Ανάλυση Κύριων Συνιστωσών & Παραγόντων

Περιεχόμενα:

1. Ανάλυση Κυριων Συνιστωσων (Principal Component Analysis (PCA))

2. Ανάλυση Παραγοντων (Factor Analysis)

Ανάλυση Κύριων Συνιστωσών (ΑΚΣ)

- Είναι τεχνική για το μετασχηματισμό των αρχικών μεταβλητών σε νέες, οι οποίες είναι ασυσχέτιστες και λαμβάνουν υπόψη φθίνουσα αναλογία διακύμανσεων στα δεδομένα.

- Οι νέες μεταβλητές αποτελούν γραμμικό συσχετισμό των αρχικών μεταβλητών.

- Ο μετασχηματισμός είναι στροφή/αντικατοπτρισμός (rotation/reflection) των αρχικών σημείων. Καμιά σημαντική στατιστική πληροφορία δεν χάνεται (ή δημιουργείται).

- Κατά τη διαδικασία αξιολογείται η σημαντικότητα των νέων συνιστωσών καθώς επίσης και πόσες συνιστώσες χρειάζονται. Αυτό επιτυγχάνεται με τη βοήθεια της Τεχνικής του Αγκώνα΄ (–screen plots etc).

- Η ανάλυση αξιολογεί τη σημαντικότητα των αρχικών μεταβλητών (εξέταση συντελεστών (loadings)).

- *Σκοπός της ΑΚΣ είναι ο μετασχηματισμός ενός συνόλου συσχετισμένων μεταβλητών σε ένα σύνολο ασυσχέτιστων μεταβλητών που αποτελούν γραμμικό συνδιασμό των αρχικών. Το σύνολο των γραμμικών συνδιασμών επιλέγονται με τρόπο που να αντικατοπτρίζουν φθίνουσα αναλογία διακύμανσεων των αρχικών μεταβλητών.*

Στόχος

- Για n μεταβλητές $x_1, \ldots, x_n$ και $X = (x_1, \ldots, x_n)$, προσπαθούμε να βρούμε ένα γραμμικό μετασχηματισμό του $X \to Y = (y_1, \ldots, y_n)$ έτσι ώστε:

  Η 1η συνιστώσα y_1 να είναι *η πιο ενδιαφέρουσα.*
  Η 2η συνιστώσα y_2 να είναι *η 2η πιο ενδιαφέρουσα.*
  Η 3η συνιστώσα y_3 να είναι *η 3η πιο ενδιαφέρουσα*, κτλ.

- Επιλέγουμε ένα σύστημα συντεταγμένων έτσι ώστε στα δεδομένα του νέου συστήματος, Y, η 1η συνιστώσα να περιέχει *τις περισσότερες πληροφορίες*, η 2η συνιστώσα να περιέχει *τις αμέσως επόμενες περισσότερες πληροφορίες*, κτλ.

- Σκοπός είναι οι πρώτες *λίγες* (2, 3 ή 4) συνιστώσες να εξηγούν *σχεδόν όλες* τις πληροφορίες των δεδομένων και οι υπολοιπόμενες 2,3,4 συνιστώσες να περιλαμβάνουν σχετικά λίγες πληροφορίες σε σημείο που να μπορεί να *αγνωηθεί* , δηλαδή, η στατιστική ανάλυση μπορεί να περιοριστεί μόνο στις πρώτες λίγες συνιστώσες. Έτσι η ανάλυση είναι ευκολότερη.

- Ο γραμμικός μετασχηματισμός $X \to Y$ δίνεται από $Y = XQ$, όπου Q είναι ένας $n \times n$ μη-ιδιόμορφος πίνακας. Αν Q είναι ένας ορθογώνιος πίνακας, δηλαδή $Q^T Q = I_n$, τότε ο μετασχηματισμός $X \to Y$ είναι ένας ορθογώνιος μετασχηματισμός.

Συγκεκριμένα

- Η βασική ιδέα είναι να βρούμε ένα σύνολο ορθογώνιων συντεταγμένων έτσι ώστε η δειγματική τους διακύμανση να βρίσκεται σε φθίνουσα σειρά σημαντικότητας, δηλαδή, η προβολή των σημείων στην 1η κύρια συνιστώσα να έχει τη μέγιστη διακύμανση μεταξύ όλων των γραμμικών προβολών. Η προβολή στη 2η συνιστώσα να έχει τη μέγιστη διακύμανση υποκείμενη στην ορθογωνικότητα με την πρώτη συνιστώσα. Η προβολή στην 3η συνιστώσα να έχει την μέγιστη διακύμανση υποκείμενη στην ορθογωνικότητα με τις άλλες δύο συνιστώσες κτλ.

- μέγιστη σημαντικότητα – μέγιστη πληροφορία – *μέγιστη διακύμανση* .

- Αυτός ο στόχος μπορεί να επιτευχθεί μέσο της ιδιοανάλυσης του πίνακα διακυμάνσεων S των δεδομένων, δηλαδή, του πίνακα συνδιακυμάνσεων μεταξύ των μεταβλητών, όπου στα διαγώνια στοιχεία υπάρχει η διακύμανση. Αν μετασχηματίσουμε τα αρχικά δεδομένα σε κύριες συνιστώσες τότε

  (α) η δειγματική διακύμανση διαδοχικών συνιστωσών ισούνται με τις ιδιοτιμές του πίνακα διακυμάνσεων S;

  (β) η συνολική μεταβολή (διασπορά) είναι η ίδια τόσο για τις αρχικές μεταβλητές, όσο και για το σύνολο των κύριων συνιστωσών. Έτσι καμιά πληροφορία δε χάνεται. Απλά δίνεται μια ανακατάταξη στη διάταξη.

- Θέτουμε S τον πίνακα $n \times n$ διακυμάνσεων-συνδιακυμάνσεων των n μεταβλητών $x_1, \ldots, x_n$. Π.χ. θέτουμε

$$X = (x_1 \ x_2 \ x_3) \quad \text{και} \quad \mathrm{Var}(X) = S = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

- Καθώς S είναι ένας συμμετρικός θετικά ορισμένος πίνακας, έχει πραγματικές ιδιοτιμές

$$\boxed{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0.}$$

- Η φασματική παραγοντοποίηση του S δίνεται από:

$$Q^T S Q = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}.$$

Π.χ.

$$\begin{pmatrix} -0.38 & 0.92 & 0 \\ 0 & 0 & 1 \\ 0.92 & 0.38 & 0 \end{pmatrix} \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} -0.38 & 0 & 0.92 \\ 0.92 & 0 & 0.38 \\ 0 & 1 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} 5.83 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0.17 \end{pmatrix}.$$

- Θέτουμε y_i μια μετασχηματισμένη μεταβλητή, όπου $Y = (y_1, \ldots, y_n)$ και $Y = Q^T X$.

Έτσι,

$$(y_1 \;\; y_2 \;\; y_3) = (x_1 \;\; x_2 \;\; x_3) \begin{pmatrix} -0.38 & 0 & 0.92 \\ 0.92 & 0 & 0.38 \\ 0 & 1 & 0 \end{pmatrix}$$

ή

$$\begin{aligned} y_1 &= -0.38x_1 + 0.92x_2 \\ y_2 &= x_3 \\ y_3 &= 0.92x_1 + 0.38x_2 \end{aligned}$$

- Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων του y δίνεται από:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}(XQ)\\ &= Q^T\mathrm{Var}(X)Q \qquad (\text{όπου } \mathrm{Var}(X)=S)\\ &= Q^TSQ\\ &= \Lambda = \begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{pmatrix}.\end{aligned}$$

Επαλήθευση

Σημειώστε ότι
$\mathrm{Var}(aX+bY) =$
$a^2\mathrm{Var}(X)+b^2\mathrm{Var}(Y)+2ab\,\mathrm{Cov}(X,Y).$

$$\begin{aligned}\mathrm{Var}(y_1) &= \mathrm{Var}(-0.38x_1+0.92x_2)\\ &= (-0.38)^2\mathrm{Var}(x_1)+(0.92)^2\mathrm{Var}(x_2)\\ &\quad +2(-0.38)(0.92)\mathrm{Cov}(x_1,x_2)\\ &= 0.15(1)+0.85(5)-0.7(-2)\\ &= 5.8=\lambda_1.\end{aligned}$$

και

$$\begin{aligned}\mathrm{Cov}(y_1, y_2) &= \mathrm{Cov}(-0.38x_1 + 0.92x_2,\ x_3) \\ &= -0.38\mathrm{Cov}(x_1, x_3) + 0.92\mathrm{Cov}(x_2, x_3) \\ &= -0.38(0) + 0.92(0) \\ &= 0.\end{aligned}$$

- $y_1, \ldots, y_n$ είναι οι κύριες συνιστώσες των $x_1, \ldots, x_n$.
- y_i έχει διακύμανση λ_i.
- y_i είναι ασυσχέτιστη με τη y_j $(i \neq j)$, έτσι Λ, δηλαδή, η συνδιακύμανση του Y, είναι διαγώνια.
- Από $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ συμπεραίνουμε ότι y_1 έχει τη μεγαλύτερη διακύμανση λ_1, y_2 έχει τη δεύτερη μεγαλύτερη διακύμανση λ_2, και ούτω καθεξής.

$$\lambda_1 = 5.83, \quad \lambda_2 = 2 \quad \lambda_3 = 0.17.$$

- Η συνολική κύμανση στα αρχικά δεδομένα δίνεται από το άθροισμα των διακυμάνσεων των αρχικών μεταβλητών $x_1, \ldots, x_n$. Δηληδή,

$$S_{11} + S_{22} + \ldots S_{nn} = \mathrm{trace}(S).$$

- Σημειώστε ότι:

$$\begin{aligned}\text{trace}(S) &= \text{trace}(Q^T \Lambda Q) \\ &= \text{trace}(\Lambda Q Q^T) \quad (\text{since trace}(AB) = \text{trace}(BA)) \\ &= \text{trace}(\Lambda) \qquad (\text{since } QQ^T = I_n.) \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_n.\end{aligned}$$

  που είναι το άθροισμα των διακυμάνσεων των n ΚΣ $y_1, \ldots, y_n$.

$$S_{11} + S_{22} + \ldots S_{nn} = 8 = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

- Έτσι, η συνολική κύμανση είναι ακριβώς η ίδια τόσο για το σύνολο των κύριων συνιστωσών, όσο και για τις αρχικές μεταβλητές. Άρα καμία πληροφορία δε χάνεται – απλά ανακατατάσσεται.

- Το άθροισμα των διακυμάνσεων $\sum_{i=1}^{n} \lambda_i$ των n ΚΣ $y_1, \ldots, y_n$ ισούνται με το άθροισμα των διακυμάνσεων $\sum_{i=1}^{n} S_{ii}$ των αρχικών μεταβλητών $x_1, \ldots, x_n$.

- Οι συνιστώσες με μικρότερη διακύμανση μπορούν να αγνοηθούν χωρίς να επηρεάζεται σημαντικά η συνολική διακύμανση. Έτσι μπορεί να μειωθεί ο αριθμός των n μεταβλητών.

- Η συνολική κύμανση των ΚΣ δίνεται από $\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} S_{ii} = \text{trace}(S)$. Αυτό ερμηνεύεται

$$\lambda_1 / \sum_{i=1}^{n} \lambda_i$$

ως το ποσοστό της συνολικής κύμανσης που εξηγείται από την πρώτη κύρια συνιστώσα. Στο παράδειγμα $\lambda_1 / \sum \lambda_i = 5.83/8 = 0.73$.

