ΚΕΦΑΛΑΙΟ IX
Η ΑΝΑΛΥΣΗ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ

Πάντα κατ' αριθμόν γίνονται
Πυθαγόρας

9.0 Γενικά

Η ανάλυση σε Κύριες Συνιστώσες (Analyse Composantes Principales, ACP) [Hotelling H. 1933] είναι η πιο παλιά από τις μεθόδους της ανάλυσης δεδομένων. Η ACP εφαρμόζεται αποκλειστικά σε πίνακες δεδομένων που περιγράφουν τις μετρήσεις p ανομοιογενών ποσοτικών μεταβλητών V_j (j=1,..p) οι οποίες παρατηρούνται στις n στατιστικές μονάδες I_i (i=1,...,n) του πληθυσμού που μελετάται.

Συμβολίζουμε με k_{ij} την τιμή της μεταβλητής V_j που μετρήθηκε στη i στατιστική μονάδα.

Ο όρος ανομοιογενείς μεταβλητές σημαίνει, πως δεν χρησιμοποιούνται στις p μεταβλητές οι ίδιες μονάδες μετρήσεις, οπότε αν πάρουμε το άθροισμα μιας γραμμής του πίνακα δεδομένων A(i,j) ο αριθμός που προκύπτει δεν έχει κανένα ουσιαστικό νόημα.

Γενικώς η ανάλυση σε κύριες συνιστώσες αφορά στην κατασκευή νέων ασυσχέτιστων μεταβλητών, που καλούνται «**παράγοντες**» οι οποίοι περιγράφουν με μεγαλύτερη ευκρίνεια τα αρχικά δεδομένα και αναδεικνύουν καταστάσεις οι οποίες δεν διαφαίνονται από την αρχική μελέτη του πίνακα δεδομένων.

Η αντικατάσταση των αρχικών μεταβλητών από τους «παράγοντες» το πλήθος των οποίων είναι p, αφήνει αναλλοίωτη την συμπεριφορά των αρχικών δεδομένων, οπότε η εικόνα του πίνακα T(nxp) {δηλαδή n αντικείμενα x p μεταβλητές) είναι ίδια και στον πίνακα C(nxp) {δηλαδή n αντικείμενα x p παράγοντες}.

Για την εύρεση των νέων αυτών μεταβλητών c_i {i=1,2,…p) προτάθηκαν αρκετά κριτήρια μεταξύ αυτών το κριτήριο της αδράνειας που πρότεινε ο K.Pearson [Pearson K.,1901] και το κριτήριο της συσχέτισης που πρότεινε ο H.Hotelling [Hotelling H.,1933].

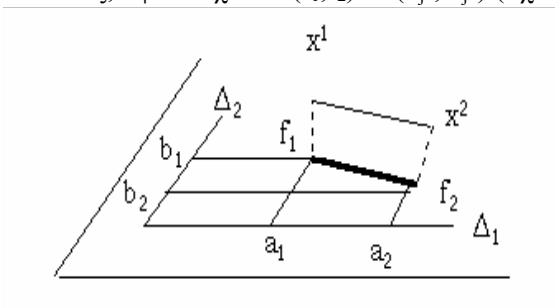
Εμείς θα αναπτύξουμε το κριτήριο της αδράνειας καθόσον αδράνεια και πληροφορία συνδέονται άμεσα όπως προαναφέρθηκε στη §2.21.

9.1 Το κριτήριο του K. Peasron

Κάθε στατιστική μονάδα I_i (i=1,...n) περιγράφεται από p ποσοτικές κανονικοποιημένες μεταβλητές, ήτοι από ένα διάνυσμα γραμμή $X_j^i=(x_1^i,x_2^i...,x_p^i)$. Το σύνολο των στοιχείων $\{X_j^i\}$ σχηματίζει ένα νέφος σημείων μέσα στο χώρο των p διαστάσεων, αφού κάθε διάνυσμα γραμμή περιγράφεται από p συντεταγμένες.

Όμως η εποπτική παρουσίαση του νέφους N(I) στον R^p δεν είναι δυνατή, έτσι αρκεί να παρουσιάσουμε γραφικά τα n διανύσματα προβάλλοντας τα σ' ένα

επίπεδο. Με την προβολή των συντεταγμένων x_j^i κάθε στατιστικής μονάδας σ' ένα επίπεδο, έχουμε κάποια παραμόρφωση της εικόνας του νέφους, την οποία θέλουμε να καταστήσουμε ελάχιστη, αφού η απόσταση $d(x_j^1, x_j^2)$ δύο στατιστικών μονάδων στο χώρο R^p, γενικώς δεν μπορεί να παραμείνει ίση με την απόσταση $d(f_1,f_2)$ των προβολών τους, αφού ισχύει $d(f_1,f_2) < d(x_j^1, x_j^2)$ (σχ. 9.1).



Σχήμα 9.1: Η προβολή του νέφους N(I) σε επίπεδο

Πρέπει λοιπόν να διαλέξουμε το κατάλληλο επίπεδο προβολής πάνω στο οποίο οι αποστάσεις $d(f_i,f_j)$ μεταξύ όλων των προβολών των σημείων X_j^i ανά δύο να διατηρούνται όσο το δυνατόν καλύτερα.

Λαμβάνοντας λοιπόν υπόψη ότι η προβολή σμικρύνει σχεδόν πάντα τις πραγματικές αποστάσεις θα καθορίσουμε ένα κριτήριο που μεγιστοποιεί το μέσο των τετραγώνων των αποστάσεων μεταξύ των προβολών $f_1, f_2,.... f_n$.

Το επίπεδο πάνω στο οποίο γίνεται η προβολή και έχει την προαναφερθείσα ιδιότητα ονομάζεται **παραγοντικό επίπεδο**. Για να προσδιοριστεί το επίπεδο αυτό αρκεί να βρούμε δύο ευθείες Δ_r, Δ_s κάθετες μεταξύ τους τέτοιες ώστε να ισχύει η σχέση

$$d^2(f_i, f_j) = d^2(a_i, a_j) + d^2(b_i, b_j) \qquad (9.1)$$

όπου a_i,a_j,b_i,b_j είναι οι προβολές των f_i,f_j πάνω στις ευθείες Δ_r,Δ_s.

Οι ευθείες αυτές ονομάζονται **παραγοντικοί άξονες** οι οποίοι είναι ανά δύο ορθογώνιοι. Προβάλλοντας τα διανύσματα X_j^i που είχαν αρχικά για συντεταγμένες τους αριθμούς $(x_1^i,x_2^i...,x_p^i)$, έχουν πλέον σε κάθε άξονα Δ_a (για a=1,...p) καινούργιες συντεταγμένες τους αριθμούς $(c^1(i),.....c^p(i))$.

Δημιουργούνται λοιπόν καινούργιες τεχνητές μεταβλητές $C^1,C^2,....,C^p$ ασυσχέτιστες μεταξύ τους, που ονομάζονται **κύριες συνιστώσες**, οι οποίες συνθέτουν τις συσχετίσεις του συνόλου των αρχικών μεταβλητών.

Για κάθε κύρια συνιστώσα C^j (j=1,2,...,p) η συντεταγμένη $c^j(i)$ των στατιστικών μονάδων, αντίστοιχη στο κύριο άξονα Δ_j, είναι γραμμικός συνδυασμός των κανονικοποιημένων συντεταγμένων x_j^i.

