Πανεπιστήμιο Μακεδονίας
Τμήμα Εφαρμοσμένης Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών

ΠΟΛΥΔΙΑΣΤΑΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ, ΜΙΑ ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ, ΠΑΡΑΔΕΙΓΜΑΤΑ

Διπλωματική Εργασία

Τσιμινάκης Γεώργιος
Μεταπτυχιακός Φοιτητής

Κοζάνη, Νοέμβριος 2013

ΠΟΛΥΔΙΑΣΤΑΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ, ΜΙΑ ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ, ΠΑΡΑΔΕΙΓΜΑΤΑ

Τσιμινάκης Γεώργιος

Πτυχίο Μαθηματικών, ΑΠΘ, 2008

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπων Καθηγητής
Πετράκης Ανδρέας

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 05/11/2013

Πετράκης Ανδρέας	Μπίσμπας Αντώνιος	Συνδουκάς Δημήτριος
.....	.....	.....

Τσιμινάκης Γεώργιος

.....

Περίληψη

Η Ανάλυση δεδομένων είναι ένας νέος σχετικά κλάδος της στατιστικής επιστήμης, μέσα στο χώρο της «Πολυδιάστατης Στατιστικής Ανάλυσης», και γνώρισε ραγδαία εξέλιξη στη Γαλλία μετά το 1970. Παρά το μικρό χρόνο ανάπτυξής της, έχει γνωρίσει μεγάλη διάδοση και απαντάται σήμερα στη διάλεκτο όλων των επιστημονικών χώρων.

Ήταν η εποχή που εμφανίστηκε στο προσκήνιο των στατιστικών ο J.-P. Benzecri για να εκτιναχθεί η μέθοδος, να βγει από την αφάνεια, να αναγνωριστεί η χρησιμότητά της και να πάρει τη θέση που της αρμόζει. Η πρώτη ολοκληρωμένη παρουσίαση της νέας στατιστικής μεθόδου που ονόμασε Παραγοντική Ανάλυση των Αντιστοιχιών (Analyse Factorielle Correspodances) έγινε από τον ίδιο τον Benzecri το 1963 στο College de France.

Δύο από τις πιο γνωστές και διαδεδομένες μεθόδους της ανάλυσης δεδομένων είναι:

- Η Ανάλυση σε Κύριες Συνιστώσες
- Η Παραγοντική Ανάλυση των Αντιστοιχιών

Στην παρούσα διπλωματική αρχικά δώσαμε κάποιους ορισμούς σχετικούς με βασικές έννοιες όπως το μέτρο διανύσματος και το εσωτερικό γινόμενο και κάποιους ορισμούς από την θεωρία πινάκων. Στη συνέχεια αναφερθήκαμε στην πολυσυγγραμμικότητα και αναλύσαμε τις μεθόδους της ανάλυσης σε κύριες συνιστώσες και της παραγοντικής ανάλυσης. Τέλος παρουσιάσαμε ένα παράδειγμα χρησιμοποιώντας τις μεθόδους αυτές.

Λέξεις Κλειδιά: Πολυδιάστατη Στατιστική Ανάλυση Δεδομένων, Ανάλυση σε Κύριες Συνιστώσες, Παραγοντική Ανάλυση των Αντιστοιχιών

Abstract

Data analysis is a relatively new branch of statistics, that belongs to the field of “Multivariate Statistical Analysis” and its rapid evolution took place in France in the 1970s. Despite its little time of growth, it is widely spread and found in almost all scientific fields nowadays.

During that period, J.-P. Benzecri appeared at the forefront of statisticians and with his contribution the method got out of the shadows, its usefulness was recognized and got the place it deserves. The first comprehensive presentation of the new statistical method called Correspondence Factor Analysis was made by Benzecri in 1963 in College de France.

Two of the most known and widely used methods of data analysis are:

- Principal Components Analysis
- Correspondence Factor Analysis

In this project we presented some definitions related to basic concepts of vector analysis and some definitions related to matrix theory. Then we analyzed the Principal Components Analysis and the Correspondence Factor Analysis methods. Finally we presented an example using the methods above.

Keywords: Multivariate Statistical Data Analysis, Principal Components Analysis, Correspondence Factor Analysis

Περιεχόμενα

1. Βασικές Έννοιες

1.1 Μέτρο διανύσματος

Μέτρο διανύσματος είναι ένας αριθμός που χαρακτηρίζει ένα διάνυσμα ή ένα βαθμωτό μέγεθος.

Στην περίπτωση του διανύσματος δεν πρόκειται για τον μοναδικό αριθμό που το χαρακτηρίζει, καθώς υπάρχουν επίσης η διεύθυνση και η φορά (που μαζί αυτά τα δύο ονομάζονται κατεύθυνση) που χαρακτηρίζουν ένα διάνυσμα. Στην περίπτωση του βαθμωτού μεγέθους, όμως, είναι ο μοναδικός αριθμός που το χαρακτηρίζει.

Το μέτρο ή μήκος ενός διανύσματος $\overrightarrow{AB}$ συμβολίζεται με $\left|\overrightarrow{AB}\right|$ και ισούται με την τετραγωνική ρίζα του αθροίσματος των συντεταγμένων τους ως προς μια ορθοκανονική βάση. Πιο συγκεκριμένα, αν $\overrightarrow{AB} = a_1\hat{e}_1 + a_2\hat{e}_2 + ... + a_n\hat{e}_n$ όπου $a_1, a_2, ..., a_n$ οι συνιστώσες του διανύσματος και $\hat{e}_1, \hat{e}_2, ..., \hat{e}_n$ τα ορθομοναδιαία διανύσματα της βάσης, τότε το μέτρο του διανύσματος θα ισούται με

$$\left|\overrightarrow{AB}\right| = \sqrt{a_1^2 + a_2^2 + ... + a_n^2}$$

Τέλος, αν το διάνυσμα $\overrightarrow{AB}$ έχει μέτρο 1 τότε ονομάζεται μοναδιαίο διάνυσμα.[15]

1.2 Εσωτερικό γινόμενο διανυσμάτων, καθετότητα, παραλληλία

Το εσωτερικό γινόμενο δύο διανυσμάτων $\vec{a}$ και $\vec{b}$ συμβολίζεται με $\vec{a}\cdot\vec{b}$ και ορίζεται ως $\vec{a}\cdot\vec{b} = \left\|\vec{a}\right\|\left\|\vec{b}\right\|\cos\theta$ όπου με $\left\|\vec{a}\right\|$ και $\left\|\vec{b}\right\|$ συμβολίζουμε τα μέτρα των διανυσμάτων $\vec{a}$ και $\vec{b}$ αντίστοιχα και θ τη γωνία που σχηματίζεται ανάμεσα στα δύο διανύσματα. Το εσωτερικό γινόμενο είναι ουσιαστικά το γινόμενο του μέτρου του πρώτου διανύσματος με το μέτρο της προβολής του δεύτερου πάνω στο πρώτο. Είναι επίσης φανερό πως το εσωτερικό γινόμενο δύο διανυσμάτων θα είναι πάντα ένας αριθμός και όχι ένα νέο διάνυσμα όπως στην πρόσθεση και την αφαίρεση διανυσμάτων.

