Ινστιτούτο Σχεδιασμού και Ανάλυσης Πειραμάτων του Γ.Π.Α.

3ος Κύκλος Διαλέξεων-Απρίλιος 2024

Πολυμεταβλητή Ανάλυση στην Πράξη

Μ. Κούτρας
Καθηγητής Τμήματος Στατιστικής & Ασφαλιστικής Επιστήμης
Δντης ΠΜΣ στην Εφαρμοσμένη Στατιστική
Πανεπιστήμιο Πειραιώς

Ανάλυση Κυρίων Συνιστωσών: η βασική ιδέα

- Είναι μια μέθοδος η οποία έχει ως στόχο να δημιουργήσει ένα μικρό αριθμό από γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να περιέχουν όσο γίνεται μεγαλύτερο μέρος της πληροφορίας που υπάρχει στις αρχικές μεταβλητές
- Τα οφέλη από την υλοποίηση αυτής της διαδικασίας είναι η οικονομία στην αποθήκευση καθώς και η ευκολία μελέτης των δεδομένων (εξόρυξη γνώσης) αφού μπορούμε να βασιστούμε σε πολύ μικρότερο αριθμό μεταβλητών απ' ότι είχαμε αρχικά

Principal Component Analysis (PCA)

Εισαγωγικά

$$X = \begin{bmatrix} 12 & 12 & 15 & 17 & 16 & 18 \\ 10 & 13 & 11 & 18 & 14 & 16 \\ 20 & 19 & 17 & 20 & 16 & 18 \\ 14 & 13 & 12 & 11 & 16 & 18 \\ 14 & 7 & 5 & 14 & 3 & 10 \end{bmatrix}$$

Παράδειγμα

Περιεκτικότητες 5 τροφίμων σε 6 συστατικά

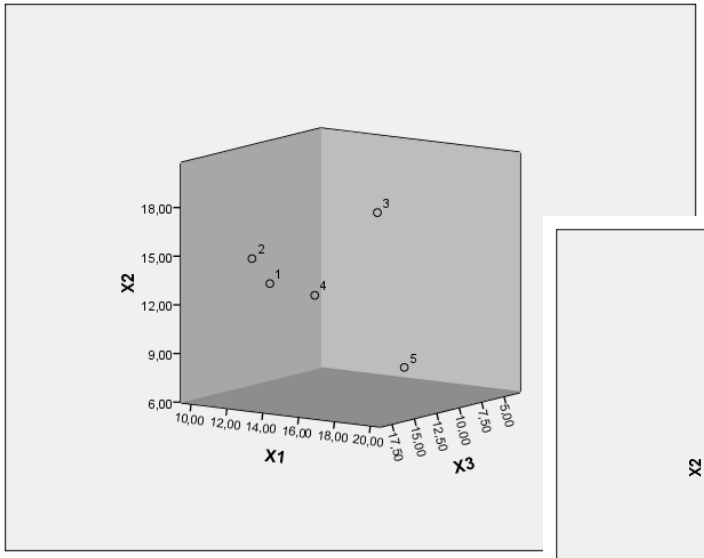
Τρόφιμο	Περιεκτικότητα στο *i* συστατικό					
	1	**2**	**3**	**4**	**5**	**6**
1	12	13	14	17	16	18
2	10	14	12	18	14	16
3	20	19	18	20	16	18
4	13	12	11	11	16	18
5	15	7	5	14	3	10

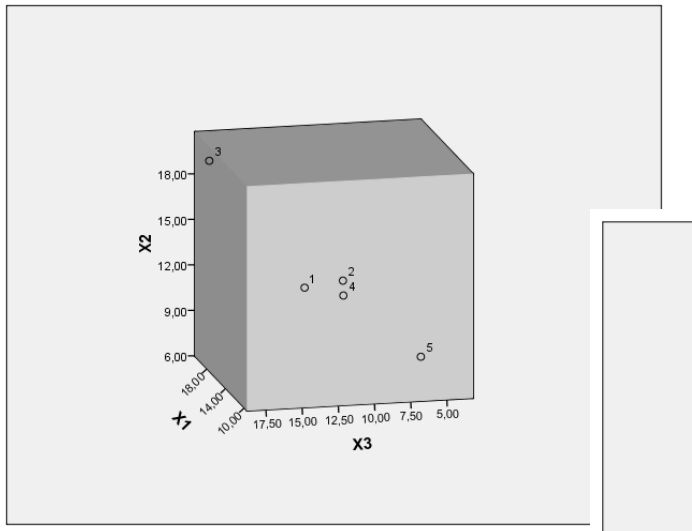
$$X = \begin{bmatrix} 12 & 13 & 14 & 17 & 16 & 18 \\ 10 & 14 & 12 & 18 & 14 & 16 \\ 20 & 19 & 18 & 20 & 16 & 18 \\ 13 & 12 & 11 & 11 & 16 & 18 \\ 15 & 7 & 5 & 14 & 3 & 10 \end{bmatrix}$$

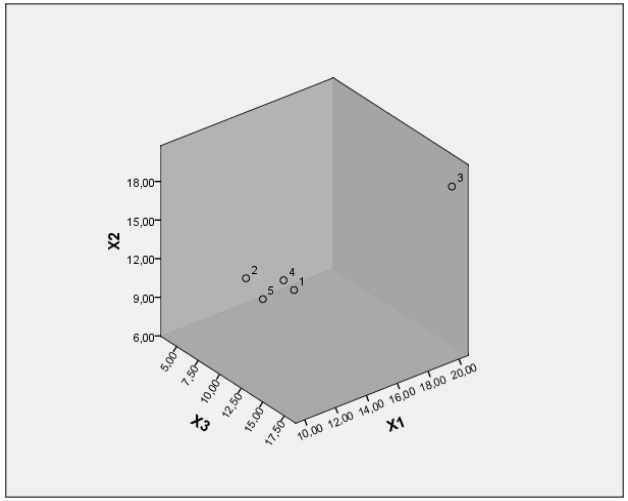
Παράδειγμα

Τρεις διαφορετικές εικόνες του ίδιου πράγματος!

Τρόφιμο i	Συστατικό j		
	1	**2**	**3**
1	12	13	14
2	10	14	12
3	20	19	18
4	13	12	11
5	15	7	5







Παράδειγμα: αποστάσεις μεταξύ των τροφίμων

Τρόφιμο *i*	Συστατικό *j*		
	1	**2**	**3**
1	12	13	14
2	10	14	12
3	20	19	18
4	13	12	11
5	15	7	5

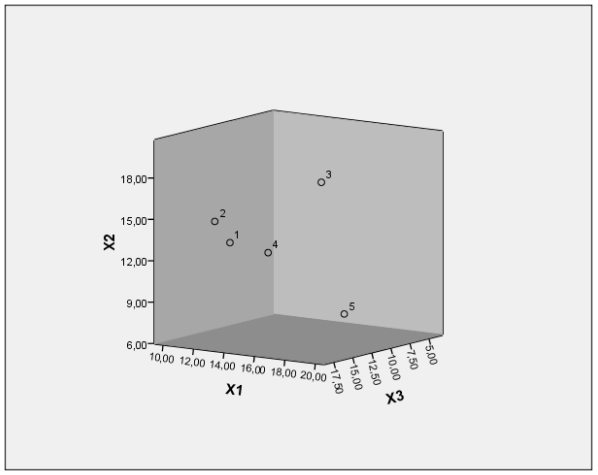
Τρόφιμο	**1**	**2**	**3**	**4**	**5**
1	0	**3**	10.77	3.32	11.22
2	3	0	12.69	3.74	11.09
3	10.77	12.69	0	12.12	**18.38**
4	3.32	3.74	12.12	0	8.06
5	11.22	11.09	18.38	8.06	0

Πίνακας (αποστάσεων)

Παράδειγμα

Πως θα μπορούσε να γίνει μια λογική ταξινόμηση των τροφίμων με βάση τις περιεκτικότητες?

Τρόφιμο	1	2	3	4	5
1	0	3	10.77	3.32	11.22
2	3	0	12.69	3.74	11.09
3	10.77	12.69	0	12.12	18.38
4	3.32	3.74	12.12	0	8.06
5	11.22	11.09	18.38	8.06	0



$$Y_1 = \frac{X_1 + X_2 + X_3}{3} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$$

$$Y_2 = \frac{8}{10}X_1 + \frac{1}{10}X_2 + \frac{1}{10}X_3 = 0.8X_1 + 0.1X_2 + 0.1X_3,$$

$$Y_3 = \frac{2}{10}X_1 + \frac{4}{10}X_2 + \frac{4}{10}X_3 = 0.2X_1 + 0.4X_2 + 0.4X_3$$

i	X_1	X_2	X_3	Y_1	Y_2	Y_3
1	12	13	14	13	12.3	13.4
2	10	14	12	12	10.6	12.4
3	20	19	18	19	19.7	18.6
4	13	12	11	12	12.7	11.6
5	15	7	5	9	13.2	7.0

Παράδειγμα

$$Y_1 = \frac{X_1 + X_2 + X_3}{3} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$$

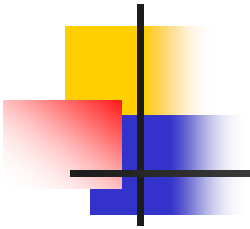
$$Y_2 = \frac{8}{10}X_1 + \frac{1}{10}X_2 + \frac{1}{10}X_3 = 0.8X_1 + 0.1X_2 + 0.1X_3,$$

$$Y_3 = \frac{2}{10}X_1 + \frac{4}{10}X_2 + \frac{4}{10}X_3 = 0.2X_1 + 0.4X_2 + 0.4X_3$$

$$Y = a_1X_1 + a_2X_2 + a_3X_3$$

i	X_1	X_2	X_3	Y_1	Y_2	Y_3
1	12	13	14	13	12.3	13.4
2	10	14	12	12	10.6	12.4
3	20	19	18	19	19.7	18.6
4	13	12	11	12	12.7	11.6
5	15	7	5	9	13.2	7.0

Ανάλυση Κυρίων Συνιστωσών



Ανάλυση Κυρίων Συνιστωσών

Είναι μια μέθοδος η οποία έχει ως στόχο να δημιουργήσει ένα μικρό αριθμό από γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί

- να είναι ασυσχέτιστοι μεταξύ τους
- να περιέχουν όσο γίνεται μεγαλύτερο μέρος της πληροφορίας που υπάρχει στις αρχικές μεταβλητές

Εύρεση της πρώτης κύριας συνιστώσας

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

Θεωρώντας τη νέα μεταβλητή

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_p X_p$$

η οποία προκύπτει ως γραμμικός συνδυασμός των $X_1, X_2, \ldots . X_p$, το ενδιαφέρον μας εστιάζεται στον προσδιορισμό των $a_1, a_2, \ldots . a_p \in \Re$ έτσι ώστε οι τιμές (scores) των n ατόμων για τη μεταβλητή Y, δηλαδή τα

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_p x_{ip}, \quad i = 1,2,\ldots,n$$

να 'διατηρούν' όσο το δυνατόν περισσότερο τις αποστάσεις που έχουν τα άτομα ως προς όλες τις αρχικές μεταβλητές.

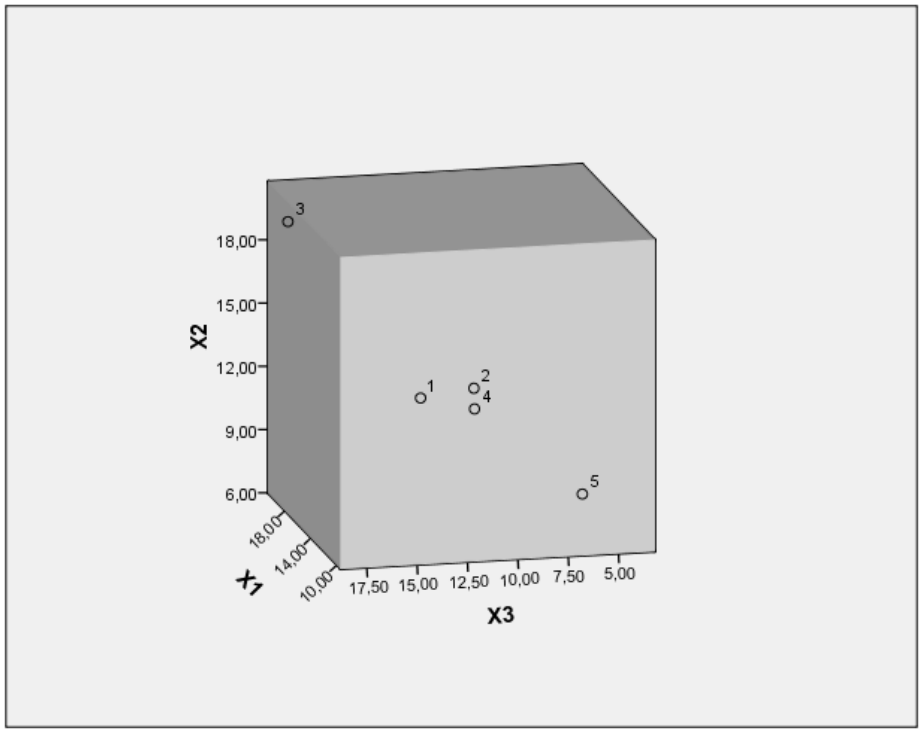
$x_1, x_2, \dots, x_n$

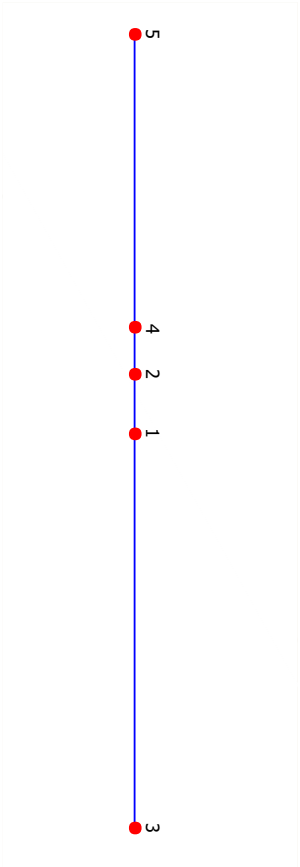
$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