- Το ποσοστό της συνολικής διακύμανσης που εξηγείται από τις πρώτες δύο ΚΣ δίνεται από $(\lambda_1 + \lambda_2) / \sum_{i=1}^{n} \lambda_i$.

$$(\lambda_1 + \lambda_2) / \sum \lambda_i = (5.83 + 2)/8 = 0.98.$$

- Γενικά το ποσοστό της συνολικής κύμανσης που εξηγείται από την πρώτη k ΚΣ δίνεται από

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}.$$

- Αν οι πρώτες ορισμένες ΚΣ εξηγούν την περισσότερη΄ κύμανση στα δεδομένα, τότε οι

τελευταίες ΚΣ είναι αμελητέες και λίγες πληροφορίες χάνονται αν αγνοηθούν. Π.χ. αν

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i} = \text{ας πούμε } 80+\ \%,$$

τότε η $(k+1)\text{h}, \ldots, n\text{h}$ συνιστώσα περιέχει σχετικά λίγες πληροφορίες και η διαστατικότητα των δεδομένων μπορεί να μειωθεί από n σε k ενώ μόνο λίγες πληροφορίες χάνονται. Είναι χρήσιμο αν $k = 1, 2, 3, 4, 5$??? Το 80% που δώθηκε πιο πάνω ορίζεται αυθαίρετα αφού εξαρτάται από τον τύπο των δεδομένων που αναλύονται – συγκεκριμένα από την εφαρμογή τους. Για κάποιες εφαρμογές αν εξηγείται το 40% της κύμανσης σε λίγες ΚΣ, είναι ικανοποιητικό. Για άλλες εφαρμογές ίσως να χρειάζεται να εξηγηθεί το 90%.

Τεχνική του Αγκώνα (Scree-plot)

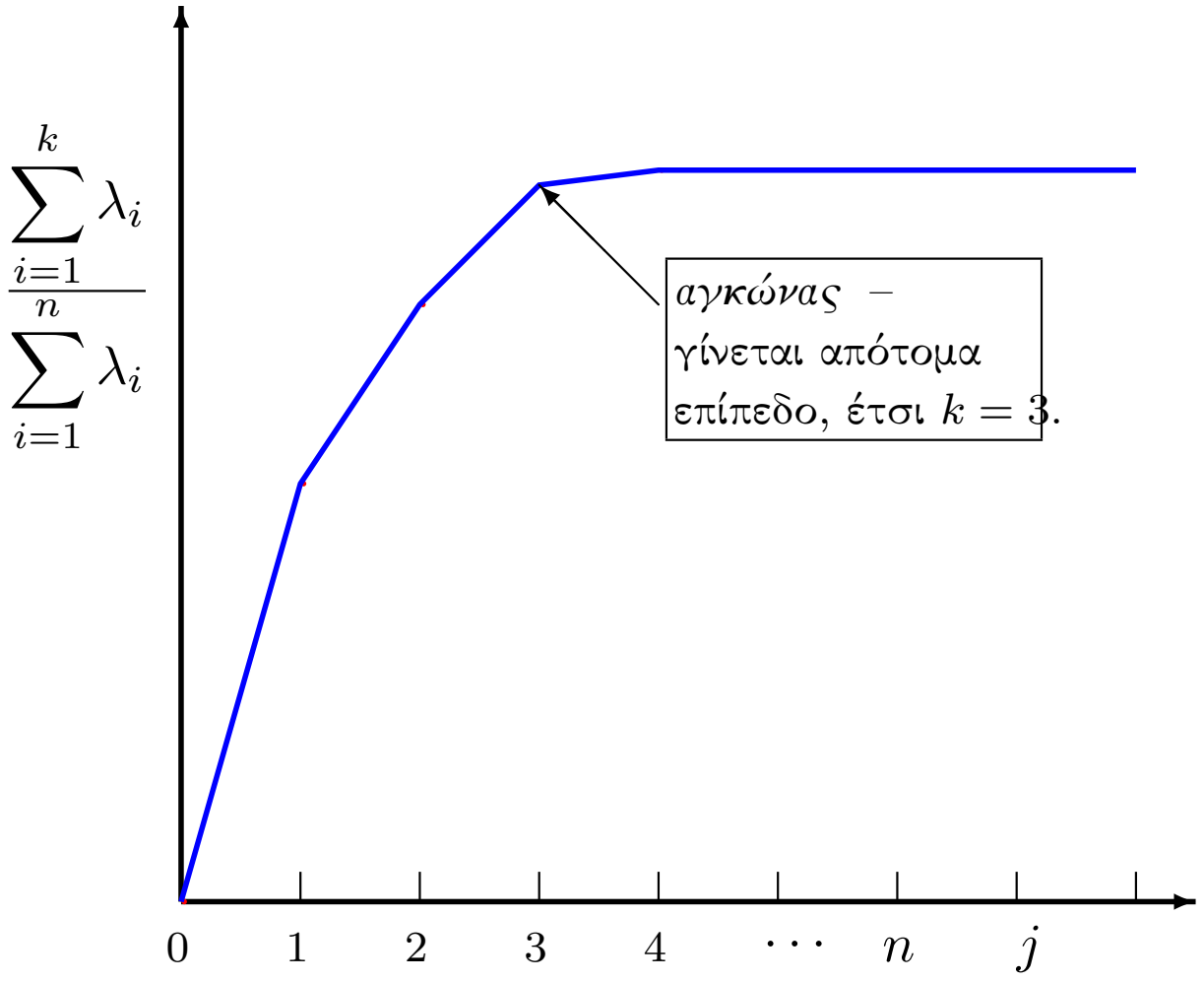
- Για να επιλεγεί το επιθυμητό ποσοστό πρέπει να ισοζυγιστεί η ωφέλεια μεταξύ ενός μικρού αριθμού k μεταβλητών και ενός μεγάλου ποσοστού συσσωρευτικής σχετικής διακύμανσης.

- Αν n είναι μεγάλο, τότε ένας πρακτικός τρόπος να επιλεγεί το k είναι μέσω της Τεχνικής του Αγκώνα (*scree-plot*).

  Αναπαριστούμε $\dfrac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$ έναντι j.

- Το γράφημα είναι μονότονο και κυρτό (convex). Ένα τέτοιο γράφημα αυξάνεται απότομα για τις πρώτες μερικές τιμές του j (δηλαδή τις πρώτες ΚΣς) και μετά σταθεροποιείται. Σε αυτό το σημείο περισσότερες κύριες συνιστώσες αυξάνουν με φθίνοντα ρυθμό το ποσοστό της διακύμανσης που εξηγείται.

- Υπάρχουν βέβαια και ορισμένοι στατιστικοί έλεγχοι $H : \lambda_i = 0$ (που όμως δεν χρησιμοποιούνται ευραίως), και ορισμοί για τα διαστήματα εμπιστοσύνης του λ_i.


$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$
αγκώνας – γίνεται απότομα επίπεδο, έτσι $k = 3$.
0 1 2 3 4 ⋯ n j

Αναλυτική Προσέγγιση

Δίνονται n μεταβλητές. Σκοπός των κύριων συνιστωσών είναι η δημιουργία n γραμμικών συνδυασμών:

$$\begin{aligned} y_1 &= w_{11}x_1 + w_{12}x_2 + \cdots + w_{1n}x_n \\ y_2 &= w_{21}x_1 + w_{22}x_2 + \cdots + w_{2n}x_n \\ &\vdots \\ y_n &= w_{n1}x_1 + w_{n2}x_2 + \cdots + w_{nn}x_n. \end{aligned}$$

Εδώ

- $y_1, y_2, \ldots y_n$ είναι οι n κύριες συνιστώσες.
- w_{ij} είναι η βαρύτητα που δίνεται στην jth μεταβλητή για την ith κύρια συνιστώσα.
- $\mathrm{Var}(y_1) > \mathrm{Var}(y_2) > \cdots > \mathrm{Var}(y_n)$.
- $\sum_{k=1}^{n} w_{ik}^2 = w_{i1}^2 + \cdots + w_{in}^2 = 1$.
- $\sum_{k=1}^{n} w_{ik}w_{jk} = w_{i1}w_{j1} + \cdots + w_{in}w_{jn} = 0$.

Παράδειγμα :

$y_1 = 0.728x_1 + 0.685x_2$ και $y_2 = -0.685x_1 + 0.728x_2$.

$0.728^2 + 0.685^2 = 1$ και $0.728 \times (-0.685) + 0.685 \times 0.728 = 0$.

Αποτελέσματα και Συντελεστές ΚΣ

- Τα αποτελέσματα των κύριων συνιστωσών είναι οι τιμές (βαθμοί) των μεταβλητών των κύριων συνιστωσών.
- Οι *Συντελεστές* είναι η συσχέτιση μεταξύ των αρχικών και των νέων (κύριων συνιστωσών) μεταβλητών. Δείχνουν το βαθμό που οι αρχικές μεταβλητές επιρεάζουν ή είναι σημαντικές στο σχηματισμό των κύριων συνιστωσών. Έτσι, όσο μεγαλύτερος είναι ο συντελεστής, τόσο πιο σημαντική είναι η μεταβλητή στο σχηματισμό των αποτελεσμάτων των κύριων συνιστωσών και το αντίθετο.
- Οι συντελεστές πηγάζουν από τη σχέση:

$$l_{ij} = \frac{w_{i,j}}{\hat{\sigma}_j}\sqrt{\lambda_i}\ ,$$

όπου

1. l_{ij} είναι ο συντελεστής της j μεταβλητής στην i κύρια συνιστώσα.
2. $w_{i,j}$ είναι το η βαρύτητα της j μεταβλητής στην i κύρια συνιστώσα.
3. $\hat{\sigma}_j$ είναι η τυπική απόκλιση της j μεταβλητής.
4. λ_i είναι η ιδιοτιμή (δηλαδή, η διακύμανση) της i κύριας συνιστώσας.

Παράδειγμα: Χρηματοοικονομικοί δείκτες X_1 και X_2

Ο πίνακας παρουσιάζει ένα μικρό δείγμα 12 παρατηρήσεων και 2 μεταβλητών X_1 και X_2 (χρηματοοικονομικοί δείκτες). Ο πίνακας μας δίνει επίσης τα διορθωμένα δεδομένα με το μέσο (που συμβολίζονται X_1^* και X_2^*), το **SSCP**, τον πίνακα συνδιακυμάνσεων **S** και τον πίνακα συσχετίσεων **R**.

Οι διορθωμένες με το μέσο μεταβλητές έχουν μετασχηματιστεί χρησιμοποιώντας τον ορθογώνιο πίνακα. Ο μετασχηματισμός (περιστροφή) έχει τη μορφή:

$$\begin{pmatrix} P_1 & P_2 \end{pmatrix} = \begin{pmatrix} X_1^* & X_2^* \end{pmatrix} \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Π.χ. αν $\theta = 43.261$, τότε $\cos(\theta) = 0.728$ και $\sin(\theta) = 0.685$.