Όταν τα διανύσματα είναι κάθετα μεταξύ τους, το εσωτερικό γινόμενο είναι ίσο με το μηδέν, ενώ όταν είναι παράλληλα (ή αντιπαράλληλα) το εσωτερικό

γινόμενο ισούται με το θετικό (ή αρνητικό αντίστοιχα) γινόμενο των μέτρων τους. Αυτό είναι φανερό γιατί $\cos 90^\circ = 0$ (το δεύτερο διάνυσμα έχει μηδενική προβολή πάνω στο άλλο) και $\cos 0^\circ = 1$ ή $\cos 180^\circ = -1$ (το δεύτερο διάνυσμα προβάλλεται ολόκληρο).

Το εσωτερικό γινόμενο ορίζεται επίσης και ως το άθροισμα των γινομένων των επιμέρους συνιστωσών των διανυσμάτων. Συγκεκριμένα, αν τα διανύσματα $\vec{a}$ και $\vec{b}$ είναι n διαστάσεων γράφονται αναλυτικά $\vec{a} = a_1\hat{e}_1 + a_2\hat{e}_2 + ... + a_n\hat{e}_n$ και $\vec{b} = b_1\hat{e}_1 + b_2\hat{e}_2 + ... + b_n\hat{e}_n$ με $\hat{e}_1, \hat{e}_2, ..., \hat{e}_n$ να είναι τα μοναδιαία διανύσματα ορθοκανονικής βάσης ενός διανυσματικού χώρου n διαστάσεων, τότε το εσωτερικό γινόμενο γράφεται ως:

$\vec{a} \cdot \vec{b} = a_1 \cdot b_1 + a_2 \cdot b_2 + ... + a_n \cdot b_n$.

Για τυχόντα $\vec{a}, \vec{b}, \vec{c} \in V_n$ και τυχόν $\lambda \in R$ ισχύουν οι ιδιότητες:

1) $\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{a}$ (αντιμεταθετική ιδιότητα)
2) $\lambda\left(\vec{a} \cdot \vec{b}\right) = \left(\lambda\vec{a}\right) \cdot \vec{b} = \vec{a} \cdot \left(\lambda\vec{b}\right)$
3) $\vec{a} \cdot \left(\vec{b} + \vec{c}\right) = \vec{a} \cdot \vec{b} + \vec{a} \cdot \vec{c}$ [4],[15]

Απόδειξη

1) Θεωρούμε μια ορθομοναδιαία βάση $B = \{\vec{\varepsilon}_1, \vec{\varepsilon}_2, ..., \vec{\varepsilon}_n\}$ του V_n. Ας είναι $\left(a_1, a_2, ..., a_n\right), \left(b_1, b_2, ..., b_n\right)$ και $\left(c_1, c_2, ..., c_n\right)$ οι συντεταγμένες των διανυσμάτων $\vec{a}, \vec{b}$ και $\vec{c}$ ως προς τη βάση B αντιστοίχως.

Είναι τότε

$$\begin{aligned}\vec{a} \cdot \vec{b} &= a_1 \cdot b_1 + a_2 \cdot b_2 + ... + a_n \cdot b_n \\ &= b_1 \cdot a_1 + b_2 \cdot a_2 + ... + b_n \cdot a_n \\ &= \vec{b} \cdot \vec{a}\end{aligned}$$

Αποδείξαμε την ισότητα 1.

2) Για να αποδείξουμε την 2 παρατηρούμε ότι οι συντεταγμένες των $\lambda\vec{a}, \lambda\vec{b}$ είναι αντιστοίχως $(\lambda a_1, \lambda a_2, ..., \lambda a_n), (\lambda b_1, \lambda b_2, ..., \lambda b_n)$.

Είναι τότε

$$\begin{aligned}\left(\lambda\vec{a}\right)\cdot\vec{b} &= (\lambda a_1)b_1 + (\lambda a_2)b_2 + ... + (\lambda a_n)b_n \\ &= \lambda\left(a_1b_1 + a_2b_2 + ... + a_nb_n\right) \\ &= \lambda\left(\vec{a}\cdot\vec{b}\right)\end{aligned}$$

Όμοια προκύπτει $\vec{a}\cdot\left(\lambda\vec{b}\right) = \lambda\left(\vec{a}\cdot\vec{b}\right)$

3) Τέλος έχουμε

$$\begin{aligned}\vec{a}\cdot\left(\vec{b}+\vec{c}\right) &= a_1(b_1+c_1) + a_2(b_2+c_2) + ... + a_n(b_n+c_n) \\ &= \left(a_1b_1 + a_2b_2 + ... + a_nb_n\right) + \left(a_1c_1 + a_2c_2 + ... + a_nc_n\right) \\ &= \vec{a}\cdot\vec{b} + \vec{a}\cdot\vec{c}\end{aligned}$$

1.3 Εξωτερικό γινόμενο δύο διανυσμάτων

Το εξωτερικό γινόμενο δύο διανυσμάτων είναι ουσιαστικά μια καινούρια πράξη μεταξύ διανυσμάτων. Διαφέρει από το εσωτερικό γινόμενο όσον αφορά τα αποτελέσματα των πράξεων. Στο μεν εσωτερικό γινόμενο λαμβάνουμε ένα βαθμωτό μέγεθος ως αποτέλεσμα, ενώ στο εξωτερικό παίρνουμε ένα διάνυσμα. Παράδειγμα εξωτερικού γινομένου είναι η ροπή δύναμης.

Το εξωτερικό γινόμενο δεν γενικεύεται για n διαστάσεις, έχει νόημα μόνο σε τρισδιάστατους χώρους. Αν $\vec{a}$ και $\vec{b}$ είναι τα δύο διανύσματα, το εξωτερικό γινόμενο συμβολίζεται ως $\vec{a}\times\vec{b}$ και ορίζεται ως $\vec{a}\times\vec{b} = \left\|\vec{a}\right\|\left\|\vec{b}\right\|\sin\theta\hat{n}$ όπου $\left\|\vec{a}\right\|\left\|\vec{b}\right\|\sin\theta$ το μέτρο και $\hat{n}$ η κατεύθυνση, $\left\|\vec{a}\right\|$ και $\left\|\vec{b}\right\|$ είναι τα μέτρα των διανυσμάτων $\vec{a}$ και $\vec{b}$, $\sin\theta$ η καθετότητα αυτών των διανυσμάτων και $\hat{n}$ είναι το μοναδιαίο διάνυσμα κάθετο στο επίπεδο που ορίζουν τα $\vec{a}$ και $\vec{b}$ έτσι ώστε $\left[\vec{a},\vec{b},\vec{n}\right] > 0$.

