

数值分类的几种多元统计方法比较

宋世德 周静芋

(西北农业大学基础科学系,陕西杨凌 712100)

摘 要 从理论和实际应用两个方面对用主成分分析、因子分析和对应分析作数值分类工作进行比较。描述了各分类方法的优缺点及其在实际应用中应注意的问题,并结合实例分析对诸方法的分类结果进行了讨论。

关键词 数值分类,主成分分析,因子分析,对应分析

中图分类号 O212.4

用于数值分类的数学方法很多,如系统聚类、动态聚类、图论法、主成分分析、因子分析、对应分析和模糊聚类等。本文着重对用主成分分析、因子分析和对应分析作数值分类工作进行比较。其基本思想是实现高维问题低维化,用样品的前几个主成分或主因子值用二维或三维空间的直观散点图来刻画类群间的关系或作其他分析,并能比系统聚类法提供更多的信息。

1 几种方法比较^[1]

主成分分析法是将分类的多指标变换成几个互不相关的综合指标,这些综合指标是原来多个指标的线性组合,选取少数的前几个综合指标就能反映原来多个指标的信息,并能克服原指标太多和指标间相关性的问题。其模型为:

$$F_i = U_i^T X \quad i = 1, 2, \dots, m$$

式中, $X = (x_1, x_2, \dots, x_m)^T$, 每个分量是一个分类指标, U_i 是协方差阵的第 i 大特征值 (λ_i) 对应的标准化特征向量。笔者只选少数前几个主成分 $F_i, i = 1, 2, \dots, q (q < m)$, 使它们能反映 X 的绝大部分变异信息,彼此间又不相关。然后以各样品的几个新的综合变量得分做为样品的新分类指标。最后,或以新分类指标值作系统聚类,或在平面(或三维空间)画出样品得分值的散点图,根据样品在几个平面(或空间)分布的综合情况进行分类。

1.1 用相关阵和协方差阵作主成分分析的比较

用相关阵和协方差阵作主成分分析时的特征值和特征向量不一样,要使前几个主因子的累计贡献率达到某个百分数所选的主成分数一般也不一样,样品的各主成分值也不一样,因此用它们作的散点图和分类结果也常不一样;分析变量对分类的重要性,性状的变异方向等结果也不一样。

1.2 因子分析法和主成分分析法的比较(R型)

因子分析是主成分分析的进一步发展,其数学模型为:

$$X = AY + X$$

这里, $X = (x_1, x_2, \dots, x_m)^T$ 是由原始变量经过标准化后得到的变量; $Y = (y_1, y_2, \dots, y_q)^T$ ($q < m$) 称为分类因子(主因子), 这是 X 中共同出现的因子, 是相互独立的不可观测的变量; 特殊因子 $X_c = (X_1, X_2, \dots, X_k)^T$ 是 X 所特有的; 各特殊因子之间以及特殊因子与所有公共因子之间都是相互独立的; A 称为因子载荷阵。

为了研究讨论对象的特征, 将公共因子表示为变量的线性组合 $Y = B^T X$, 其中 Y 称为因子得分, B 为回归系数矩阵, 由最小二乘法求出, 即

$$\hat{Y} = A^T R^{-1} X$$

这里 R 为原变量的相关系数矩阵。将每个样品的 m 个变量值代入可算出因子得分, 以下步骤同上。因子分析的前提是对变量的相关阵进行标准化, 求出特征值和特征向量。故分类时要使前几个主因子的累计贡献率达到某个百分数所选的主因子数和用相关阵进行主成分分析时所选的主成分的个数一样(各个因子的贡献率和累计贡献率都是一样的)。因子分析法中的因子载荷阵为

$$A = (U_1 \lambda_1, U_2 \lambda_2, \dots, U_q \lambda_q)$$

其中, U_q 是从相关阵出发求出的特征值 (λ_q) 对应的标准化的特征向量, 算出的因子得分(因子分析法中计算样品的因子得分时的回归系数阵是根据最小二乘法求出)和从相关阵出发用主成分法算出的各样品的相应主成分值(这时 $F = U^T X$)是不一样的。在主成分分析的实用程序中也常有规格化的主成分得分(这时的因子载荷阵就和因子分析法算出的因子载荷阵是完全一样), 但是样品的主成分值和因子分析法中的因子得分仍不一样, 因此分类情况也不一致。