Ήτοι $$C^j(i)=u_i^j x_1^i+u_2^j x_2^i+.....+u_p^j x_p^i \qquad (9.2)$$

Οι συντελεστές $(u_i^j, u_2^j,....u_p^j)$ αποτελούν τον j-ιοστό **κύριο παράγοντα u_j** (j=1.,,,p).

Η προσέγγιση μιας στατιστικής μονάδας I_i στο επίπεδο που σχηματίζουν δύο παραγοντικοί άξονες Δ_r και Δ_s μετράται με το συν$^2\theta$ της γωνίας που σχηματίζει το διάνυσμα X_i με τη προβολή του P(i) στο επίπεδο των παραγοντικών αξόνων.

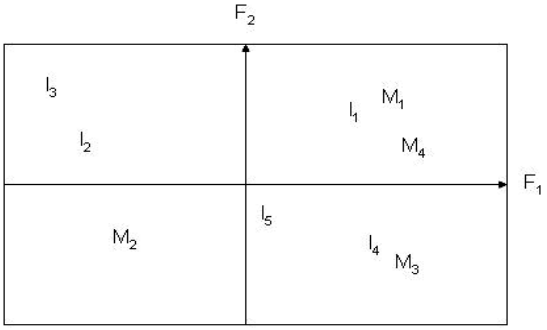
$$\text{συν}^2\theta = \frac{\|P(i)\|^2}{\|X(i)\|^2} \qquad (9.3)$$

όπου ||P(i)|| το μέτρο της προβολής του διανύσματος X(i).

Επειδή η καλύτερη απεικόνιση του νέφους πραγματοποιείται όταν προβάλλεται στο επίπεδο των δύο πρώτων παραγοντικών αξόνων, καθόσον οι χαρακτηριστικές ρίζες βαίνουν κατά φθίνουσα σειρά οι στατιστικές μονάδες που εμφανίζουν να έχουν την μεγαλύτερη επίδραση στο σχηματισμό του νέφους παρουσιάζουν την σημαντικότερη ποιότητα προβολής στο επίπεδο αυτό, με άλλα λόγια βρίσκονται πλησίον του κύριου παραγοντικού επιπέδου 1x2.

Τα σημεία αυτά εντοπίζονται εύκολα, αφού έχουν το υψηλότερο άθροισμα συν$^2\,\theta_1$+συν$^2\,\theta_2$ από τα υπόλοιπα σημεία [Cibois Ph,1983].

Στο παραγοντικό επίπεδο δύο αντικείμενα με υψηλή συσχέτιση μεταξύ των θα είναι πλησίον το ένα με το άλλο, λ.χ τα αντικείμενα I_2 και I_3 (σχήμα 9.2)



Σχήμα 9.2: Παραγοντικό επίπεδο 1x2

Πρέπει να τονιστεί ιδιαίτερα ότι στην ανάλυση σε Κύριες Συνιστώσες ενδιαφερόμαστε πρωτίστως για την συμπεριφορά των μεταβλητών, δηλαδή ποιες παρουσιάζουν τις πλέον υψηλές συντεταγμένες, καθόσον η συντεταγμένη μιας μεταβλητής στον παραγοντικό άξονα Δ_α είναι ο συντελεστής συσχέτισης της μεταβλητής αυτής με τον άξονα αυτόν.

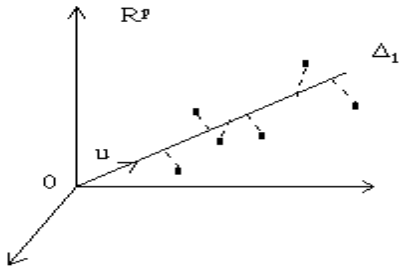
Έτσι θα ερμηνευτούν οι παραγοντικοί άξονες με βάση την συγκέντρωση ορισμένων τέτοιων μεταβλητών σε συνδυασμό με εκείνες που βρίσκονται σε αντίθεση.

9.2 Διαδικασία κατασκευής και ανάλυσης του νέφους των στατιστικών μονάδων N(I)

Έστω n στατιστικές μονάδες οι οποίες μελετώνται ως προς την συμπεριφορά p ποσοτικών μεταβλητών, δημιουργώντας τον πίνακα δεδομένων A(i,j) με i=1,...n και j=1,...p.

Η ανάλυση προβλέπει να εντοπιστεί αρχικά ο 1ος παραγοντικός άξονας του νέφους N(I).Ψάχνουμε δηλαδή να βρούμε μία ευθεία Δ_1 που να περνά από την αρχή των αξόνων O, κέντρο μάζας του νέφους των κανονικοποιημένων δεδομένων $\{X_1^i|i=1\ldots,n)$ και στη συνέχεια να υπολογιστεί η αδράνεια $I(N,\Delta_1)$ των ορθογώνιων προβολών $\{c^1(i)|i=1,...n\}$ των σημείων του νέφους πάνω στη ευθεία Δ_1 (σχ. 9.3) ως προς το κέντρο μάζας $G\{=O\}$.

Στη συνέχεια επαναλαμβάνουμε την ίδια διαδικασία προς εντοπισμό και των άλλων παραγοντικών αξόνων Δ_a των οποίων το πλήθος είναι ίσο με τη διάσταση p του χώρου των μεταβλητών V_j.



Σχήμα 9.3: Εντοπισμός του παραγοντικού άξονα Δ_1

Έτσι, λοιπόν, σε κάθε γραμμή i του πίνακα A(nxp) αντιστοιχούμε το διάνυσμα $A_j^i=(k_{i1},...k_{ip})$,όπου το k_{ij} παριστάνει την τιμή της μεταβλητής V_j για την i στατιστική μονάδα.

Στη συνέχεια τροποποιούμε τα διανύσματα A_j^i πραγματοποιώντας τον εξής μετασχηματισμό:

$$x_j^i = x_{ij} - \bar{k}_j \qquad (i=1,,n),(j=1..,p) \tag{9.4}$$

$$\bar{k}_j = \frac{1}{n}\sum_i k_j \tag{9.5}$$

Επομένως ο πίνακας A(nxp) μετατρέπεται σ' ένα πίνακα X(nxp) του οποίου οι γραμμές δημιουργούν ένα νέφος σημείων N(I) που ανήκουν στο χώρο R^p. Η μετρική που χρησιμοποιείται είναι η D_p (nxn) και δίνεται από την σχέση 9.6.

$$\underset{(nxn)}{Dp} = \begin{bmatrix} p_1 & & 0 \\ 0 & \ldots & 0 \\ 0 .. & p_i .. & .0 \\ & & \\ 0 . .. . & & 0 \\ 0 \ldots\ldots\ldots & & p_n \end{bmatrix} \tag{9.6}$$

όπου $$p_i = \frac{1}{n} \tag{9.7}$$

Ακολούθως υπολογίζουμε τον πίνακα V(pxp), ο οποίος αποτελεί τον πίνακα

διακυμάνσεων - συνδιακυμάνσεων των p μεταβλητών χρησιμοποιώντας την σχέση 9.8.

$$\underset{(pxp)}{V} = X' \cdot D_p \cdot X \qquad (9.8)$$

Μετά υπολογίζουμε τον πίνακα των συσχετίσεων μεταξύ των p μεταβλητών χρησιμοποιώντας τη σχέση 9.9.

$$\underset{(pxp)}{R} = D_{1/s} \cdot V \cdot D_{1/s} \qquad (9.9)$$

Ο πίνακας $D_{1/s}$ διαστάσεων pxp είναι διαγώνιος με στοιχεία τις τιμές $1/s_j$, όπου s_j είναι η τυπική απόκλιση των τιμών κάθε μεταβλητής j.

$$s_j = \sqrt{\frac{\sum\left(k_{ij} - \overline{k}_j\right)^2}{n}} \qquad (9.10)$$

Παρατήρηση: Το γενικό στοιχείο r_{ij} του πίνακα R είναι ο συντελεστής συσχέτισης μεταξύ των μεταβλητών i και j.