Ισχύει ότι $\vec{b}\times\vec{a}=-\vec{a}\times\vec{b}$, δηλαδή στην πράξη του εξωτερικού γινομένου δεν ισχύει η αντιμεταθετική ιδιότητα. Τα φυσικά μεγέθη (διανύσματα) που παράγουν το νέο διάνυσμα γράφονται με συγκεκριμένη σειρά και η κατεύθυνση του εξωτερικού γινομένου προκύπτει με τον κανόνα του δεξιού χεριού. Για παράλληλα ή αντιπαράλληλα διανύσματα το εξωτερικό γινόμενο δίνει το μηδενικό διάνυσμα, εφόσον $\sin 0^\circ = 0$ και $\sin 180^\circ = 0$.

Για τα μοναδιαία διανύσματα του Καρτεσιανού συστήματος αξόνων ισχύει

$$i\times j = k$$
$$j\times k = i$$
$$k\times i = j$$

Για το εξωτερικό γινόμενο ισχύουν οι εξής ιδιότητες:

1) $\vec{a}\times\vec{b}=-\vec{b}\times\vec{a}$
2) $\vec{a}\times\left(\vec{b}+\vec{c}\right)=\vec{a}\times\vec{b}+\vec{a}\times\vec{c}$ άρα και $\left(\vec{b}+\vec{c}\right)\times\vec{a}=\vec{b}\times\vec{a}+\vec{c}\times\vec{a}$
3) $\left(\lambda\vec{a}\right)\times\vec{b}=\lambda\left(\vec{a}\times\vec{b}\right)=\vec{a}\times\left(\lambda\vec{b}\right)$, για κάθε $\lambda\in R$
4) $\left(\vec{a}\times\vec{b}\right)\times\vec{c}=\left(\vec{a}\cdot\vec{c}\right)\vec{b}-\left(\vec{b}\cdot\vec{c}\right)\vec{a}$ και $\vec{a}\times\left(\vec{b}\times\vec{c}\right)=\left(\vec{a}\cdot\vec{c}\right)\vec{b}-\left(\vec{a}\cdot\vec{b}\right)\vec{c}$ [4],[15]

Απόδειξη

1) Έστω ότι τα δύο διανύσματα είναι μη μηδενικά. Ισχύει ότι $\left|\vec{a}\times\vec{b}\right|=\left|\vec{a}\right|\left|\vec{b}\right|\sin\left(\vec{a},\vec{b}\right)=\left|\vec{b}\right|\left|\vec{a}\right|\sin\left(\vec{b},\vec{a}\right)=\left|\vec{b}\times\vec{a}\right|$. Η διεύθυνση των δύο είναι η ίδια, κάθετα πάνω στο επίπεδο των $\vec{a},\vec{b}$. Σύμφωνα με τον κανόνα του δεξιού χεριού τα δύο εξωτερικά γινόμενα είναι αντίρροπα. Επομένως, τα δύο εξωτερικά γινόμενα είναι αντίθετα. Στην περίπτωση που ένα από τα δύο είναι το μηδενικό διάνυσμα η σχέση γίνεται $\vec{0}=-\vec{0}$ που ισχύει.

3) Πρώτα σημειώνουμε ότι το διάνυσμα $\left(\vec{a}\times\vec{b}\right)\times\vec{c}$ είναι κάθετο προς τα διανύσματα $\vec{a}\times\vec{b}$ και $\vec{c}$, οπότε θα είναι συνεπίπεδο προς τα $\vec{a},\vec{b}$ και συνεπώς μπορεί να γραφεί

ως γραμμικός συνδυασμός των $\vec{a}$ και $\vec{b}$. Επομένως υπάρχουν $\lambda, \mu \in \mathbb{R}$ τέτοια ώστε $\left(\vec{a}\times\vec{b}\right)\times\vec{c} = \lambda\vec{a} + \mu\vec{b}$.

Αν $\vec{a} = \left(a_1, a_2, a_3\right)$ και $\vec{b} = \left(b_1, b_2, b_3\right)$, τότε με απευθείας υπολογισμό βρίσκουμε ότι: $\lambda = -\left(\vec{b}\cdot\vec{c}\right)$ και $\mu = \left(\vec{a}\cdot\vec{c}\right)$. Πράγματι έχουμε

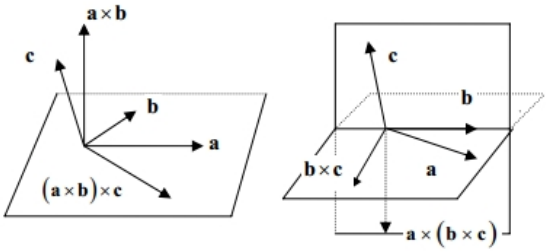
$$\left(\vec{a}\times\vec{b}\right)\times\vec{c} = \begin{vmatrix} i & j & k \\ a_2b_3 - a_3b_2 & a_3b_1 - a_1b_3 & a_1b_2 - a_2b_1 \\ c_1 & c_2 & c_3 \end{vmatrix}$$

ή μετά από πράξεις

$$\left(\vec{a}\times\vec{b}\right)\times\vec{c} = \left(a_1c_1 + a_2c_2 + a_3c_3\right)\vec{b} - \left(b_1c_1 + b_2c_2 + b_3c_3\right)\vec{a} = \left(\vec{a}\cdot\vec{c}\right)\vec{b} - \left(\vec{b}\cdot\vec{c}\right)\vec{a}$$

ή ισοδύναμα $\left(\vec{a}\times\vec{b}\right)\times\vec{c} = \left(\vec{a}\cdot\vec{c}\right)\vec{b} - \left(\vec{b}\cdot\vec{c}\right)\vec{a}$.

Τέλος $\vec{a}\times\left(\vec{b}\times\vec{c}\right) = -\left[\left(\vec{b}\times\vec{c}\right)\times\vec{a}\right] = -\left[\left(\vec{b}\cdot\vec{a}\right)\vec{c} - \left(\vec{c}\cdot\vec{a}\right)\vec{b}\right] = \left(\vec{a}\cdot\vec{c}\right)\vec{b} - \left(\vec{a}\cdot\vec{b}\right)\vec{c}$



1.4 Διανυσματικός χώρος με νορμ (norm)

Θεωρούμε τη γενική περίπτωση ενός διανυσματικού χώρου V πάνω στο σώμα K.

Αν τώρα ορίσουμε επί του V την απεικόνιση

$\|\cdot\| : V \to \mathbb{R}$ με τύπο $x \to \|x\|$,

όπου για κάθε $\lambda \in K$, $x, y \in V$ ισχύουν οι ιδιότητες

$\left(N_1\right)$ $\|x\| \geq 0$ και $\|x\| = 0 \Leftrightarrow x = 0$

$\left(N_2\right)$ $\|\lambda x\| = |\lambda|\|x\|$

(N_3) $\|x+y\| \le \|x\| + \|y\|$,

τότε ο διανυσματικός χώρος V λέγεται χώρος με νορμ (norm), η δε συνάρτηση $\|\cdot\|$ λέγεται συνάρτηση μέτρου επί του V. Όπως γνωρίζουμε το μήκος ενός διανύσματος a το συνδέουμε με το εσωτερικό γινόμενο μέσω της σχέσης $\|a\| = \sqrt{a \cdot a}$.