$x_1 = (x_{11}, x_{12}, \dots, x_{1p})$

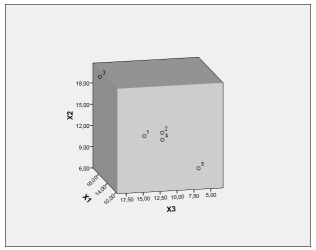
$x_2 = (x_{21}, x_{22}, \dots, x_{2p})$

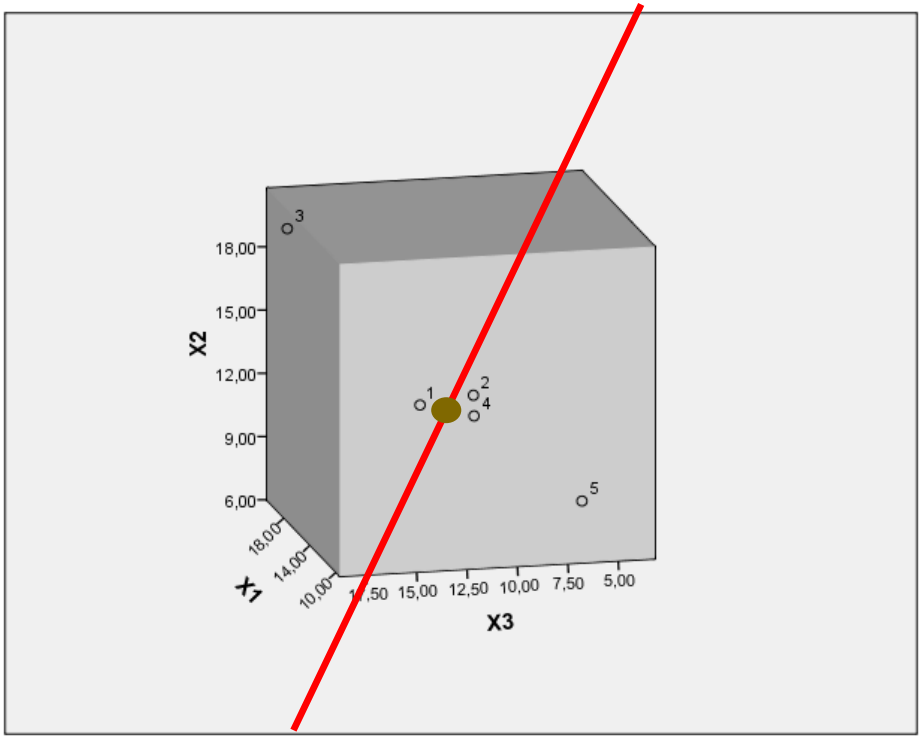
$$y_1 = a_1\ x_{11} + a_2x_{12} + \cdots + a_px_{1p}$$
$$y_2 = a_1\ x_{21} + a_2x_{22} + \cdots + a_px_{2p}$$
$$\dots$$
$$y_n = a_1\ x_{n1} + a_2x_{n2} + \cdots + a_px_{np}$$

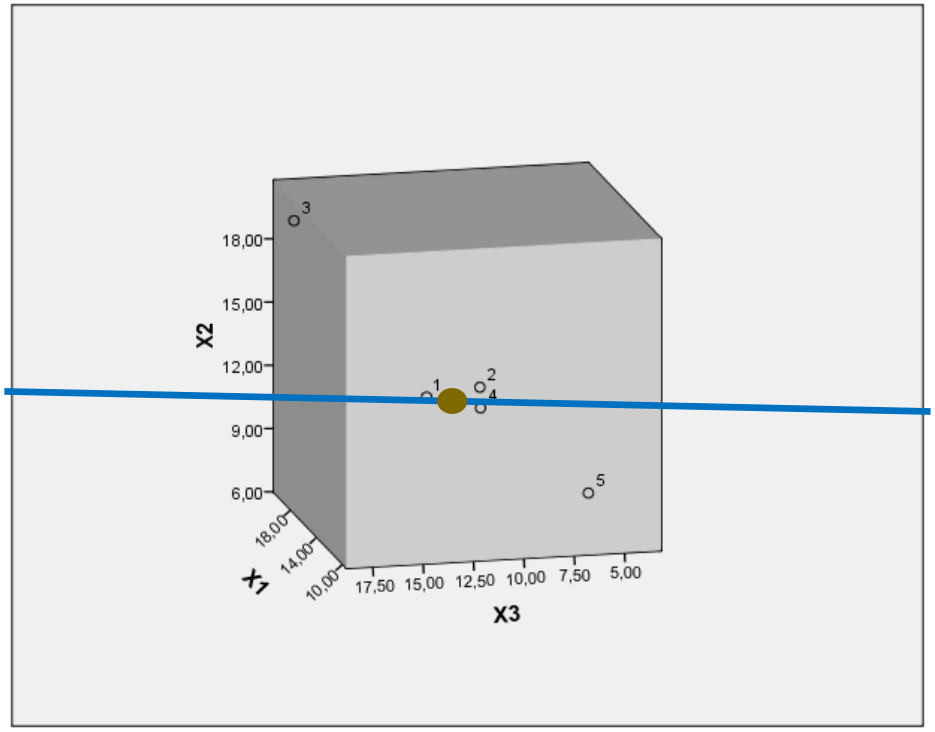




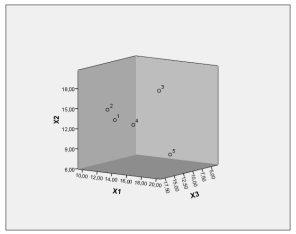
Εύρεση της πρώτης κύριας συνιστώσας

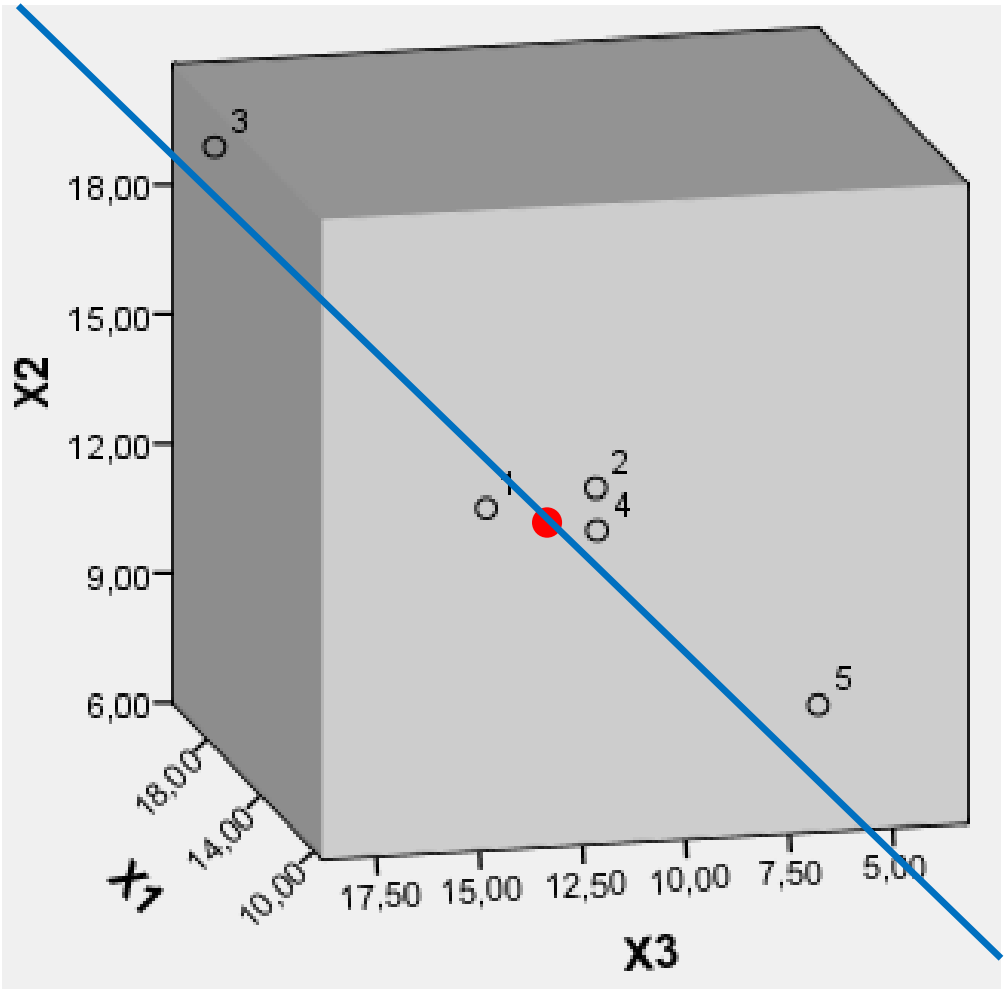






Εύρεση της πρώτης κύριας συνιστώσας

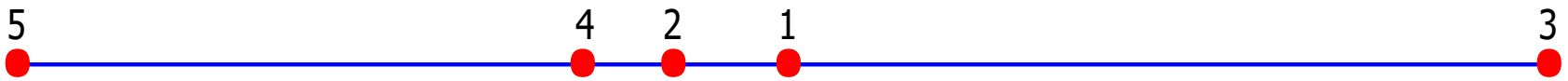




Εύρεση της πρώτης κύριας συνιστώσας

$$y_1 = a_1\, x_{11} + a_2 x_{12} + \cdots + a_p x_{1p}$$
$$y_2 = a_1\, x_{21} + a_2 x_{22} + \cdots + a_p x_{2p}$$
$$\ldots$$
$$y_n = a_1\, x_{n1} + a_2 x_{n2} + \cdots + a_p x_{np}$$

$$Dis(Y) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



Το πρόβλημα της εύρεσης των «βέλτιστων» συντελεστών ανάγεται στη μεγιστοποίηση της παραπάνω ποσότητας η οποία θα συμβολίζεται και ως

$$Dis_{\boldsymbol{a}}(N) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Έκφραση της $Dis(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$ μέσω του πίνακα X

$$y_1 = a_1\, x_{11} + a_2 x_{12} + \cdots + a_p x_{1p}$$
$$y_2 = a_1\, x_{21} + a_2 x_{22} + \cdots + a_p x_{2p}$$
$$\cdots$$
$$y_n = a_1\, x_{n1} + a_2 x_{n2} + \cdots + a_p x_{np}$$

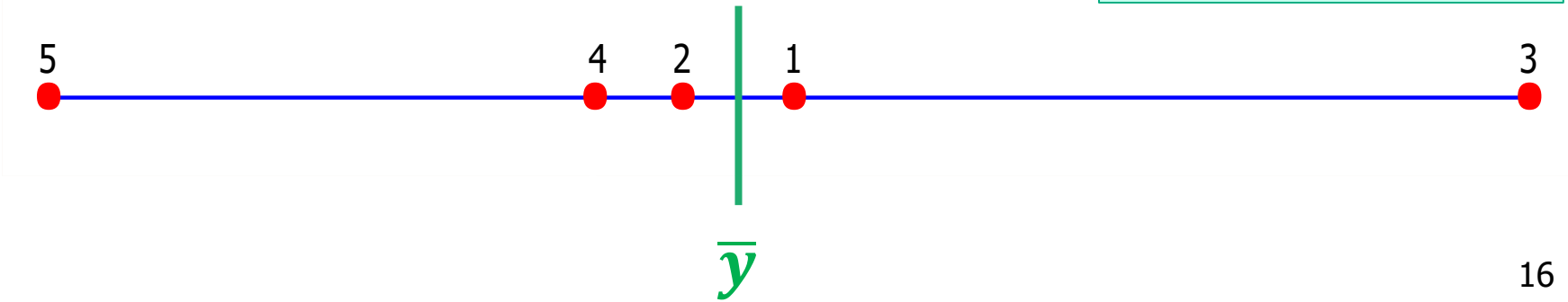
$$f_i = y_i - \bar{y}$$

$$Dis(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} f_i^2$$

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

$$\boldsymbol{f} = \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{pmatrix}$$

$$Dis(Y) = \sum_{i=1}^{n} f_i^2 = \boldsymbol{f'f}$$



Ο πίνακας Z

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \dots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

$\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_p$

Πίνακας των κεντρικοποιημένων δεδομένων

$-\bar{x}_1$ $-\bar{x}_2$ $-\bar{x}_p$

$$Z = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1j} - \bar{x}_j & \cdots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} - \bar{x}_1 & \cdots & x_{ij} - \bar{x}_j & \dots & x_{ip} - \bar{x}_p \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nj} - \bar{x}_j & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