Διαγωνοποιώντας τον τετραγωνικό πίνακα R(pxp),βρίσκουμε τις χαρακτηριστικές ρίζες λ_a και τα αντίστοιχα χαρακτηριστικά διανύσματα U_a για a=1,...p.

Έχοντας βρει τα χαρακτηριστικά διανύσματα U_a έχουμε τη δυνατότητα να υπολογίσουμε τις κύριες συνιστώσες C^J (J=1,..p) των στατιστικών μονάδων σε κάθε παραγοντικού άξονα Δ_a, σύμφωνα με την διανυσματική σχέση (9.11).

$$\underset{(nxp)}{C^J} = \underset{(nxp)}{X} \cdot \underset{(pxp)}{U_a} \qquad (9.11)$$

όπου X(nxp) ο κανονικοποιημένος πίνακας του πίνακα A(nxp) και U_a είναι το χαρακτηριστικό διάνυσμα που συνδέεται με τον παραγοντικό άξονα Δ_a (a=1,..p).

Π.χ. οι συντεταγμένες της $1^{ης}$ κύριας συνιστώσας για κάθε στατιστική μονάδα i είναι ίση με

$$c^1(i) = \sum_{j=1}^{p} \frac{(x_j^i - \overline{x}_j)}{s_j\sqrt{n}} \cdot u_1^j \qquad (i = 1,...,n) \qquad (9.12)$$

Με βάση λοιπόν τη σχέση 9.11 ο αρχικός πίνακας δεδομένων A(nxp) μετασχηματίζεται σε πίνακα C(nxp), ο οποίος περιγράφει τις στατιστικές μονάδες σε σχέση με τις κύριες συνιστώσες.

Πίνακας 9.1:Ο πίνακας C(nxp):Οι κύριες συνιστώσες των n στατιστικών μονάδων

	C^1 .. C^2.... .C^J…..... .C^p
1	$c^1(1)$. $c^j(1)$ . .
.	. . .
i	$c^1(i)$..... $c^j(i)$…..$c^p(i)$
.	. . .
n	..... $c^j(n)$….....

Η διακύμανση της $1^{ης}$ κύριας συνιστώσας C^1 κατά μήκος του $1^{ου}$ παραγοντικού άξονα Δ_1 είναι ίση με την αδράνεια λ_1 των προβαλλόμενων σημείων του νέφους των κανονικοποιημένων δεδομένων πάνω στην ευθεία Δ_1 ως προς το κέντρο μάζας G={O}. Ήτοι

$$Var[C^1(i)]=\frac{1}{n}\sum_{i=1}^{n} [C^1(i)]^2]=\frac{1}{n}\sum_{i=1}^{n} d^2\ (C^1(i),O)=I(\{c^1(1),..c^1(n)\},O)=\lambda_1 \qquad (9.13)$$

Η ιδιότητα αυτή ισχύει για κάθε κύρια συνιστώσα.

Επομένως κάθε χαρακτηριστική ρίζα λ_a είναι η αδράνεια που ερμηνεύει ο αντίστοιχος παραγοντικός άξονας Δ_a.

Αν επιθυμούμε να προσδιορίσουμε το μέρος της ολικής αδράνειας (σ' αυτήν την περίπτωση και της διακύμανσης) που ερμηνεύεται από κάθε κύρια συνιστώσα (και κατ' επέκταση κάθε παραγοντικού άξονα Δ_a) χρησιμοποιούμε τη σχέση 9.14

$$\tau_\alpha = \frac{\lambda_\alpha}{\sum_\alpha \lambda_\alpha} \qquad (9.14)$$

Από τη σχέση 9.12 σε συνδυασμό με τη σχέση 9.13 έχουμε ότι η συνολική αδράνεια του νέφους I(N,G) είναι ίση με

$$I(N,G)=\sum_{a=1}^{p}\lambda_a = p \qquad (9.15)$$

Για κάθε κύρια συνιστώσα C^J (όπου J=1,..p) των στατιστικών μονάδων και σε κάθε παραγοντικό άξονα Δ_a ισχύουν οι παρακάτω σχέσεις:

$$\sum_{i=1}^{n} c^j(i)=0 \qquad (9.16) \qquad \frac{1}{n}\sum_{i=1}^{n}[c^j(i)]^2=\lambda_a \qquad (9.16a)$$

$$E[c^s(i),c^r(i)]=0 \ \text{ για } s<r \qquad (9.17)$$

Για να υπολογίσουμε την ποιότητα προβολής κάθε στατιστικής μονάδος i (i=1,...n) στον παραγοντικό άξονα Δ_a, χρησιμοποιούμε ως κριτήριο το τετράγωνο του συνημίτονου της γωνίας που σχηματίζει το διάνυσμα X_j^i με τον παραγοντικό άξονα Δ_a. (δηλαδή το $\cos^2(X_j^i,\Delta_a)$ το οποίο συμβολίζεται με (COS_a (i)). Ήτοι

$$COS_a(i) = \frac{\left[c^a(i)\right]^2}{\sum_a \left[c^a(i)\right]^2} \qquad (a=1,...p) \qquad (9.18)$$

Έτσι όσο το COS_a (i) πλησιάζει τη μονάδα τόσο καλύτερη ποιότητα προβολής της στατιστικής μονάδας έχουμε στον παραγοντικό άξονα Δ_a.

Η συμβολή κάθε στατιστικής μονάδος i (i=1,...n) στη διαμόρφωση του παραγοντικού άξονα Δ_a υπολογίζεται από την σχέση 9.19

$$CTR_a(i) = \frac{\left[c^a(i)\right]^2}{n\cdot\lambda_a} \qquad (a=1,...p) \qquad (9.19)$$

Ο συνδυασμός υψηλών τιμών $COS_a(i)$ και $CTR_a(i)$ για την i στατιστική μονάδα είναι ένδειξη πως συμβάλλει στη διαμόρφωση και ερμηνεία του άξονα Δ_a.

9.2.1 Η ανάλυση του νέφους N(J) των μεταβλητών

Οι μεταβλητές V_j μετά τη διαδικασία της κανονικοποίησης, στο χώρο των κυρίων συνιστωσών αποτελούν διανύσματα Y^j των οποίων οι συντεταγμένες είναι οι συντελεστές συσχέτισης των μεταβλητών αυτών με τις κύριες συνιστώσες C^J (J=1,....,p) [Tenenhaus M,1993,σ.125].