Αυτό που κάνουμε είναι να ορίσουμε τη συνάρτηση νορμ επί του Δ_3 ως εξής:

$$\|\cdot\| : \Delta_3 \to \mathbb{R} \text{ με τύπο } x \to \|x\| := \sqrt{x \cdot x}. \qquad (1)$$

(Το σύνολο των διανυσμάτων του χώρου χωρίζεται σε κλάσεις ισοδυναμίας: κάθε κλάση ισοδυναμίας περιέχει όλα τα διανύσματα τα οποία είναι ίσα μεταξύ τους και μόνον αυτά. Ο «καλύτερος» αντιπρόσωπος (διάνυσμα) από μια κλάση ισοδυναμίας είναι (πολλές φορές) εκείνο το διάνυσμα που έχει αρχή το σημείο αναφοράς Ο. Η ένωση όλων των κλάσεων ισοδυναμίας μας δίνει το σύνολο όλων των διανυσμάτων. Κάθε κλάση ισοδυναμίας ονομάζεται ελεύθερο διάνυσμα και το σύνολο όλων των κλάσεων ισοδυναμίας είναι ο χώρος των ελεύθερων διανυσμάτων Δ_3).

Δηλαδή αυτό που κάνει η συνάρτηση νορμ είναι να αντιστοιχεί σε κάθε διάνυσμα το μήκος του. Όπως έχουμε αναφέρει ένα διάνυσμα x με $\|x\| = 1$, μήκος 1, λέγεται μοναδιαίο διάνυσμα και ότι το μήκος διανύσματος ικανοποιεί τις ιδιότητες $(N_1) - (N_3)$. Με άλλα λόγια, το εσωτερικό γινόμενο ορίζει μια νορμ στον αντίστοιχο διανυσματικό χώρο, μέσω του τύπου (1), και η ιδέα αυτή μπορεί να γενικευτεί σε κάθε χώρο V, ο οποίος είναι εφοδιασμένος με ένα εσωτερικό γινόμενο.

Πραγματικά, για κάθε διάνυσμα $x \in V$, ισχύει $x \cdot x \ge 0$ και επομένως υπάρχει η $\sqrt{x \cdot x}$. Εύκολα αποδεικνύεται η ιδιότητα (N_1), ενώ για την (N_2) έχουμε

$$\|\lambda x\| = \sqrt{(\lambda x) \cdot \lambda x} = \sqrt{\lambda \overline{\lambda}(x \cdot x)} = \sqrt{|\lambda|^2}\sqrt{x \cdot x} = |\lambda|\|x\|.$$

Μπορούμε τώρα να χρησιμοποιήσουμε την (N_2) για να διαπιστώσουμε ότι για κάθε διάνυσμα $x \ne 0 \in V$ το $\frac{1}{\|x\|}x = \hat{x}$ είναι ένα μοναδιαίο διάνυσμα στη

διεύθυνση του x. Στην περίπτωση αυτή λέμε ότι το διάνυσμα έχει κανονικοποιηθεί. Πράγματι έχουμε $\left\| \frac{1}{\|x\|} x \right\| = \left| \frac{1}{\|x\|} \right| \|x\| = \frac{1}{\|x\|} \|x\| = 1$. [5]

1.5 Απόσταση στο χώρο n διαστάσεων

Έστω τα $\mathrm{A}(a_1, a_2, ..., a_n)$ και $\mathrm{B}(b_1, b_2, ..., b_n)$ είναι δύο σημεία στον n-διάστατο χώρο. Τότε η απόσταση d μεταξύ τους είναι η νόρμα του διανύσματος $\overrightarrow{\mathrm{AB}}$. Εφόσον $\overrightarrow{\mathrm{AB}} = (b_1 - a_1, b_2 - a_2, ..., b_n - a_n)$ παίρνουμε ότι

$d(\mathrm{A}, \mathrm{B}) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + ... + (b_n - a_n)^2}$. [4]

1.5.1 Μέσο του ΑΒ

Το μέσο των σημείων $\mathrm{A}(a_1, a_2, ..., a_n)$ και $\mathrm{B}(b_1, b_2, ..., b_n)$ στον n-διάστατο έχει συντεταγμένες:

$\left(\frac{a_1 + b_1}{2}, \frac{a_2 + b_2}{2}, \frac{a_3 + b_3}{2}, ..., \frac{a_n + b_n}{2} \right)$. [4]

1.6 Κέντρο βάρους

Για να είναι δυνατή η άμεση σύγκριση μεταξύ των γραμμών ή των στηλών ενός πίνακα συμπτώσεων, πρέπει να ληφθούν υπόψη τα αντίστοιχα αθροίσματα γραμμών και στηλών. Στο πλαίσιο της Παραγοντικής Ανάλυσης των Αντιστοιχιών, οι απόλυτες συχνότητες στα κελία του πίνακα συμπτώσεων μετασχηματίζονται σε ποσοστά των αντίστοιχων αθροισμάτων γραμμών και στηλών.

Το προφίλ μιας γραμμής i αποτελείται από τα στοιχεία:

$$\left\{ \frac{n_{i1}}{n_{i+}}, \frac{n_{i2}}{n_{i+}}, ..., \frac{n_{il}}{n_{i+}} \right\}, \text{ με } n_{i+} = \sum_j n_{ij} .$$

Ανάλογα ορίζεται και το προφίλ της j στήλης που είναι το σύνολο με στοιχεία:

$$\left\{ \frac{n_{1j}}{n_{+j}}, \frac{n_{2j}}{n_{+j}}, ..., \frac{n_{kj}}{n_{+j}} \right\}, \text{ με } n_{i+} = \sum_j n_{ij} .$$

Έτσι, κατασκευάζονται δύο διαφορετικοί πίνακες σχετικών συχνοτήτων, ο ένας με τα προφίλ των γραμμών (Πίνακας Α) και ο άλλος με τα προφίλ των στηλών (Πίνακας Β). Το άθροισμα κάθε γραμμής του πίνακα Α και κάθε στήλης του πίνακα Β είναι ίσο με τη μονάδα. Η περιθώρια γραμμή του πίνακα Α ονομάζεται «μέσο προφίλ» των γραμμών ή αλλιώς «κέντρο βάρους» των γραμμών. Αντίστοιχα, η περιθώρια γραμμή του πίνακα Β ορίζεται ως το μέσο προφίλ των στηλών ή διαφορετικά το κέντρο βάρους των στηλών. [9]

Ο πίνακας Α με τα προφίλ των Γραμμών (στις γραμμές)