Έκφραση της $Dis(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$ μέσω του πίνακα Z

$$Z = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1j} - \bar{x}_j & \cdots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots & \square & \vdots \\ x_{i1} - \bar{x}_1 & \cdots & x_{ij} - \bar{x}_j & \cdots & x_{ip} - \bar{x}_p \\ \vdots & \square & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nj} - \bar{x}_j & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

$$\boldsymbol{f} = \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{pmatrix} = Z\boldsymbol{a}$$

$$Dis(Y) = \boldsymbol{a}'(Z'Z)\boldsymbol{a}$$

Διασπορά του συνόλου Ν

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p}$$

- Πίνακας κεντρικοποιημένων δεδομένων

$$Z = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1j} - \bar{x}_j & \cdots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} - \bar{x}_1 & \cdots & x_{ij} - \bar{x}_j & \cdots & x_{ip} - \bar{x}_p \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nj} - \bar{x}_j & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

- Διασπορά του συνόλου $N = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$ κατά μήκος του διανύσματος $\boldsymbol{\alpha}$

$$Dis(Y) = Dis_{\boldsymbol{a}}(N) = \boldsymbol{a}'(Z'Z)\boldsymbol{a}$$

- Scores των *n* ατόμων

$$\boldsymbol{f} = Z\boldsymbol{a}$$

$$X = \begin{bmatrix} 12 & 13 & 14 \\ 10 & 14 & 12 \\ 20 & 19 & 18 \\ 13 & 12 & 11 \\ 15 & 7 & 5 \end{bmatrix}$$

i	X_1	X_2	X_3	Y_1	Y_2	Y_3
1	12	13	14	13	12.3	13.4
2	10	14	12	12	10.6	12.4
3	20	19	18	19	19.7	18.6
4	13	12	11	12	12.7	11.6
5	15	7	5	9	13.2	7.0

Παράδειγμα

$$\bar{x}_1 = \frac{12+10+\ldots+15}{5} = 14,\ \bar{x}_2 = \frac{13+14+\ldots+7}{5} = 13,\ \bar{x}_3 = \frac{14+12+\ldots+5}{5} = 12$$

$$Z = \begin{bmatrix} 12-14 & 13-13 & 14-12 \\ 10-14 & 14-13 & 12-12 \\ 20-14 & 19-13 & 18-12 \\ 13-14 & 12-13 & 11-12 \\ 15-14 & 7-13 & 5-12 \end{bmatrix} = \begin{bmatrix} -2 & 0 & 2 \\ -4 & 1 & 0 \\ 6 & 6 & 6 \\ -1 & -1 & -1 \\ 1 & -6 & -7 \end{bmatrix}.$$

$$Z'Z = \begin{bmatrix} -2 & -4 & -6 & -1 & -1 \\ 0 & 1 & 6 & -1 & -6 \\ 2 & 0 & 6 & -1 & -7 \end{bmatrix} \begin{bmatrix} -2 & 0 & 2 \\ -4 & 1 & 0 \\ 6 & 6 & 6 \\ -1 & -1 & -1 \\ 1 & -6 & -7 \end{bmatrix} = \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix}$$

$$X = \begin{bmatrix} 12 & 13 & 14 \\ 10 & 14 & 12 \\ 20 & 19 & 18 \\ 13 & 12 & 11 \\ 15 & 7 & 5 \end{bmatrix}$$

i	X_1	X_2	X_3	Y_1	Y_2	Y_3
1	12	13	14	13	12.3	13.4
2	10	14	12	12	10.6	12.4
3	20	19	18	19	19.7	18.6
4	13	12	11	12	12.7	11.6
5	15	7	5	9	13.2	7.0

Παράδειγμα

$$Z'Z = \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix}$$

$$Dis_{\boldsymbol{\alpha}_1}(N) = \boldsymbol{\alpha}_1'(Z'Z)\boldsymbol{\alpha}_1 = (\frac{1}{3},\frac{1}{3},\frac{1}{3}) \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix} \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} = 54\,.$$

Όμοια, για τα διανύσματα $\boldsymbol{\alpha}_2 = (8/10,1/10,1/10)'$ και $\boldsymbol{\alpha}_3 = (2/10,4/10,4/10)'$ έχουμε

$$Dis_{\boldsymbol{\alpha}_2}(N) = 48.8\,, \qquad Dis_{\boldsymbol{\alpha}_3}(N) = 69\,.$$

Δεδομένου ότι οι ποσότητες αυτές μας ενδιαφέρει να έχουν όσο το δυνατόν μεγαλύτερη τιμή, ανάμεσα στις τρεις περιπτώσεις θα επιλέγαμε το $\boldsymbol{\alpha}_3 = (2/10,4/10,4/10)'$.

Διασπορά του συνόλου σημείων *N*

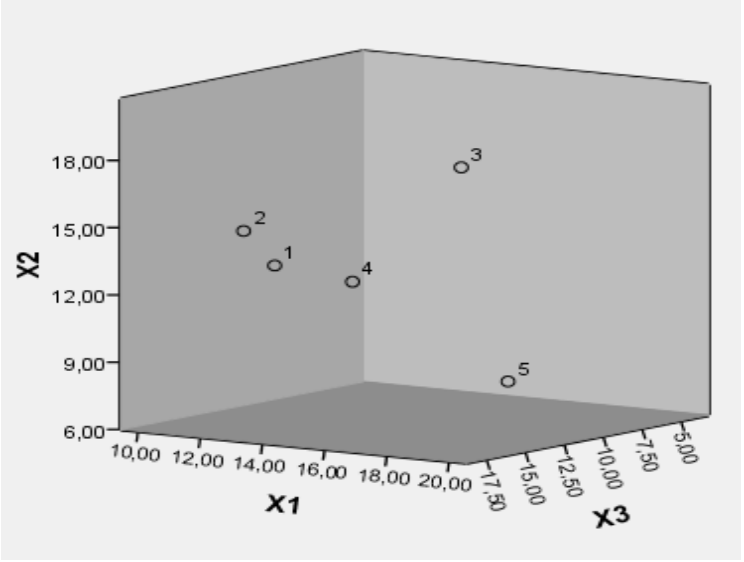
$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

Για τα αρχικά δεδομένα έχουμε

$$Dis(N) = \sum_{i=1}^{n} d^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}})$$

και η διασπορά του συνόλου σημείων *N* υπολογίζεται από τον τύπο

$$\boldsymbol{Dis(N) = trace(Z'Z)}$$



Εύρεση της πρώτης κύριας συνιστώσας

$$Z = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1j} - \bar{x}_j & \cdots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots & \square & \vdots \\ x_{i1} - \bar{x}_1 & \cdots & x_{ij} - \bar{x}_j & \cdots & x_{ip} - \bar{x}_p \\ \vdots & \square & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nj} - \bar{x}_j & \cdots & x_{np} - \bar{x}_p \end{bmatrix}$$

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

$$\boldsymbol{\alpha} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$$

$$f = \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{pmatrix} = Za$$

$$Dis(N) = trace(Z'Z)$$

$$Dis(Y) = a'(Z'Z)a$$

$$Dis(Y) = a'(Z'Z)a \leq Dis(N) = trace(Z'Z)$$

$$Dis(N) = tr(Z'Z) = \sum_{j=1}^{p} \lambda_j = \lambda_1 + \lambda_2 + ... + \lambda_p$$

Εύρεση της πρώτης κύριας συνιστώσας (καιρός ήταν!)

ΠΡΟΤΑΣΗ 2.2.1. Ο πίνακας $Z'Z$ έχει μη αρνητικές ιδιότιμες, έστω $\lambda_1, \lambda_2, ..., \lambda_p$. Ας θεωρήσουμε ότι $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ και ας συμβολίσουμε $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_p$ τα αντίστοιχα μοναδιαία ιδιοδιανύσματα. Τότε

i. το διάνυσμα $\boldsymbol{\alpha}$ που μεγιστοποιεί την $\boldsymbol{\alpha}'(Z'Z)\boldsymbol{\alpha}$ είναι το μοναδιαίο διάνυσμα $\mathbf{u}_1$ που αντιστοιχεί στη μεγαλύτερη ιδιότιμη λ_1

ii. η μέγιστη τιμή της τετραγωνικής μορφής είναι ίση με λ_1, δηλαδή ισχύει

$$\max_{||\mathbf{a}||=1} Dis_{\mathbf{a}}(N) = Dis_{\mathbf{u}_1}(N) = \lambda_1 .$$

Η μεταβλητή Y για την οποία επιτυγχάνεται η προαναφερθείσα μεγιστοποίηση λέγεται **πρώτη κύρια συνιστώσα** (first principal component).

i	X_1	X_2	X_3
1	12	13	14
2	10	14	12
3	20	19	18
4	13	12	11
5	15	7	5

$$Z'Z = \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix}$$

Παράδειγμα

$$|Z'Z - \lambda I_3| = 0 \Rightarrow \begin{vmatrix} 58-\lambda & 27 & 26 \\ 27 & 74-\lambda & 79 \\ 26 & 79 & 90-\lambda \end{vmatrix} = 0 \Rightarrow -\lambda^3 + 222\lambda^2 - 8526\lambda + 19854 = 0$$

$$\lambda_1 = 173.51,\ \lambda_2 = 46.04,\ \lambda_3 = 2.45$$

$$(Z'Z)\mathbf{u} = \lambda\mathbf{u} \Leftrightarrow \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \lambda\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

με τον περιορισμό $x^2 + y^2 + z^2 = 1$

Ιδιοδιανύσματα:

$\lambda = \lambda_1 = 173.5$	$\lambda = \lambda_2 = 46.0$	$\lambda = \lambda_3 = 2.45$
$\mathbf{u}_1 = (0.31, 0.64, 0.70)'$	$\mathbf{u}_2 = (-0.95, 0.16, 0.27)'$	$\mathbf{u}_3 = (0.06, -0.75, 0.66)'$

i	X_1	X_2	X_3
1	12	13	14
2	10	14	12
3	20	19	18
4	13	12	11
5	15	7	5

$$Z'Z = \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix}$$

Παράδειγμα 1
(συνέχεια)

Ιδιοδιανύσματα:

$\lambda = \lambda_1 = 173.5$	$\lambda = \lambda_2 = 46.0$	$\lambda = \lambda_3 = 2.45$
$\mathbf{u}_1 = (0.31, 0.64, 0.70)'$	$\mathbf{u}_2 = (-0.95, 0.16, 0.27)'$	$\mathbf{u}_3 = (0.06, -0.75, 0.66)'$

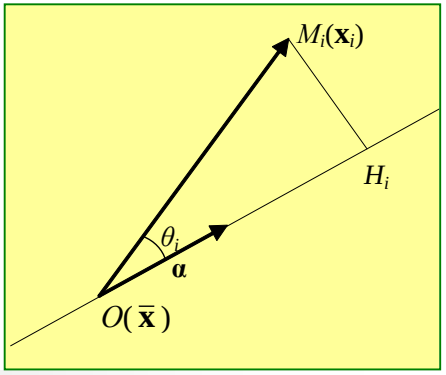
Επομένως η **καλύτερη στάθμιση** της βαθμολογίας (με την έννοια της 'πιστότερης' διατήρησης της μεταξύ των ατόμων αποστάσεων) είναι η

$$Y = 0.31x_1 + 0.64x_2 + 0.70x_3 .$$

$$Dis(Y) = Dis_{\mathbf{u}_1}(Y) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 173.51 .$$

Αξίζει να σημειωθεί ότι η τιμή αυτή είναι **αρκετά κοντά στη συνολική διασπορά** των $n = 5$ τρισδιάστατων σημείων, η οποία έχει βρεθεί ίση με $Dis(N) = tr(Z'Z) = 222$. Συνήθως λέμε ότι *η μεταβλητή Y εξηγεί το* 173.51/222=78% *της διασποράς των (αρχικών) δεδομένων.*

Χρήσιμες ποσότητες για την ερμηνεία της πρώτης κύριας συνιστώσας



Ποσοστό της συνολικής διασποράς που εξηγείται από τον πρώτο κύριο άξονα

$$\frac{Dis_{\alpha}(N)}{Dis(N)} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \leq 1$$

Ποιότητα αναπαράστασης της *i*-οστής παρατήρησης

$$Q(\mathbf{x}_i) = \frac{f_i^2}{OM_i^2} = \frac{f_i^2}{\|z_i\|^2} = \frac{f_i^2}{\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}, \qquad i = 1,2,\ldots,n$$

Τιμές του $Q(x_i)$ μεγαλύτερες από 90% αντιστοιχούν σε γωνίες θ_i μικρότερες των 18.5^o ενώ τιμές μεγαλύτερες από 95% σε γωνίες μικρότερες από 13^o.

