对主成分分析法三个问题的剖析

许淑娜 李长坡
(许昌学院城市与环境学院,许昌,461000)

摘 要 从主成分分析法的基本原理入手,针对教学过程中学生对主成分分析法感到费解的三个问题进行了逐一剖析:1. 为什么主成分系数是经标准差标准化后原始变量的协方差矩阵的特征向量? 2. 特征向量正负号如何选取? 对进一步的研究如计算综合得分和聚类分析有何影响? 3. 主成分载荷值是如何得来的? 同时指出有些教材在计算主成分得分时混淆了主成分载荷和特征向量的概念,以致造成错误的结果.

关键词 主成分分析法 特征值 特征向量 主成分载荷 主成分得分

Dissection to Three Typical Issues of Principal Component Analysis

Xu Shuna Li Changpo
(College of Urban Planning and Environmental Science, XuChang University, Xuchang, China, 461000)

Abstract Starting from the basic principles of Principal Component Analysis(PCA), dissected the three issues which always puzzling students in the process of teaching one by one. The first one is, why the principal component coefficients is the eigenvectors of the covariance matrix of normalized original variables? And the second one, How to select the sign of eigenvectors? What is the impact on further studies such as the calculation of composite scores and cluster analysis? The third one, How the principal component loading values come from? Besides, confusion of the concept of principal component loading and eigenvectors in the process of calculating the principal component scores from some materials was pointed out, which would cause erroneous results.

Key words Principal Component Analysis Eigenvalue Eigenvectors Principal component loading Principal component scores

1 前言

主成分分析法(Principle Component Analysis)是一种重要的多元统计分析方法,已被广泛地应用与经济学、生物学、地球科学等领域. 然而,介绍主成分分析方法的诸多教材中,存在介绍过于简单、思路不清,甚至还有错误之处[1-2],这给教师的教学和学生的学习带来了困扰. 在近几年的教学实践中发现,主成分分析法这一节内容是教学过程中的难点,它的基本原理是很

收稿日期:2011年10月28日

容易理解的，然而它的求解过程并非同样简单易懂，学生往往感到费解. 在查阅了大量与该方法相关的线性代数资料，加上多年的教学经验，对该方法中的几个难点进行一一剖析，为深刻地理解该方法提供支持.

2 主成分分析法的原理

主成分分析法的原理是比较容易理解的，且在多本教材中都有较详细的介绍. 然而为了保持内容的连贯性，仍需对其做简要介绍.

我们在研究某一个问题时，为了研究地更全面、详尽而不遗漏重要信息，总是选取尽可能多的指标. 这就会带来这样的问题: 选取的指标过多，给研究带来一定困难，并且众多的指标之间可能存在一定的相关性，这样就造成了信息的重叠，给研究结果带来影响. 那么，能否通过原始众多指标之间的线性组合，用较少几个综合指标(主成分) 代替原来众多的原始指标，并且能解释原始指标大部分信息？ 这就是主成分分析法的基本原理.

设有 n 个样本，涉及到 m 个指标，用 $x_1,x_2,\cdots,x_m$ 表示; 它们的综合指标用 $z_1,z_2,\cdots,z_p(p\leqslant m)$ 来表示. 新的综合指标(设 $p=m$) 可由原始指标的线性组合表示.

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1m}x_m \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2m}x_m \\ \cdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mm}x_m \end{cases}$$

用矩阵形式表示:

$$X=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \qquad Z=\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix}$$

$$L=\begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_n \end{bmatrix}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nm} \end{bmatrix}$$

则
$$Z = LX \tag{1}$$

问题的关键在求出系数 l_{ij}，由下列原则决定:

① $z_i,z_j(i\neq j,\ i,j=1,2,\cdots,p)$ 互相无关;

② z_1 是 $x_1,x_2,\cdots,x_m$ 的所有线性组合中方差最大的; z_2 是与 z_1 不相关的 $x_1,x_2,\cdots,x_m$ 的所

有线性组合中方差最大的;z_m 是与 $z_1, z_2, \cdots, z_{m-1}$ 不相关的 $x_1, x_2, \cdots, x_m$ 的所有线性组合中方差最大的.

3 系数 L 的求解过程

设 X 为经过标准差标准化的值,即 x_j 的平均值 $\overline{x_j}(j = 1, 2, \cdots, m) = 0$;则第 j 个综合指标 Z_j 的平均值 $\overline{Z_j} = 0$(证: $\frac{1}{n}\sum_{i=1}^{n} z_{ji} = \frac{1}{n}(l_{j1}\sum_{i=1}^{n} x_{1i} + l_{j2}\sum_{i=1}^{n} x_{2i} + \cdots + l_{jm}\sum_{i=1}^{n} x_{mi} = 0(j = 1, 2, \cdots, m))$.