Για παράδειγμα,

$$\begin{pmatrix} 9.25 & -1.84 \end{pmatrix} = \begin{pmatrix} 8 & 5 \end{pmatrix} \begin{pmatrix} 0.728 & -0.685 \\ 0.685 & 0.728 \end{pmatrix}.$$

	Αρχικά		Διορθωμένα Μέσου		Περιστροφή 43.26^o	
	X_1	X_2	X_1^*	X_2^*	P_1	P_2
	16	8	8	5	9.25	-1.84
	12	10	4	7	7.71	2.36
	13	6	5	3	5.70	-1.24
	11	2	3	-1	1.50	-2.78
	10	8	2	5	4.88	2.27
	9	-1	1	-4	-2.01	-3.60
	8	4	0	1	0.69	0.73
	7	6	-1	3	1.33	2.87
	5	-3	-3	-6	-6.29	-2.31
	3	-1	-5	-4	-6.38	0.51
	2	-3	-6	-6	-8.48	-0.26
	0	0	-8	-3	-7.88	3.30
Μέσος:	8	3	0	0	0	0
Διακύ.:	23.09	21.09	23.09	21.09	38.58	5.60

Αρχικές Μεταβλητές:

$$\mathbf{SSCP} = \begin{pmatrix} 254 & 181 \\ 181 & 232 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 23.09 & 16.46 \\ 16.46 & 21.09 \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{pmatrix}$$

Νέες Μεταβλητές:

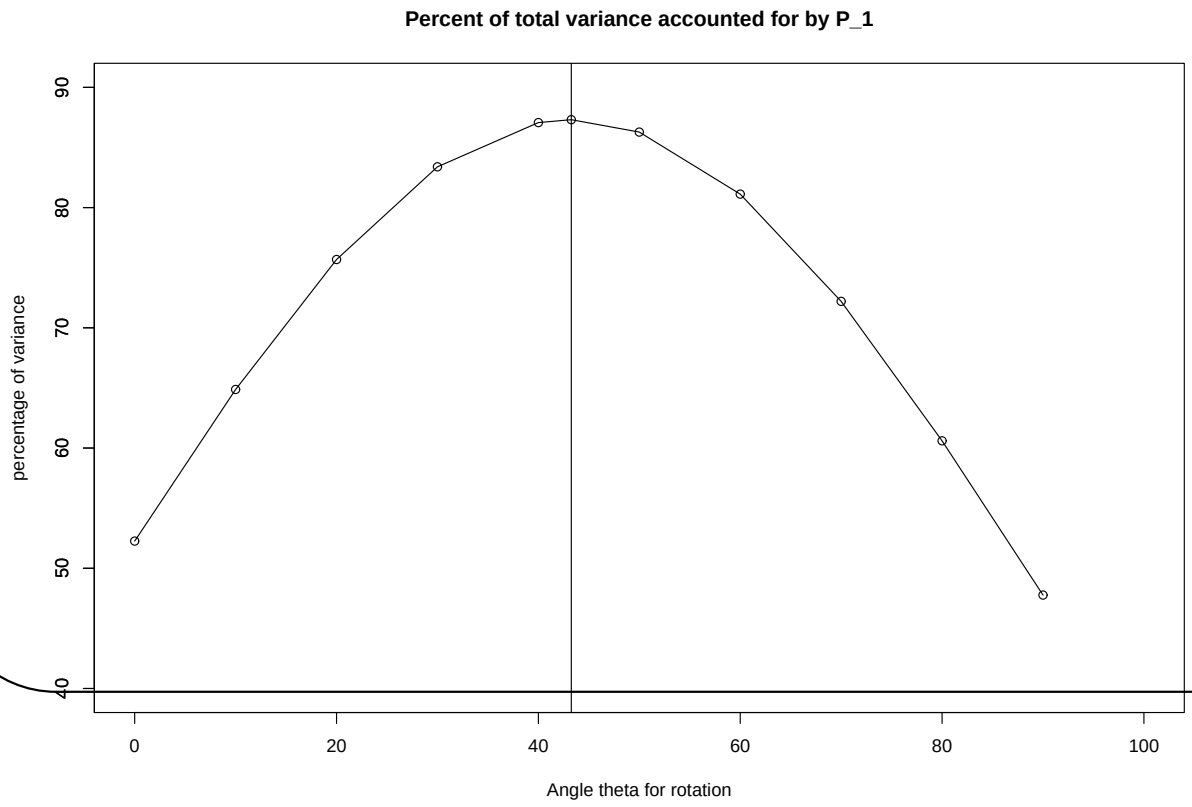
$$\mathbf{SSCP} = \begin{pmatrix} 424.33 & 0.0 \\ 0.0 & 61.67 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 38.58 & 0.0 \\ 0.0 & 5.60 \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$

Το ποσοστό της διακύμανσης που εξηγείται από τις νέες μεταβλητές P_1 και P_2 για διάφορες γωνιές (στροβιλισμούς).

Στροφή Γωνιά θ	Συνολική Διακύμανση	Διακύμανση του P_1	Ποσοστό %
0	44.12	23.09	52.26
10	44.12	28.66	64.87
20	44.12	33.43	75.68
30	44.12	36.84	83.39
40	44.12	38.47	87.07
43.261	44.12	38.58	87.31
50	44.12	38.12	86.28
60	44.12	35.84	81.12
70	44.12	31.90	72.20
80	44.12	26.78	60.60
90	44.12	21.09	47.77



Αποτελέσματα υπολογιστή: Χρημ. δείκτες X_1 και X_2

Simple Statistics

	X_1	X_2
Mean	8.00	3.00
St. Dev	4.81	4.59

Total Variance = 44.18

Covariances

	X_1	X_2
X_1	23.09	16.45
X_2	16.45	21.09

Importance of Components:

	Comp.1	Comp.2
Standard deviation	6.21	2.37
Proportion of Variance	0.87	0.13
Cumulative Proportion	0.87	1.00

Eigenvectors:

	Comp.1	Comp.2
X_1	0.728	0.685
X_2	0.685	-0.728

Correlation Coefficients:

	X_1	X_2	Comp.1	Comp.2
X_1	1.00	0.75	0.94	0.34
X_2	0.75	1.00	0.93	-0.38
Comp.1	0.94	0.93	1.00	-9.9e-17
Comp.2	0.34	-0.38	-9.9e-17	1.00

Scores:

X_1	X_2	Comp.1	Comp.2
16	8	9.25	1.84
12	10	7.71	-2.36
13	6	5.70	1.24
11	2	1.50	2.78
10	8	4.88	-2.27
9	-1	-2.01	3.60
8	4	0.69	-0.73
7	6	1.33	-2.87
5	-3	-6.30	2.31
3	-1	-6.38	-0.51
2	-3	-8.48	0.26
0	0	-7.88	-3.30

Simple Statistics:

	Mean	St.Dev
X_1	8.00	4.81
X_2	3.00	4.59
Comp.1	7.4e-17	6.21
Comp.2	-1.5e-16	2.36

Σύνοψη :

- Η συνολική διακύμανση του X_1 και X_2 είναι 44.18 (i.e. $23.09 + 21.09$).

- Οι μεταβλητές X_1 και X_2 έχουν συντελεστή συσχέτισης 0.746.

- Το ποσοστό της συνολικής διακύμανσης που αναλογεί στο X_1 και X_2 δίνεται αντίστοιχα, 52.26% και 47.74%.

- Κάθε νέα μεταβλητή (δηλαδή, η κύριες συνιστώσες P_1 και P_2) είναι γραμμικοί συνδιασμοί των αρχικών μεταβλητών και παραμένουν διορθωμένες με το μέσο. Έτσι, ο μέσος τους είναι μηδέν.

- Το συνολικό SS (Άθροισμα Τετραγώνων (Sum of Squares)) του P_1 και P_2 ισούνται με το συνολικό SS των αρχικών μεταβλητών ($424.33 + 61.67 = 486$).

- Οι διακυμάνσεις των P_1 και P_2 είναι, αντίστοιχα, 38.58 και 5.61. Η συνολική διακύμανση των κύριων συνιστωσών είναι 44.18 και ισούνται με τη συνολική διακύμανση των αρχικών μεταβλητών X_1 και X_2.

- Το ποσοστό της συνολικής διακύμανσης που αναλογεί στο P_1 και P_2 είναι, αντίστοιχα, 87.31% ($= 38.58/44.18$) και 12.69% ($= 5.61/44.18$).

- Η διακύμανση που αναλογεί στην πρώτη κύρια συνιστώσα P_1 είναι μεγαλύτερη από την διακύμανση που αναλογεί σε κάθε μια από τις αρχικές μεταβλητές.

- Η δεύτερη κύρια συνιστώσα P_2 αναλογεί σε διακύμανση που δεν έχει εξηγηθεί από την P_1. Οι δύο κύριες συνιστώσες αναλογούν στη συνολική διακύμανση των δεδομένων.

- Η συσχέτιση μεταξύ των κύριων συνιστωσών είναι μηδέν, έτσι, P_1 και P_2 είναι ασυσχέτιστες.

Πως επηρεάζει ο τύπος των δεδομένων την ΑΚΣ

Θεωρείστε τα πιο κάτω δεδομένα τα οποία δείχνουν τις *Εκτιμημένες Χονδρικές Τιμές ανα Πόλη, Μάρτιος 1973, U.S. Department of Labour, Boureau of Labor Statistics, pp1-8.*

	Μέση Τιμή (σε σεντ ανα λίρα)				
Πόλη	Ψωμί	Μπιφτέκι	Γάλα	Πορτοκ.	Ντομάτες
Atlanta	24.5	94.5	73.9	80.1	41.6
Baltimore	26.5	91.0	67.5	74.6	53.3
Boston	29.7	100.8	61.4	104.0	59.6
Buffalo	22.8	86.6	65.3	118.4	51.2
Chicago	26.7	86.7	62.7	105.9	51.2
Cincinnati	25.3	102.5	63.3	99.3	45.6
Cleveland	22.8	88.8	52.4	110.9	46.8
Dallas	23.3	85.5	62.5	117.9	41.8
Detroit	24.1	93.7	51.5	109.7	52.4
Honolulu	29.3	105.9	80.2	133.2	61.7
Houston	22.3	83.6	67.8	108.6	42.4
Kansas City	26.1	88.9	65.4	100.9	43.2
Los Angeles	26.9	89.3	56.2	82.7	38.4
Milwaukee	20.3	89.6	53.8	111.8	53.9
Minneapolis	24.6	92.2	51.9	106.0	50.7
New York	30.8	110.7	66.0	107.3	62.6
Philadelphia	24.5	92.3	66.7	98.0	61.7
Pittsburgh	26.2	95.4	60.2	117.1	49.3
St. Louis	26.5	92.4	60.8	115.1	46.2
San Diego	25.5	83.7	57.0	92.8	35.4
San Francisco	26.3	87.1	58.3	101.8	41.5
Seattle	22.5	77.7	62.0	91.1	44.9
Washington DC	24.2	93.8	66.0	81.6	46.2
Μέσος	25.3	91.9	62.3	103.0	48.8
Διακύμανση	6.3	57.1	48.3	202.8	57.8
% of total Variance:	1.7	15.3	13.0	54.5	15.5

Συνολική Διακύμανση: 372.22

Σκοπός είναι η δημιουργία ενός Δείκτη Καταναλωτικών Τιμών (Consumer Price Index (CPI)). Συγκεκριμένα, θέλουμε να δημιουργήσουμε ένα σταθμισμένο άθροισμα τιμών διαφόρων καταναλωτικών προιόντων που θα δείχνει πόσο ακριβά ή φθηνά είναι τα προιόντα κάθε πόλης. Η ΑΚΣ αποτελεί κατάλληλο εργαλείο για τη δημιουργία κάποιου δείκτη.