$$\text{συν}\,\theta_i = \frac{f_i}{OM_i}$$

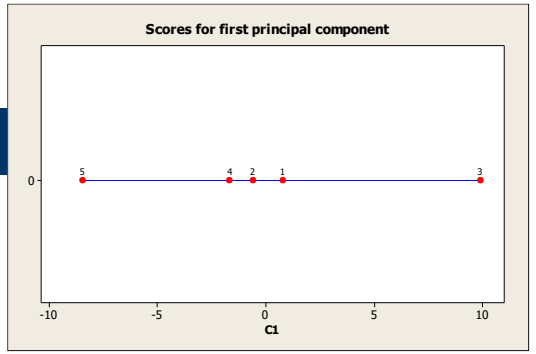
i	X_1	X_2	X_3
1	12	13	14
2	10	14	12
3	20	19	18
4	13	12	11
5	15	7	5

$$\lambda = \lambda_1 = 173.5$$

$$\mathbf{u}_1 = (0.31, 0.64, 0.70)'$$

$$Z'Z = \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix}$$

Παράδειγμα



$$y_i = 0.31x_{i1} + 0.64x_{i2} + 0.70x_{i3}, \quad i = 1,2,\ldots,5$$

$$f_i = y_i - \bar{y} = y_i - (0.31\bar{x}_1 + 0.64\bar{x}_2 + 0.70\bar{x}_3) = 0.31z_{i1} + 0.64z_{i2} + 0.71z_{i3}$$

i	z_{i1}	z_{i2}	z_{i3}	z_{i1}^2	z_{i1}^2	z_{i1}^2	$\| \mathbf{z}_i \|^2$	f_i	$Q(\mathbf{x}_i)$
1	-2	0	2	4	0	4	8	0.78	0.08
2	-4	1	0	16	1	0	17	-0.60	0.02
3	6	6	6	36	36	36	108	9.91	0.91
4	-1	-1	-1	1	1	1	3	-1.65	0.91
5	1	-6	-7	1	36	49	86	-8.4	0.83
Άθροισμα	**0**	**0**	**0**	**58**	**74**	**90**	**222**	**0**	-

i	X_1	X_2	X_3
1	12	13	14
2	10	14	12
3	20	19	18
4	13	12	11
5	15	7	5

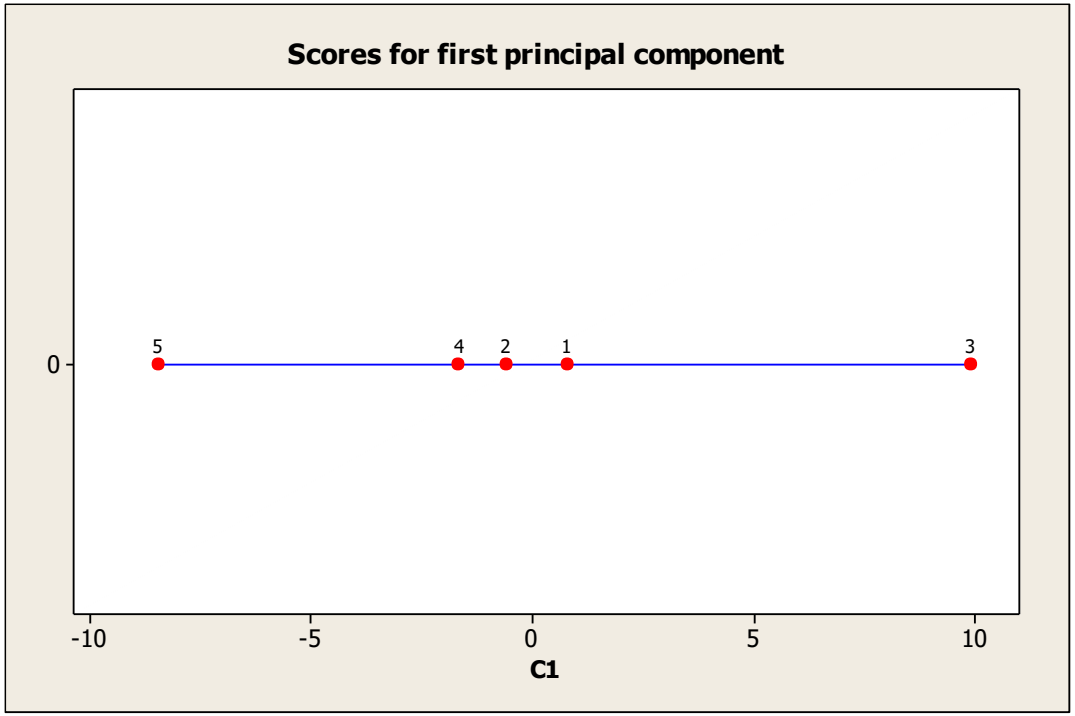
$$\lambda = \lambda_1 = 173.5$$

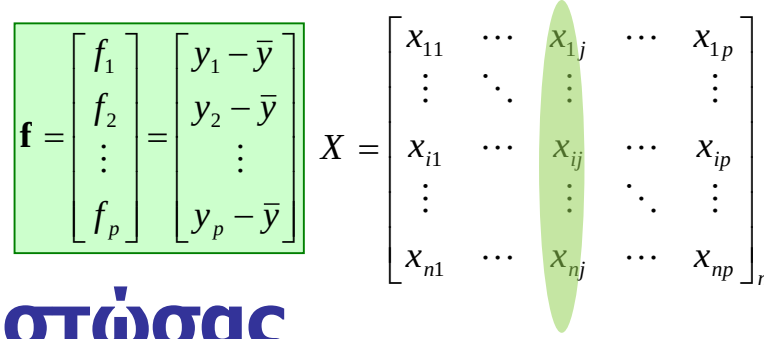
$$\mathbf{u}_1 = (0.31, 0.64, 0.70)'$$

$$Z'Z = \begin{bmatrix} 58 & 27 & 26 \\ 27 & 74 & 79 \\ 26 & 79 & 90 \end{bmatrix}$$

Παράδειγμα

i	f_i	$Q(\mathbf{x}_i)$
1	0.78	0.08
2	-0.60	0.02
3	9.91	0.91
4	-1.65	0.91
5	-8.4	0.83
Άθροισμα	**0**	-





Στατιστική ερμηνεία της πρώτης κύριας συνιστώσας

Δειγματικός συντελεστής συσχέτισης μεταξύ της 1ης κύριας συνιστώσας και της j αρχικής μεταβλητής X_j

$$r_{1j} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

i	X_1	X_2	X_3
1	12	13	14
2	10	14	12
3	20	19	18
4	13	12	11
5	15	7	5

$$\lambda = \lambda_1 = 173.5$$

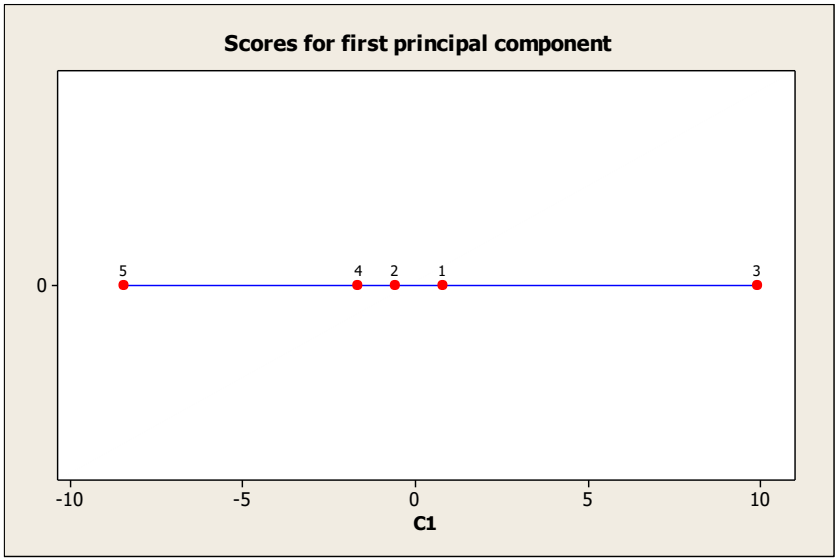
$$\mathbf{u}_1 = (0.31, 0.64, 0.70)'$$

Παράδειγμα

$$r_{11} = \frac{53.46}{\sqrt{172.77}\sqrt{58}} = 53\%$$

$$r_{12} = \frac{11.03}{\sqrt{172.77}\sqrt{74}} = 98\%$$

$$r_{13} = \frac{121.62}{\sqrt{172.77}\sqrt{90}} = 97\%$$



Οι υπόλοιπες κύριες συνιστώσες

Οι υπόλοιπες κύριες συνιστώσες

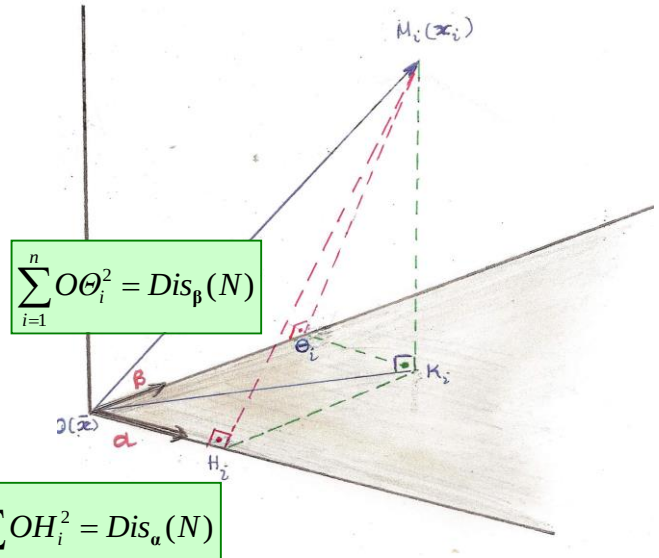
$$Dis(N) = \sum_{i=1}^{n} OH_i^2 + \sum_{i=1}^{n} O\Theta_i^2 + \sum_{i=1}^{n} K_i M_i^2$$

Η ποσότητα

$$Dis_{\alpha,\beta}(N) = Dis_{\alpha}(N) + Dis_{\beta}(N) \qquad (2.6.2)$$

θα λέγεται **διασπορά του συνόλου *N* στο επίπεδο** που καθορίζεται από τα μοναδιαία (και κάθετα μεταξύ τους) διανύσματα $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

Την ποσότητα αυτή θα πρέπει να μεγιστοποιήσουμε.



$$Dis(N) = Dis_{\alpha}(N) + Dis_{\beta}(N) + \sum_{i=1}^{n} K_i M_i^2$$

Οι υπόλοιπες κύριες συνιστώσες

$$Dis_{\alpha,\beta}(N) = Dis_{\alpha}(N) + Dis_{\beta}(N)$$

Αυτό μπορεί να γίνει επιλέγοντας αρχικά το $\boldsymbol{\alpha}$ έτσι ώστε να μεγιστοποιείται η

$$Dis_{\alpha}(N) = \boldsymbol{\alpha}' Z' Z \boldsymbol{\alpha}$$

και στη συνέχεια το $\boldsymbol{\beta}$ έτσι ώστε να είναι κάθετο στο $\boldsymbol{\alpha}$ και να μεγιστοποιείται η

$$Dis_{\beta}(N) = \boldsymbol{\beta}' Z' Z \boldsymbol{\beta}.$$

Επομένως μπορούμε να πάρουμε το $\boldsymbol{\alpha}$ ίσο με το $\mathbf{u}_1$ και θα έχουμε

$$\max_{\|\boldsymbol{\alpha}\|=1} Dis_{\alpha}(N) = Dis_{\mathbf{u}_1}(N) = \lambda_1$$

και το $\boldsymbol{\beta}$ θα πρέπει να επιλεγεί ως το μοναδιαίο ιδιοδιάνυσμα $\mathbf{u}_2$ που αντιστοιχεί στη **δεύτερη μεγαλύτερη ιδιοτιμή** λ_2 του πίνακα $Z'Z$ και θα ισχύει

$$Dis_{\mathbf{u}_2}(N) = \lambda_2.$$

Οι υπόλοιπες κύριες συνιστώσες

$$Dis_{\alpha,\beta}(N) = Dis_{\alpha}(N) + Dis_{\beta}(N)$$

Στο διάνυσμα $\mathbf{u}_1 = (\mathbf{u}_{11}, \mathbf{u}_{12}, ..., \mathbf{u}_{1p})'$ αντιστοιχεί ο γραμμικός συνδυασμός

$$Y_1 = u_{11}X_1 + u_{12}X_2 + ... + u_{1p}X_p$$

(τον οποίο ονομάσαμε **πρώτη κύρια συνιστώσα**), ενώ το διάνυσμα $\mathbf{u}_{21} = (\mathbf{u}_{21}, \mathbf{u}_{22}, ..., \mathbf{u}_{2p})'$ θα δημιουργεί ένα δεύτερο γραμμικό συνδυασμό

$$Y_2 = u_{21}X_1 + u_{22}X_2 + ... + u_{2p}X_p$$

που θα λέγεται **δεύτερη κύρια συνιστώσα**.