Κλάσεις της μεταβλητής Y							
Κλάσεις της μεταβλητής X	1	2	…	j	…	l	Άθροισμα
1	$\frac{n_{11}}{n_{1+}}$	$\frac{n_{12}}{n_{1+}}$	…	$\frac{n_{1j}}{n_{1+}}$	…	$\frac{n_{1l}}{n_{1+}}$	1
2	$\frac{n_{21}}{n_{2+}}$	$\frac{n_{22}}{n_{2+}}$	…	$\frac{n_{2j}}{n_{2+}}$	…	$\frac{n_{2l}}{n_{2+}}$	1
…	…	…	…	…	…	…	…
i	$\frac{n_{i1}}{n_{i+}}$	$\frac{n_{i2}}{n_{i+}}$	…	$\frac{n_{ij}}{n_{i+}}$	…	$\frac{n_{il}}{n_{i+}}$	1
…	…	…	…	…	…	…	…
k	$\frac{n_{k1}}{n_{k+}}$	$\frac{n_{k2}}{n_{k+}}$	…	$\frac{n_{kj}}{n_{k+}}$	…	$\frac{n_{kl}}{n_{k+}}$	1
Μέσο προφίλ γραμμών Κέντρο βάρους γραμμών	$\frac{n_{+1}}{n}$	$\frac{n_{+2}}{n}$	…	$\frac{n_{+j}}{n}$	…	$\frac{n_{+l}}{n}$	

Ο πίνακας Β με τα προφίλ των Στηλών (στις στήλες)

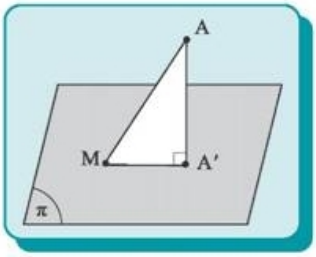
Κλάσεις της μεταβλητής Y							
Κλάσεις της μεταβλητής X	1	2	…	j	…	l	Μέσο προφίλ στηλών ή Κέντρο βάρους στηλών
1	$\frac{n_{11}}{n_{+1}}$	$\frac{n_{12}}{n_{+2}}$	…	$\frac{n_{1j}}{n_{+j}}$	…	$\frac{n_{1l}}{n_{+l}}$	$\frac{n_{1+}}{n}$
2	$\frac{n_{21}}{n_{+1}}$	$\frac{n_{22}}{n_{+2}}$	…	$\frac{n_{2j}}{n_{+j}}$	…	$\frac{n_{2l}}{n_{+l}}$	$\frac{n_{2+}}{n}$
…	…	…	…	…	…	…	…
i	$\frac{n_{i1}}{n_{+1}}$	$\frac{n_{i2}}{n_{+2}}$	…	$\frac{n_{ij}}{n_{+j}}$	…	$\frac{n_{il}}{n_{+l}}$	$\frac{n_{i+}}{n}$
…	…	…	…	…	…	…	…

k	$\frac{n_{k1}}{n_{+1}}$	$\frac{n_{k2}}{n_{+2}}$	…	$\frac{n_{kj}}{n_{+j}}$	…	$\frac{n_{kl}}{n_{+l}}$	$\frac{n_{k+}}{n}$
Άθροισμα	1	1	…	1	…	1	

1.7 Προβολή σημείου στο επίπεδο

Ορισμός

Ορθή προβολή ή προβολή A' σημείου A στο επίπεδο π λέγεται το σημείο τομής του επιπέδου π με την κάθετο από το στο A επίπεδο π.



Αν το σημείο A βρίσκεται εκτός επιπέδου π, A' είναι η προβολή του A στο π και M τυχαίο σημείο του π, τότε από το ορθογώνιο τρίγωνο $\mathrm{AA'M}$ προκύπτει ότι η κάθετη πλευρά AA' είναι μικρότερη από την υποτείνουσα AM. Δηλαδή, το τμήμα AA' είναι το μικρότερο από τα τμήματα με αρχή το σημείο A και τέλος το τυχαίο σημείο M του επιπέδου π. [6]

1.8 Απλός λόγος τριών σημείων

Θεωρούμε τρία σημεία P_1, P_2, P πάνω σε μία ευθεία. Τα διανύσματα $\overrightarrow{P_1P}$ και $\overrightarrow{PP_2}$ είναι συγγραμμικά και εάν $P \neq P_2$ υπάρχει ένας αριθμός μ τέτοιος ώστε $\overrightarrow{P_1P} = \mu\overrightarrow{PP_2}$.

Ο αριθμός μ ονομάζεται απλός λόγος των τριών σημείων, και συμβολίζεται (P_1P_2P). Παρατηρούμε ότι $(P_1P_2P) = \frac{\overrightarrow{P_1P}}{\overrightarrow{PP_2}}$, και ότι η αλλαγή του προσανατολισμού της ευθείας δεν επηρεάζει τον απλό λόγο. Εάν θεωρήσουμε τα P_1, P_2 σταθερά και το P να κινείται πάνω στην ευθεία, η τιμή του απλού λόγου μεταβάλλεται όπως στο παρακάτω σχήμα 1.

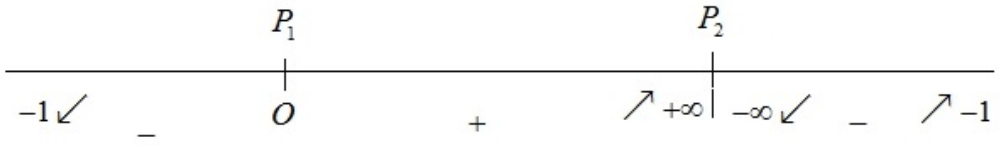
Εάν τα σημεία έχουν συντεταγμένες $P_1(x_1, y_1), P_2(x_2, y_2)$ και $P(x, y)$, τότε

$$x - x_1 = \mu(x_2 - x) \text{ και } y - y_1 = \mu(y_2 - y),$$

απ' όπου προκύπτουν οι τύποι

$$x = \frac{x_1 + \mu x_2}{1 + \mu} \text{ και } y = \frac{y_1 + \mu y_2}{1 + \mu}. \quad (1)$$

Όταν $\mu > 0$, το P βρίσκεται μεταξύ των P_1 και P_2, και οι τύποι δίδουν τις συντεταγμένες του σημείου που χωρίζει το διάστημα P_1P_2 σε δύο τμήματα με λόγο $\mu:1$. [14]



Σχήμα 1 : Η τιμή του απλού λόγου καθώς το σημείο P κινείται πάνω στην ευθεία P_1P_2.

2. Πίνακες

2.1 Εισαγωγή

Είναι γνωστό ότι πίνακας A είναι μια ορθογώνια διάταξη από στοιχεία $a_{ij}(i = 1, 2, ..., n, j = 1, 2, ..., m)$ ενός σώματος K ($K = \mathbb{R}$ ή $\mathbb{C}$), δηλαδή

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}.$$

Λέμε ότι ο παραπάνω πίνακας είναι διάστασης $n \times m$, ή τύπου $n \times m$, δηλαδή έχει n γραμμές και m στήλες.