Η ανάλυση κύριων συνιστωσών μπορεί να εφαρμοστεί είτε σε διορθωμένα με το μέσο είτε σε τυποποιημένα δεδομένα. Κάθε σύνολο δεδομένων μπορεί να δώσει διαφορετική λύση ανάλογα με το βαθμό στον οποίο οι διακυμάνσεις των μεταβλητών διαφέρουν. Έτσι, οι διακυμάνσεις των μεταβλητών πιθανό να έχουν αντίκτυπο στην ΑΚΣ.

ΑΚΣ :

Απλή Στατιστική

	Ψωμί	Μπιφτέκι	Γάλα	Πορτοκάλια	Ντομάτες
Μέσος	25.29	91.86	62.30	102.99	48.77
Τυπ. Απόκλ.	2.51	7.55	6.95	14.24	7.60

Πίνακα Διακυμάνσεων

	Ψωμί	Μπιφτέκι	Γάλα	Πορτοκάλια	Ντομάτες
Ψωμί	6.284	12.91	5.7191	1.3104	7.285
Μπιφτέκι	12.911	57.08	17.5075	22.6919	36.295
Γάλα	5.719	17.51	48.3059	-0.2750	13.443
Πορτοκάλια	1.310	22.69	-0.2750	202.7563	38.762
Ντομάτες	7.285	36.29	13.4435	38.7624	57.801

Σημαντικότητα Συνιστωσών:

	ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5
Τυπική Απόκλιση	14.799	9.577	6.137	4.5619	1.74047
Ποσοστό Διακύμανσης	0.588	0.246	0.101	0.0559	0.00814
Συσσωρευτικό Ποσοστό	0.588	0.835	0.936	0.9919	1.00000

Ιδιοδιανύσματα:

	ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5
Ψωμί	0.028	0.165	-0.021	0.190	-0.967
Μπιφτέκι	0.200	0.632	-0.254	0.659	0.249
Γάλα	0.042	0.442	0.889	-0.108	0.036
Πορτοκαλια	0.939	-0.314	0.121	0.069	-0.015
Ντομάτες	0.276	0.528	-0.361	-0.717	-0.034

Συντελεστής Συσχέτισης:

	Ψωμί	Μπιφτέκι	Γάλα	Πορτοκάλια	Τομάτες
Π̃1	0.168	0.39	0.089	0.976	0.54
Π̃2	0.632	0.80	0.609	-0.211	0.67
Π̃2	-0.052	-0.21	0.785	0.052	-0.29

Αποτελέσματα:

	ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5
Βαλτιμορε	-25.33	13.3	-0.27	-6.106	-0.92
Λος Ανγελες	-22.63	-3.1	-3.52	5.307	-1.75
Ατλαντα	-22.48	10.1	9.47	3.897	2.44
Ωασηινγτον Δ̃	-20.28	8.1	1.15	1.036	2.09
Σεαττλε	-15.15	-7.8	3.35	-7.872	-0.52
⋮	⋮	⋮	⋮	⋮	⋮
Δαλλας	10.76	-12.6	6.16	1.436	0.36
Νεω Ψορκ	11.94	20.4	-6.09	3.437	-1.05
Πιττσβυργη	14.04	-2.7	-1.26	3.323	-0.31
Βυφφαλο	14.14	-6.0	5.05	-4.940	0.89
Ηονολυλυ	35.60	14.8	11.25	0.896	-0.64

Η πρώτη κύρια συνιστώσα ΚΣ1 δίνεται από:

$$\begin{aligned}\text{PC1} =& 0.028\,\Psi\omega\mu\acute{\iota} \ + 0.2\,\text{Μπιφτέκι} \ + 0.042\,\Gamma\acute{\alpha}\lambda\alpha \\ &+ 0.939\,\Pi\text{ορτοκάλια} \ + 0.276\,\text{Ντομάτες}.\end{aligned}$$

Η διακύμανση της ΚΣ1 είναι 218.99 και αναλογεί στο 58.8% της συνολικής διακύμανσης των αρχικών δεδομένων. Η ΚΣ1 αποτελεί το άθροισμα όλων των τιμών τροφήμων και επιρεάζεται από τις τιμές των πορτοκαλιών.

Μιας και όλα τα βάρη (weights) της ΚΣ1 είναι θετικά, ένας μεγάλος αριθμός υπονοεί ότι οι καταναλωτικές τιμές είναι ψηλές και αντίστροφα. Έτσι, από τα αποτελέσματα (τιμές) της ΚΣ1 υποδειλώνει ότι η *Honolulu* είναι η πιο ακριβή πόλη και η *Baltimore* είναι η πιο φθηνή πόλη.

Ο κυριότερος λόγος που οι τιμές των πορτοκαλιών υπερισχύουν έναντι άλλων στην δημιουργία της ΚΣ1 είναι η ύπαρξη μεγάλης κύμανσης στις τιμές των πορτοκαλιών μεταξύ των χωρών. Έτσι, η διακύμανση των τιμών των πορτοκαλιών είναι πολύ ψηλή σε σχέση με τη διακύμανση των τιμών άλλων τροφήμων.

Γενικά, το βάρος που αναλογεί σε μια μεταβλητή επιρεάζεται από τη σχετική διακύμανση των μεταβλητών. Εάν δεν επιθυμούμε η σχετική διακύμανση να επιρεάζει τα βάρη, τότε τα δεδομένα πρέπει να τυποποιηθούν έτσι ώστε η διακύμανση για κάθε μεταβλητή να είναι η ίδια (δηλαδή, ένα).

ΑΚΣ με τυποποίηση δεδομένων :

Πίνακας συσχέτισης

	Ψωμί	Μπιφτέκι	Γάλα	Πορτοκάλια	Ντομάτες
Ψωμί	1.000	0.68	0.3282	0.0367	0.38
Μπιφτέκι	0.682	1.00	0.3334	0.2109	0.63
Γάλα	0.328	0.33	1.0000	-0.0028	0.25
Πορτοκάλια	0.037	0.21	-0.0028	1.0000	0.36
Ντομάτες	0.382	0.63	0.2544	0.3581	1.00

Σημασία Συνιστωσών:

	ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5
Τυπική Απόκλιση	1.556	1.051	0.859	0.7026	0.4907
Ποσοστό Διακύμανσης	0.484	0.221	0.148	0.0987	0.0481
Συσσωρευτικό Ποσοστό	0.484	0.705	0.853	0.9518	1.0000

Ιδιοδιανύσματα:

	ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5
Ψωμί	0.496	-0.3086	-0.3864	0.5093	0.49990
Μπιφτέκι	0.576	-0.0438	-0.2625	-0.0281	-0.77264
Γάλα	0.340	-0.4308	0.8346	0.0491	-0.00788
Πορτοκάλια	0.225	0.7968	0.2916	0.4790	0.00597
Ντομάτες	0.506	0.2870	-0.0123	-0.7127	0.39120

Συντελεστές Συσχέτισης:

	Ψωμί	Μπιφτέκι	Γάλα	Πορτοκάλια	Ντομάτες
ΚΣ1	0.772	0.896	0.529	0.350	0.788
ΚΣ2	-0.324	-0.046	-0.453	0.837	0.302

Αποτελέσματα:

	ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5
Seattle	-2.14	-0.376	0.6639	-0.567	0.703
San Diego	-1.93	-0.741	-0.5847	0.967	0.194
Houston	-1.32	0.152	1.5685	0.253	-0.085
Cleveland	-1.24	1.336	-0.5450	-0.117	-0.277
Los Angeles	-1.21	-1.362	-1.3190	0.595	0.048
⋮	⋮	⋮	⋮	⋮	⋮
Pittsburgh	0.62	0.825	-0.2319	0.594	-0.149
Philadelphia	0.89	0.032	0.5239	-1.546	0.466
Boston	2.30	-0.075	-1.1192	-0.129	0.535
New York	3.78	-0.259	-1.0153	-0.079	-0.122
Honolulu	4.17	0.505	1.6790	0.708	0.022

Το δεδομένα είναι κανονικοποιημένα έτσι η διακύμανση είναι ένα και κάθε μεταβλητή αναλογεί στο 20% της συνολικής διακύμανσης. Η πρώτη κύρια συνιστώσα εξηγεί το 48.84% $(= 1.556^2/5)$ της συνολικής διακύμανσης και δίνεται από:

$$\begin{aligned}\text{ΚΣ1} \ =& 0.496\,\text{Ψωμί} \ + 0.576\,\text{Μπιφτέκι} \ + 0.340\,\text{Γάλα} \\ &+ 0.225\,\text{Πορτοκάλια} \ + 0.506\,\text{Ντομάτες} \ .\end{aligned}$$

Η δεύτερη κύρια συνιστώσα εξηγεί το 22.1% $(= 1.051^2/5)$ της συνολικής διακύμανσης και δίνεται από:

$$\begin{aligned}\text{ΚΣ2} \ =& -0.309\,\text{Ψωμί} \ - 0.044\,\text{Μπιφτέκι} \ - 0.431\,\text{Γάλα} \\ &+ 0.797\,\text{Πορτοκάλια} \ + 0.287\,\text{Ντομάτες} \ .\end{aligned}$$

Η ΚΣ1 αποτελεί το σταθμισμένο άθροισμα όλων των καταναλωτικών τιμών, ενώ δεν υπάρχει προϊόν που να υπερισχύει στη δημιουργία των αποτελεσμάτων.