Οι υπόλοιπες κύριες συνιστώσες

$$Dis_{\alpha,\beta}(N) = Dis_{\alpha}(N) + Dis_{\beta}(N)$$

Από τη συνολική διασπορά του συνόλου N (δηλαδή των διαθέσιμων δεδομένων) η οποία είναι ίση με $Dis(N) = tr(Z)$, η πρώτη κύρια συνιστώσα εξηγεί ποσοστό ίσο με

$$\frac{Dis(Y_1)}{Dis(N)} = \frac{Dis_{\mathbf{u}_1}(N)}{Dis(N)} = \frac{\lambda_1}{tr(Z'Z)} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \ldots + \lambda_p},$$

η δεύτερη κύρια συνιστώσα εξηγεί ποσοστό ίσο με

$$\frac{Dis(Y_2)}{Dis(N)} = \frac{Dis_{\mathbf{u}_2}(N)}{Dis(N)} = \frac{\lambda_2}{tr(Z'Z)} = \frac{\lambda_2}{\lambda_1 + \lambda_2 + \ldots + \lambda_p},$$

ενώ και οι δύο μαζί ποσοστό ίσο με

$$\frac{Dis(Y_1, Y_2)}{Dis(N)} = \frac{Dis_{\mathbf{u}_1,\mathbf{u}_2}(N)}{Dis(N)} = \frac{\lambda_1 + \lambda_2}{tr(Z'Z)} = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}.$$

Οι υπόλοιπες κύριες συνιστώσες

Τα scores $f_{i1},\ i=1,2,...,n$ και $f_{i2},\ i=1,2,...,n$ των n ατόμων στις δύο κύριες συνιστώσες θα δίνονται από τους τύπους

$$f_{i1}=u_{11}z_{i1}+u_{12}z_{i2}+...+u_{1p}z_{ip},\qquad f_{i2}=u_{21}z_{i1}+u_{22}z_{i2}+...+u_{2p}z_{ip}\ \text{για}\ i=1,2,...,n$$

ή ισοδύναμα, με χρήση πινάκων

$$\mathbf{f}_1=\begin{bmatrix} f_{11}\\ f_{21}\\ \vdots\\ f_{n1}\end{bmatrix}=Z\mathbf{u}_1,\qquad \mathbf{f}_2=\begin{bmatrix} f_{12}\\ f_{22}\\ \vdots\\ f_{n2}\end{bmatrix}=Z\mathbf{u}_2.$$

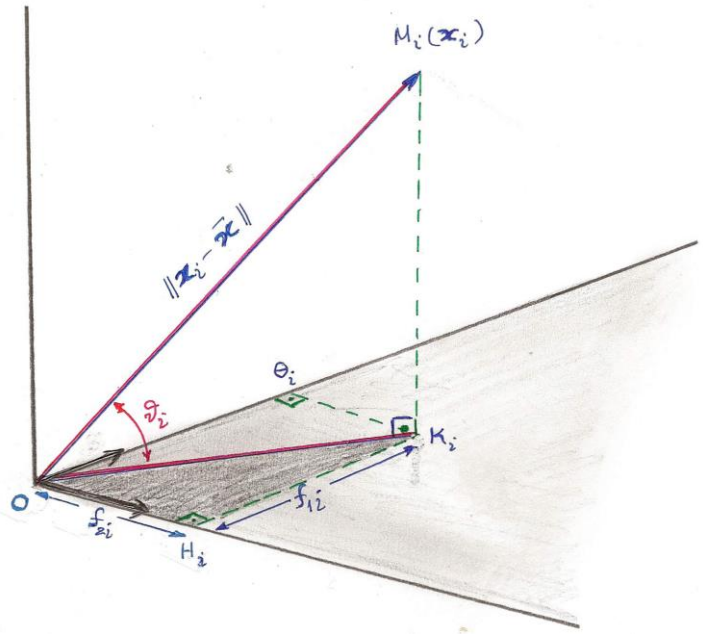
Οι υπόλοιπες κύριες συνιστώσες

Η ποιότητα αναπαράστασης της i-οστής παρατήρησης $\mathbf{x}_i$ στο επίπεδο είναι ίση με

$$Q(\mathbf{x}_i) = \text{συν}^2\theta_i = \frac{OK_i^2}{OM_i^2} = \frac{f_{1i}^2 + f_{2i}^2}{\| \mathbf{z}_i \|^2}$$

δηλαδή

$$Q(\mathbf{x}_i) = \frac{f_{1i}^2 + f_{2i}^2}{\| \mathbf{x}_i - \bar{\mathbf{x}} \|^2}, \qquad i = 1,2,\ldots,n$$



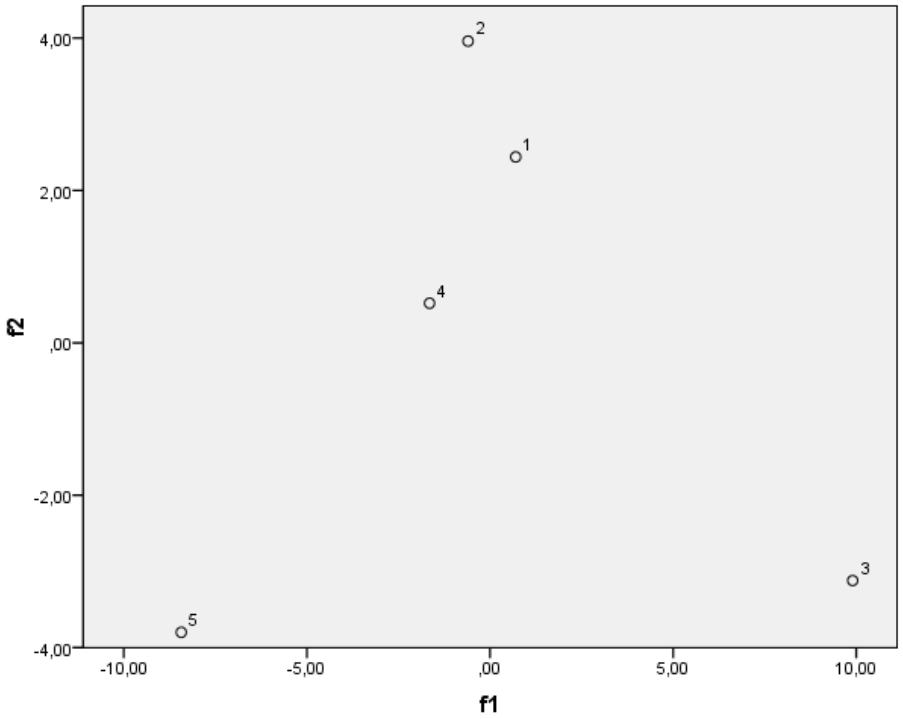
Παράδειγμα

$$\lambda_1 = 173.51, \quad \lambda_2 = 46.04$$

$$\mathbf{u}_1 = (0.31, 0.64, 0.70)', \mathbf{u}_2 = (-0.95, 0.16, 0.27)'$$

$$f_{i1} = 0.31 z_{i1} + 0.64 z_{i2} + 0.70 z_{i3},$$

$$f_{i2} = -0.95 z_{i1} + 0.16 z_{i2} - 0.27 z_{i3}.$$



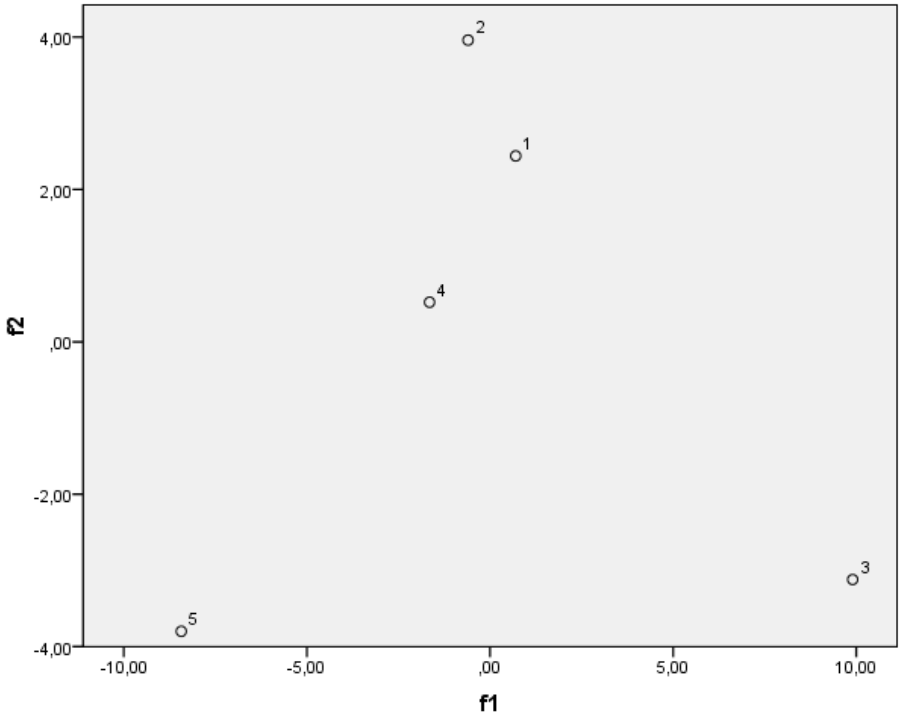
$$Dis_{\mathbf{u}_1,\mathbf{u}_2}(N) = \lambda_1 + \lambda_2 = 219.55$$

δηλαδή ποσοστό $219.55 / 222 = 98.9\%$,

i	f_{1i}	f_{2i}
1	0.78	2.44
2	-0.60	3.96
3	9.90	-3.12
4	-1.65	0.52
5	-8.43	-3.8
Άθροισμα	**0**	**0**

Παράδειγμα

i	f_{1i}	f_{2i}
1	0.78	2.44
2	-0.60	3.96
3	9.90	-3.12
4	-1.65	0.52
5	-8.43	-3.8
Άθροισμα	**0**	**0**



i	$Q(\mathbf{x}_i)$
1	0.820
2	0.944
3	0.998
4	0.997
5	0.994
Άθροισμα	-

$$r_{21} = \frac{-43.76}{\sqrt{46.08}\sqrt{58}} = -0.846$$

$$r_{22} = \frac{7.52}{\sqrt{46.08}\sqrt{74}} = 0.129$$

$$r_{23} = \frac{12.24}{\sqrt{46.08}\sqrt{90}} = 0.193$$

Επιλογή του πλήθους κυρίων συνιστωσών που θα διατηρήσουμε

Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

Σύμφωνα με αυτό το κριτήριο βάζουμε κάποιο όριο (π.χ. 75%) και επιλέγουμε τον αριθμό των συνιστωσών έτσι ώστε όλες μαζί (αθροιστικά) να εξηγούν μεγαλύτερο ποσοστό από το όριο που βάλαμε. Ως κριτήριο είναι πολύ απλό και εύκολο αλλά στην πράξη δεν δίνει πάντα καλά αποτελέσματα, ιδίως αν ο στόχος είναι αρκετά υψηλός (οπότε μπορεί να χρειαστεί να διατηρήσουμε ιδιαίτερα μεγάλο πλήθος κυρίων συνιστωσών).

Κριτήριο του Kaiser

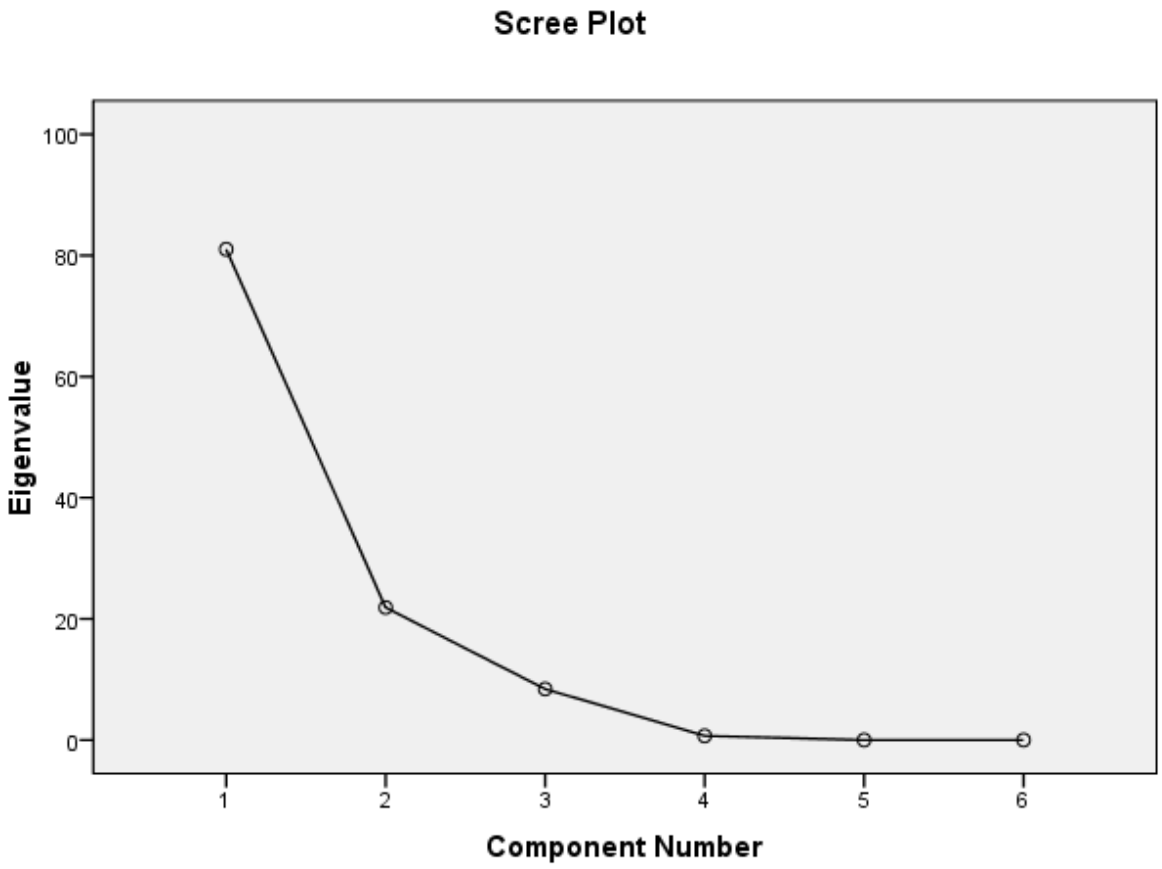
Ο Kaiser προτείνει να διατηρούμε μόνο τις ιδιοτιμές που είναι μεγαλύτερες από τη μέση τιμή των ιδιοτιμών $\lambda_1, \lambda_2, ..., \lambda_p$, δηλαδή από την ποσότητα

$$\bar{\lambda} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_p}{p} = \frac{Dis(N)}{p} = \frac{tr(Z'Z)}{p}.$$

Στην περίπτωση που εργαζόμαστε με τον πίνακα συσχετίσεων ισχύει ότι $tr(Z'Z) = p$ οπότε η μέση τιμή των $\lambda_1, \lambda_2, ..., \lambda_p$ είναι ίση με $\bar{\lambda} = 1$.