Συμβολίζουμε με $M_{n,m}(K)$ το σύνολο των $n \times m$ πινάκων με στοιχεία από το σώμα K. Το σύνολο $M_{n,n}(K)$ το γράφουμε συντμημένα ως $M_n(K)$.

Θεωρούμε γνωστούς τους ορισμούς των πράξεων της πρόσθεσης και του πολλαπλασιασμού πινάκων και σημειώνουμε ότι

1) Αν $A, B \in M_{n,m}(K)$ και $C \in M_{m,k}(K)$ τότε $(A+B)C = AC + BC$

2) Αν $A \in M_{n,m}(K)$, $B \in M_{m,k}(K)$ και $C \in M_{k,r}(K)$, τότε $A(BC) = (AB)C$

3) Αν $A \in M_{n,m}(K)$, τότε υπάρχει ένας πίνακας $O \in M_{n,1}(K)$ τέτοιος ώστε

$$AO = 0$$

Ο πίνακας O έχει όλα τα στοιχεία του μηδέν.

4) Αν $A \in M_n(K)$ τότε το μοναδιαίο στοιχείο του πολλαπλασιασμού πάντα υπάρχει, το συμβολίζουμε με I και είναι ένας $n \times n$ της μορφής

$$I = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Ο ορισμός επεκτείνεται κατά προφανή τρόπο και για μη πεπερασμένο n.[1]

2.2 Ιδιοτιμές και ιδιοδιανύσματα

Ένα από τα πλέον χρήσιμα εργαλεία στη θεωρία πινάκων είναι το σύνολο των ιδιοτιμών και ιδιοδιανυσμάτων ενός πίνακα.

Ορισμός

Εάν $\mathrm{A} \in M_n$ και $x \in K^n$, θεωρούμε την εξίσωση

$$\mathrm{A}x = \lambda x, \; x \neq 0,$$

όπου $\lambda \in K$ ($K = \mathbb{R}$ ή $\mathbb{C}$). Εάν ένα $\lambda \in K$ και ένα μη μηδενικό διάνυσμα x ικανοποιούν αυτήν την εξίσωση, τότε το λ ονομάζεται ιδιοτιμή (eigenvalue) του A και το x ονομάζεται (δεξιό ή από δεξιά) ιδιοδιάνυσμα (eigenvector) του A που αντιστοιχεί στην ιδιοτιμή λ δηλαδή αν

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ και } \mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix},$$

τότε το λ είναι μια ιδιοτιμή του A με αντίστοιχο ιδιοδιάνυσμα x εάν

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = \lambda x_1$$
$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = \lambda x_2$$
$$.....$$
$$a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = \lambda x_n$$

Εάν θεωρήσουμε ότι $y*\mathrm{A} = \lambda y*, y \neq 0,$ το y καλείται αριστερό (ή από αριστερά) ιδιοδιάνυσμα του A που αντιστοιχεί στην ιδιοτιμή λ.

Το σύνολο όλων των $\lambda \in K$ που είναι ιδιοτιμές του A ονομάζεται φάσμα του A και συμβολίζεται με $\sigma(\mathrm{A})$.

Η φασματική ακτίνα του A είναι ο μη αρνητικός πραγματικός αριθμός $\rho(\mathrm{A})$ τέτοιος ώστε

$$\rho(\mathrm{A}) = \max\{|\lambda| : \lambda \in \sigma(\mathrm{A})\}.$$ [1]

2.3 Είδη πινάκων

Εάν $\mathrm{A} \in M_{m,n}(K)$, ονομάζουμε **ανάστροφό** του τον πίνακα $\mathrm{A}' \in M_{m,n}(K)$ ο οποίος έχει ως γραμμές τις αντίστοιχες στήλες του A και ως στήλες τις αντίστοιχες γραμμές του A. Είναι φανερό ότι $(\mathrm{A}')' = \mathrm{A}$.Εάν $\mathrm{A} \in M_{m,n}(K)$, συμβολίζουμε με A^* τον ανάστροφο συζυγή του A και με $\overline{\mathrm{A}}$ τον συζυγή του A.

Ένας πίνακας A ο οποίος έχει τον ίδιο αριθμό γραμμών και στηλών λέγεται **τετραγωνικός** πίνακας. Αν ο A είναι τύπου $n \times n$, λέμε ότι είναι τάξης n. Στο σύνολο $M_{nn}(K)$ των τετραγωνικών πινάκων τάξης n με στοιχεία από το K ($K = \mathbb{R}$ ή $\mathbb{C}$), η πράξη του πολλαπλασιασμού πινάκων ορίζεται πάντοτε και το αποτέλεσμα ανήκει στο ίδιο σύνολο $M_{nn}(K)$. Στον τετραγωνικό πίνακα $\mathrm{A} = [a_{ij}], i, j = 1,...,n$ ορίζουμε την κύρια διαγώνιο ως τα διατεταγμένα στοιχεία $a_{11}, a_{22}, ..., a_{nn}$, δηλαδή τα στοιχεία a_{ij} του A για τα οποία $i = j$. Το άθροισμα των στοιχείων της κύριας διαγωνίου του A λέγεται ίχνος (trace) του A, δηλαδή

$$tr(\mathrm{A}) = a_{11} + a_{22} + ... + a_{nn} = \sum_{i=1}^{n} a_{ii}\ .$$

Ο τετραγωνικός πίνακας τάξης n ο οποίος έχει κάθε στοιχείο της κύριας διαγωνίου ίσο με 1 και όλα τα υπόλοιπα στοιχεία του ίσα με μηδέν, δηλαδή ο πίνακας

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

λέγεται **μοναδιαίος** πίνακας και αποτελεί το ουδέτερο στοιχείο για τον πολλαπλασιασμό στο $M_n(K)$, αφού ισχύει

$I_n \mathrm{A} = \mathrm{A} I_n = \mathrm{A}$, για κάθε $\mathrm{A} \in M_n(K)$.

Ένας πίνακας $\mathrm{A} \in M_n(K)$ αντιστρέφεται (ή είναι **αντιστρέψιμος** ή **μη-ιδιάζων** ή **μη-ιδιόμορφος**), όταν και μόνο όταν υπάρχει πίνακας $\mathrm{B} \in M_n(K)$ τέτοιος ώστε $\mathrm{AB} = \mathrm{BA} = I_n$.

Είναι εύκολο να αποδειχθεί ότι ο πίνακας αυτός B, όταν υπάρχει, είναι μοναδικός. Πραγματικά, ας υποθέσουμε ότι ο C έχει και αυτός την ιδιότητα $\mathrm{A}C = C\mathrm{A} = I_n$.

Τότε $C = CI_n = C(\mathrm{AB}) = (C\mathrm{A})\mathrm{B} = I_n\mathrm{B} = \mathrm{B}$.

Ο μοναδικός αυτός πίνακας B λέγεται αντίστροφος του A και γράφεται $\mathrm{B} = \mathrm{A}^{-1}$.

Προφανώς $\left(\mathrm{A}^{-1}\right)^{-1} = \mathrm{A}$.