Οι συσχετίσεις των Y^j με κάθε συνιστώσα C^j υπολογίζεται βάσει της σχέσης 9.20.

$$COR(Y^j,C^a)=\sqrt{\lambda_a}\cdot U_a \quad \forall \ j=1,...p \text{ και } a=1,....p \qquad (9.20)$$

όπου U_a ο κύριος παράγοντας που συνδέεται με τον παραγοντικό άξονα Δ_a και λ_a η αντίστοιχη χαρακτηριστική ρίζα του παραγοντικού άξονα Δ_a.

Ο πίνακας 9.2 παρουσιάζει τις μεταβλητές Y^j στο χώρο R^p που δημιουργούν οι κύριες συνιστώσες C^j (j=1,...p).

Πίνακας 9.2:Συσχετίσεις μεταβλητών και κυρίων συνιστωσών

	C^1	C^2	.....	C^j	.....	C^p
Y^1				.		
.				.		
Y^k			.....	$COR(Y^k, C^j)$	.....	
.				.		
Y^p				.		

Με βάση τη σχέση 9.20 ισχύει

$$\sum_{j=1}^{p} COR_a^2(Y^j,C^a)=\lambda_a \qquad (9.21)$$

αφού ως γνωστόν

$$\sum_{r}^{p}(u_r^j)^2=1 \quad \text{για} \quad \text{κάθε} \quad j=1,....p \qquad (9.22)$$

Ισχύει επίσης

$$\sum_{a=1}^{p} COR_a^2(Y^j,C^a)=1 \qquad (9.23)$$

Τέλος επειδή η χαρακτηριστική ρίζα λ_a είναι η αδράνεια που εξηγείται από τον άξονα Δ_a, η συμβολή κάθε μεταβλητής στη διαμόρφωση του παραγοντικού άξονα Δ_a υπολογίζεται από την σχέση 9.24

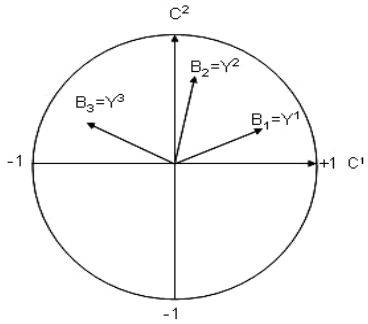
$$CTR_a(j)=\frac{COR_a^2(Y^j,C^j)}{\lambda_a} \quad \text{για } j=1,..,p \qquad (9.24)$$

9.3 Ο κύκλος συσχέτισης

Ορίζοντας με N(J) το νέφος των σημείων Y^j, αυτό βρίσκεται εντός υπερσφαίρας ακτίνας r=1,καθόσον η κάθε συντεταγμένη των σημείων αυτών είναι ένας συντελεστής συσχέτισης.

Συνεπώς όταν προβάλουμε το νέφος N(J) σε οποιοδήποτε παραγοντικό επίπεδο οι προβολές των σημείων θα βρίσκονται στο εσωτερικό ενός κύκλου (0,1),που καλείται **κύκλος συσχέτισης** (σχήμα 9.4).

Θεωρώντας, λοιπόν, το παραγοντικό επίπεδο 1x2,(αφού προσφέρει τη μεγαλύτερη πληροφορία) η κάθε μεταβλητή Y^j ορίζεται στο επίπεδο αυτό με τη βοήθεια του σημείου B_j=(COR(Y^j,C^1),COR(Y^j,C^2)).



Σχήμα 9.4: Ο κύκλος συσχέτισης

Το μήκος R_j των διανυσμάτων-μεταβλητών Bj αντιπροσωπεύει το συντελεστή πολλαπλής συσχέτισης R(Y^j,C^1,C^2) μεταξύ της μεταβλητής Y^j με τις δύο πρώτες κύριες συνιστώσες C^1,C^2.

Πράγματι

$$||B_j||^2=COR^2\ (Y^j,C^1)+COR^2\ (Y^j,C^2)=R(Y^j,C^1,C^2) \qquad (9.25)$$

αφού οι μεταβλητές C^1 και C^2 είναι ασυσχέτιστες μεταξύ τους.

Τα σημεία του νέφους N(J) όσο πλησιάζουν την περιφέρεια του κύκλου, τόσο καλύτερα αντιπροσωπεύονται στο παραγοντικό επίπεδο.

9.4 Τα παραγοντικά επίπεδα

Τα σημεία των νεφών N(I) και N(J) έχουμε τη δυνατότητα να τα προβάλουμε στο ίδιο παραγοντικό επίπεδο. Η ταυτόχρονη προβολή τους στο παραγοντικό επίπεδο $\Delta_r x \Delta_s$ (r=1,..p,s=2,...p με r<s) χρησιμεύει στην εξαγωγή των παρακάτω συμπερασμάτων:

Η προβολή των μεταβλητών στο επίπεδο επιτρέπει να χαρακτηρίσουμε τους παραγοντικούς άξονες του επιπέδου, δηλαδή να τους δώσουμε μία φυσική σημασία (ιδιότητα),ανάλογα με το ποιες μεταβλητές συσχετίζονται περισσότερο με τον καθένα χωριστά.

Στη συνέχεια εντοπίζονται οι μεταβλητές που παρουσιάζουν μεταξύ τους υψηλές συσχετίσεις. Ο εντοπισμός αυτός δίνει την δυνατότητα να παραλείψουμε ορισμένες απ' αυτές, κρατώντας για την ανάλυση τις υπόλοιπες q (q<p),αφού προκαλούν περίπου τις ίδιες συνέπειες στις στατιστικές μονάδες.

Έτσι οι q μεταβλητές θεωρούνται, η κάθε μια χωριστά, ως αντιπρόσωπος μιας ομάδας άλλων μεταβλητών που είναι έντονα συσχετισμένες με αυτές.

Με τον εντοπισμό, λοιπόν, των q μεταβλητών που συμβάλλουν περισσότερο στην ερμηνεία του υπό μελέτη φαινομένου, η ACP δίνει την δυνατότητα στον ερευνητή να αναπαραστήσει πλέον το νέφος N(I) στο χώρο R^q (χρησιμοποιώντας τις q πρώτες κύριες συνιστώσες).Ο μελετητής γνωρίζοντας το ποσοστό απώλειας της αδράνειας (πληροφορίας) που παρουσιάζει η ανάλυση με τις q μεταβλητές προχωρά στην ερμηνεία του φαινομένου, διατηρώντας ταυτόχρονα τις βασικές σχέσεις εξάρτησης που παρουσιάζουν τόσο οι στατιστικές μονάδες μεταξύ τους, όσο και οι μεταβλητές.

Η προβολή των στατιστικών μονάδων σε συνδυασμό με την ερμηνεία των παραγοντικών αξόνων επιτρέπει στον αναλυτή να εντοπίσει τις ιδιότητες των στατιστικών μονάδων και να τις ομαδοποιήσει κατά τέτοιο τρόπο, ώστε κάθε ομάδα να περιλαμβάνει στατιστικές μονάδες με παρεμφερή συμπεριφορά, ως προς τις q μεταβλητές.

Για την ταυτόχρονη παρουσίαση των νεφών N(I) και N(J) στο ίδιο παραγοντικό επίπεδο, πρέπει να ληφθούν υπόψη οι εξής παρατηρήσεις:

α) Οι συντεταγμένες των μεταβλητών επειδή προσδιορίζουν συντελεστές συσχέτισης η αριθμητική τιμή της κάθε μιας είναι μικρότερη της μονάδος. Για τον λόγο αυτό είναι σκόπιμο να εξετάζεται αρχικά το παραγοντικό επίπεδο μόνο με τις μεταβλητές.

