Phương pháp phân tích thành phần chính (Principal component analysis – PCA) được giới thiệu lần đầu tiên bởi Hotelling (1933).

PCA là là một công cụ, là quy trình toán học dùng để giảm dữ liệu nhiều biến số thành số chiều ít hơn nhưng không mất mát thông tin trước đó. Trong đó, thành phần chính thứ nhất đại diện cho biến thiên lớn nhất của dữ liệu. Các thành phần tiếp theo cho độ biến thiên ít hơn thành phần chính thứ nhất. phân tích PCA được thao tác thông qua các hàm chính factoextra thông qua ngôn ngữ R. Hai giá trị eigenvector (hướng)và eigenvalue (giá trị) sẽ biểu thị sự khác biệt của các dữ liệu trong mỗi hướng. Tham số eigenvector với giá trị eigenvalue lớn nhất luôn luôn là thành phần thứ nhất trong các thành phần chính.

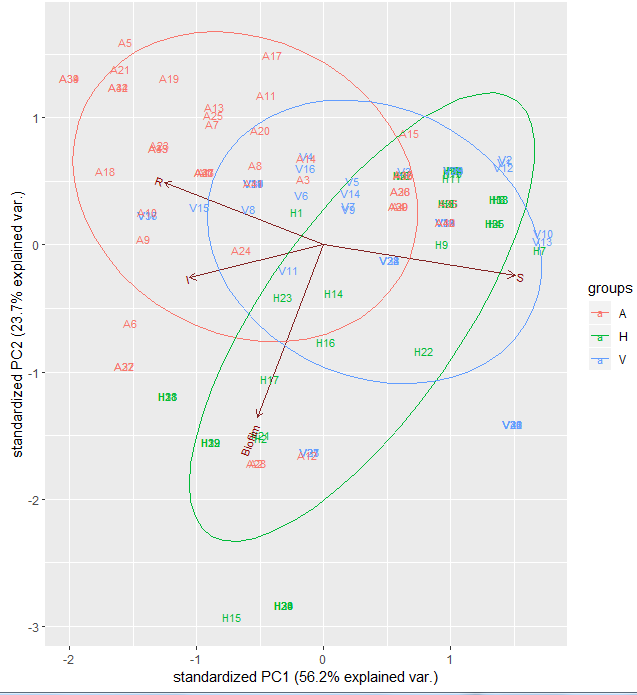
Bảng 1

|  |  |  |  |
| --- | --- | --- | --- |
| Thành phần | Sự khác biệt | % | % tích lũy |
| 1  2  3  4 | 2.248991e+00  9.495936e-01  8.014152e-01  1.320379e-30 | 5.622478e+01  2.373984e+01  2.003538e+01  3.300947e-29 | 56.22478  79.96462  100.00000  100.00000 |

Bảng 1, giải thích sự khác biệt của mỗi thành phần dựa vào giá trị eigenvalue được thể hiện ở cột thứ 2. Cụ thể, 2.248991(2.248991e+00) chia cho 4 sẽ có kết quả 56.22478% (5.622478e+01%) biểu thị sự khác biệt của thành phần thứ nhất. Giá trị phần trăm tích lũy được tính bằng cách cộng thêm các tỷ lệ ở các thành phần tiếp theo. Cụ thể, 56.22478%% cộng với 23.73984% sẽ là 79.96462%. Có thể hiểu, với 2 thành phần đầu tiên, chiếm 79.96462% trong tổng số các thành phần.

Giá trị eigenvalues có thể được dùng để xác định số thành phần còn lại sau PCA (Kaiser 1961). Cụ thể:

* Nếu eigenvalue > 1 chỉ ra rẳng, thành phần (PC) đó có ý nghĩa khác biệt thực sự khi so sánh với giá trị giá trị chuẩn (standardized) ban đầu. Giá trị này được dùng để giảm bớt số thành phần đối với trường hợp dữ liệu đầu vào có quá nhiều cột.
* Trong trường hợp giá trị eigenvalue <1 hoặc có quá ít thành phần phân tích, chúng ta có thể sử dụng tổng số sự khác biệt làm tiêu chí để giảm bớt số thành phần. Cụ thể, nếu có hơn 70% tổng số sự khác biệt tích lũy thì số thành phần tương ứng được coi là chấp nhận được.



Nhìn vào hình, có thể nói rằng S độc lập vớ 3 chỉ tiêu còn lại (với mức ý nghĩa p\_value<0.05) được trình bày ở bảng sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 |
| R | 0.8239838 | -0.3209429 | 0.466954409 | 6.475341e-16 |
| I | 0.6920135 | 0.1698996 | -0.701606356 | 4.336787e-16 |
| S | -0.9872987 | 0.1588590 | 0.002253489 | 8.443940e-16 |
| Biofilm | 0.3411753 | 0.8902175 | 0.301847970 | -2.391998e-32 |

Mối tương quan giữa biến (I,S,R,Biofilm) và thành phần tích lũy (PC) dùng để xác định số nhóm (Abdi and Williams 2010).

TLTK

Abdi, Hervé, and Lynne J. Williams. 2010. “Principal Component Analysis.” John Wiley and Sons, Inc. WIREs Comp Stat 2: 433–59. <http://staff.ustc.edu.cn/~zwp/teach/MVA/abdi-awPCA2010.pdf>.

Kaiser, Henry F. 1961. “A Note on Guttman’s Lower Bound for the Number of Common Factors.” British Journal of Statistical Psychology 14: 1–2.

Hotelling, H. (1933). The analysis of complex statistical variables into principal components. Journal of Educational Psychology 24, 417–441; 498–520.