Είναι δυνατόν να αποδειχθεί ότι :

Αν $\mathrm{A}, \mathrm{B} \in M_n(K)$, τότε ισχύει: $\mathrm{AB} = I_n$, όταν και μόνο όταν $\mathrm{BA} = I_n$.

Ένας πραγματικός τετραγωνικός πίνακας A λέγεται **συμμετρικός**, όταν και μόνο όταν $\mathrm{A} = \mathrm{A}^T$.

Αν ο $\mathrm{A} = \left[a_{ij}\right], i, j = 1, 2, ..., n$ είναι συμμετρικός, τότε $\mathrm{A} = \mathrm{A}^T$ ή $\left[a_{ij}\right] = \left[a_{ji}\right]$ ή $a_{ij} = a_{ji}$, για κάθε $i, j = 1, 2, ..., n$. Είναι φανερό ότι ισχύει και το αντίστροφο. Οι συμμετρικοί πίνακες (και μόνο αυτοί) έχουν τα στοιχεία τους που είναι συμμετρικά ως προς την κύρια διαγώνιο ίσα.

Ένας πραγματικός τετραγωνικός πίνακας A λέγεται **αντισυμμετρικός**, όταν και μόνο όταν $-\mathrm{A}=\mathrm{A}^T$.

Αν ο $\mathrm{A}=\left[a_{ij}\right], i,j=1,2,...,n$ είναι αντισυμμετρικός, τότε $-\mathrm{A}=\mathrm{A}^T$ ή $\left[-a_{ij}\right]=\left[a_{ji}\right]$ ή $-a_{ij}=a_{ji}$, για κάθε $i,j=1,2,...,n$. Και εδώ το αντίστροφο είναι προφανές. Οι αντισυμμετρικοί πίνακες έχουν τα στοιχεία τους που είναι συμμετρικά ως προς την κύρια διαγώνιο αντίθετα. Όταν $i=j$, έχουμε $-a_{ii}=a_{ii}$ ή $2a_{ii}=0$ ή $a_{ii}=0$. Όλα τα στοιχεία της κύριας διαγωνίου ενός αντισυμμετρικού πίνακα είναι μηδέν. Οι δύο ιδιότητες μαζί εξασφαλίζουν ότι ο A είναι αντισυμμετρικός.

Ένας πραγματικός τετραγωνικός πίνακας A λέγεται **ορθογώνιος** όταν και μόνο όταν $\mathrm{A}^T=\mathrm{A}^{-1}$. Οι ορθογώνιοι πίνακες είναι υποχρεωτικά αντιστρέψιμοι και ικανοποιούν τη σχέση $\mathrm{A}\mathrm{A}^T=\mathrm{A}^T\mathrm{A}=I_n$.

Οι πραγματικοί πίνακες A με την ιδιότητα $\mathrm{A}\mathrm{A}^T=\mathrm{A}^T\mathrm{A}$ λέγονται **κανονικοί.** Ένας συμμετρικός ή αντισυμμετρικός ή ορθογώνιος είναι κανονικός, αλλά οι τύποι αυτοί των πινάκων δεν εξαντλούν τους κανονικούς.

Θεωρούμε τώρα το σύνολο των μιγαδικών τετραγωνικών πινάκων $M_{nn}\left(\mathbb{C}\right)$. Ας ανακαλέσουμε στη μνήμη μας το γεγονός ότι σε κάθε μιγαδικό αριθμό $z=a+bi, a,b\in\mathbb{R}$ αντιστοιχεί ο συζυγής μιγαδικός $\overline{z}=a-bi$. Η εικόνα του $\overline{z}$ στο μιγαδικό επίπεδο Oxy είναι η συμμετρική του z ως προς τον $x-$άξονα. Ο $z=a+bi$ είναι πραγματικός αριθμός, όταν και μόνο όταν ικανοποιεί τη σχέση $z=\overline{z}$. Αν για δύο μιγαδικούς αριθμούς ισχύει $z_1=\overline{z}_2$, τότε $\overline{z}_1=\overline{\overline{z}}_2=z_2$, δηλαδή οι z_1,z_2 είναι συζυγείς: αν ο ένας έχει τη μορφή $x+yi, x,y\in\mathbb{R}$, τότε ο άλλος είναι ο $x-yi$. Σε έναν πίνακα με μιγαδικά στοιχεία $\mathrm{A}=\left[a_{ij}\right]$ αντιστοιχεί ο συζυγής $\overline{\mathrm{A}}=\left[\overline{a}_{ij}\right]$. Λόγω του ορισμού του μιγαδικού εσωτερικού γινομένου, το ρόλο του ανάστροφου A^T αντικαθιστά τώρα ο συζυγής ανάστροφος $\overline{\left(\mathrm{A}^T\right)}$. Οι διαδικασίες της αναστροφής και της συζυγίας αντιμετατίθενται και για τον συζυγή ανάστροφο του A θα χρησιμοποιούμε τον συμβολισμό $\mathrm{A}^*=\overline{\left(\mathrm{A}^T\right)}=\left(\overline{\mathrm{A}}\right)^T$.

Είναι φανερό ότι όταν ο A είναι πραγματικός, τότε ο A^* ταυτίζεται με τον A^T.

Ένας μιγαδικός τετραγωνικός πίνακας A λέγεται **ερμιτιανός** (Hermitian), όταν και μόνο όταν $\mathrm{A} = \mathrm{A}^*$.

Αν ο $\mathrm{A} = \left[a_{ij} \right], i, j = 1, 2, ..., n$ είναι ερμιτιανός, τότε $\mathrm{A} = \mathrm{A}^*$ ή $\left[a_{ij} \right] = \overline{\left[a_{ji} \right]}$ ή $a_{ij} = \overline{a}_{ji}$. Στην περίπτωση $i = j$, έχουμε $a_{ii} = \overline{a}_{ii}$, ή $a_{ii} \in \mathbb{R}$, για κάθε $i = 1, 2, ..., n$. Τα στοιχεία της κύριας διαγωνίου ενός ερμιτιανού πίνακα είναι πραγματικά. Για $i \neq j$, τα στοιχεία τα οποία είναι συμμετρικά ως προς την κύρια διαγώνιο είναι συζυγείς αριθμοί. Οι δύο ιδιότητες μαζί χαρακτηρίζουν τους ερμιτιανούς πίνακες και μόνο.

Προφανώς, ένας ερμιτιανός και πραγματικός πίνακας είναι συμμετρικός, και αντίστροφα. Με άλλα λόγια, οι ερμιτιανοί αποτελούν τη γενίκευση των συμμετρικών πινάκων στη μιγαδική περίπτωση. Σε αντιστοιχία με τους αντισυμμετρικούς πίνακες βρίσκονται οι **αντιερμιτιανοί** (skew-Hermitian) οι οποίοι ορίζονται από την ιδιότητά τους $-\mathrm{A} = \mathrm{A}^*$. Τέλος το μιγαδικό ανάλογο των ορθογώνιων πινάκων είναι οι ορθομοναδιαίοι.