在因子分析实用程序中, 因子分析通常还要进行方差最大旋转变换, 这时前几个因子的贡献率和累计贡献率是不变的, 所以进行分类时所选用因子数相同, 但因子载荷阵发生变化。因子载荷阵中每列元素尽量向 0 和 1 靠拢, 这是因子分析的一个优点, 便于作有关的生物学解释。由于这个原因求出的因子得分数也不一样, 各样品的散点图与分类结果也就含有差异。理论上, 主成分分析也可作方差最大旋转变换。

1.3 对应分析法和主成分分析法(因子分析)的比较

对应分析法是一个对变量和样品同时进行主成分分析的方法, 也可以说是在 R 型和 Q 型因子分析方法基础上发展起来的一种多元分析方法。它的关键一步是将原始数据阵 X 变换成 Z , 使得 $Z^T Z$ 和 ZZ^T 分别起到变量和样品协方差阵的作用, 并且 $Z^T Z$ 和 ZZ^T 具有相同的非零特征值。变换的第一步是把 X 中的每个元素除以 X 所有元素总和 T , 然后通过概率处理分别得到 m 个变量和 n 个样品间的协方差阵。所得到的协方差阵分别与 R 型和 Q 型的因子分析的协方差阵不同, 因而得到的特征值和特征向量也不同, 当选取累计贡献率为某个百分数的主因子时, 选取的主因子个数自然常不同, 相应的得到的样品的各因子得分也不一样(因子载荷阵不同)。因而用对应分析作的分类也与主成分分析和因子分析不同。

对应分析法的优点是: 可将变量与样品同时反映在相同因子轴的同一平面图上, 因为它可以用共同的因子轴同时表示变量和样品, 根据它们在图的位置, 比较直观的解释变量和样品二者之间的关系。但应用时应注意^[2], 由于在对 X 进行变换时先要求所有数据之

和 T , 若各变量的量纲或数量级截然不同时, 求和一般是不能进行的。

通常用多元统计分析法进行数值分类的主要目的有: 样品与性状的分类, 变量重要性的判别, 讨论性状变异的方向和对每一类的特点作出解释等。实际应用中选用哪个方法, 首先要能满足上面的要求, 其最重要的标准就是分析所得结果是否符合生物学意义, 或者说哪一个分析更具生物学内涵。另外, 用几个分类性状来表征一个分类单元时, 分类单元间的差异是分类的主要依据, 根据分类要求, 做到形神兼备最好。

2 应用举例

用 20 个分类指标对金翅夜蛾亚科的 18 个属 (分类单元, 或称样品) 进行数值分类 (为便于比较, 分类单元和指标变量的编号仍用文献 [3] 的编号, 这里不再叙述)。

用主成分分析时, 特征值累计百分比大于 80% 时的主因子个数分别为 5 (协方差阵) 和 6 (相关阵), 选取前 3 个主因子 [累计贡献率分别为 68.85% (协方差阵) 和 56.44% (相关阵)] 进行聚类。从样品的散点图 (协方差阵见图 1, 相关阵图略; 图中 I、II 和 III 分别代表第 1、2 和 3 主因子, 下同) 中的几个虚线圈可以看出, 所有样品的分类和用系统聚类法 (数据不变换, 相似性尺度用欧氏距离, 聚类方法用类平均法) 完全一致。其中 7 号葫芦夜蛾属和 19 号淡银纹夜蛾属的分类和传统分类不一致, 在文献 [3] 和 [4] 中已详细论述了原因, 提出了值得探讨和重视的观点。两种主成分分析法得到的聚类结果基本相同 (用相关阵时的散点图中 7 号也可归入 8~9 号, 19 号归入 12~20 号, 这时就和传统分类完全一致), 但在考虑各变量对分类作用的大小以及对性状变异方向的分析时, 得到了用协方差阵比用相关阵好的结论 [4~5]。