Επομένως τότε, σύμφωνα με το κριτήριο του Kaiser, διαλέγουμε τόσες συνιστώσες όσες ιδιοτιμές μεγαλύτερες της μονάδας έχουμε.

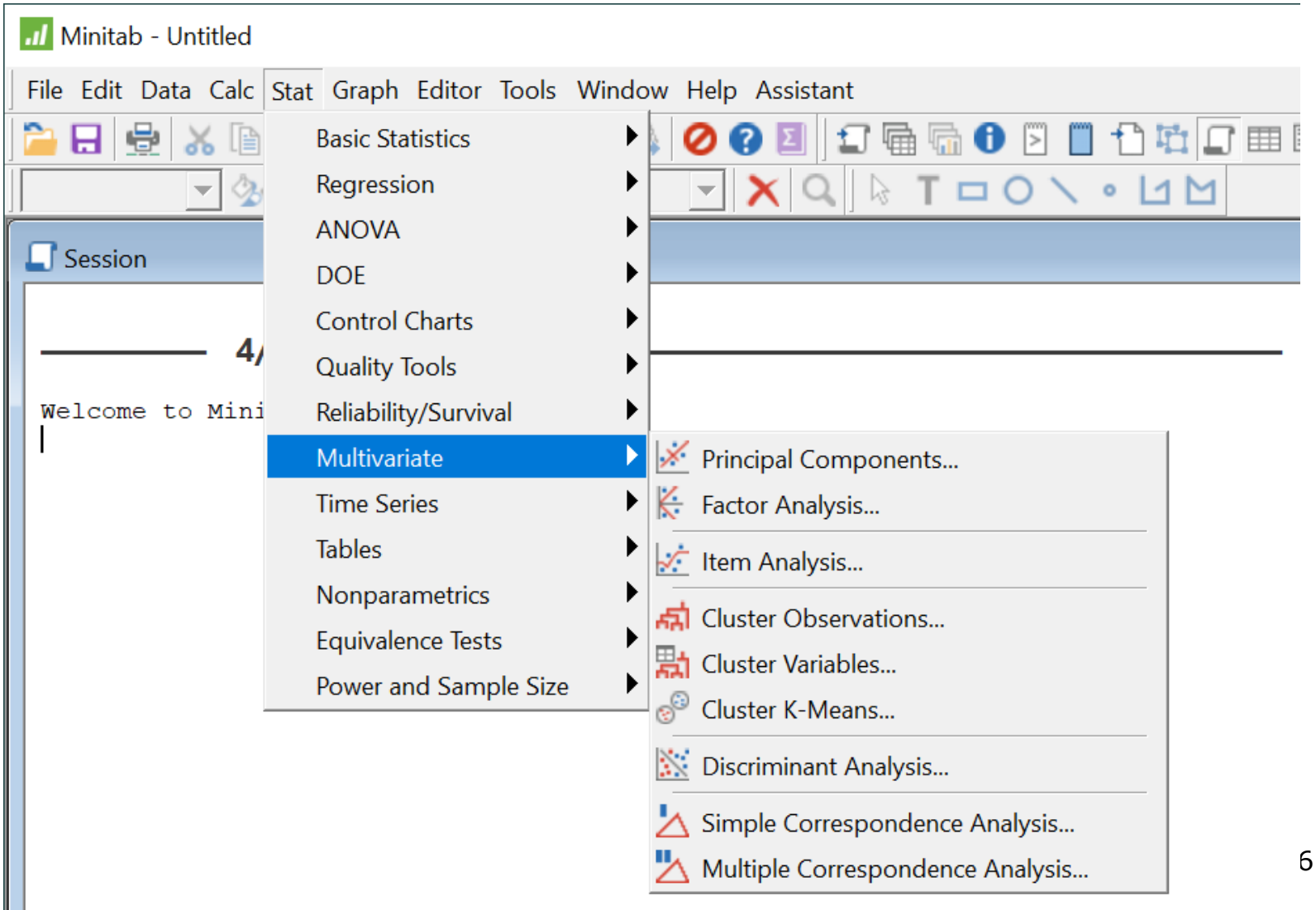
Scree plot


Scree Plot
Eigenvalue
100
80
60
40
20
0
1
2
3
4
5
6
Component Number

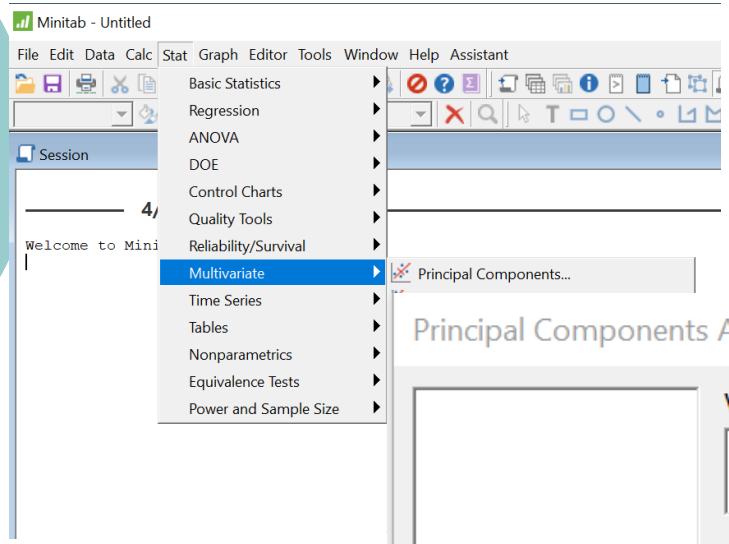
Ανάλυση κυρίων συνιστωσών

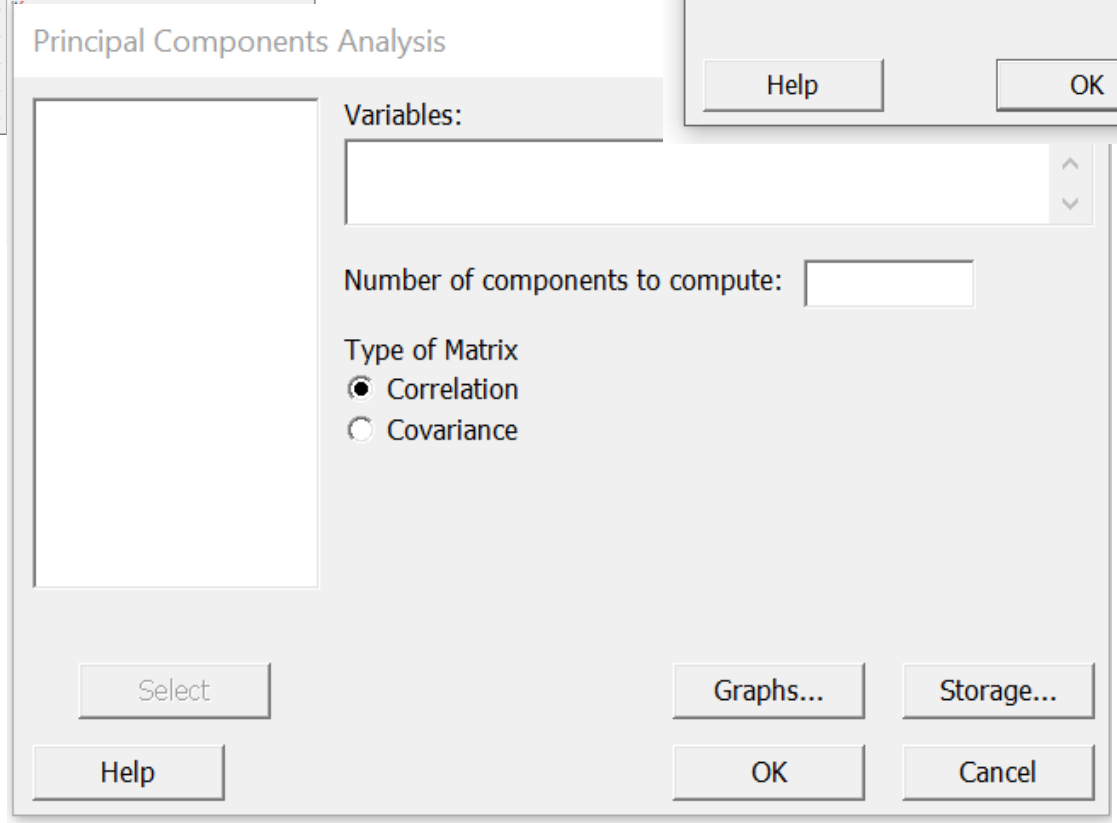
case studies

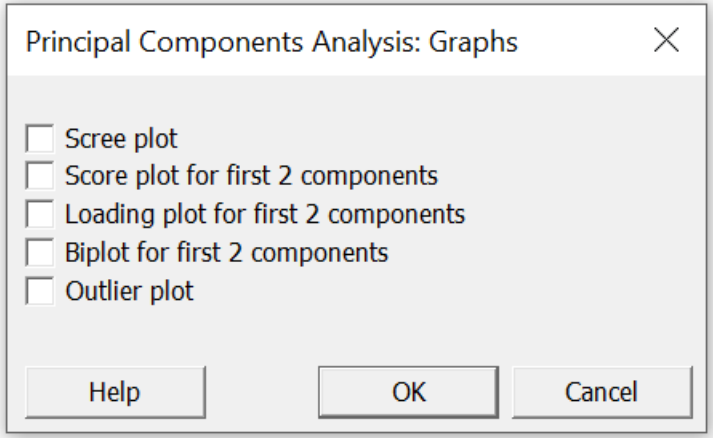
Principal Component Analysis



Principal Component Analysis





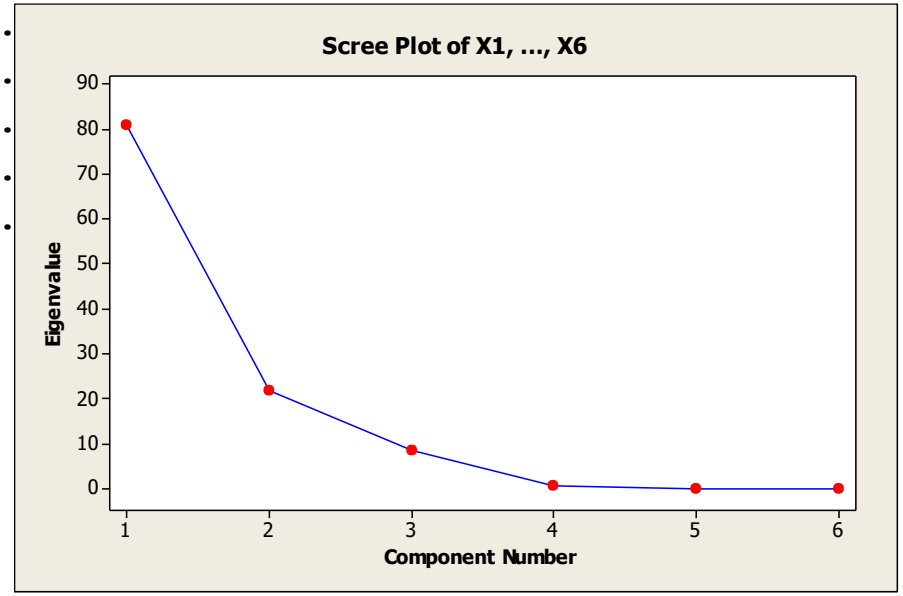


Principal Component Analysis

Eigenanalysis of the **Covariance** Matrix

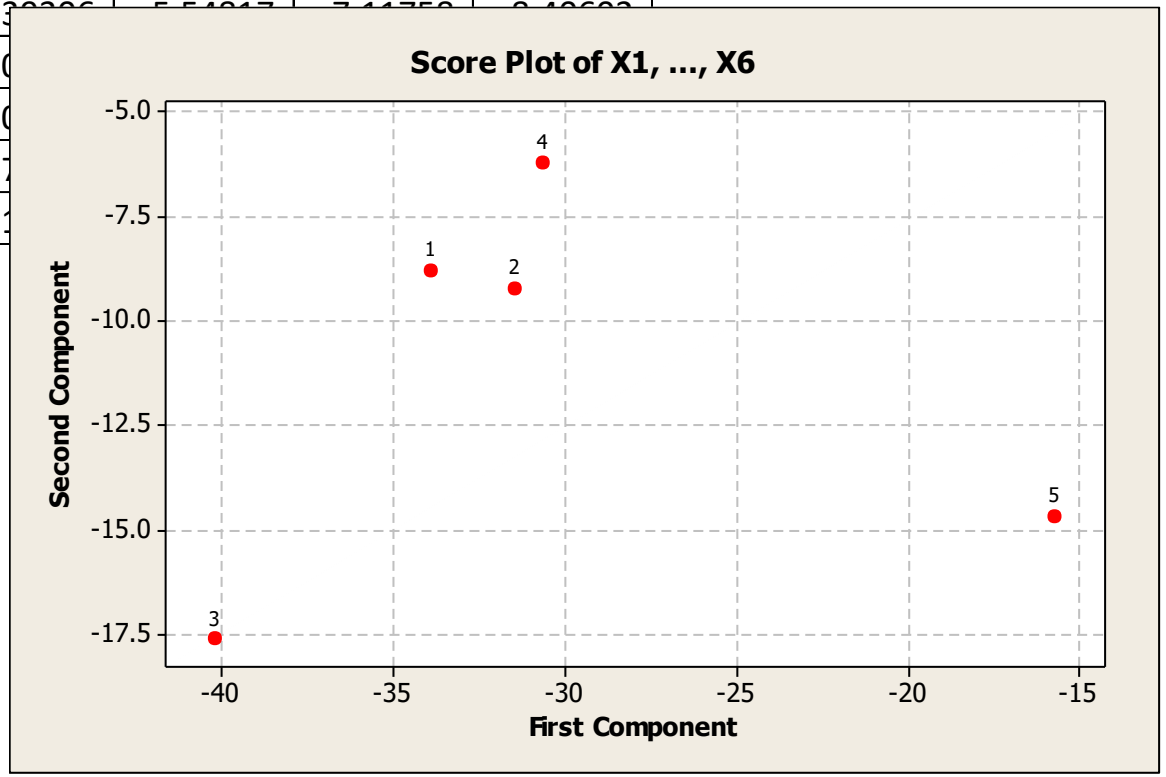
Eigenvalue	81.026	21.871	8.415	0.687	0.000	0.000
Proportion	0.723	0.195	0.075	0.006	0.000	0.000
Cumulative	0.723	0.919	0.994	1.000	1.000	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
X1	-0.100	-0.681	-0.647	0.139	-0.052	0.292
X2	-0.455	-0.235	0.			
X3	-0.519	-0.163	0.			
X4	-0.225	-0.425	0.			
X5	-0.581	0.453	-0.			
X6	-0.355	0.260	-0.			



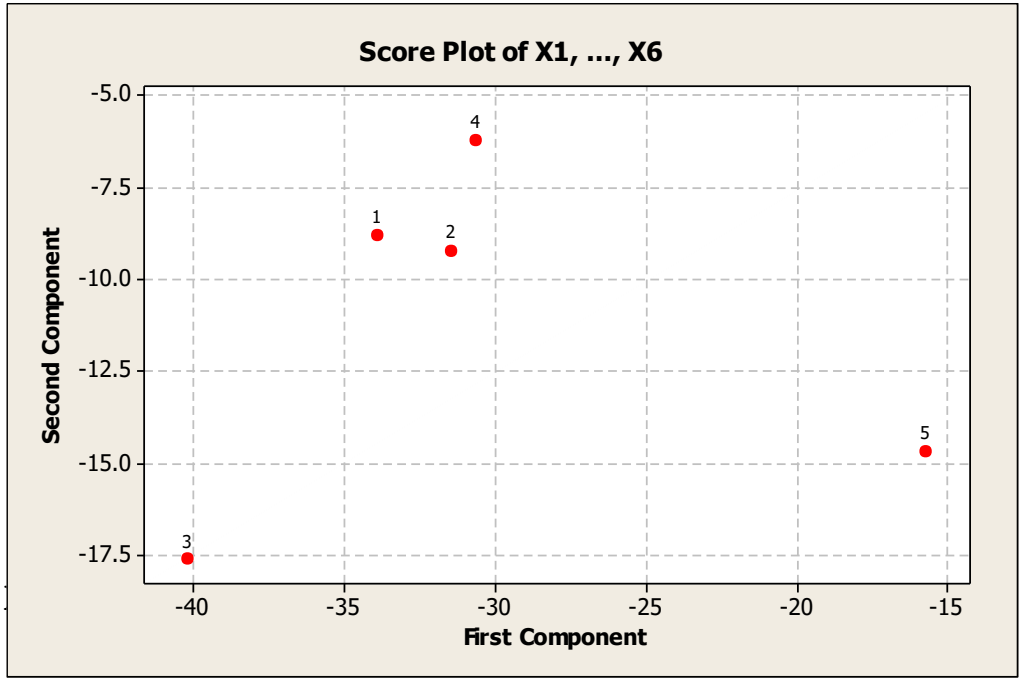
Principal Component scaterplot

y1	y2	y3	y4	y5	y6
-33.8846	-8.8078	0.3			
-31.4553	-9.2059	3.(			
-40.1666	-17.5967	-2.(			
-30.6245	-6.2123	-4.7			
-15.723	-14.6699	-1.1			



PCA: Ερμηνεία κυρίων αξόνων

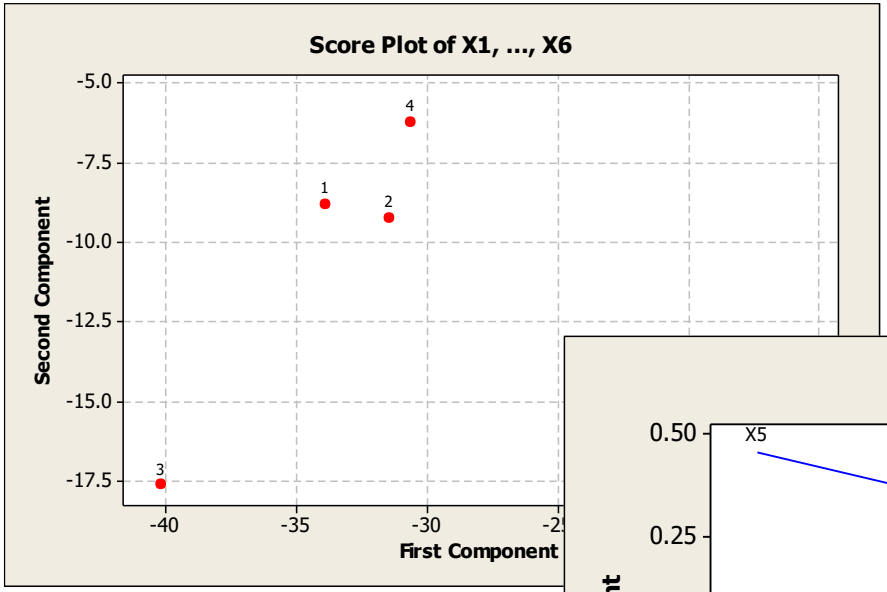
	X1	X2	X3	X4	X5	X6
1	12	13	14	17	16	18
2	10	14	12	18	14	16
3	20	19	18	20	16	18
4	13	12	11	11	16	18
