降维算法

1. SNE 算法

- 当把数据映射到低维空间后, 高维数据点之间的相似性也应该在低维空间的数据点上体现出来. 假设$\boldsymbol{X}_i,\boldsymbol{X}_j$映射到低维空间后对应 $\boldsymbol{Y}_i,\boldsymbol{Y}_j$, 那么$\boldsymbol{Y}_i,\boldsymbol{Y}_j$ 邻域的条件概率为 $Q_{j|i}$:

$$Q_{j|i}=\frac{\exp\left(-\left\|\boldsymbol{Y}_j-\boldsymbol{Y}_i\right\|^2\right)}{\sum_{k\neq i}\exp\left(-\left\|\boldsymbol{Y}_k-\boldsymbol{Y}_i\right\|^2\right)}.$$

  - 低维空间中的方差直接设置为 $\sigma_i=\frac{1}{\sqrt{2}}$. 同样 $Q_{i|i}=0$.

- 如果条件概率 $Q_{j|i}$ 反映了高维数据点 $\boldsymbol{X}_i$, $\boldsymbol{X}_j$ 之间的关系, 那么我们希望条件概率 $P_{j|i}$ 与 $Q_{j|i}$ 应该完全相等. 若给定 $\boldsymbol{X}_i$ 与其他所有点之间的条件概率, 则可构成一个条件概率分布 $\mathcal{P}_i$. 同理在低维空间存在一个条件概率分布 $\mathcal{Q}_i$ 与之对应, 那么我们希望条件概率分布 $\mathcal{Q}_i$ 与 $\mathcal{P}_i$ 完全一样.

2.4.2 *t*-SNE 非线性降维算法

1. SNE 算法

- 为了衡量两个分布之间的相似性, 采用 KL 散度最小化低维与高维下两个条件概率分布的差异, SNE 最终目标就是对所有数据点最小化 KL 距离, 可以使用梯度下降算法最小化如下代价函数:

$$C=\sum_{i}\mathrm{KL}\left(\mathcal{P}_{i}\|\mathcal{Q}_{i}\right)=\sum_{i}\sum_{j}P_{j|i}\log\frac{P_{j|i}}{Q_{j|i}}.$$

- 但由于 KL 距离是一个非对称的度量, 这意味着当 $P_{j|i}$ 较大, $Q_{j|i}$ 较小时, 代价较高; 而 $P_{j|i}$ 较小, $Q_{j|i}$ 较大时, 代价较低. 即高维空间中两个数据点距离较近时, 若映射到低维空间后距离较远, 那么将得到一个很高的惩罚, 这符合我们的初衷. 反之, 高维空间中两个数据点距离较远时, 若映射到低维空间距离较近, 将得到一个很低的惩罚值, 我们的初衷是这里也应得到一个较高的惩罚. 即 SNE 的代价函数更关注局部结构, 而忽视了全局结构.

2.4.2 *t*-SNE 非线性降维算法

2. *t*-SNE 算法

- 在 SNE 中, 高维空间中条件概率 $P_{j|i}$ 不等于 $P_{i|j}$, 低维空间中 $Q_{j|i}$ 不等于 $Q_{i|j}$, 于是为简化计算提出对称 SNE, 使得 $P_{ij}=P_{ji}$, $Q_{ij}=Q_{ji}$, 优化$P_{i|j}$ 和 $Q_{i|j}$ 的 KL 散度的一种替换思路是, 使用联合概率分布来替换条件概率分布, 即 $\mathcal{P}$ 和 $\mathcal{Q}$ 分别是高维空间和低维空间里各个点的联合概率分布, 此时目标函数为

$$C=\mathrm{KL}\left(\mathcal{P}\|\mathcal{Q}\right)=\sum_i\sum_j P_{ij}\log\frac{P_{ij}}{Q_{ij}},$$

  - 其中, 在高维空间中定义

$$P_{ij}=\frac{P_{i|j}+P_{j|i}}{2}.$$

- 为了使得高维度下的中高等距离在映射后距离更大, 以减轻拥挤问题, 在低维空间中使用更加偏重长尾分布的方式将距离转换为概率分布.

2.4.2 *t*-SNE 非线性降维算法

2. *t*-SNE 算法

- 因此, 对于高维数据点 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 的低维对应点 $\boldsymbol{Y}_i$ 和 $\boldsymbol{Y}_j$, 采用自由度为 1 的归一化的 t 核:

$$Q_{ij}=\frac{\left(1+\left\|\boldsymbol{Y}_i-\boldsymbol{Y}_j\right\|^2\right)^{-1}}{\sum_{k\neq 1}\left(1+\left\|\boldsymbol{Y}_k-\boldsymbol{Y}_l\right\|^2\right)^{-1}},\quad Q_{ij}=0.$$

- 因此 t-SNE 是通过仿射变换将数据点映射到概率分布, 将两个数据点之间的距离转换为以一个点为中心一定范围内另一个点出现的条件概率. 由于在嵌入空间中的目标函数是非凸的, 因此, 可以用梯度下降最小化目标函数,此时

$$\frac{\partial C}{\partial \boldsymbol{Y}_i}=4\sum_{k\neq 1}\left(P_{ij}-Q_{ij}\right)Q_{ij}Z\left(\boldsymbol{Y}_i-\boldsymbol{Y}_j\right),$$

  - 其中, 在高维空间中定义 $Z=\sum_{k\neq 1}\left(1+\left\|\boldsymbol{Y}_k-\boldsymbol{Y}_l\right\|^2\right)^{-1}$.

2.4.2 *t*-SNE 非线性降维算法

2. *t*-SNE 算法

- 因此, *t*-SNE 的主要优势在于通过 *t* 分布与正态分布的差异, 解决了样本分布拥挤、边界不明显的问题, 使得相似的样本能够聚集在一起, 而差异大的样本能够有效地分开. 但它同时也有计算复杂度高, 目标函数非凸, 容易得到局部最优解; 对参数敏感、结果具有随机性等缺陷.

- *t*-SNE 非线性降维算法在实际问题中可用于处理高维数据集, 并应用于自然语言处理、图像处理、基因组数据和语音处理等问题中.

2.5　主成分分析实践

实践代码