对主成分分析中综合得分方法的质疑

王学民

（发表于《统计与决策》，2007 年 4 月下）

摘要：在作主成分分析时，国内近年来流行一种通过建立综合评价函数来对各样品进行综合排名的方法。本文对这一方法的不科学性作了阐述，并指出在综合评价函数中对各主成分使用贡献率加权是错中加错。

关键词：主成分；信息量；综合评价函数；综合得分

一、问题的提出

在多元数据分析中，近年来国内流行一种通过建立综合评价函数来对所有样品进行综合排名的方法。该方法是这样的：对 p 个原始变量 $x_1,x_2,\cdots,x_p$，通过主成分分析，取前 m 个主成分 $y_1,y_2,\cdots,y_m$，其方差分别为 $\lambda_1,\lambda_2,\cdots,\lambda_m$，以每个主成分 y_i 的贡献率 $\alpha_i=\lambda_i\Big/\sum_{i=1}^{p}\lambda_i$ 作为权数，构造综合评价函数

$$F=\alpha_1 y_1+\alpha_2 y_2+\cdots+\alpha_m y_m$$

计算出每个样品的（ F ）综合得分，然后依这个得分的大小对所有样品进行综合排名。对这种用线性组合的方式来综合各主成分的方法，笔者从未在国外的有关多元统计分析的文献中见过。该方法粗看起来似乎有一定道理且很有吸引力（似乎可以综合排名了），但仔细推敲之后就会发现这一方法是对主成分思想和方法的误解，是不科学的，没有什么理论和应用上的价值。该综合排名方法在我国的多元数据分析应用中已得到了比较普遍的误用，笔者曾在参考文献[1]中的 253 页上简略地谈到过这一问题，现觉得很有必要针对这一问题作一具体阐述，谈谈自己的观点，供大家参考和讨论。

二、主成分的基本思想

除了将主成分法用于聚类或回归分析或寻找变量之间的共线性关系等目的之外，主成分分析的一般目的由两点组成：（1）将多个有相关关系的变量压缩成少数几个不相关的主成分（综合变量），并保留绝大部分信息；（2）给出各主成分的具有实际背景和意义的解释。

这里我们只讨论主成分分析的这种一般目的。主成分的价值就在于它的信息量（可用方差来度量）达到最大化，即使前少数几个主成分能使累计贡献率达到一个较大的百分数，这几个主成分能不能用还得看它们是否都能得到符合实际意义的解释。

例 1　在 1984 年洛杉矶奥运会 IAAF/ATFS 田径统计手册中，有 55 个国家和地区的如下八项男子径赛运动记录：

x_1：100 米（单位：秒）　　x_5：1500 米（单位：分）

x_2：200 米（单位：秒）　　x_6：5000 米（单位：分）

x_3：400 米（单位：秒）　　x_7：10000 米（单位：分）

x_4：800 米（单位：秒）　　x_8：马拉松（单位：分）

经计算，$x_1,x_2,\cdots,x_8$ 的样本相关矩阵 R 列于表 1。R 的前两个特征值、特征向量及贡献率列于表 2，其中 x_i^* 是 x_i 经标准化得到的，即 x_i^* 的均值和标准差分别为 0 和 1。

表 1

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1.000							
x_2	0.923	1.000						
x_3	0.841	0.851	1.000					
x_4	0.756	0.807	0.870	1.000				
x_5	0.700	0.775	0.835	0.918	1.000			
x_6	0.619	0.695	0.779	0.864	0.928	1.000		
x_7	0.633	0.697	0.787	0.869	0.935	0.975	1.000	
x_8	0.520	0.596	0.705	0.806	0.866	0.932	0.943	1.000

表 2

特征向量	t_1	t_2
x_1^*：100 米	0.318	0.567

x_2^*：200 米	0.337	0.462
x_3^*：400 米	0.356	0.248
x_4^*：800 米	0.369	0.012
x_5^*：1500 米	0.373	-0.140
x_6^*：5000 米	0.364	-0.312
x_7^*：10000 米	0.367	-0.307
x_8^*：马拉松	0.342	-0.439
特征值	6.622	0.878
贡献率	0.828	0.110
累计贡献率	0.828	0.937

由表 2 知，前两个主成分的累计贡献率已高达 93.7%，第一主成分 y_1 在所有变量上有几乎相等的正载荷，可称为在径赛项目上的强弱成分。第二主成分 y_2 在 $x_1^*,x_2^*,\cdots,x_8^*$ 上的载荷基本上逐个递减，反映了速度与耐力成绩的对比[①]。前两个主成分 y_1 和 y_2 虽然得到了很好的符合实际意义的解释，但这种解释毕竟带有一定程度的模糊性，这是主成分分析的一个特点，这种解释的模糊性也是变量降维需要付出的代价。体育径赛项目方面的专家也许能制定出实际意义更清楚、更能反映各国在径赛项目上强弱的指标 z_1（例如，在系数平方和为 1 的前提下，取 $z_1=\frac{1}{2\sqrt{2}}x_1^*+\frac{1}{2\sqrt{2}}x_2^*+\cdots+\frac{1}{2\sqrt{2}}x_8^*\approx 0.354x_1^*+0.354x_2^*+\cdots+0.354x_8^*$）和反映速度与耐力成绩对比的指标 z_2，但 z_1、z_2 这两个指标合起来所包含的信息量不如或明显不如 y_1、y_2 所包含的信息量大。这两个主成分的优势就在于它们合在一起能拥有最大的信息量，而不是它们各自能多么准确地反映各国在径赛项目上的强弱和速度与耐力成绩的对比。

