第 25 卷 第 2 期
2010 年 4 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.25 No. 2
Apr. 2010

基于主成分分析中的几类问题的计算公式及结论

张绍璞
（天津科技大学理学院，天津 300457）

摘 要：讨论了主成分分析中的几类问题的计算方法，重点揭示在协方差矩阵中对角元素相同时的第一主成分及贡献率和对角元素不同时的主成分贡献率. 经讨论后给出有关计算公式，并说明这些计算公式在多元统计分析中的实际作用.

关键词：主成分分析；协方差矩阵；特征值；贡献率
中图分类号：O212.4 **文献标志码**：A **文章编号**：1672-6510(2010)02-0076-03

Based on Principal Component Analysis of the Types of Problems in the Formulas for Calculating and Conclusions

ZHANG Shao-pu
(College of Science, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract: Several types of problems and calculation methods in the analysis of principal components were discussed. It especially focused on the first principal component and its contribution rates when the diagonal elements in covariance matrix are the same as well as the contribution rates of principal components when the diagonal elements are different. The discussion is given in relation to the calculation formula and the formula for calculating these in the multivariate statistical analysis of the actual effect .

Keywords: principal component analysis; covariance matrix; eigenvalue; contribution rate

随着科学技术的发展,多元统计分析在很多领域得到了越来越广泛的应用. 在多元统计分析中,所研究的多项随机变量指标之间常有一定的相关性,因而在一定程度上增加了问题的复杂性. 为了使复杂的问题简化,常利用主成分分析,把原来多个随机变量转化为不相关的,并反映了原来多个变量信息的若干个主成分综合变量. 例如在气象预报中,常将复杂的天气状况分解为温度、气压、湿度、风力等若干个主成分变量指标. 主成分分析正是研究如何通过原来便利的少数几个线性组合来解释原来变量绝大多数信息的一种多元统计方法. 因此,主成分分析是多元分析中的重要工具之一. 对于主成分分析的理论与应用的讨论与研究,在当前多元分析的理论研究中非常重要,并得到广泛的重视[1–2].

1 主成分分析的有关定理及推论

定理[3] 设 $X=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是 n 维随机变量,且其协方差阵 $\boldsymbol{V}=D(X)$,$\boldsymbol{V}$ 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n\geqslant 0$,$\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n$ 为相应的单位正交特征向量,则 X 的第 i 主成分为

$$Z_i=a_i^{\mathrm{T}}X\quad(i=1,2,\cdots,n)$$

推论 设 $Z=(Z_1,Z_2,\cdots,Z_n)^{\mathrm{T}}$ 是 n 维随机变量,则其分量 $Z_i(i=1,2,\cdots,n)$ 依次是 X 的第 i 主成分的充分必要条件是:

(1) $Z=\boldsymbol{A}^{\mathrm{T}}X$,$\boldsymbol{A}$ 为正交矩阵;

(2) $\boldsymbol{D}(\boldsymbol{Z})=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$,即随机向量 $\boldsymbol{Z}$ 的

收稿日期：2009–06–29；**修回日期**：2009–09–20
作者简介：张绍璞（1954—），男，山东人，教授，zsp888@tust.edu.cn.

协方差阵为对角矩阵；

(3) $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$.

上述定理和推论给出了主成分分析的主要计算方法，另外通过定理也可推导出一些主成分分析的重要性质.

2　主成分分析的性质

记 $\boldsymbol{V}=(\sigma_{ij})$，$\Lambda = diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$，其中 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ 为 $\boldsymbol{V}$ 的特征值，$\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n$ 为相应的单位正交特征向量，记正交矩阵 $\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n)$. 则总体主成分 $Z=(Z_1,Z_2,\cdots Z_n)^{\mathrm{T}}$ 有如下性质[4]：

性质 1　$\boldsymbol{D}(Z)=\Lambda$，即 n 个主成分的方差为：$Var(Z_i)=\lambda_i,(i=1,2,\cdots,n)$，且它们是互不相关的.