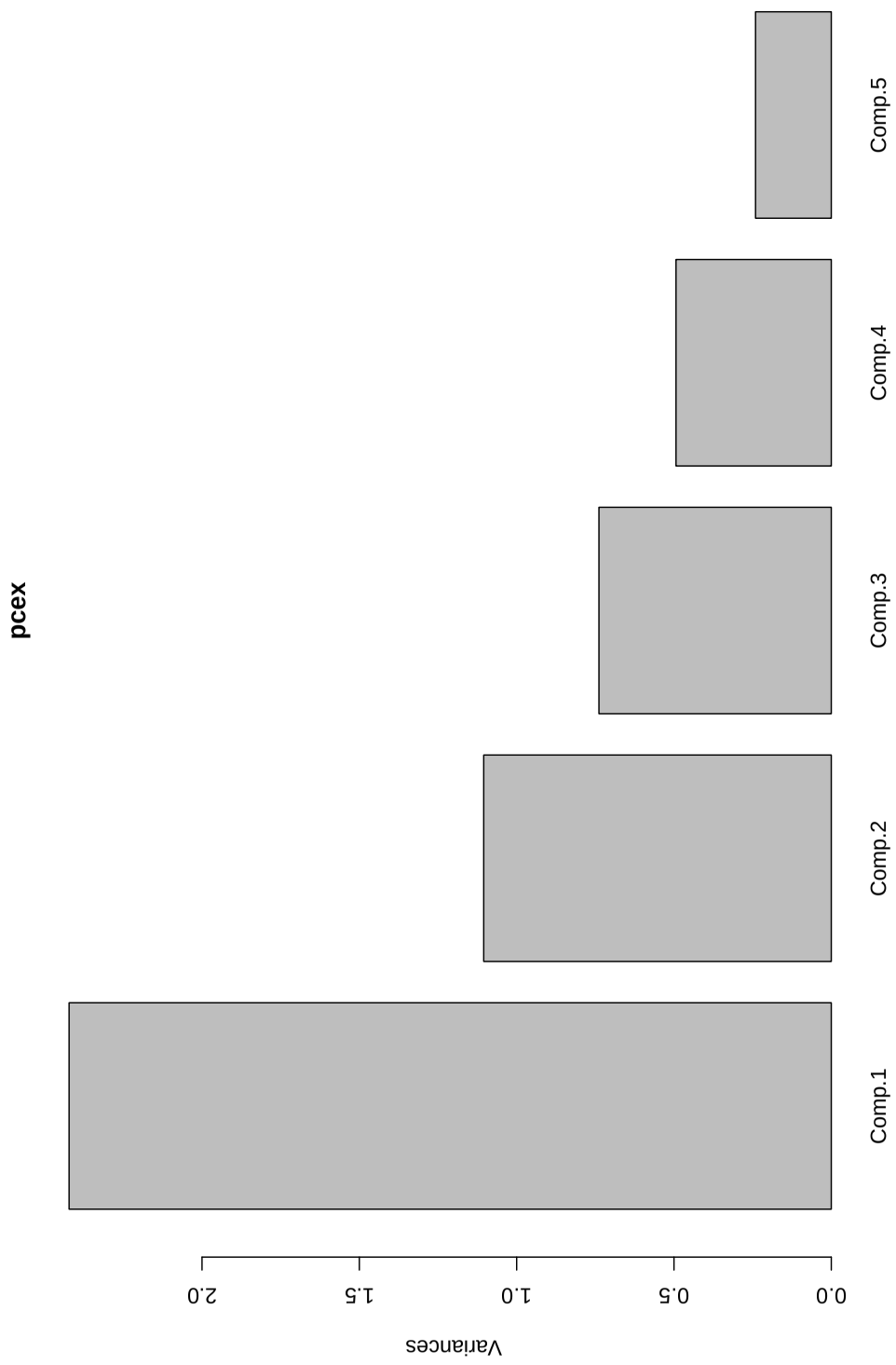
Η τιμή της ΚΣ1 υπονοεί ότι η *Honolulu* είναι η πιο ακριβή πόλη και το *Seattle* είναι τώρα η πιο φθηνή πόλη, όταν αυτή συγκριθεί με την *Baltimore* χωρίς την κανονικοποίηση των δεδομένων. Με λίγα λόγια, τα βάρη (weights) που χρησιμοποιήθηκαν για την δημιουργία του CPI επηρεάζονται από τις σχετικές διακυμάνσεις των μεταβλητών.

Για τον αριθμό των κύριων συνιστωσών που θα επιλεγούν αρκεί να αποφασίσουμε πόση πληροφόρηση, δηλαδή, ανεξήγητη διακύμανση, είμαστε διαθετημένοι να θυσιάσουμε. Η απόφαση είναι φυσικά υποκειμενική. Υπάρχουν δύο εναλλακτικές:

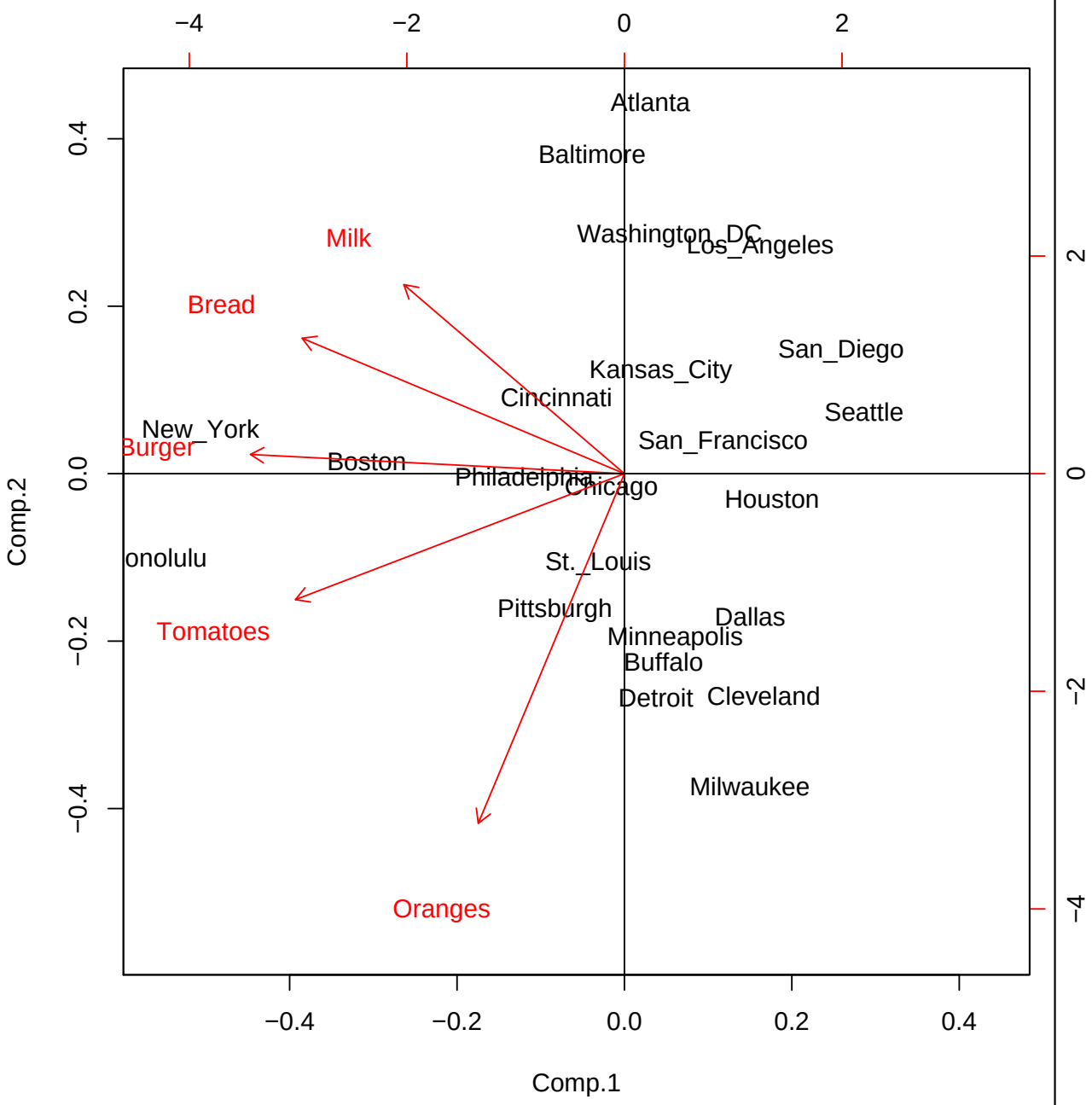
- Χρησιμοποιείστε την Τεχνική του Αγκώνα (Screen Plot). Μπορεί να χρησιμοποιηθεί για διορθωμένα με το μέσο δεδομένα και για κανονικοποιημένα.

- Στην περίπτωση των κανονικοποιημένων δεδομένων διατηρείστε μόνο τις συνιστώσες των οποίων οι ιδιοτιμές (διακύμανση) είναι μεγαλύτερη της μονάδας. Αυτό αναφέρεται ως *ιδιοτιμή-μεγαλύτερη-από-ένα*. Αυτός ο κανόνας επιλέγεται αυτόματα στα περισσότερα στατιστικά πακέτα (SAA, SPSS). Το σκεπτικό αυτού του κανόνα είναι ότι το ποσό της διακύμανσης που εξηγείται από κάθε συνιστώσα πρέπει να είναι τουλάχιστον ίσο με τη διακύμανση μιας μεταβλητής. Αυτό ισχύει για κανονικοποιημένα δεδομένα.

