ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
AGRICULTURAL UNIVERSITY OF ATHENS

Ανάλυση Κύριων Συνιστωσών

Κατσιλέρος Αναστάσιος

2017

ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

Η ανάλυση κύριων συνιστωσών είναι μία στατιστική διαδικασία που έχει σκοπό από ένα πλήθος μεταβλητών, τον σχηματισμό νέων μεταβλητών, οι οποίες καλούνται κύριες συνιστώσες και έχουν την ιδιότητα να είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και ταυτόχρονα να μη συσχετίζονται μεταξύ τους.

Ο αριθμός των νέων μεταβλητών που προκύπτει είναι ίσος ή μικρότερος από τον αριθμό των αρχικών μεταβλητών. Η πρώτη συνιστώσα εξηγεί τη μέγιστη δυνατή διακύμανση που αναπτύσσεται μεταξύ των αρχικών μεταβλητών, η δεύτερη, μη συσχετιζόμενη με την πρώτη, εξηγεί ένα σημαντικό μέρος αυτής αλλά πάντα μικρότερο της πρώτης κοκ.

Τα στάδια της ανάλυσης των κύριων συνιστωσών έχουν ως εξής:

1. Τυποποίηση των αρχικών μεταβλητών X_1, X_2,... X_p.

2. Υπολογισμός του πίνακα των συνδιακυμάνσεων ή των συσχετίσεων των αρχικών μεταβλητών X_1, X_2,... X_p.

3. Εκτίμηση των χαρακτηριστικών ριζών (eigenvalue) λ_1, λ_2... λ_p, και των διανυσμάτων (eigenvector) v_1, v_2 ... v_P.

4. Οι συνιστώσες που έχουν $\lambda < 1$ και εξηγούν μικρό ποσοστό της ολικής μεταβλητότητας, απορρίπτονται.

Αρχικά υπολογίζεται η μήτρα C των συνδιακυμάνσεων των αρχικών μεταβλητών, όπου τα διαγώνια στοιχεία είναι οι διακυμάνσεις της X_i και τα υπόλοιπα στοιχεία οι συνδιακυμάνσεις των μεταβλητών X_i και X_j.

$$C = \begin{bmatrix} s_{11}^2 & s_{12}^2 & . & s_{1p}^2 \\ s_{21}^2 & s_{22}^2 & . & s_{2p}^2 \\ . & . & . & . \\ s_{p1}^2 & s_{p2}^2 & . & s_{pp}^2 \end{bmatrix}$$

Με την τυποποίηση των αρχικών μεταβλητών η μήτρα των συνδιακυμάνσεων μεταπίπτει στη μήτρα των συσχετίσεων. Ουσιαστικά, η ανάλυση των κύριων συνιστωσών εκτελείται με βάση τη μήτρα των συσχετίσεων.

$$R = \begin{bmatrix} 1 & r_{12} & . & r_{1p} \\ r_{21} & 1 & . & r_{2p} \\ . & . & . & . \\ r_{p1} & r_{p2} & . & 1 \end{bmatrix}$$

Οι χαρακτηριστικές ρίζες – ιδιοτιμές (λ) και διανύσματα – ιδιοδιανύσματα (v) υπολογίζονται από την σχέση:

$$(R - \lambda_i I)v_i = 0$$

Η πρώτη κύρια συνιστώσα, με χαρακτηριστική ρίζα λ_1, προκύπτει από το γραμμικό συνδυασμό των p μεταβλητών:

$$Z_1 = v_{11}X_1 + v_{12}X_2 + v_{13}X_3 + \ldots + v_{1p}X_p$$

όπου ισχύει $v^2_{11} + v^2_{12} + v^2_{13} + \ldots + v^2_{1p} = 1$

…

Η p κύρια συνιστώσα, με χαρακτηριστική ρίζα λ_p, προκύπτει από το γραμμικό συνδυασμό των p μεταβλητών:

$$Z_p = v_{p1}X_1 + v_{p2}X_2 + v_{p3}X_3 + \ldots + v_{pp}X_p$$

όπου ισχύει $v^2_{p1} + v^2_{p2} + v^2_{p3} + \ldots + v^2_{pp} = 1$

Σημαντική ιδιότητα των χαρακτηριστικών ριζών είναι ότι το άθροισμά τους ισοδυναμεί με το άθροισμα των διακυμάνσεων των τυποποιημένων αρχικών μεταβλητών: $\lambda_1+\lambda_2+\ldots\lambda_p = c_{11}+c_{22}+\ldots+c_{pp} = p$