β) Η ανάλυση του νέφους N(I) γίνεται σε σχέση με το κέντρο βάρους του νέφους, ενώ δε συμβαίνει το αυτό με την ανάλυση του νέφους N(J). Λόγω της ιδιαιτερότητας αυτής για να επιτύχουμε καλύτερη γραφική απεικόνιση των νεφών N(J) και N(I) στα διάφορα παραγοντικά επίπεδα, πολλαπλασιάζουμε τις συντεταγμένες του νέφους N(J) με τον αριθμό $\sqrt{n/p}$ ώστε να εξασφαλίσουμε στις προβολές των σημείων των δύο νεφών μία σχετική συμβατότητα.

Και αυτό επειδή η ανάλυση του νέφους N(I) γίνεται σε σχέση με το κέντρο βάρους του νέφους, ενώ δε συμβαίνει το ίδιο με την ανάλυση του νέφους N(J). Οι κύριες συνιστώσες των μεταβλητών είναι συντελεστές συσχέτισης, άρα πολύ πιθανόν όλες οι μεταβλητές να παρουσιάζουν θετική ή αρνητική συσχέτιση μ' ένα από τους άξονες, οπότε θα βρίσκονται προς την ίδια μεριά του άξονα και όχι εκατέρωθεν της αρχής των αξόνων όπως συμβαίνει με τις συνιστώσες των στατιστικών μονάδων.

γ) Το πρόσημο των συντεταγμένων των σημείων των δύο νεφών εξαρτάται από τον αλγόριθμο που χρησιμοποιείται για τη διαγωνοποίηση του πίνακα R (pxp).

δ) Με την ανάλυση σε κύριες συνιστώσες ερμηνεύονται οι γραμμικές σχέσεις μεταξύ των μεταβλητών. Έτσι ένας μικρός συντελεστής συσχέτισης μεταξύ δύο μεταβλητών παρέχει την ένδειξη ότι οι μεταβλητές αυτές είναι

ανεξάρτητες, μπορεί βέβαια να υφίσταται μεταξύ τους μία **μη γραμμική σχέση**, η οποία δεν αποκαλύπτεται στη παρούσα φάση της ανάλυσης.

9.4.1 Ερμηνεία του παραγοντικού επιπέδου

Οι γραφικές παραστάσεις μεταξύ γραμμών και στηλών ενός πίνακα δεδομένων πάνω στο παραγοντικό επίπεδο επιτρέπουν να οπτικοποιήσουμε τις προσεγγίσεις μεταξύ αντικειμένων (γραμμών) και μεταβλητών (στηλών) (σχήμα 9.2).

- Στον χώρο R^p των σημείων-αντικειμένων $\{I_k: (k=1,2,\ldots n)\}$ όταν δύο σημεία-αντικείμενα είναι γειτονικά, σημαίνει ότι τα δύο αυτά αντικείμενα χαρακτηρίζονται από τις ίδιες περίπου τιμές για κάθε μία από τις p μεταβλητές, οι δε αποστάσεις μεταξύ των σημείων υπολογίζονται με βάση την Ευκλείδεια απόσταση.
- Στον χώρο R^n των σημείων-μεταβλητών $\{J_m : (m=1,2,\ldots p)\}$ όταν δύο σημεία μεταβλητές είναι κοντά το ένα από το άλλο, αυτό σημαίνει πως το σύνολο των αντικειμένων έδωσε στις μεταβλητές αυτές περίπου τις ίδιες τιμές.
- Όταν ένα σημείο-αντικείμενο και ένα σημείο-μεταβλητή είναι γειτονικά σημαίνει ότι το αντικείμενο χαρακτηρίζεται από την ιδιότητα της μεταβλητής.
- Όταν δύο μεταβλητές βρίσκονται σε αντίθετα τεταρτημόρια του κύκλου συσχέτισης, σημαίνει ότι ο συντελεστής πολλαπλής συσχέτισης των δύο μεταβλητών είναι αρνητικός.
- Βέβαια οι μονάδες μέτρησης των μεταβλητών μπορεί να διαφέρουν μεταξύ τους, για τον λόγο αυτό είναι απαραίτητο να υπάρξουν μετατροπές των στοιχείων του πίνακα δεδομένων, χρησιμοποιώντας τον μετασχηματισμό z, καθιστώντας τα δεδομένα του πίνακα **κανονικοποιημένα** με βάση τη σχέση $z = (x_{ij} - \bar{x}) / s_i$

9.5 Ανασύσταση των αρχικών δεδομένων από τις q κύριες συνιστώσες

Αν λοιπόν οι q πρώτοι παραγοντικοί άξονες ερμηνεύουν ένα ποσοστό αδράνειας που ικανοποιεί τον αναλυτή, τότε μπορεί ο αναλυτής να διατηρήσει τις πρώτες q κύριες συνιστώσες, με τις οποίες μπορεί να ανακατασκευάσει προσεγγιστικά τα αρχικά δεδομένα.

Η προσέγγιση ισούται με το ποσοστό της αδράνειας των q διατηρηθέντων κυρίων παραγοντικών αξόνων. Έτσι αν επιθυμούμε να ανασυστήσουμε τις συσχετίσεις μεταξύ δύο μεταβλητών, χρησιμοποιούμε το εσωτερικό γινόμενο των διανυσμάτων-μεταβλητών B_i και B_j.' Ήτοι

$$COR(Y^i,Y^j)=\langle B_i,B_j\rangle \qquad (9.26)$$

π.χ αν θεωρήσουμε το παραγοντικό επίπεδο 1x2 θα έχουμε

$$\langle B_i,B_j\rangle=COR(Y^i,C^1)\cdot COR(Y^j,C^1)+\ldots\ldots\ldots+COR(Y^i,C^p)\cdot COR(Y^j,C^p)$$

Αν θελήσουμε να ανασυστήσουμε τις στατιστικές μονάδες οι οποίες περιγράφονται από τις κανονικοποιημένες μεταβλητές $\{X_j^i | i=1,..n$ και $j=1,..p\}$, για κάθε μετα-

βλητή Y^j χρησιμοποιούμε τη σχέση 9.27

$$X_j^i = \sum_{a=1}^{q} \frac{c^a(i)}{\sqrt{\lambda_a}} COR(Y^J, C^a) \qquad (9.27)$$

9.6 Συμπληρωματικές στατιστικές μονάδες

Θεωρούμε ως συμπληρωματική στατιστική μονάδα κάθε μονάδα I_s η οποία δεν συμμετείχε στην ανάλυση,(δηλαδή δεν συμπεριλαμβανόταν ως γραμμή του αρχικού πίνακα δεδομένων A(nxp)).