三、综合评价函数存在的问题

在许多实际问题中，我们确实非常需要一个综合指标来对所有样品进行排序，但这个综合指标不应想当然地从前几个主成分的线性组合来产生。设作主成分分析时取前 m 个主成分 $y_1,y_2,\cdots,y_m$ 是合适的，则综合评价函数为 $F=\alpha_1y_1+\alpha_2y_2+\cdots+\alpha_my_m$，它存在以下一些

[①]此例用因子分析法效果更好，见参考文献[1]，本文这里只是作为一个说明性的例子。

问题：

（1）F 到底包含有原始变量 $x_1,x_2,\cdots,x_p$ 的多少信息，应用此方法者都未作说明。当然，F 所含的信息量不会超过第一主成分 y_1。

（2）F 到底具有什么样的实际含义，应用者都没有解释或作不出解释，只是笼统地理解为所谓的“综合”指标，用这种不知其具体含义的指标来对所有样品进行排序又有何实际意义呢？这样的排序说明不了什么问题。

（3）$y_1,y_2,\cdots,y_m$ 的首要价值就在于它们合在一起拥有最大量的信息，这种信息对原始的 p 个变量绝对不是包罗万象的（如并不含有关于原始变量均值等的信息），而仅是体现在数据的变异性上。把反映数据变异性信息的前 m 个主成分线性组合起来将会瓦解主成分在变异性信息上的优势，主成分分析一旦离开了反映变异性的信息量，也就没有价值和意义了。

（4）由于

$$\begin{aligned}\mathrm{Var}(F)&=\alpha_1^2\,\mathrm{Var}(y_1)+\alpha_2^2\,\mathrm{Var}(y_2)+\cdots+\alpha_m^2\,\mathrm{Var}(y_m)\\&=\frac{\lambda_1^3}{\left(\sum_{i=1}^{p}\lambda_i\right)^2}+\frac{\lambda_2^3}{\left(\sum_{i=1}^{p}\lambda_i\right)^2}+\cdots+\frac{\lambda_m^3}{\left(\sum_{i=1}^{p}\lambda_i\right)^2}\end{aligned}$$

故第 i 个主成分 y_i 对 F 的方差贡献所占的比例（容易证明，该比例就是 $\rho^2(y_i,F)$，其中 $\rho(y_i,F)$ 是 y_i 与 F 的相关系数）为

$$\frac{\lambda_i^3}{\sum_{i=1}^{m}\lambda_i^3},\quad i=1,2,\cdots,m$$

在主成分分析中 λ_1 一般会远大于其他的 $\lambda_i(i=2,\cdots,m)$，以致 y_1 对 F 的方差贡献所占的比例通常是很大的，而其他 y_i 对 F 的方差贡献所占的比例通常都很小，因此 F 未能对 $y_i(i=2,\cdots,m)$ 起到什么“综合”作用。在许多实际问题中，作主成分分析时常常会出现 $\lambda_1>2\lambda_2$，若取前两个主成分 y_1 和 y_2，则 y_1 对 F 的方差贡献所占的比例为

$$\frac{\lambda_1^3}{\lambda_1^3+\lambda_2^3}>\frac{(2\lambda_2)^3}{(2\lambda_2)^3+\lambda_2^3}=\frac{8}{9}=88.89\%$$

而 y_2 对 F 的方差贡献所占的比例为

$$\frac{\lambda_2^3}{\lambda_1^3+\lambda_2^3}<\frac{\lambda_2^3}{(2\lambda_2)^3+\lambda_2^3}=\frac{1}{9}=11.11\%$$

在例 1 中

$$\frac{\lambda_1^3}{\lambda_1^3+\lambda_2^3}=\frac{0.828^3}{0.828^3+0.110^3}=99.77\%$$

$$\frac{\lambda_2^3}{\lambda_1^3+\lambda_2^3}=\frac{0.110^3}{0.828^3+0.110^3}=0.23\%$$

因此，通常影响 F 的主要是第一主成分 y_1，而其他主成分对 F 的影响一般都很小。在例 1 中，综合评价函数为

$$\begin{aligned}F&=\alpha_1y_1+\alpha_2y_2=0.828y_1+0.110y_2\\&=0.326x_1^*+0.330x_2^*+0.322x_3^*+0.307x_4^*+0.293x_5^*+0.267x_6^*+0.270x_7^*+0.235x_8^*\end{aligned}$$

将各系数均除以这些系数的平方和的平方根（以使调整后的系数平方和为 1，便于与主成分的载荷进行比较），得

$$F'=0.390x_1^*+0.395x_2^*+0.385x_3^*+0.367x_4^*+0.351x_5^*+0.320x_6^*+0.323x_7^*+0.281x_8^*$$

将变量前的各系数与表 2 的主成分载荷比较，可以发现 F' 与 y_1 较接近，而与 y_2 相差很远。F' 既没有 y_1 的信息量大，又不如 y_1 易解释，看不出构造 F' 有什么实际价值。

（5）在综合评价函数中，对各主成分 $y_1,y_2,\cdots,y_m$ 分别使用权数 $\alpha_1,\alpha_2,\cdots,\alpha_m$ 是错中加错，实际上各主成分的方差不同，具有自动加权的功能。也就是说，使用 $F=\alpha_1y_1+\alpha_2y_2+\cdots+\alpha_my_m$ 比使用 $F^*=y_1+y_2+\cdots+y_m$ 更糟糕。y_i 对 F^* 的方差贡献所占的比例为 $\frac{\lambda_i}{\sum_{i=1}^{m}\lambda_i}(i=1,2,\cdots,m)$，与 y_i 的贡献率成正比。

在因子分析中，对因子得分建立类似综合评价函数的方法同样也是错误的。

[参考文献]

[1]王学民.应用多元分析（第二版）[M].上海：上海财经大学出版社，2004.