5	15	7	5	14	3	10



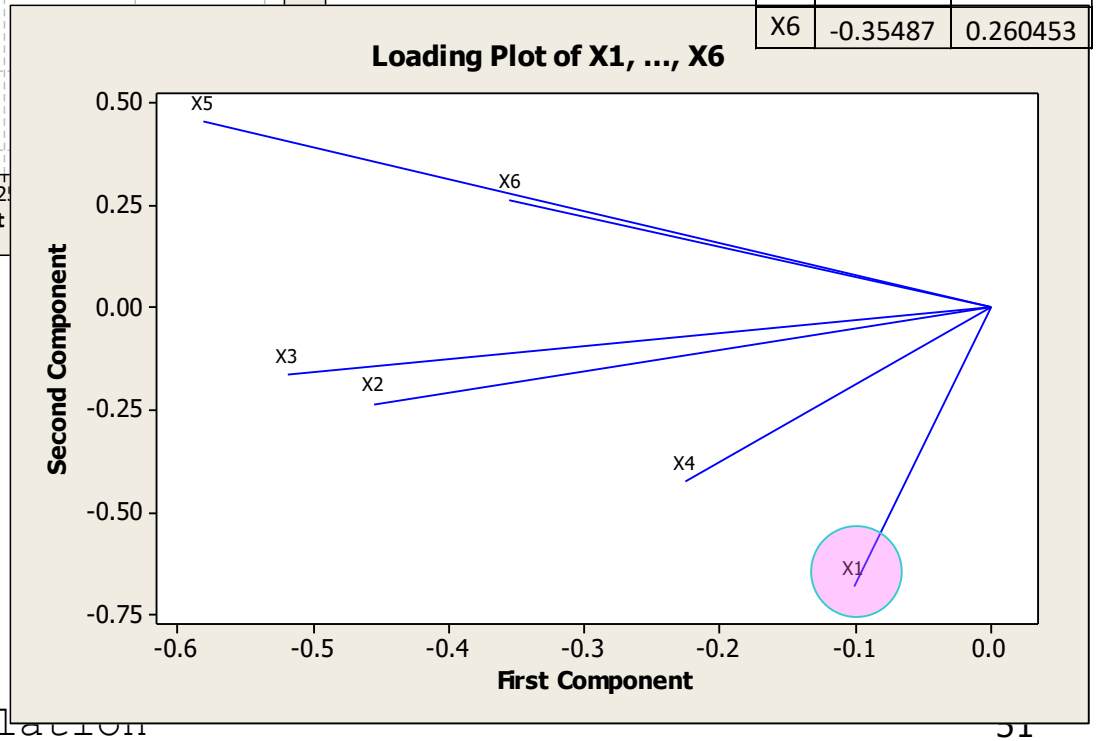
	y1	y2	X1				
y2	-0.000						
X1	-0.237	**-0.837**					
X2	-0.952	**-0.255**	0.412				
X3	-0.984	**-0.161**	0.360	0.968			
X4	-0.573	**-0.563**	0.316	0.707	0.671		
X5	-0.924	**0.375**	-0.058	0.781	0.848	0.275	
X6	-0.922	**0.352**	0.000	0.772	0.852	0.245	0.995

Cell Contents: Pearson correlation

PCA: Ερμηνεία κυρίων αξόνων



	u1	u2
X1	-0.1003	-0.68124
X2	-0.45508	-0.23495
X3	-0.51861	-0.16339
X4	-0.22491	-0.42527
X5	-0.58084	0.453136
X6	-0.35487	0.260453



	y1	y2	X1
y2	-0.000		
X1	-0.237	-0.837	
X2	-0.952	-0.255	0.412
X3	-0.984	-0.161	0.360
X4	-0.573	-0.563	0.316
X5	-0.924	0.375	-0.058
X6	-0.922	0.352	0.000

Cell Contents: Pearson correlation

Τιμές ειδών διατροφής σε 13 πολιτείες των ΗΠΑ

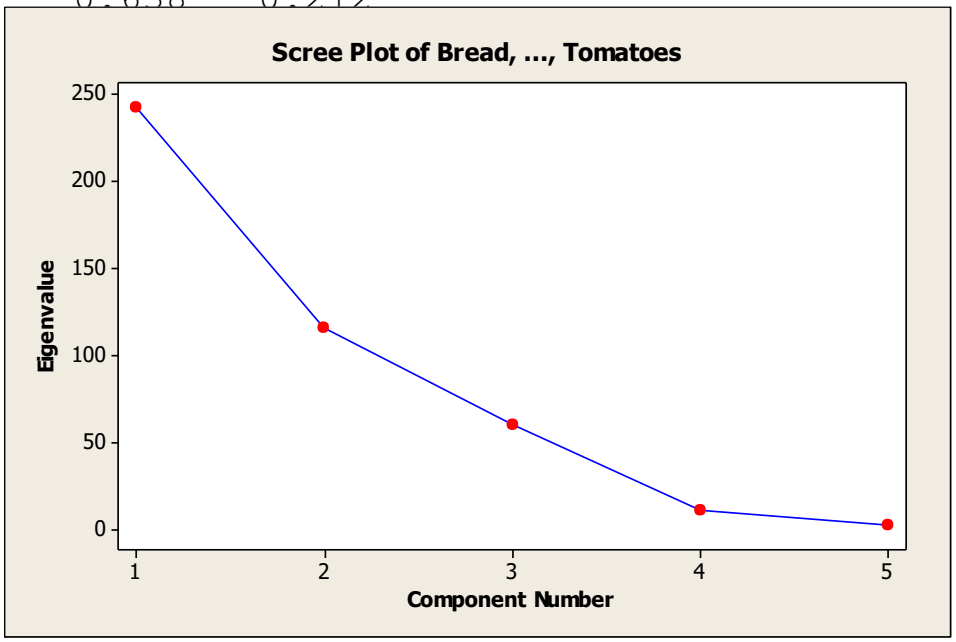
	City	Bread	Burger	Milk	Oranges	Tomatoes
1	**Atlanta**	24.5	94.5	73.9	80.1	41.6
2	**Baltimore**	26.5	91	67.5	74.6	53.3
3	**Buffalo**	22.8	86.6	65.3	118.4	51.2
4	**Chicago**	26.7	86.7	62.7	105.9	51.2
5	**Cleveland**	22.8	88.8	52.4	110.9	46.8
6	**Dallas**	23.3	85.5	62.5	117.9	41.8
7	**Detroit**	24.1	93.7	51.5	109.7	52.4
8	**Honolulu**	29.3	105.9	80.2	133.2	61.7
9	**Houston**	22.3	83.6	67.8	108.6	42.4
10	**Kansas city**	26.1	88.9	65.4	100.9	43.2
11	**Milwaukee**	20.3	89.6	53.8	111.8	53.9
12	**New York**	30.8	110.7	66	107.3	62.6
13	**Minneapolis**	24.6	92.2	51.9	106	50.7