Επιθυμία μας είναι να προσδιορίσουμε τις ιδιότητες της I_s σε συνδυασμό με τις ιδιότητες των στατιστικών μονάδων που συμμετείχαν στην ανάλυση. Προς τούτο τοποθετούμε την I_s στα διάφορα παραγοντικά επίπεδα που δημιουργήθηκαν στα προηγούμενα στάδια της ανάλυσης ώστε να προσδιορίσουμε τη συμπεριφορά της. Προφανώς γνωρίζουμε το διάνυσμα

$$A(i_s)=\{V_1(i_s),.....,V_p(i_s)\} \qquad (9.28)$$

όπου $V_j(i_s)$ είναι η μέτρηση της μεταβλητής V_j στη στατιστική μονάδα i_s

Δημιουργώντας το κανονικοποιημένο διάνυσμα $X(i_s)=\{x_1(i_s),.,x_p(i_s)\}$ πολλαπλασιάζουμε τις συντεταγμένες του διανύσματος $X(i_s)$ με τους συντελεστές κάθε κύριου παράγοντα u^j και βρίσκουμε τις αντίστοιχες κύριες συνιστώσες $C^J(i_s)$ οι οποίες προσδιορίζουν τη θέση της στατιστικής μονάδας i_s στα παραγοντικά επίπεδα που δημιουργούνται από την ανάλυση [Lebart, Piron, Steines,2003].

9.7 Κανόνες διατηρητέων συνιστωσών

Ένα συνηθισμένο ερώτημα που τίθεται είναι το εξής: «Πόσες Κύριες Συνιστώσες (παράγοντες) οφείλει να διατηρήσει κάποιος σε μία ανάλυση;» Στο ερώτημα αυτό υπάρχουν ποικίλες απαντήσεις ανάλογα με τα κριτήρια που θα χρησιμοποιηθούν. Τα πλέον γνωστό είναι το κριτήριο του Kaiser (1960).

- **Το κριτήριο του Kaiser**

Όπως είναι γνωστό οι διακυμάνσεις των μεταβλητών σε μία Ανάλυση σε Κύριες Συνιστώσες θεωρούνται ίσες με την μονάδα. Εφόσον στην ανάλυση συμμετέχουν p μεταβλητές, η συνολική διακύμανση είναι ίση με p.

Η διακύμανση αυτή θα κατανεμηθεί σε διαφορετικές **κύριες συνιστώσες**. Το ποσοστό που ερμηνεύει κάθε κύρια συνιστώσα αντιπροσωπεύεται από την αντίστοιχη χαρακτηριστική ρίζα.

Το κριτήριο του Kaiser λέει ότι πρέπει να διατηρήσουμε στην ανάλυση όσες κύριες συνιστώσες έχουν χαρακτηριστική ρίζα μεγαλύτερη του 1.

Έστω ότι από η ανάλυση ενός πίνακα συσχετίσεων 4x4 μεταβλητών έδωσε τα παρακάτω αποτελέσματα.

Πίνακας 9.3:Πίνακας με τις χαρακτηριστικές ρίζες

αύξων	**χαρ. Ρίζα**	**ερμηνείας**	**άθροισμα**	**| ιστόγραμμα χαρακτηρ. ριζών**
1	2,1497646	53,74	53,74	|**********
2	1,3703070	34,26	88,00	|**********
3	0,4283667	10,71	98,71	|*********
4	0,0515616	1,29	100,00	|**

Στο παράδειγμά μας μπορούμε να κρατήσουμε τις δύο πρώτες κύριες συνιστώσες.

Παρατηρούμε αρχικά ότι σ' ένα πίνακα με 4 μεταβλητές, το άθροισμα των χαρακτηριστικών ριζών είναι ίσο με 4. Επίσης η πρώτη χαρακτηριστική ρίζα έχει τιμή 2,149, η δεύτερη έχει τιμή 1,37 και η τρίτη 0,428. Αυτό σημαίνει ότι η πρώτη κύρια συνιστώσα «εξηγεί» 2,149 μονάδες διακύμανσης, ενώ η δεύτερη 1,37 και η τρίτη μόνο 0,428 της διακύμανσης.

Συνεπώς είναι ανώφελο να χρησιμοποιηθούν περισσότερες των δύο πρώτων κυρίων συνιστωσών, αφού ερμηνεύουν το 88% της συνολικής διακύμανσης.

Από τον παρακάτω πίνακα προκύπτουν οι συμπιεσμένες κυρίες συνιστώσες με παραμέτρους τις τέσσερις μεταβλητές TEST1,TEST2,TEST3 και TEST4

Πίνακας 9.4:Συντεταγμένες, προβολές και συνεισφορές των μεταβλητών

IND	#G1	COR	CTR	#G2	COR	CTR	#G3	COR	CTR	#G4	COR	CTR
ΤΕΣΤ1	954	911	424	-210	44	32	138	19	45	160	26	499
ΤΕΣΤ2	-707	499	232	686	471	344	-94	9	20	144	21	404
ΤΕΣΤ3	-742	550	256	-450	202	147	497	247	577	32	1	20
ΤΕΣΤ4	-435	189	88	-809	654	477	-392	153	358	63	4	77

Έτσι για τις δύο πρώτες κύριες συνιστώσες προκύπτουν οι ακόλουθες συναρτήσεις

$$c_1 = 0,954 \cdot TEST_1 - 0,707 \cdot TEST_2 - 0,742 \cdot TEST_3 - 0,435 \cdot TEST_4$$
$$c_2 = -0,21 \cdot TEST_1 + 0,686 \cdot TEST_2 - 0,45 \cdot TEST_3 - 0,809 \cdot TEST_4$$

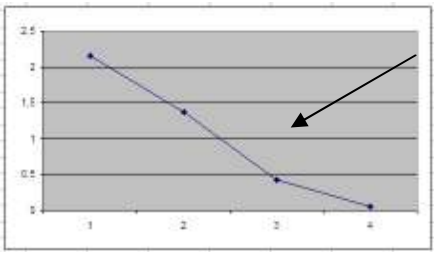
- **Το κριτήριο του Cattell**

Ο Cattell πρότεινε μία γραφική λύση και μία αλγεβρική.

- **Η γραφική λύση**

Με βάση τα δεδομένα του πίνακα 9.3 σχεδιάζουμε το γράφημα των χαρακτηριστικών ριζών σε συνάρτηση της τάξης των συνιστωσών ($1^η$,$2^η$,.. κ.λ.π)

Η ένδειξη μέχρι πόσες κύριες συνιστώσες θα διατηρήσουμε στην ανάλυση, αποτελεί το σημείο όπου εκδηλώνεται έντονη αλλαγή της κλίσης του γραφήματος.



Σχήμα 9.5: Γραφική λύση του κριτηρίου Cattell

- **Η αλγεβρική λύση**

Επειδή η γραφική λύση δεν είναι απολύτως καθαρή για τον προσδιορισμό της αλλαγής της κλίσης του γραφήματος των χαρακτηριστικών ριζών, χρησιμοποιείται ο επιταχυντής FA(λ), ο οποίος προσδιορίζει ακριβώς την συνιστώσα στη οποία αντιστοιχεί το επίμαχο σημείο αλλαγής της κλίσης.

Ο τύπος του επιταχυντή FA(λ) είναι ο εξής:

$$FA(\lambda_i) = \lambda_{i+1} - 2 \cdot \lambda_i + \lambda_{i-1} \qquad (9.29)$$

Ο επιταχυντής FA(λ), με την προϋπόθεση ότι ισχύει λ_{i-1} >1,00 , οδηγεί στο συμπέρασμα να κρατήσει ο αναλυτής τόσες κύριες συνιστώσες όσες υποδηλώνει το λ_{i-1} για το οποίο το FA(λ_i) είναι μέγιστο.