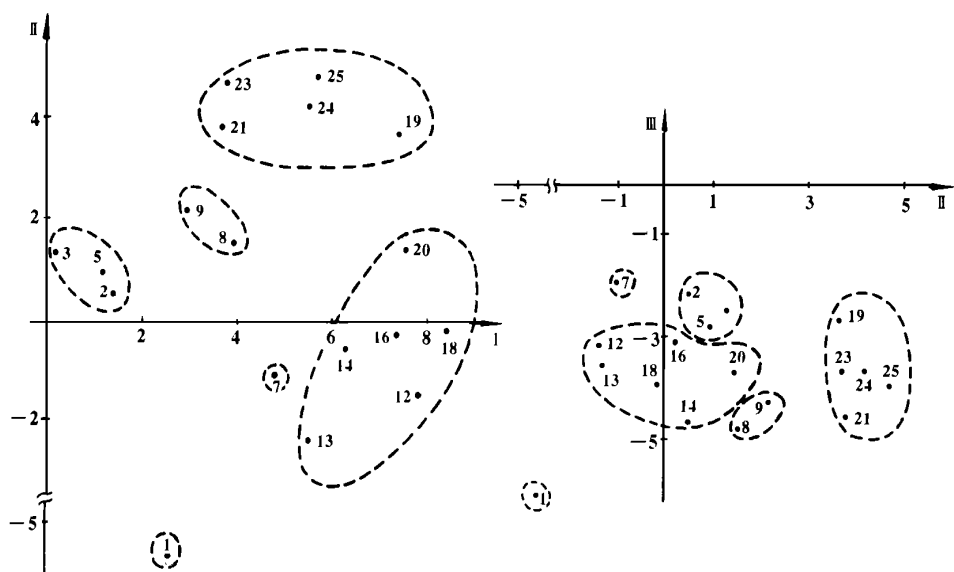


图 1 用主成分分析作的散点图 (协方差阵)

用因子分析 (R 型) 时, 特征值累计百分比大于 80% 时的主因子个数为 6, 选取前 3 个主因子 (累计贡献率为 56.44%) 进行聚类, 从样品的散点图 (图略) 中可以看出: 图中的样品点较之其他图要分散, 但大多数样品的分类与前面相同。7 号葫芦夜蛾属没有单独为一

类,而和 8 号金弧夜蛾属, 9 号金翅夜蛾属距离近可聚为一类,这与传统分类结果一致。20 号珠纹夜蛾属和 8 号金弧夜蛾属的距离在 I- II 和 II- III 主因子图上都近,分析原因是由于它们在第 1 和第 2 主因子上贡献最大的前 4 个变量的值一样,在第 3 主因子上贡献最大的前 2 个变量的值一样(实际它们区别很大)。19 号淡银纹夜蛾属和 21、23、24、25 号的距离较远(但 I- II 图上它们都在左边, II- III 图上它们都在下边),传统分类将它分在 12~20 号一类(银纹夜蛾族)。

由方差极大变换后的公因子方差中可知,对分类影响最大的前几个变量依次是腹足,前翅竖立鳞,中室下纹,毛刷和前翅臀齿栉。但在文献 [4] 中这已被认为是不妥的。这也是此法对此例不太合适的原因。

用对应分析时,特征值的累计百分比大于 85% 时的主因子个数为 6,选取前 3 个主因子(累计贡献率为 68.04%)进行聚类,从样品和变量的散点图(图 2)中的几个虚线圈可以看出大多数样品的分类和用主成分分类类似。7 号没有单独为一类,这里它和 8 号金弧夜蛾属, 9 号金翅夜蛾属距离近可聚归为一类。7 号和 8 号的距离在 II- III 主因子图上尤为近,分析原因是 7 号和 8 号样品与 5 号基腹弧和 20 号囊导管变量的距离特别近,这 3 个样品的这两个变量值是一样的,但 5 号变量和 20 号变量在前 3 个主因子上的载荷都不在前 5 位,这说明它们对分类的作用不是最大。9 号金翅夜蛾属和 21 号金斑夜蛾属和变量 4 号内突、13 号 A₁ 节 SV 毛及 16 号 A₉₍₂₎ 节 SD 毛距离特别近,这几个变量对分类的作用不大。这两个样品在主要分类性状上差异很大,不可能在一类。由于篇幅关系,有关本例的其他分析略。