Principal Component Analysis

Eigenanalysis of the Covariance Matrix

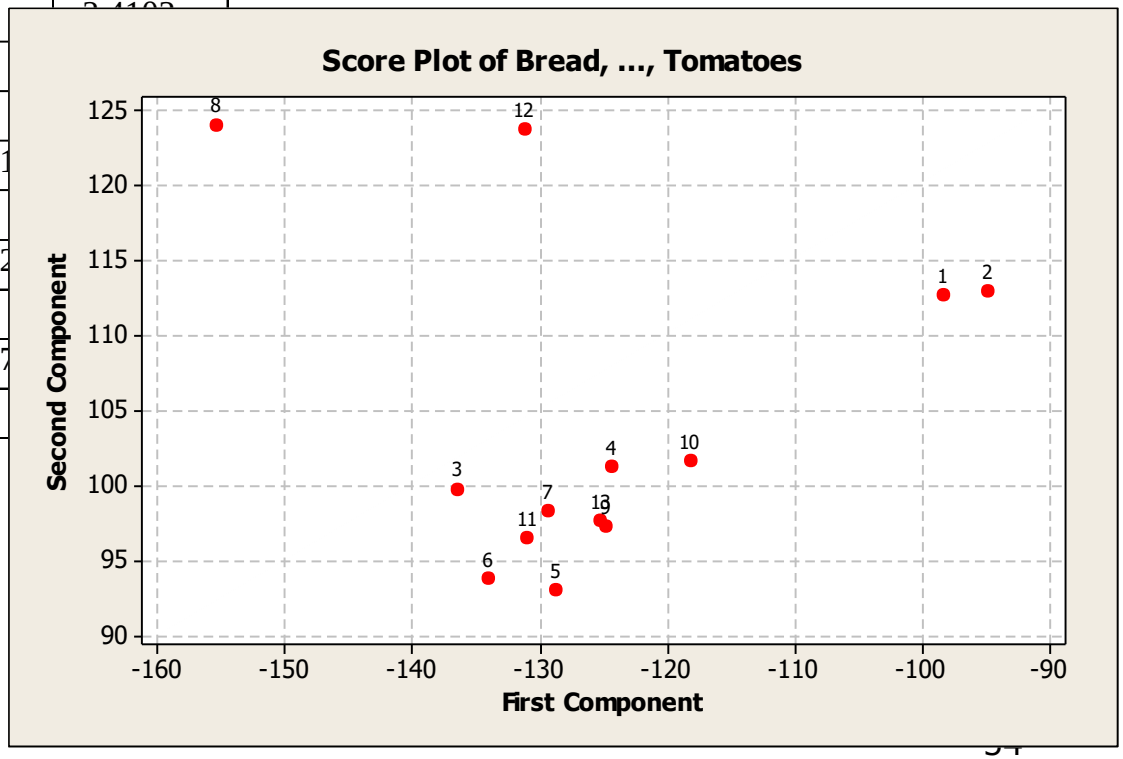
Eigenvalue	241.71	115.29	60.04	10.26	2.40
Proportion	0.562	0.268	0.140	0.024	0.006
Cumulative	0.562	0.831	0.971	0.994	1.000

Variable	PC1	PC2	PC3	PC4	PC5
Bread	-0.016	0.228	0.041	0.071	0.970
Burger	-0.134	0.630	0.363	0.638	-0.212
Milk	0.009	0.593	-0.781		
Oranges	-0.971	-0.168	-0.155		
Tomatoes	-0.195	0.413	0.482		



Principal Component scaterplot

	City	Y1	Y2	Y3
1	Atlanta	-98.323	112.689	-14.7972
2	Baltimore	-94.887	112.900	-4.4951
3	Buffalo	-136.392	99.739	-12.3303
4	Chicago	-124.351	101.253	-8.1628
5	Cleveland	-128.661	92.921	
6	Dallas	-133.957	93.701	
7	Detroit	-129.276	98.287	
8	Honolulu	-155.376	124.071	
9	Houston	-124.721	97.232	
10	Kansas city	-118.192	101.642	
11	Milwaukee	-130.976	96.467	
12	New York	-131.192	123.747	
13	Minneapolis	-125.153	97.613	

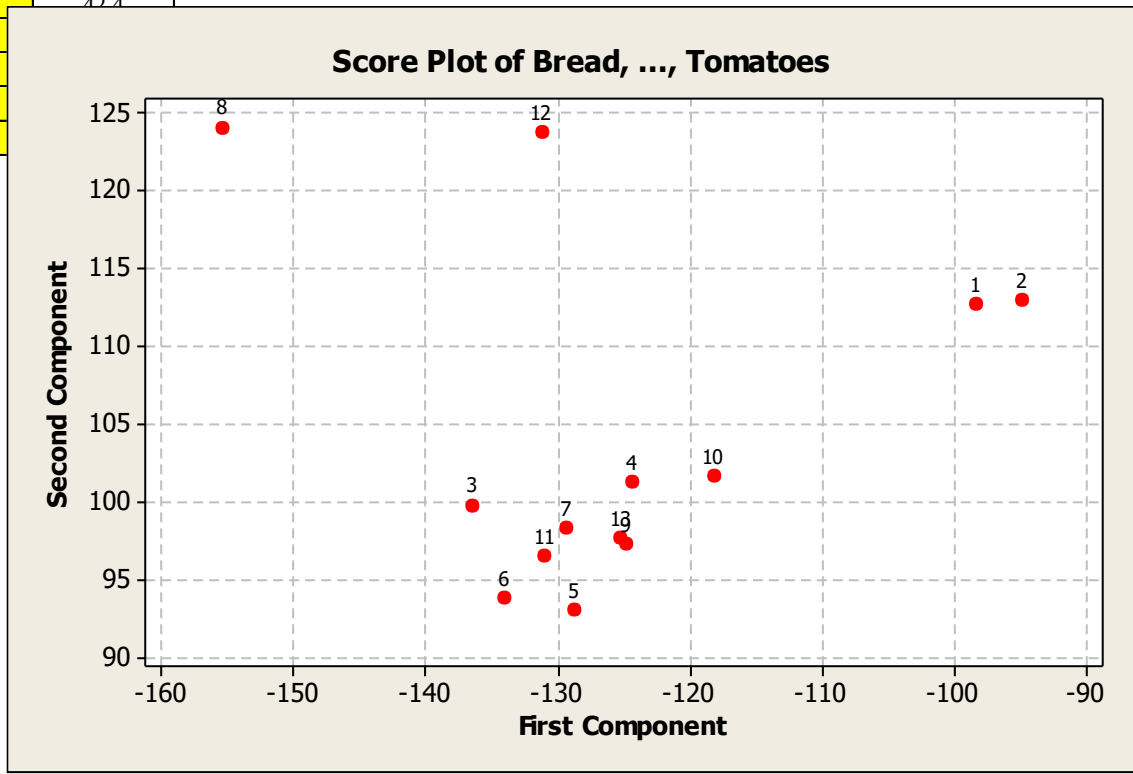


PCA: Ερμηνεία κυρίων αξόνων

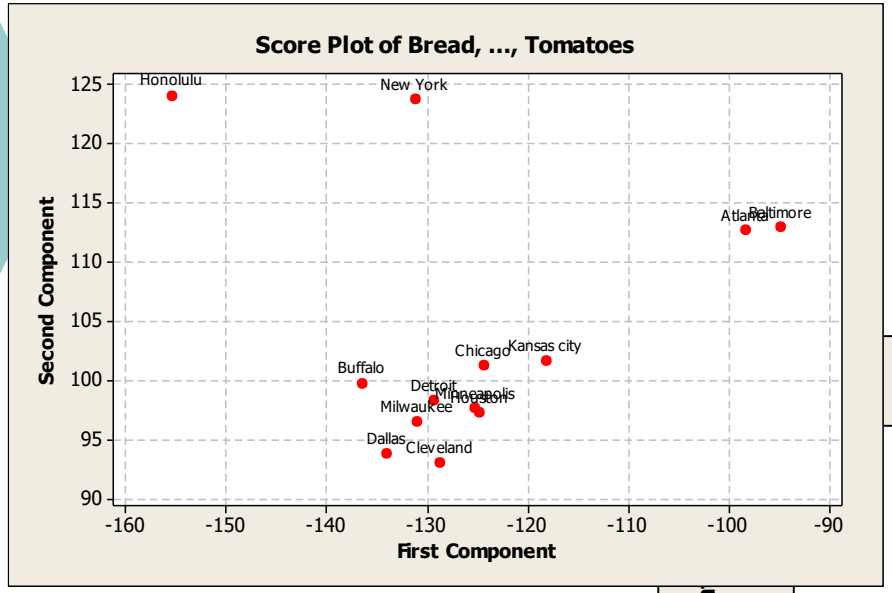
	City	Bread	Burger	Milk	Oranges	Tomatoes
1	Atlanta	24.5	94.5	73.9	80.1	41.6
2	Baltimore	26.5	91	67.5	74.6	53.3
3	Buffalo	22.8	86.6	65.3	118.4	51.2
4	Chicago	26.7	86.7	62.7	105.9	51.2
5	Cleveland	22.8	88.8	52.4	110.9	46.8
6	Dallas	23.3	85.5	62.5	117.9	41.8
7	Detroit	24.1	93.7	51.5	109.7	52.4
8	Honolulu	29.3	105.9	80.2	133.2	61.7
9	Houston	22.3	83.6	67.8 ?	108.6	42.4
10	Kansas city	26.1	88.9	65.4	100.9	
11	Milwaukee	20.3	89.6	53.8	111.8	
12	New York	30.8	110.7	66	107.3	
13	Minneapolis	24.6	92.2	51.9	106	

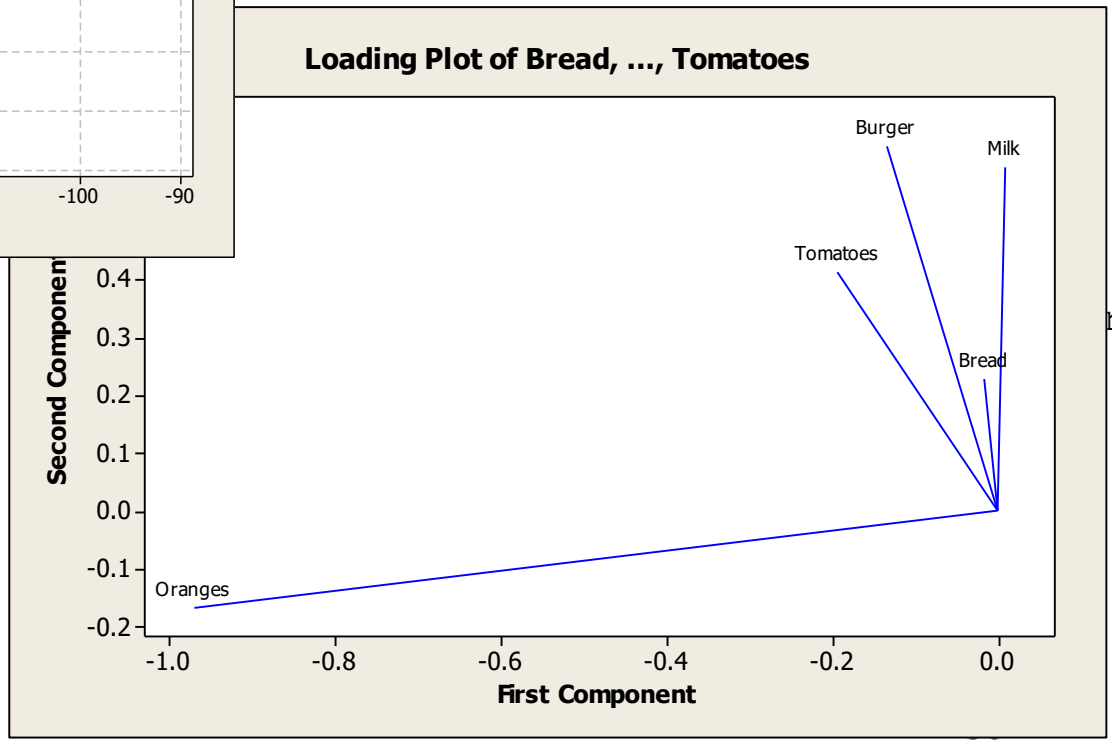
	y1	y2
y2	0.000	
y3	0.000	0.000
Bread	-0.087	0.841
Burger	-0.264	0.857
Milk	0.016	0.723
Oranges	-0.990	-0.118
Tomatoes	-0.436	0.637

Pearson correlation



PCA: Ερμηνεία κυρίων αξόνων





	y1	y2
y2	0.000	
y3	0.000	0.000
Bread	-0.087	0.841
Burger	-0.264	0.857
Milk	0.016	0.723
Oranges	-0.990	-0.118
Tomatoes	-0.436	0.637

Pearson correlation

PCA: Ποιότητες αναπαράστασης

	City	$Q_{12}(\mathbf{x}_i)$
1	**Atlanta**	0.908109
2	**Baltimore**	0.963989
3	**Buffalo**	0.757577
4	**Chicago**	0.262047
5	**Cleveland**	0.824277
6	**Dallas**	0.722584
7	**Detroit**	0.307365
8	**Honolulu**	0.965541
9	**Houston**	**0.280215**
10	**Kansas city**	0.601244
11	**Milwaukee**	0.557384
12	**New York**	0.778791
13	**Minneapolis**	0.323813