Γράφημα Διακύμανσης

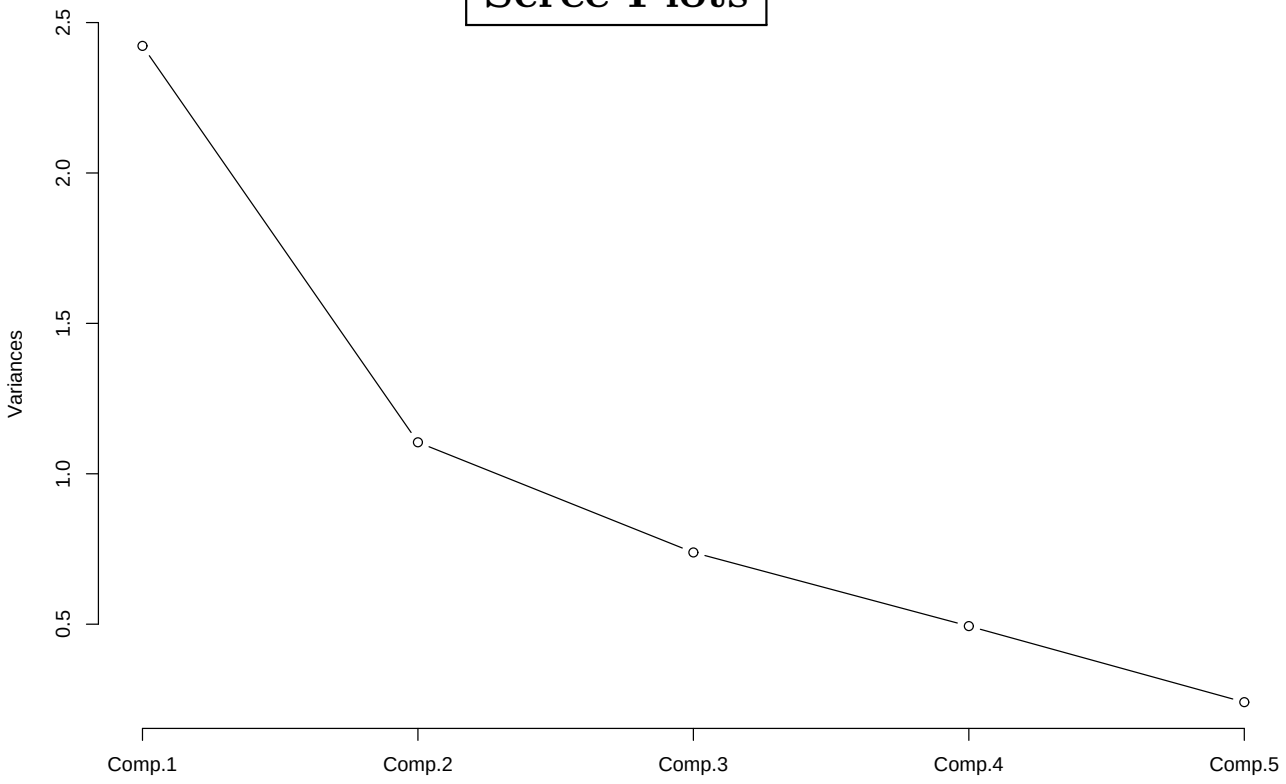


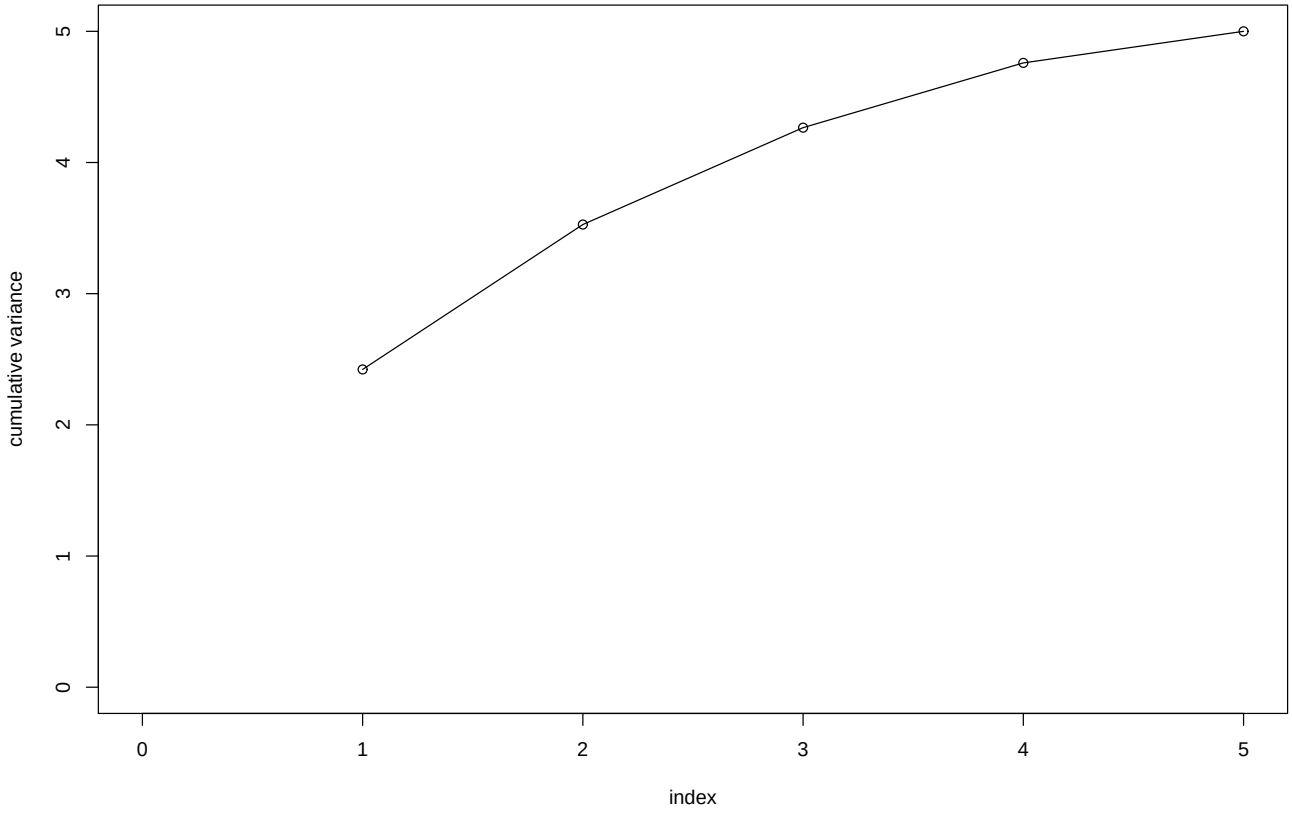
Bilot


−4
−2
0
2
0.4
0.2
0.0
−0.2
−0.4
Comp.2
Atlanta
Baltimore
Milk
Washington_DC
Los_Angeles
Bread
San_Diego
Kansas_City
Cincinnati
Seattle
New_York
Burger
San_Francisco
Boston
Philadelphia
Chicago
Houston
onolulu
St._Louis
Pittsburgh
Tomatoes
Dallas
Minneapolis
Buffalo
Detroit
Cleveland
Milwaukee
Oranges
−0.4
−0.2
0.0
0.2
0.4
Comp.1
2
0
−2
−4

Scree Plots

pcex





Τραπεζογραμμάτια της Swiss Bank

Έξι μεταβλητές αξιολογήθηκαν βάση 100 αληθινών και 100 πλαστών παλαιών τραπεζογραμματίων της Swiss 1000-φράγκα. Το δεδομένα προέρχονται από το βιβλίο "Multivariate Statistics A practical approach", του Bernhard Flury and Hans Riedwyl, Chapman and Hall, 1988, Tables 1.1 and 1.2 pp. 5-8. Οι στήλες αντιστοιχούν στις ακόλουθες 6 μεταβλητές:

X_1:	Μήκος γραμματίου,
X_2:	Ύψος γραμματίου, από αριστερά
X_3:	Ύψος γραμματίου, από αριστερά
X_4:	Απόσταση εσωτερικού πλαισίου από κατώτατο άκρο
X_5:	Απόσταση εσωτερικού πλαισίου από ανώτατο άκρο
X_6:	Μήκος διαγωνίου.

Οι παρατηρήσεις 1-100 αντιπροσωπεύουν τα αληθινά χρεόγραφα και οι άλλες 100 παρατηρήσεις τα πλαστά.

Μήκος	Ύψος (αριστερ.)	Ύψος (δεξιά)	Εσωτ. Πλ. (κάτω.)	Εσωτ. Πλ. (άνω.)	Διαγών.
x_1	x_2	x_3	x_4	x_5	x_6
214.8	131.0	131.1	9.0	9.7	141.0
214.6	129.7	129.7	8.1	9.5	141.7
214.8	129.7	129.7	8.7	9.6	142.2
214.8	129.7	129.6	7.5	10.4	142.0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
214.9	130.3	130.5	11.6	10.6	139.8
215.0	130.4	130.3	9.9	12.1	139.6
215.1	130.3	129.9	10.3	11.5	139.7
214.8	130.3	130.4	10.6	11.1	140.0
214.7	130.7	130.8	11.2	11.2	139.4
214.3	129.9	129.9	10.2	11.5	139.6

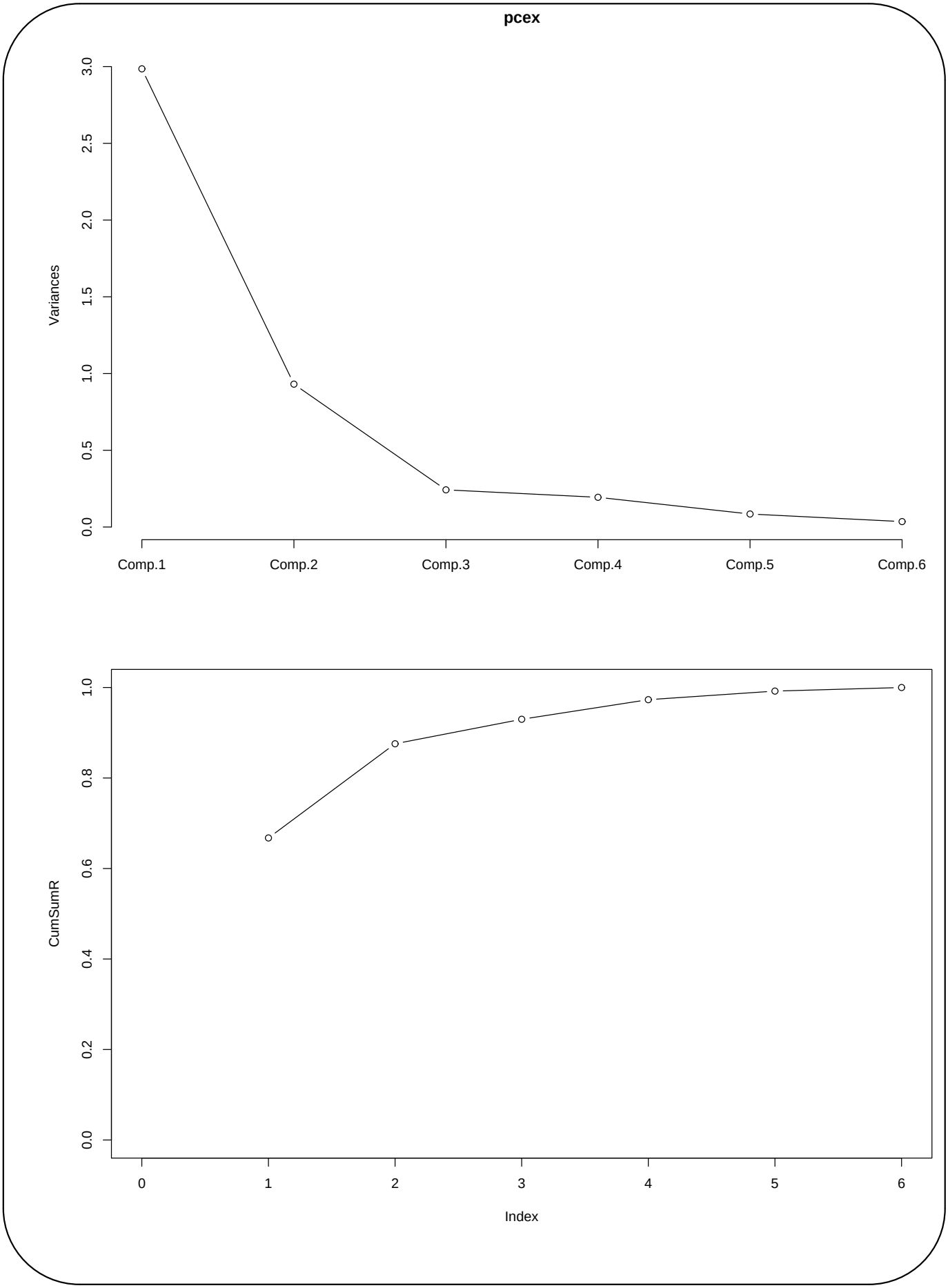
Οι διακυμάνσεις και οι συνδιακυμάνσεις των μεταβλητών είναι:

	Μήκος	Αριστερ.	Δεξιά	Κάτω	Πάνω	Διαγ.
Μήκος	0.14	0.03	0.02	-0.1	-0.02	0.08
Αριστερ.	0.03	0.13	0.11	0.2	0.11	-0.21
Δεξιά	0.02	0.11	0.16	0.3	0.13	-0.24
Κάτω	-0.10	0.22	0.28	2.1	0.16	-1.04
Πάνω	-0.02	0.11	0.13	0.2	0.64	-0.55
Διαγ.	0.08	-0.21	-0.24	-1.0	-0.55	1.33

Σημασία συνιστωσών:

	ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5	ΚΣ6
Τυπ. Αποκλ.	1.73	0.96	0.49	0.44	0.29	0.19
Ποσοστ. Διακύμ.	0.67	0.21	0.05	0.04	0.02	0.01
Συσσωρευτ. Ποσοστ.	0.67	0.88	0.93	0.97	0.99	1.00

Συντελεστές:						
ΚΣ1	ΚΣ2	ΚΣ3	ΚΣ4	ΚΣ5	ΚΣ6	
Μήκος			-0.33	0.56	0.75	
Αριστ.	0.11		-0.26	0.46	-0.35	-0.77
Δεξιά	0.14		-0.35	0.42	-0.54	0.63
Κάτω	0.77	-0.56	-0.22	-0.19		
Πάνω	0.20	0.66	-0.56	-0.45	0.10	
Διαγ.	-0.58	-0.49	-0.59	-0.26		


pcex
Variances
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
CumSumR
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
Index

Η βασική ιδέα του μετασχηματισμού κύριων συνιστωσών είναι να βρεθούν οι πιο πληροφοριακές προβολές που μεγιστοποιούν τη διακύμανση. Η περισσότερη πληροφόρηση δίνεται από το πρώτο ιδιοδιάνυσμα. Συγκεκριμένα:

$$y_1 = -0.0x_1 + 0.11x_2 + 0.14x_3 + 0.77x_4 + 0.20x_5 - 0.58x_6$$
$$y_2 = -0.56x_4 + 0.66x_5 - 0.49x_6$$

Έτσι, η πρώτη ΚΣ ειναι η διαφορά μεταξύ του κάτω πλαισίου και της διαγωνίου. Η δεύτερη ΚΣ δίνεται καλύτερα από τη διαφορά μεταξύ του πάνω πλαισίου και του αθροίσματος του κάτω πλαισίου και των διαγώνιων μεταβλητών.