原始指标 $x_1, x_2, \cdots, x_m$ 之间的协方差矩阵为实对称阵 C,因为新的综合指标 Z 之间互不相关,所以它们之间的协方差矩阵应为对角阵 Λ:

$$C = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1m} \\ C_{21} & C_{22} & \cdots & C_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mm} \end{bmatrix} \qquad \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix}$$

由线性代数知识可知[3-4],若 C 为 m 阶实对称阵,则一定可以对角化,即有正交阵 P,使

$$P^{-1}CP = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix} \tag{2}$$

其中,Λ 对角线上的元素 $\lambda_1, \lambda_2, \cdots, \lambda_m$ 为 C 的特征值,P 的列向量是 C 的 m 个线性无关的特征向量.

证明:P 用列向量可表示为$(p_1, p_2, \cdots, p_m)$,由 $P^{-1}CP = \Lambda$ 可得:$CP = P\Lambda$,即

$$C(p_1, p_2, \cdots, p_m) = (p_1, p_2, \cdots, p_m)\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix} = (\lambda_1 p_1, \lambda_2 p_2, \cdots, \lambda_m p_m)$$

于是有:

$$Cp_i = \lambda_i p_i (i = 1, 2, \cdots, m)$$

根据方阵特征值与特征向量的定义可知,λ_i 是 C 的特征值,P 的列向量 p_i 就是对应于特征值 λ_i 的特征向量.

因为:

$$XX^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \times \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}\, x_{mi} \\ \sum_{i=1}^{n} x_{2i}\, x_{1i} & \sum_{i=1}^{n} x_{2i}^2 & \cdots & \sum_{i=1}^{n} x_{2i}\, x_{mi} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{mi}\, x_{1i} & \sum_{i=1}^{n} x_{mi}\, x_{2i} & \cdots & \sum_{i=1}^{n} x_{mi}^2 \end{bmatrix} = nC \tag{3}$$

同理:

$$ZZ^T = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix} \times \begin{bmatrix} z_{11} & z_{21} & \cdots & z_{m1} \\ z_{12} & z_{22} & \cdots & z_{m2} \\ \vdots & \vdots & \cdots & \vdots \\ z_{1n} & z_{2n} & \cdots & z_{mn} \end{bmatrix} = n\Lambda \tag{4}$$

将式(3)、**式**(4) **代入式**(2) **有:**

$$P^{-1} XX^T P = ZZ^T \tag{5}$$

又因 P **为正交阵,则有** $P^T = P^{-1}$**,因此有**

$$P^T XX^T P = ZZ^T \tag{6}$$

令 $L = P^T$**,则有**

$$LXX^T L^T = ZZ^T\text{,即}(LX)(LX)^T = ZZ^T \tag{7}$$

所以, $LX = Z$**,**L **的行向量是** C **的** m **个线性无关的特征向量. 求解** L **的问题转化为求** X **的协方差矩阵的特征向量问题,**X **经过标准差标准化后再求协方差矩阵,相当于直接求原始数据的相关系数矩阵,也等价于对标准差标准化后的数据求相关系数矩阵. 从几何的角度来理解,** $l_1, l_2, \cdots, l_m$ **是** m **维空间** V_m **的** m **个相互垂直的坐标轴,主成分** $z_i(i = 1, 2, \cdots, m)$ **是原始变量** $(x_1, x_2, \cdots, x_m)$ **在** $l_i(i = 1, 2, \cdots, m)$ **坐标轴的投影.**

4 特征向量的正负号问题

根据特征值和特征向量的定义可得: 对于 m **阶方阵** C**,如果存在数** λ **和** m **维非零列向量** p **满足:** $Cp = \lambda p$**,则** $C(-p) = \lambda(-p)$ **也成立,因此根据协方差矩阵** C **求出的特征向量矩阵** P **每**

一列向量取 P_i 还是 $-P_i(i=1,2,\cdots,m)$，是一个亟待解决的问题. 特征向量正负号不同，主成分得分也不同，得到的主成分在应用于进一步研究的时候也可能得出不同的结论: 比如得到的各主成分得分用于进一步计算综合得分(各主成分乘以相应的贡献率再相加求和) 时，正负号取值不当将会影响到最终的结论.

事实上，对于同一个协方差矩阵，利用不同的软件(比如 SPSS，MATLAB 或 DPS) 求其特征向量，有时会得到的每一列向量值正负号是相反的. 那么针对一个具体的专业问题怎样正确地选择特征向量的正负号呢? 笔者认为应该根据主成分载荷矩阵再结合具体的专业知识判断，因为主成分载荷矩阵每一列反应了各个新的主成分与原始各变量之间的相关性，也就是新的主成分能够解释原始各变量的程度; 应该使每一个主成分主要解释的几个原始变量(重要性按主成分载荷值排队) 系数取正号，尽量从正相关的意义上去解释.