Έστω ο παρακάτω πίνακας ο οποίος περιέχει τεχνικά δεδομένα 15 πολυτελών αυτοκινήτων. Οι μεταβλητές είναι οι εξής: Κυβισμός: σε cm^3 , Βάρος: σε Kgr, Επιτάχυνση 0-100: σε Km/h, Τελική ταχύτητα : σε Km/h, Τιμή: σε Ευρώ

Πίνακας 9.5: Τεχνικά δεδομένα 15 αυτοκινήτων

Μάρκα	**Ind**	Κυβισμός	Βάρος	Επιτάχυνση 0-100	Ταχύτητα	Τιμή
Aston Martin V12	Asto	5935	1885	5,4	306	236.460
Bentley	Bent	6750	2585	5,8	273	263.068
BMW Z8	BMW	4941	1650	5,2	250	122.700
Jaguar XKR	Jagu	3996	1840	5,9	250	92.120
Maserati Spyder	Mase	4244	1730	5,2	283	98.800
Ferrari Engo	Fer1	5998	1365	3,3	350	645.000
Ferrari 612	Fer2	5748	1840	4,2	320	218.000
Ferrari F430	Fer3	4961	1640	4,2	309	148.480
Ford GT	Ford	5409	1578	3,3	320	177.000
Lamborghini Murci	Lam1	6192	1825	4,0	330	223.880
Lamborghini 575M	Lam2	5748	1795	4,7	325	196.000
Mercedes CLK AMG	Merc	5439	1748	4,1	320	236.060
Porsche GT2	Por1	3600	1440	4,0	319	184.674
Porsche Carrera 2.7	Por2	2687	1037	6,3	240	33.000
Saleen S7	Sale	6999	1250	3,6	350	366.000

Η ανάλυση του πίνακα με την ACP μας δίνει τις παρακάτω χαρακτηριστικές ρίζες.

Πίνακας 9.6: Ιστόγραμμα χαρακτηριστικών ριζών

αύξων	χαρ. Ρίζα	ερμηνείας	άθροισμα	| ιστόγραμμα χαρακτηρ. ριζών
1	2,9419517	58,84	58,84	|**********
2	1,3779707	27,56	86,40	|**********
3	0,4136925	8,27	94,67	|*****
4	0,1770378	3,54	98,21	|***
5	0,0893473	1,79	100,00	|*

Εφαρμόζοντας τον τύπο 9.29 προκύπτει ο πίνακας 9.7

Πίνακας 9.7: Εξαγωγή του επιταχυντή FA(λ) του Cattell

Κύρια Συνιστώσα	λ_i	%	FA(λ)
1	2,941952	58,84	
2	1,377971	86,40	0,599703
3	0,413693	94,67	**0,727624**
4	0,177038	98,21	0,148964
5	0,089347	100,00	

Από τα δεδομένα των πινάκων 9.6 και 9.7, συνάγεται ότι πρέπει να διατηρηθούν οι δύο πρώτες κύριες συνιστώσες καθόσον η μέγιστη τιμή $FA(\lambda_i)$=0,727624 αντιστοιχεί στην τιμή i=3, οπότε i-1=2, για την οποία ισχύει η προϋπόθεση $\lambda_{i-1}>1$ (λ_2=1,377971).

- **Το κριτήριο του Horn**

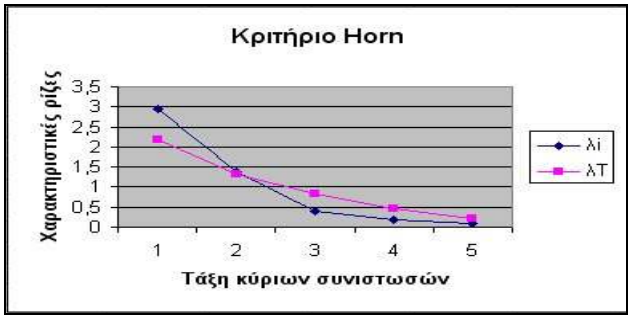
Ο **Horn** προτείνει να πραγματοποιηθεί μία Ανάλυση σε Κύριες Συνιστώσες σε μία μήτρα συσχετίσεων που δημιουργήθηκε από ένα πίνακα που έχει τον ίδιο αριθμό μεταβλητών και αντικειμένων που σχηματίστηκαν βάσει τυχαίων αριθμών. Οι κύριες συνιστώσες που θα προκύψουν από μία τέτοια ανάλυση, θα συγκριθούν με τις κύριες συνιστώσες του πραγματικού πίνακα δεδομένων. Σύμφωνα με τον **Horn** πρέπει να διατηρηθούν τόσες κύριες συνιστώσες όσες έχουν τιμή μεγαλύτερη από την αντίστοιχη κύρια συνιστώσα που προέκυψε με τυχαίο τρόπο.

Με βάση τα δεδομένα του πίνακα 9.5 δημιουργείται ένας πίνακας δεδομένων διαστάσεων 15x5 με τυχαίους αριθμούς, ο οποίος μετά από ανάλυση με την ACP έδωσε τις παρακάτω χαρακτηριστικές τιμές λ_T

Πίνακας 9.8: Χαρακτηριστικές τιμές κανονικές λ_i και τυχαίες λ_T

λ_i	λ_T
2,941952	2,167485
1,377971	1,325574
0,413693	0,830216
0,177038	0,458513
0,089347	0,218212

Το διάγραμμα των τιμών του πίνακα 9.8 παρουσιάζεται στο σχήμα 9.6



Σχήμα 9.6: Διάγραμμα χαρακτηριστικών ριζών κανονικού και τυχαίου δείγματος

Με βάση το κριτήριο Horn πρέπει να διατηρηθούν οι δύο πρώτες κύριες συνιστώσες.

9.8 Διαστήματα εμπιστοσύνης του Anderson

Ο **Anderson** το 1963 διατύπωσε τα όρια μέσα στα οποία πρέπει να κινούνται οι χαρακτηριστικές ρίζες ενός πίνακα δεδομένων που αναλύεται με την μέθοδο της Ανάλυσης σε Κύριες Συνιστώσες [Lebart,Piron,Steines, 2003].

Το εύρος των διαστημάτων εμπιστοσύνης παρέχει μία ένδειξη για το κατά πόσο μπορεί να είναι σταθερές οι τιμές των χαρακτηριστικών ριζών, λόγω των διακυμάνσεων που οφείλονται στις δειγματοληψίες.

Η επικάλυψη δύο διαστημάτων εμπιστοσύνης συνεχόμενων χαρακτηριστικών ριζών υποδηλώνει πιθανώς ισότητα των τιμών αυτών των χαρακτηριστικών ριζών. Το γεγονός αυτό μας απαλλάσσει να ερμηνεύσουμε έναν ασταθή παραγοντικό άξονα.

Σε επίπεδο σημαντικότητας 5% το διάστημα εμπιστοσύνης κάθε θεωρητικής χαρακτηριστικής ρίζας λ_θ (θ=1,….p) δίδεται για κάθε δειγματική χαρακτηριστική ρίζα λ_α από την σχέση (9.29).

$$\lambda_\alpha[1-1,96\sqrt{\frac{2}{n-1}}] \le \lambda_\theta \le \lambda_\alpha[1+1,96\sqrt{\frac{2}{n-1}}] \quad (9.29)$$

Παρατήρηση: Για να λειτουργήσει ικανοποιητικά η σχέση 9.29 πρέπει το δείγμα να είναι αρκετά μεγάλο (τουλάχιστον n>800) σε συνδυασμό με μεγάλο πλήθος μεταβλητών (p>30).