性质 2　$\sum_{i=1}^{n}\sigma_{ii}=\sum_{i=1}^{n}\lambda_i$，通常称 $\sum_{i=1}^{n}\sigma_{ii}$ 为原总体 X 的总方差.

性质 3　主成分 Z_k 与原始变量 X_i 的相关系数 $\rho(Z_k,X_i)$ 为 $\rho(Z_k,X_i)=\sqrt{\lambda_k}a_{ik}\big/\sqrt{\sigma_{ii}}\,(k,i=1,2,\cdots,n)$，并把主成分 Z_k 与原始变量 X_i 的相关系数称为因子负荷量.

主成分分析的目的之一是为了简化数据结构，故在实际应用中一般不用 n 个主成分，而选用 $m(m<n)$ 个主成分. 为此，引入贡献率的概念.

定义[5]　记 $\lambda_k\Big/\sum_{i=1}^{n}\lambda_i$ 为主成分 Z_k 的贡献率；$\sum_{k=1}^{m}\lambda_k\Big/\sum_{i=1}^{n}\lambda_i$ 为主成分 $Z_1,Z_2,\cdots,Z_m(m<n)$ 的累计贡献率.

3　关于主成分分析一些计算问题的研究及结论

在生产领域、科研领域、经济学领域和日常生活中经常需要根据观测到的数据资料，以主成分分析为工具，对所研究的对象进行分析或定性. 另外一些很复杂的问题也需用主成分分析简化，可以参考文献[6]. 下面针对某几类主成分分析问题给予讨论和分析，并通过计算和证明给出有关新的性质和结论，推导出解决此类问题和简化计算的方法.

问题 1　设 $X=(X_1,X_2,\cdots X_n)^{\mathrm{T}}$ 是 n 维随机变量，且其协方差阵

$$\boldsymbol{V}=\sigma^2\begin{pmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & \cdots & \rho\\ \vdots & \vdots & & \vdots\\ \rho & \rho & \cdots & 1\end{pmatrix},\ \text{其中}\ \sigma^2>0,\ 0<\rho<1.$$

(1) 计算 X 的第一主成分 Z_1；

(2) 计算主成分 Z_1 的贡献率；计算主成分 Z_1，$Z_2,\cdots,Z_m(m<n)$ 的累计贡献率.

解：(1) 先求 $\boldsymbol{V}$ 的特征值，由

$$|\lambda\boldsymbol{E}-\boldsymbol{V}|=\begin{vmatrix}\lambda-\sigma^2 & -\sigma^2\rho & \cdots & -\sigma^2\rho\\ -\sigma^2\rho & \lambda-\sigma^2 & \cdots & -\sigma^2\rho\\ \vdots & \vdots & & \vdots\\ -\sigma^2\rho & -\sigma^2\rho & \cdots & \lambda-\sigma^2\end{vmatrix}=\begin{vmatrix}(\lambda-\sigma^2)-(n-1)\sigma^2\rho & (\lambda-\sigma^2)-(n-1)\sigma^2\rho & \cdots & (\lambda-\sigma^2)-(n-1)\sigma^2\rho\\ -\sigma^2\rho & \lambda-\sigma^2 & \cdots & -\sigma^2\rho\\ \vdots & \vdots & & \vdots\\ -\sigma^2\rho & -\sigma^2\rho & \cdots & \lambda-\sigma^2\end{vmatrix}=$$

$$\left[\lambda-\sigma^2-(n-1)\sigma^2\rho\right]\begin{vmatrix}1 & 1 & \cdots & 1\\ -\sigma^2\rho & \lambda-\sigma^2 & \cdots & -\sigma^2\rho\\ \vdots & \vdots & & \vdots\\ -\sigma^2\rho & -\sigma^2\rho & \cdots & \lambda-\sigma^2\end{vmatrix}=\left[\lambda-\sigma^2-(n-1)\sigma^2\rho\right]\begin{vmatrix}1 & 1 & \cdots & 1\\ 0 & \lambda-\sigma^2+\sigma^2\rho & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \lambda-\sigma^2+\sigma^2\rho\end{vmatrix}=$$

$$[\lambda-\sigma^2-(n-1)\sigma^2\rho]\left(\lambda-\sigma^2+\sigma^2\rho\right)^{n-1}$$