值得一提的是，主成分分析的结果即主成分得分 $[Z_{1i},Z_{2i},\cdots Z_{mi}](i=1,2,\cdots,n)$ 往往进一步应用于聚类分析，聚类分析可分为 Q 型聚类(针对样品的聚类) 和 R 型聚类(针对变量的聚类) . Q 型聚类法在衡量样品之间关系远近时选用的统计量是距离系数，在计算距离系数时，各主成分得分的正负号 ($+Z_{ji}$ 或 $-Z_{ji}$) 并不影响最终的距离矩阵，会得出相同的结果，这时如果主成分分析的目的是为了进一步的 Q 型聚类分析，可以不考虑特征向量正负号的问题. R 型聚类分析方法衡量变量之间关系亲疏远近的统计量是相似系数，包括夹角余弦和相关系数，由于经过主成分变换后新的主成分之间是相互独立的，则它们之间的夹角余弦和相关系数都为 0. 因此对新的主成分再做 R 型聚类分析是没有意义的，事实上，主成分分析法和 R 型聚类分析法的功能是类似的，从某种意义上来说，都是把相关的变量聚在一起，不同的是，主成分变量是所有原始变量的线性组合，而 R 型聚类只是把部分原始变量聚在一起.

5 主成分载荷问题

5.1 主成分载荷的概念

主成分载荷表征的是主成分 z_k 与原始变量 x_i 之间的相关系数，用 $p(z_k,x_i)$ 表示. 主成分载荷的大小表明主成分能解释原始变量的程度.

5.2 主成分载荷的计算

两个变量之间的相关系数等于它们之间的协方差除以它们标准差的积，即 $p(z_k,x_i)=\dfrac{Cov(z_k,x_i)}{\sqrt{\lambda_k}\sqrt{s_i}}$，其中 $\sqrt{\lambda_k}$ 表示主成分 z_k 的标准差，$\sqrt{s_i}$ 表示原始变量 x_i 的标准差，其值为 1.

因为 $Z=LX$，则 $L^TZ=L^TLX$，又因 L 为正交阵($L=P^T$，P 为正交阵) ，所以 $L^TL=1$，因此 $X=L^TZ$.

则 $x_i=l_{1i}z_1+l_{2i}z_2+\cdots+l_{ki}z_k+\cdots+l_{mi}z_m$

$$\begin{aligned} Cov(z_k, x_i) &= Cov(z_k, l_{1i}z_1 + l_{2i}z_2 + \cdots + l_{ki}z_k + \cdots + l_{mi}z_m) \\ &= Cov(z_k, l_{1i}z_1) + Cov(z_k, l_{2i}z_2) + \cdots + Cov(z_k, l_{ki}z_k) + \cdots + Cov(z_k, l_{mi}z_m) \\ &= l_{1i}\,Cov(z_k, z_1) + l_{2i}Cov(z_k, z_2) + \cdots + l_{ki}\,Cov(z_k, z_k) + \cdots + l_{mi}\,Cov(z_k, z_m) \end{aligned}$$

因为新的主成分 $z_1, z_2, \cdots z_m$ 之间完全不相关，所以 $Cov(z_i, z_j) = 0(i, j = 1, 2, \cdots, m$ 且 $i \neq j)$.

则 $Cov(z_k, x_i) = l_{ki}\,Cov(z_k, z_k) = l_{ki}\lambda_k$

因此 $p(z_k, x_i) = \dfrac{Cov(z_k, x_i)}{\sqrt{\lambda_k}\sqrt{s_i}} = \dfrac{l_{ki}\lambda_k}{\sqrt{\lambda_k}\sqrt{s_i}} = l_{ki}\sqrt{\lambda_k}$

6 主成分得分问题

主成分得分可利用 $Z = LX$ 计算，其中 X 为经标准差标准化后的原始变量，L 的行向量为 X 协方差矩阵 C 的 m 个线性无关的特征向量. 有些教材将主成分载荷和特征向量混淆[1-2]，计算主成分得分时，利用 $Z = p(z_k, x_i) \times X$ 求取，这显然是不正确的. 另外，利用 $SPSS$ 软件进行主成分分析时，结果只给出了主成分载荷矩阵而没有特征向量矩阵，在计算主成分得分时，需要利用 $l_{ki} = p(z_k, x_i) / \sqrt{\lambda_k}$ 求出特征向量矩阵.

7 小结

在实践教学中发现，很多教材对主成分分析法的介绍过于简单，只给出了一些结论性的东西，比如求主成分的系数等价于求原始变量相关系数矩阵的特征向量，主成分载荷等于特征向量乘以对应特征值的开方，但为什么是这样却并未给予解释; 另外特征向量正负号的问题也是主成分分析法在实际应用中经常碰到的; 有些教材在计算主成分得分时将主成分载荷与特征向量混淆. 针对上述问题，本文做了逐一的剖析和详细的阐述，对于学生或初学者更好、更深刻地理解主成分分析法提供帮助.

参考文献

[1] 徐建华. 计量地理学 [M]. 北京: 高等教育出版社，2008: 96 - 97.

[2] 何晓群. 现代统计分析方法与应用 [M]. 第 2 版. 北京: 中国人民大学出版社，2007: 335 - 349.

[3] 同济大学数学教研室. 线性代数 [M]. 第 2 版. 北京: 高等教育出版社，2000: 115 - 125.

[4] 沈大庆，沈长源，李峰. 线性代数与线性规划 [M]. 北京: 北京邮电大学出版社，2001: 154 - 179.