Εισαγωγή στην Ανάλυση Παραγόντων

- Ανάλυση παραγόντων (Factor analysis (FA)) ονομάζονται οι πολυμεταβλητές στατιστικές μέθοδοι των οποίων πρωταρχικός σκοπός είναι ο ορισμός της δομής ενός πίνακα δεδομένων.

- Ασχολείται με την ανάλυση της συσχέτισης (δομής) μεγάλου αριθμού μεταβλητών μέσω ενός αριθμού παραγόντων (διαστάσεων).

- Στην ανάλυση παραγόντων ο ερευνητής προσδιορίζει πρώτα τις διαστάσεις και στην συνέχεια καθορίζει το βαθμό στον οποίο κάθε διάσταση εξηγεί κάθε μεταβλητή. Όταν καθοριστούν οι διαστάσεις και η επεξήγηση κάθε μεταβλητής, τότε μπορεί να επιτευχθεί η ανάλυση παραγόντων. Κύριος της σκοπός είναι η περιληπτική παρουσίαση και η μείωση των δεδομένων.

- Στην περιληπτική παρουσίαση των δεδομένων, η FA περιγράφει τα δεδομένα χρησιμοποιώντας πολύ λιγότερες έννοιες από ότι οι αρχικές μεταβλητές.

- Η μείωση (αναγωγή, απλοποίηση) δεδομένων μπορεί να επιτευχθεί με τον υπολογισμό των αποτελεσμάτων για κάθε διάσταση και την αντικατάσταση αυτών με τις αρχικές μεταβλητές.

- Η FA διαφέρει από τις τεχνικές εξάρτησης (π.χ. πολλαπλή παλινδρόμηση) στην οποία μία ή περισσότερες μεταβλητές θεωρούνται εξαρτημένες μεταβλητές και όλες οι άλλες είναι ανεξάρτητες ή προβλέπουσες.

- Η FA είναι μια τεχνική αλληλοεξάρτησης όπου όλες οι μεταβλητές θεωρούνται ότι συσχετίζονται ταυτόχρονα μεταξύ τους. Οι μεταβλητές (παράγοντες) σχηματίζονται με σκοπό τη μεγιστοποίηση της επεξηγηματικότητας όλων των μεταβλητών. Δεν χρησιμοποιούνται για την πρόβλεψη εξαρτημένης μεταβλητής(τών).

- Οι τεχνικές ανάλυσης παραγόντων επιτυγχάνουν το σκοπό τους επεξηγηματικά ή επικυρωτικά. Με τον όρο επεξηγηματικά εννοούμε την εξέταση της δομής ενός συνόλου μεταβλητών ή την μείωση της

διάστασης των δεδομένων.

Σε αυτή την περίπτωση δεν υπάρχουν εκ των προτέρων περιορισμοί στην εκτίμηση των συνιστωσών ή του αριθμού των συνιστωσών που εξάγονται. Αυτή είναι η πιο κατάλληλη FA στις περισσότερες περιπτώσεις.

Στην επικυρωτική περίπτωση θέλουμε να εξετάσουμε την υπόθεση για το ποιες μεταβλητές πρέπει να κατηγοριοποιηθούν στον ίδιο παράγοντα και πόσους παράγοντες να επιλέξουμε. Σε αυτή την περίπτωση η FA μπορεί να χρησιμοποιηθεί για να αξιολογηθεί εάν τα δεδομένα προσαρμόζονται στην αναμενόμενη δομή.

Παράδειγμα

Υποθέστε ότι μέσω μιας ποσοτικής μελέτης μια εμπορική εταιρεία εντόπισε 80 διαφορετικά χαρακτηριστικά και υπηρεσίες που επιρεάζουν την απόφαση των αγοραστών να στραφούν σε διαφορετικά καταστήματα.

Ο λιανοπωλητής θέλει να κατανοήσει πως οι καταναλωτές παίρνουν αποφάσεις, όμως δε μπορεί να αξιολογήσει ή να πάρει μέτρα για 80 διαφορετικά χαρακτηριστικά (μεταβλητές) μιας και είναι πολύ συγκεκριμένα. Εναλλακτικά, θέλει να εντοπίσει κατά πόσο οι καταναλωτές αξιολογούν με βάση πιο γενικευμένες παραμέτρους. Για να προσδιορίσει αυτές τις παραμέτρους, ο λιανοπωλητής, συντάσσει ένα ερωτηματολόγιο για συγκεκριμένα χαρακτηριστικά.

Η Ανάλυση Παραγόντων μπορεί να χρησιμοποιηθεί στη συνέχεια για τον εντοπισμό των σημαντικών παραμέτρων. Συγκεκριμένα χαρακτηριστικά που συσχετίζονται σε μεγάλο βαθμό ανήκουν σε μια ευρύτερη κατηγορία γενικευμένων παραμέτρων.

Π.χ. τιμές αγαθού, ποιότητα αγαθού, κατάταξη αγαθού, προσωπικό καταστήματος, υπηρεσίες και ατμόσφαιρα καταστήματος μπορεί να καθοριστούν από την FA ως οι πιο σημαντικές διαστάσεις. Κάθε διάσταση περιλαμβάνει συγκεκριμένα στοιχεία για την αξιολόγηση αυτών των διαστάσεων. Από τα αποτελέσματα ο λιανοπωλητής μπορεί να χρησιμοποιήσει τις διαστάσεις (παράγοντες) για προγραμματισμό και λήψη μέτρων.

Σκοπός της Ανάλυσης Παραγόντων

Ο γενικός σκοπός των αναλυτικών τεχνικών παραγόντων είναι η σύνοψη (περίληψη) των πληροφοριών που περιλαμβάνουν οι αρχικές μεταβλητές σε ένα μικρότερο αριθμό νέων παραγόντων, με την ελάχιστη απώλεια πληροφοριών. Συγκεκριμένα, αυτές οι τεχνικές ικανοποιούν ένα από τους πιο κάτω στόχους: (1) προσδιορισμό δομής μέσω σύνοψης δεδομένων, ή (2) μείωση διαστάσεων δεδομένων.

1. Η FA προσδιορίζει τη δομή της σχέσης μεταξύ είτε μεταβλητών είτε παρατηρήσεων εξετάζοντας τη συσχέτιση μεταξύ ειτε των μεταβλητών είτε τη συσχέτιση μεταξύ των παρατηρήσεων. Για παράδειγμα υποθέστε ότι έχουμε 100 παρατηρήσεις για 10 χαρακτηριστικά. Αν σκοπός της ανάλυσης είναι η σύνοψη των χαρακτηριστικών, τότε για την FA θα εφαρμοστεί ο πίνακας συσχετίσεων των μεταβλητών. Αυτή είναι και η πιο σύνηθες μορφή FA και αναφέρεται ως **R factor analysis**.

R FA αναλύει ένα σύνολο μεταβλητών για τον εντοπισμό των διαστάσεων που δεν μπορούν εύκολα να παρατηρηθούν (latent).

Η FA που εφαρμόζεται στον πίνακα συσχετίσεων των ατομικών παρατηρήσεων βάση των χαρακτηριστικών τους ονομάζεται **Q factor analysis**. Αυτή η τεχνική δεν χρησιμοποιείται συχνά. Εναλλακτικά, χρησιμοποιείται η *ανάλυση ομάδων* για την κατηγοριοποίηση παρατηρήσεων.

2. Η FA μπορεί επίσης

- Να προσδιορίσει ένα αντιπροσωπευτικό δείγμα μεταβλητών από ένα μεγαλύτερο σύνολο μεταβλητών το οποίο μπορεί να χρησιμοποιηθεί σε ακόλουθη πολυμεταβλητή ανάλυση.

- Να δημιουργήσει ένα εντελώς νέο σύνολο μεταβλητών, πολύ μικρότερο σε αριθμό, για μερική ή πλήρη αντικατάσταση του αρχικού συνόλου μεταβλητών. Το νέο σύνολο μπορεί να χρησιμοποιηθεί και από άλλες μεθόδους.

  Και στις δύο περιπτώσεις, σκοπός είναι η διατήρηση της δομής και του χαρακτήρα των

αρχικών μεταβλητών, με τη μείωση του αριθμού τους για την απλοποίηση ακόλουθων πολυμεταβλητών αναλύσεων.

Ένας ερευνητής προσπαθεί να εξασφαλήσει το πιο λιτό σύνολο μεταβλητών για να συμπεριλάβει στην ανάλυση.

Παράδειγμα

Θεωρείστε ένα απλό παράδειγμα ανάλυσης παραγόντων το οποίο περιλαμβάνει τα ακόλουθα 9 στοιχεία καταστημάτων:

V_1	Επίπεδο τιμών	V_6	Βάθος κατάταξης
V_2	Προσωπικό	V_7	Πλάτος κατάταξης
V_3	Πολιτική επιστροφών	V_8	Υπηρεσίες στο κατάστημα
V_4	Διαθεσιμότητα αγαθών	V_9	Ατμόσφαιρα
V_5	Ποιότητα αγαθών		

Ο πίνακας συσχέτισης των μεταβλητών δίνεται από:

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
V_1	1.0								
V_2	0.427	1.0							
V_3	0.302	0.771	1.0						
V_4	0.470	0.497	0.427	1.0					
V_5	0.765	0.406	0.307	0.472	1.0				
V_6	0.281	0.445	0.423	0.713	0.325	1.0			
V_7	0.354	0.490	0.471	0.719	0.378	0.724	1.0		
V_8	0.242	0.719	0.733	0.428	0.240	0.311	0.435	1.0	
V_9	0.372	0.737	0.774	0.479	0.326	0.429	0.466	0.710	1.0

Ερώτηση

Υπάρχει για κάθε ένα από τα στοιχεία ξεχωριστή επεξηγηματική αξία ή μπορούν κάποιες από αυτές να ομαδοποιηθούν και να δώσουν γενικευμένες αξιολογήσεις;

Οι αρχικές συσχετίσεις δεν υποδηλώνουν κάποιο συγκεκριμένο υπόδειγμα. Ενώ υπάρχουν ψηλές συσχετίσεις, δεν είναι προφανείς οποιαδήποτε ομαδοποίηση. Η ανάλυση παραγόντων οδηγεί στην κατηγοριοποίηση των μεταβλητών ως ακολούθως:

	V_3	V_8	V_9	V_2	V_6	V_7	V_4	V_1	V_5
V_3	1.0								
V_8	0.733	1.0							
V_9	0.774	0.710	1.0						
V_2	0.741	0.719	0.787	1.0					
V_6	0.423	0.311	0.429	0.445	1.0				
V_7	0.471	0.435	0.468	0.490	0.724	1.0			
V_4	0.427	0.428	0.479	0.497	0.713	0.719	1.0		
V_1	0.302	0.242	0.372	0.427	0.281	0.354	0.470	1.0	
V_5	0.307	0.240	0.326	0.406	0.325	0.378	0.472	0.765	1.0

- Οι πρώτες τέσσερις μεταβλητές που σχετίζονται με τις εντυπώσεις των καταναλωτών μέσα στο κατάστημα κατηγοριοποιούνται μαζί.
- Οι τρεις μεταβλητές που περιγράφουν την κατάταξη και τη διαθεσιμότητα των αγαθών ομαδοποιούνται μαζί.
- Τέλος, η ποιότητα των αγαθών και το επίπεδο τιμών ομοδοποιούνται.