9.9 Ο Παραγοντικός δείκτης απλότητας (IFS) του H. F. Kaiser

Όσον αφορά την εξέταση κάθε μιας μεταβλητής χωριστά, αν πρέπει να συμμετέχει ή όχι στην ανάλυση, διευκολύνεται πάρα πολύ εφόσον χρησιμοποιηθεί το κριτήριο του Henry F. Kaiser (1974) το οποίο ονομάζεται παραγοντικός δείκτης απλότητας (Index of Factorial Simplicity IFSi).Ο παραγοντικός δείκτης απλότητας (IFS) υπολογίζεται βάσει του παρακάτω τύπου

$$IFS_i = \frac{p\left[\sum_{k=1}^{p} a_{ik}^4 - \left(\sum_{k=1}^{p} a_{ik}^2\right)^2\right]}{(p-1)\cdot\left(\sum_{k=1}^{p} a_{ik}^2\right)^2} \qquad (9.30)$$

Ο δείκτης αυτός υπολογίζεται για κάθε μία μεταβλητή και παίρνει τιμές από 0 έως 1. Για να διατηρηθεί μία μεταβλητή σε μία ACP πρέπει η μεταβλητή να έχει δείκτη με τιμή μεγαλύτερη του 0,5. H. F. Kaiser πρότεινε την παρακάτω διαβάθμιση

Πίνακας 9.10: Δείκτης IFS των επτά μεταβλητών

Πίνακας 9.9: Διαβαθμίσεις του δείκτη IFS	Χαρακτηρισμός
IFS<0,5	Απαράδεκτη
0,50<IFS≤0,60	Άθλια
0,60<IFS≤0,70	Μέτρια
0,70<IFS≤0,80	Ενδιάμεση
0,80<IFS≤0,90	Άξια
0,90<IFS	Εξαιρετική

Πίνακας 9.10: Δείκτης IFS των επτά μεταβλητών	Χαρακτηρισμός
Τιμή	0,78512
Ποιότητα	0,55894
Αλκοόλ	0,64103
Κύρος	0,74289
Χρώμα	0,59006
Άρωμα	0,79444
Γεύση	0,67012

9.10 Ο συντελεστής Alpha του Lee J. Cronbach

Ο συντελεστής Alpha του Lee J. Cronbach υπολογίζεται με βάση την παρακάτω σχέση:

$$a = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} s_i^2}{(\sum_{I=1}^{n} s_i)^2}\right) \qquad (9.31)$$

Όπου

n είναι το πλήθος των μεταβλητών Yi

s_i^2 η διακύμανση των τιμών κάθε μεταβλητής Y

s_i η τυπική απόκλιση κάθε μεταβλητής Yi

Το a του Cronbach μεταβάλλεται συνήθως μεταξύ 0≤a≤1

Όταν το a ισούται με 1 τότε υφίσταται μεγάλη εσωτερική συνέπεια στις απαντήσεις μεταξύ των μεταβλητών

Τιμές	**Χαρακτηρισμός**
a<0,50	Απαράδεκτη
0,50<a≤0,60	Φτωχή
0,60<a≤0,70	Αμφισβητήσιμη
0,70<a≤0,80	Αποδεκτή
0,80<a≤0,90	Καλή
0,90<a	Εξαιρετική

9.11 Test Bartlett

Το τεστ Bartlett ελέγχει αν ο πίνακας συσχέτισης $R=(\rho_{ij})$ διαστάσεων (pxp) αποκλίνει σημαντικά από τον πίνακα ταυτότητας του οποίου η τιμή της ορίζουσας είναι ίση με |R|=1

Ο έλεγχος που πραγματοποιείται με βάση τη κατανομή X^2 έχει την μορφή:

Ho: |R|=1

Hε: |R|<1

Χρησιμοποιείται ο παρακάτω τύπος:

$$x^2 = -\left(n - 1 - \frac{2 \cdot p + 5}{6}\right) \cdot \ln|R| \qquad (9.32)$$

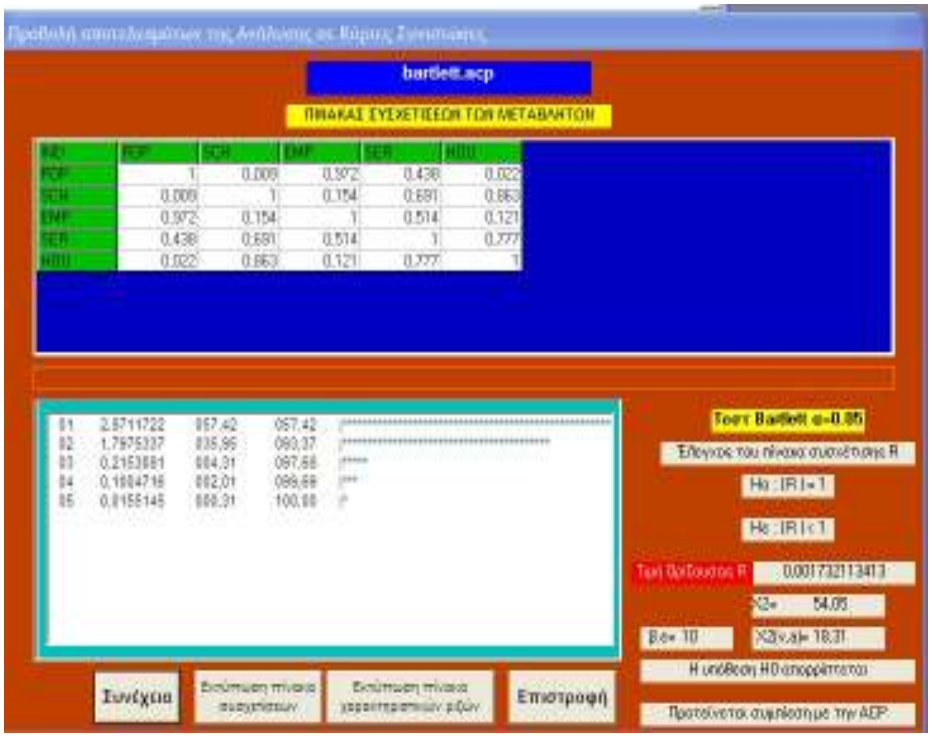
Παράδειγμα από οθόνη του MAD

Όταν ισχύει $X^2 > X^2(\nu, 0.05)$ με $\nu = \frac{p \times p}{2}$

Τότε είναι εφικτή η συμπίεση των p μεταβλητών σε m μεταβλητές όπου m=1,2,3…<p ανάλογα με το πόσοι παράγοντες ερμηνεύουν περισσότερο από μια διακύμανση.

Στην αντίθετη περίπτωση δεν κρίνεται απαραίτητο να γίνει συμπίεση των p μεταβλητών.

Στο παράδειγμα έχουμε 2 παράγοντες τον c1 με 2,87 και τον c2 με 1,79, όπου η πληροφορία που αντλείται από τους 2 πρώτους παράγοντες ανέρχεται σε 93,37% της συνολικής διακύμανσης.



Εικόνα 9.1: Οθόνη με το αποτέλεσμα του τεστ **Bartlett**