令 $|\lambda\boldsymbol{E}-\boldsymbol{V}|=0$，得特征值 $\lambda_1=\sigma^2+(n-1)\sigma^2\rho$，$\lambda_2=\lambda_3=\cdots=\lambda_n=(1-\rho)\sigma^2$.

由 $(\lambda_1\boldsymbol{E}-\boldsymbol{V})\begin{pmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{pmatrix}=\sigma^2\rho\begin{pmatrix}n-1 & -1 & \cdots & -1\\ -1 & n-1 & \cdots & -1\\ \vdots & \vdots & & \vdots\\ -1 & -1 & \cdots & n-1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{pmatrix}=0$，

得特征值 $\lambda_1=\sigma^2+(n-1)\sigma^2\rho$ 的单位特征向量为 $a_1=\frac{1}{\sqrt{n}}(1,\ 1,\ \cdots,\ 1)^{\mathrm{T}}$，根据定理得 X 的第一主成分

$$Z_1=a_1^{\mathrm{T}}X=\frac{1}{\sqrt{n}}(X_1+X_2+\cdots+X_n).$$

(2) 主成分 Z_1 的贡献率为

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\cdots+\lambda_n}=\frac{\sigma^2\left[1+(n-1)\rho\right]}{\sigma^2\left[1+(n-1)\rho+(n-1)(1-\rho)\right]}=$$

$\frac{1+(n-1)\rho}{n}$；主成分 $Z_1,Z_2,\cdots,Z_m\,(m<n)$ 的累计贡献率为 $\frac{\lambda_1+\lambda_2+\cdots+\lambda_m}{\lambda_1+\lambda_2+\cdots+\lambda_n}=\frac{\sigma^2\left[1+(n-1)\rho+(m-1)(1-\rho)\right]}{\sigma^2\left[1+(n-1)\rho+(n-1)(1-\rho)\right]}=\frac{m+(n-1)\rho-(m-1)\rho}{n}=\frac{m+(n-m)\rho}{n}$

主成分分析是一种通过降维技术把多个变量化为少数几个主成分的统计分析方法，这些主成分能够反映原始变量的绝大部分信息. 在实际应用中一般选用 $m(m<n)$ 个主成分，使之能够反映 n 个原始变量的70%或80%以上的信息. 这种反映原始变量信息量的程度，往往用累计贡献率来表示.

问题 2　设 $X=(X_1,X_2,\cdots,X_5)^{\mathrm{T}}$ 是5维随机变量，且其协方差阵

$$\boldsymbol{V}=\sigma^2\begin{pmatrix}1&\rho&0&0&0\\ \rho&1&\rho&0&0\\ 0&\rho&1&\rho&0\\ 0&0&\rho&1&\rho\\ 0&0&0&\rho&1\end{pmatrix}$$

其中：$\sigma^2>0,\ 0<\rho\leqslant\frac{1}{2}$.

(1) 计算 $\boldsymbol{V}$ 的特征值；

(2) 当 $\rho=\frac{1}{2}$，问 m 为何值时，才能使主成分 $Z_1,Z_2,\cdots,Z_m\,(m<5)$ 的累计贡献率大于85%.