Αριθμητικό Παράδειγμα: 3 μεταβλητές 10 δειγμάτων

Τυποποίηση αρχικών μεταβλητών

Δείγμα	**X1**	**X2**	**X3**	**X1**	**X2**	**X3**
1	2	10	5	-1,796	-1,038	0,707
2	5	12	3	-0,482	-0,389	-0,236
3	7	17	1	0,394	1,232	-1,179
4	8	14	4	0,832	0,259	0,236
5	6	11	8	-0,044	-0,713	2,121
6	4	9	5	-0,920	-1,362	0,707
7	8	18	2	0,832	1,556	-0,707
8	10	16	1	1,708	0,908	-1,179
9	5	14	3	-0,482	0,259	-0,236
10	6	11	3	-0,044	-0,713	-0,236
Mean	6,10	13,20	3,50			
SD	2,28	3,08	2,12			
var(X)	5,21	9,51	4,50			

Υπολογισμός μήτρας C διακυμάνσεων - συνδιακυμάνσεων αρχικών μεταβλητών

	X1	X2	X3
X1	5,21	5,31	-2,5
X2	5,31	9,51	-4,77
X3	-2,5	-4,77	4,5

$$\mathrm{var}(X_i) = \frac{\Sigma (X_i - \overline{X}_i)^2}{n-1} \qquad \mathrm{cov}(X_i, X_j) = \frac{\Sigma (X_i - \overline{X}_i)(X_j - \overline{X}_j)}{n-1}$$

Υπολογισμός μήτρας R συσχετίσεων τυποποιημένων αρχικών μεταβλητών

	X1	X2	X3
X1	1	0,7544	-0,516
X2	0,754	1	-0,730
X3	-0,516	-0,730	1

Από τον πίνακα συσχετίσεων R

$$R = \begin{bmatrix} 1 & 0,75 & -0,52 \\ 0,75 & 1 & -0,73 \\ -0,52 & -0,73 & 1 \end{bmatrix}$$

και την σχέση

$$\left| R - \lambda I \right| = 0$$

υπολογίζουμε τις χαρακτηριστικές ρίζες - ιδιοτιμές (eigenvalue) λ_i

$$\left\| \begin{bmatrix} 1 & 0,75 & -0,52 \\ 0,75 & 1 & -0,73 \\ -0,52 & -0,73 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\| = 0$$

το οποίο μας δίνει λ_1 = 2,3393, λ_2 = 0,4843 και λ_3 = 0,1764

Για την χαρακτηριστική ρίζα λ_1= 2,339 και την σχέση

$$(R-\lambda_i I)v_i = 0 \Rightarrow$$

$$\left(\begin{bmatrix} 1 & 0{,}75 & -0{,}52 \\ 0{,}75 & 1 & -0{,}73 \\ -0{,}52 & -0{,}73 & 1 \end{bmatrix} - 2{,}339\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right)\begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow$$

$$\begin{bmatrix} -1{,}339 & 0{,}75 & -0{,}52 \\ 0{,}75 & -1{,}339 & -0{,}73 \\ -0{,}52 & -0{,}73 & -1{,}339 \end{bmatrix}\begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

υπολογίζουμε τα χαρακτηριστικά διανύσματα v_{11} = 0,56, v_{12} = 0,61 και v_{13}= -0,55

Επαναλαμβάνοντας την ίδια διαδικασία για τις χαρακτηριστικές ρίζες λ_2 = 0,484 και λ_3 = 0,176 , υπολογίζουμε και τα υπόλοιπα διανύσματα.