Κάθε ομάδα αντιπροσωπεύει ένα σύνολο μεταβλητών με ψηλό δείκτη συσχέτισης που απεικονίζει μια πιο γενικευμένη διάσταση αξιολόγησης. Αυτές οι ομάδες μπορούν να ονομαστούν *εντυπώσεις στο κατάστημα*, *προσφορά αγαθών* και *αξία*.

Ανάλυση Παραγόντων & Κύριων Συνιστωσών

- Υπολογείστε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα διακυμάνσεων-συνδιακυμάνσεων (ή συσχέτισης). Οι ιδιοτιμές αντιστοιχούν στη διακύμανση των παραγόντων.

- Υπολογείστε το *Ποσοστό Διακύμανσης (Proportion of Variance)* για κάθε παράγοντα. Μεγάλο ποσοστό διακύμανσης υπονοεί ότι ο παράγοντας είναι σημαντικός.

- Υπολογείστε το *Συσσωρευτικό Ποσοστό Διακύμανσης (Cumulative Proportion of Variance)* και βρέστε των αριθμό των παραγόντων που θα κρατήσετε.

- Χρησιμοποιείστε Τεχνική του Αγκώνα (screeplot) (ιδιοτιμές έναντι συσσωρευτικού ποσοστού διακύμανσης) για αναπαράσταση του σημείου τομής.

- Υπολογείστε τα ιδιοδιανύσματα για την εξαγωγή των συνδυασμών μεταβλητών σε κάθε παράγοντα (συντελεστές).

- Αν χρησιμοποιείται ο πίνακας συσχέτισης τότε τα βάρη υπολογίζονται μέσω μετασχηματισμού.

Παράδειγμα

- Ο πίνακας συσχέτισης R:

```
1.0   0.427 0.302 0.470 0.765 0.281 0.354 0.242 0.372
0.427 1.0   0.771 0.497 0.406 0.445 0.490 0.719 0.737
0.302 0.771 1.0   0.427 0.307 0.423 0.471 0.733 0.774
0.470 0.497 0.427 1.0   0.472 0.713 0.719 0.428 0.479
0.765 0.406 0.307 0.472 1.0   0.325 0.378 0.240 0.326
0.281 0.445 0.423 0.713 0.325 1.0   0.724 0.311 0.429
0.354 0.490 0.471 0.719 0.378 0.724 1.0   0.435 0.466
0.242 0.719 0.733 0.428 0.240 0.311 0.435 1.0   0.710
0.372 0.737 0.774 0.479 0.326 0.429 0.466 0.710 1.0
```

- Εφορμογή PCA στην R:

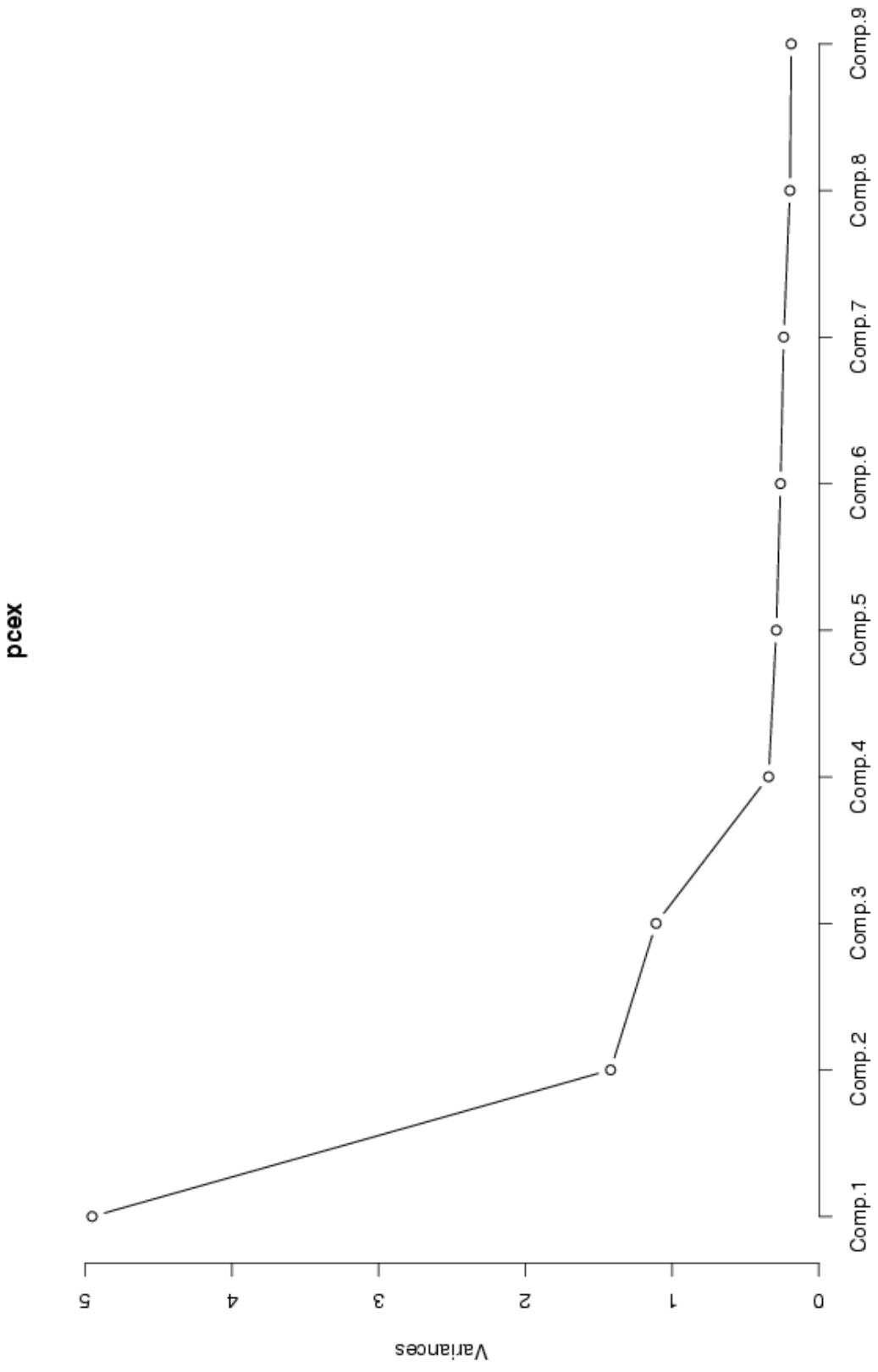
```
> pcex <-princomp(covmat=R); summary(pcex)
Importance of components:
Component   1    2    3     4    5    6    7    8    9
SD         2.22 1.19 1.05 0.59 0.54 0.51 0.49 0.44 0.43
Propor Var 0.55 0.16 0.12 0.04 0.03 0.03 0.03 0.02 0.02
Cum Propor 0.55 0.71 0.83 0.87 0.90 0.93 0.96 0.98 1.00
```

- Συντελεστές παραγόντων (ιδιοδιανύσματα):

```
Loadings:
Com.1  Com.2 Com.3 Com.4 Com.5 Com.6 Com.7 Com.8 Com.9
-0.27 -0.44  0.47  0.01  0.21 -0.24 -0.42  0.14  0.46
-0.38  0.24  0.16  0.15 -0.16  0.47 -0.60  0.02 -0.39
-0.36  0.36  0.10  0.31 -0.24 -0.06  0.18 -0.59  0.45
-0.35 -0.27 -0.30 -0.36  0.53  0.26  0.04 -0.47 -0.08
-0.27 -0.47  0.42  0.02 -0.33  0.18  0.54  0.00 -0.31
-0.31 -0.20 -0.52  0.47  0.01  0.27  0.14  0.45  0.28
-0.34 -0.18 -0.42 -0.24 -0.50 -0.54 -0.21 -0.04 -0.17
-0.33  0.41  0.10 -0.63 -0.06  0.15  0.20  0.42  0.27
-0.37  0.30  0.12  0.27  0.48 -0.48  0.21  0.18 -0.39
```

• Screeplot:


pcex
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7
Comp.8
Comp.9
0
1
2
3
4
Variances


pcex
Variances
0
1
2
3
4
5
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7
Comp.8
Comp.9

- Ανάλυση Παραγόντων πίνακα συσχετίσεων R (4 παράγοντες)

```
factanal(covmat = R, factors = 4)
Loadings:
      Factor1 Factor2 Factor3 Factor4
 [1,]  0.172   0.163   0.934
 [2,]  0.770   0.285   0.266
 [3,]  0.872   0.248   0.122
 [4,]  0.247   0.752   0.325   0.129
 [5,]  0.170   0.243   0.745
 [6,]  0.227   0.838   0.114
 [7,]  0.293   0.764   0.190
 [8,]  0.816   0.186           0.506
 [9,]  0.782   0.273   0.203

                Factor1 Factor2 Factor3 Factor4
SS loadings       2.890   2.189   1.713   0.291
Proportion Var    0.321   0.243   0.190   0.032
Cumulative Var    0.321   0.564   0.755   0.787
```

- Ανάλυση Παραγόντων πίνακα συσχετίσεων R (3 παράγοντες)

```
factanal(covmat = R, factors = 3)
Loadings:
     Factor1 Factor2 Factor3
 [1,] 0.172   0.160   0.955
 [2,] 0.786   0.275   0.260
 [3,] 0.852   0.242   0.123
 [4,] 0.270   0.748   0.319
 [5,] 0.173   0.249   0.728
 [6,] 0.227   0.824   0.116
 [7,] 0.304   0.769   0.187
 [8,] 0.806   0.201
 [9,] 0.795   0.261   0.201

                Factor1 Factor2 Factor3
SS loadings       2.902   2.160   1.722
Proportion Var    0.322   0.240   0.191
Cumulative Var    0.322   0.562   0.754
```

- Ανάλυση Παραγόντων πίνακα συσχετίσεων R (2 παράγοντες)

```
Loadings:
     Factor1 Factor2
[1,] 0.260   0.443
[2,] 0.794   0.354
[3,] 0.851   0.271
[4,] 0.258   0.839
[5,] 0.235   0.469
[6,] 0.214   0.785
[7,] 0.290   0.780
[8,] 0.792   0.233
[9,] 0.800   0.319

                       Factor1 Factor2
SS loadings              2.940   2.700
Proportion Var           0.327   0.300
Cumulative Var           0.327   0.627
```

- Ανάλυση Παραγόντων πίνακα συσχετίσεων R (1 παράγοντας)

```
Loadings:
     Factor1                          Factor1
[1,] 0.466          SS loadings        4.369
[2,] 0.871          Proportion Var     0.485
[3,] 0.858
[4,] 0.627
[5,] 0.455
[6,] 0.563
[7,] 0.625
[8,] 0.792
[9,] 0.850
```