解：先求 $\boldsymbol{V}$ 的特征值，记 D_n 为相应的 n 阶行列式，则

$$D_5=|\lambda\boldsymbol{E}-\boldsymbol{V}|=\begin{vmatrix}\lambda-\sigma^2&-\sigma^2\rho&0&0&0\\ -\sigma^2\rho&\lambda-\sigma^2&-\sigma^2\rho&0&0\\ 0&-\sigma^2\rho&\lambda-\sigma^2&-\sigma^2\rho&0\\ 0&0&-\sigma^2\rho&\lambda-\sigma^2&-\sigma^2\rho\\ 0&0&0&-\sigma^2\rho&\lambda-\sigma^2\end{vmatrix}$$

根据行列式按行展开法则，可得递推公式：

$$D_n=(\lambda-\sigma^2)D_{n-1}-(\sigma^2\rho)^2D_{n-2}$$

根据递推公式可得：

$$D_1=\lambda-\sigma^2$$

$$D_2=(\lambda-\sigma^2)-\sigma^4\rho^2$$

$$D_3=(\lambda-\sigma^2)\left[(\lambda-\sigma^2)^2-2\sigma^4\rho^2\right]$$

$$D_4=(\lambda-\sigma^2)^4-3(\lambda-\sigma^2)^2\sigma^4\rho^2+\sigma^8\rho^4$$

$$D_5=(\lambda-\sigma^2)\cdot\left[(\lambda-\sigma^2)^4-4\sigma^4\rho^2(\lambda-\sigma^2)^2+3\sigma^8\rho^4\right]$$

令 $D_5=0$，可得 $\boldsymbol{V}$ 的 5 个由大到小顺序排列的特征值为：$\lambda_1=\sigma^2(1+\sqrt{3}\rho)$；$\lambda_2=\sigma^2(1+\rho)$；$\lambda_3=\sigma^2$；$\lambda_4=\sigma^2(1-\rho)$；$\lambda_5=\sigma^2(1-\sqrt{3}\rho)$.

显然 $\lambda_1+\lambda_2+\cdots+\lambda_5=5\sigma^2$

$$\frac{\lambda_1+\lambda_2+\lambda_3}{\lambda_1+\lambda_2+\cdots+\lambda_5}=\frac{\sigma^2\left[3+(\sqrt{3}+1)\rho\right]}{5\sigma^2}$$

当 $\rho=\frac{1}{2}$ 时，知

$$\frac{\lambda_1+\lambda_2+\lambda_3}{\lambda_1+\lambda_2+\cdots+\lambda_5}=\frac{\sigma^2\left[3+(\sqrt{3}+1)\,1/2\right]}{5\sigma^2}=\frac{7+\sqrt{3}}{10}=0.873\,205.$$

所以取 $m=3$，主成分 Z_1,Z_2,Z_3 的累计贡献率大于85%. 即用前三个主成分就能够反映5个原始变量的绝大部分信息. 此类问题结论的有关计算方法也可推广到相应的 n 维情况.

4　结　语

多元分析应用非常广泛，其中有许多问题涉及到主成分分析的计算. 通过上述两类问题的计算和讨论可知，对于本文所讨论的计算公式，经过对其性质的有关研究和证明，给出了简化计算相对应主成分和累计贡献率的有关方法，同时确定了能够反映原始变量的绝大部分信息的主成分个数，可以大幅度简化有关复杂计算，在实际应用中起到很重要的作用.

参考文献：

[1] 赵杰辉，葛少云，刘自发. 基于主成分分析的径向基函数神经网络在电力系统负荷预测中的应用[J]. 电网技术，2004，28(5)：35–37.

[2] 陈蕊，王华统. 营销问题的主成份分析[J]. 合肥学院学报：自然科学版，2004，14(1)：18–22.

[3] 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005：266–269.

[4] 何晓群. 多元统计分析[M]. 北京：中国人民大学出版社，2004：143–146.

[5] 王学民. 应用多元分析[M]. 上海：上海财经大学出版社，2004：236–237.

[6] 柳进，唐降龙. 基于主成分分析 L-M 神经网络高峰负荷预测研究[J]. 继电器，2004，32(13)：24–27.