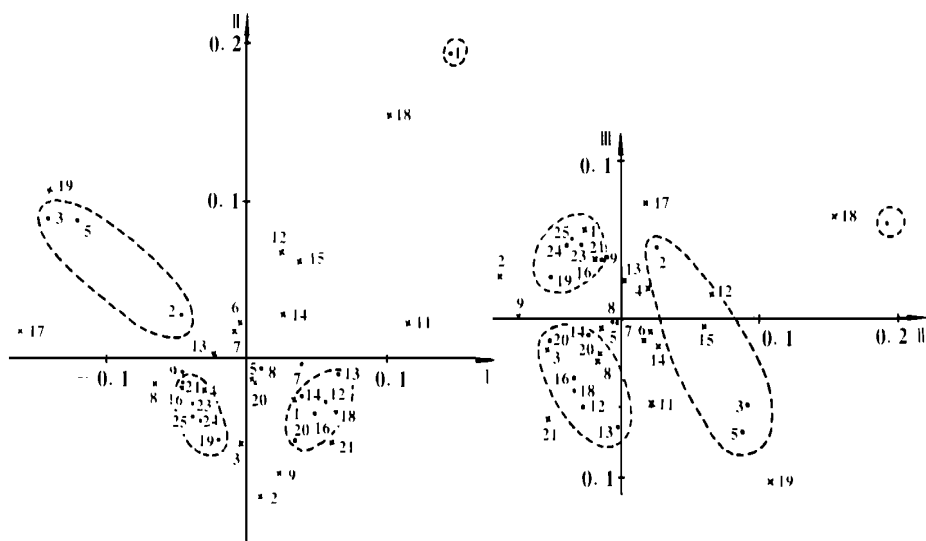


图 2 用对应分析作的散点图

● 表示分类单元; × 表示分类指标

综上所述,几种方法的分类结果基本大同小异,对于此例比较而言用协方差阵作主成分分析的效果较好(7 号葫芦夜蛾属的分类还需深入讨论),但绝不是对任何问题都是用此法好,而要根据数据的特点,综合分析用多种方法得到的各方面的结果信息,将几种方

法的共性取出来,对有不同分类结果的样品再根据分类问题本身的知识来决定才能得到结论

3 小 结

1)目前从理论上难以判断用哪种方法较好。但主成分分析从理论到实用程序都比较成熟,且能提供较多的有用信息;而因子分析和对应分析在这两个方面都不太成熟。从这个角度上讲,用主成分分析较好。

2)用散点图分类虽然必须兼顾几个平面情况,但仍可能有较大的主观性,因而要仔细考虑多方面的原因,特别当前几个主因子的贡献率不够大时更需如此。在实际分类问题中,前 2个(或前 3个)主因子的累计贡献率并不很高(小于 75%甚至更小),也能提供较多的有用信息。可参阅文献 [3, 4]

3)一般在做这几个多元统计分析时,皆要求 $n > m$ (样品数大于变量数)。这样实际上可保证协方差阵的正定性,但不能理论证明。作者在实用中(不少论文中也有此情况),发现虽然 $n < m$,只要协方差阵是非负的,就不影响进行数值分类工作。这时,关于前几个主成分的贡献率以及累计贡献率等有关结论是很容易证明是对的

参 考 文 献

- 1 裴鑫德.多元统计分析及其应用.北京:北京农业大学出版社,1991
- 2 孟德顺,郭满才.对应分析方法的注记.西北林学院学报,1996,11(3): 78~ 81
- 3 周静芋,宋世德.金翅夜蛾亚科(鳞翅目:夜蛾科)数值分类的研究.西北农业大学学报,1995,23(6): 28~ 32
- 4 周静芋,宋世德,袁志发.用主成分分析法分析金翅夜蛾亚科各性状对分类的重要性和性状变异方向.生物数学学报,1996,11(4): 75~ 82
- 5 周静芋,宋世德,袁志发.用主成分分析法进行数值分类的一些问题.西北农业大学学报,1996,24(5): 95~ 98

Comparison of Some Multivariate Statistical Methods Used in Numerical Taxonomy

Song Shide Zhou Jingyu

(Department of Basic Science, Northwestern Agricultural University, Yangling, Shaanxi 712100)

Abstract The principle component analysis, factor analysis and correspondence analysis used in numerical taxonomy are compared both theoretically and practically. The advantage and disadvantage of all the methods and the problems existing in practice were discussed. Then, the results of the numerical taxonomy were dealt with practically.

Key words numerical taxonomy, principal component analysis, factor analysis, correspondence analysis