$$V = \begin{bmatrix} 0{,}56 & 0{,}68 & -0{,}46 \\ 0{,}61 & 0{,}03 & 0{,}78 \\ -0{,}52 & 0{,}72 & 0{,}40 \end{bmatrix}$$

Οι κύριες συνιστώσες είναι οι εξής:

$Z_1 = 0{,}56X_1 + 0{,}61X_2 - 0{,}52X_3$ με $\lambda_1 = 2{,}339$

$Z_2 = 0{,}68X_1 + 0{,}03X_2 + 0{,}72X_3$ με $\lambda_2 = 0{,}484$

$Z_3 = -0{,}46X_1 + 0{,}78X_2 + 0{,}40X_3$ με $\lambda_3 = 0{,}176$

P	Eigenvalue	Percent	Cum Percent
1	2,3393	77,97	77,97
2	0,4843	16,14	94,12
3	0,1764	5,88	100

```
> attach(pca.example)
> pca.example
```

	X1	X2	X3
1	2	10	5
2	5	12	3
3	7	17	1
4	8	14	4
5	6	11	8
6	4	9	5
7	8	18	2
8	10	16	1
9	5	14	3
10	6	11	3

```
> Scale=scale(pca.example)
> Scale
              X1         X2         X3
[1,]  -1.79605111 -1.0376105  0.7071068
[2,]  -0.48186737 -0.3891039 -0.2357023
[3,]   0.39425512  1.2321624 -1.1785113
[4,]   0.83231637  0.2594026  0.2357023
[5,]  -0.04380612 -0.7133572  2.1213203
[6,]  -0.91992862 -1.3618638  0.7071068
[7,]   0.83231637  1.5564157 -0.7071068
[8,]   1.70843886  0.9079092 -1.1785113
[9,]  -0.48186737  0.2594026 -0.2357023
[10,] -0.04380612 -0.7133572 -0.2357023

attr(,"scaled:center")
 X1   X2   X3
6.1 13.2  3.5
attr(,"scaled:scale")
      X1       X2       X3
2.282786 3.084009 2.121320
```

```
> cov=cov(pca.example)
> cov

          X1        X2        X3
X1  5.211111  5.311111 -2.500000
X2  5.311111  9.511111 -4.777778
X3 -2.500000 -4.777778  4.500000

> cor=cor(pca.example)
> cor

           X1         X2         X3
X1  1.0000000  0.7544051 -0.5162601
X2  0.7544051  1.0000000 -0.7303046
X3 -0.5162601 -0.7303046  1.0000000

> eigen=eigen(cor)
> eigen
eigen() decomposition
$values
[1] 2.3393060   0.4842657    0.1764283

$vectors
           [,1]        [,2]        [,3]
[1,] -0.5604809   0.68424760   -0.4665473
[2,] -0.6169608   0.03082708    0.7863899
[3,]  0.5524677   0.72859791    0.4048760
```

Πειραματικοί Σχεδιασμοί

```
> Vector=eigen$vectors
> Vector
           [,1]        [,2]        [,3]
[1,] -0.5604809  0.68424760 -0.4665473
[2,] -0.6169608  0.03082708  0.7863899
[3,]  0.5524677  0.72859791  0.4048760

> Score= Scale %*% Vector
> Score
            [,1]         [,2]         [,3]
[1,]   2.03747095   -0.7457336   0.3082670
[2,]   0.37992144   -0.5134437  -0.1766036
[3,]  -1.63225781   -0.5509088   0.3078705
[4,]  -0.49632078    0.7492393  -0.0888932
[5,]   1.63662687    1.4936246   0.3183324
[6,]   1.74647262   -0.1562447  -0.3554751
[7,]  -1.81739857    0.1022937   0.5493441
[8,]  -2.16878110    0.3383225  -0.5602479
[9,]  -0.02018169   -0.4934521   0.3333753
[10,]  0.33444806   -0.2236971  -0.6359694

> par(mfrow=c(2,2))
> biplot(Score[,c(1,2)], Vector[,c(1,2)])
> biplot(Score[,c(1,3)], Vector[,c(1,3)])
```