Ένας μιγαδικός τετραγωνικός πίνακας A λέγεται **ορθομοναδιαίος** ή **ορθοκανονικός** (Unitary) όταν και μόνο όταν $\mathrm{A}^{-1} = \mathrm{A}^*$.

Οι ορθομοναδιαίοι πίνακες είναι οι αντιστρέψιμοι μιγαδικοί πίνακες A οι οποίοι έχουν για αντίστροφο το συζυγή ανάστροφό τους, δηλαδή ικανοποιούν τη σχέση $\mathrm{A}\mathrm{A}^* = \mathrm{A}^*\mathrm{A} = I_n$.

Ιδιότητες ορθομοναδιαίων πινάκων

Για κάθε ορθομοναδιαίο πίνακα $\mathrm{A} \in M_n\left(F\right)$, ισχύουν οι επόμενες ιδιότητες:

1) Ο πίνακας A είναι αντιστρέψιμος με $\mathrm{A}^{-1} = \mathrm{A}^*$ και $\left|\det \mathrm{A}\right| = 1$
2) Ο πίνακας A^* είναι ορθομοναδιαίος πίνακας και ισχύει $\left(\mathrm{A}^*\right)^* = \mathrm{A}$
3) Οι γραμμές (στήλες) του A είναι μία ορθοκανονική βάση του
4) Το γινόμενο ορθομοναδιαίων πινάκων είναι ορθομοναδιαίος πίνακας.

Στην περίπτωση που $A \in M_n(\mathbb{R})$, οι παραπάνω ιδιότητες ισχύουν, μόνο που δεν υπάρχει η συζυγία των πινάκων.

Ιδιότητες ορθογώνιων πινάκων

1) Ο $n \times n$ πίνακας A είναι ορθογώνιος, τότε και μόνο όταν οι γραμμές του αποτελούν ορθοκανονική βάση του $\mathbb{R}^n$.
2) Σ' έναν ευκλείδειο χώρο ο πίνακας μετάβασης από ορθοκανονική σε ορθοκανονική βάση, είναι πάντα ορθογώνιος. Αντίστροφα, κάθε ορθογώνιος πίνακας μπορεί να θεωρηθεί σαν ο πίνακας μια τέτοιας μετάβασης.
3) Ένας ενδομορφισμός ενός n-διάστατου ευκλείδειου χώρου E είναι ισομετρία, τότε και μόνο όταν ο πίνακάς του σε μια ορθοκανονική βάση του E είναι ορθογώνιος.
4) Η ορίζουσα κάθε ορθογώνιου πίνακα είναι είτε 1 ή -1.

Ιδιότητες ερμιτιανά ορθογώνιων πινάκων

1) Ένας πίνακας $A \in M_n(\mathbb{C})$ είναι ερμιτιανά ορθογώνιος τότε και μόνο όταν οι γραμμές του αποτελούν ορθοκανονική βάση του ερμιτιανού χώρου $\mathbb{C}_n$.
2) Το γινόμενο δύο ερμιτιανά ορθογώνιων πινάκων είναι πάλι ερμιτιανά ορθογώνιος και ο αντίστροφος ερμιτιανά ορθογωνίου πίνακα είναι πάλι ερμιτιανά ορθογώνιος.
3) Η ορίζουσα ενός ερμιτιανά ορθογώνιου πίνακα έχει μέτρο ίσο με 1.
4) Ο πίνακας μετάβασης από ορθοκανονική σε ορθοκανονική βάση είναι πάντα ερμιτιανά ορθογώνιος.
5) Ο πίνακας μια ισομετρίας $T : H \to H$ ως προς μία ορθοκανονική βάση του H είναι ερμιτιανά ορθογώνιος και αντίστροφα. [1], [5]

2.4 Jordan κανονική μορφή πίνακα

Αν υποθέσουμε ότι $A \in M_n$, τότε οι στοιχειώδεις διαιρέτες του A έχουν τη μορφή

$$(\lambda - \lambda_1)^{p_1}, (\lambda - \lambda_2)^{p_2}, ..., (\lambda - \lambda_k)^{p_k}. \quad (1)$$

Αντιστοιχούμε στον τυχαίο στοιχειώδη διαιρέτη $e_i(\lambda) = (\lambda - \lambda_i)^{p_i}$ τον $p_i \times p_i$ πίνακα

$$\begin{pmatrix} \lambda_i & 1 & 0 & ... & 0 \\ 0 & \lambda_i & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \\ 0 & 0 & 0 & ... & \lambda_i \end{pmatrix} \quad (2)$$

Ο παραπάνω πίνακας συμβολίζεται J_{p_i} ή συντομότερα J_i και καλείται Jordan μπλοκ που αντιστοιχεί στον στοιχειώδη διαιρέτη $(\lambda - \lambda_1)^{p_1}$. Έτσι, στους στοιχειώδεις διαιρέτες (1) αντιστοιχούν τα Jordan μπλοκ $J_1, J_2, ..., J_k$. Το τυχαίο μπλοκ J_i μπορεί να γραφεί στη μορφή

$$J_i = \lambda_i I + H_{p_i} \quad (3)$$

όπου $H_{p_i} : p_i \times p_i$ πίνακας με 1 στην υπερδιαγώνιο και 0 αλλού.

Μπορεί εύκολα να δειχτεί ότι ο πίνακας (2) έχει μοναδιαίο στοιχειώδη διαιρέτη το $e_i(\lambda) = (\lambda - \lambda_i)^{p_i}$. Τότε, ο μπλοκ-διαγώνιος πίνακας

$$J = diag\{J_1, J_2, ..., J_k\}, \quad (4)$$

έχει ως μοναδικούς στοιχειώδεις διαιρέτες τα πολυώνυμα (1). Επειδή οι πίνακες A και J έχουν τους ίδιους στοιχειώδεις διαιρέτες, είναι όμοιοι. Ο πίνακας (4) καλείται Jordan κανονική μορφή του A ή απλά Jordan μορφή του A.

Σε έναν στοιχειώδη διαιρέτη $(\lambda - \lambda_i)^{p_i}$ θα μπορούσε να αντιστοιχηθεί αντί για τον πίνακα (2) ο πίνακας

$$\begin{pmatrix} \lambda_1 & 0 & 0 & ... & 0 & 0 \\ 1 & \lambda_i & 0 & ... & 0 & 0 \\ 0 & 1 & \lambda_i & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \lambda_i & 0 \\ 0 & 0 & 0 & ... & 1 & \lambda_i \end{pmatrix}. \quad (5)$$

Προς διάκριση των δύο περιπτώσεων καλούμε τον πίνακα (2) “άνω Jordan μπλοκ”, ενώ τον πίνακα (5) “κάτω Jordan μπλοκ”. [1]

2.4.1 Η πραγματική Jordan κανονική μορφή

Στην παράγραφο αυτή θα αναζητήσουμε μια κανονική μορφή πίνακα ανάλογη της Jordan κανονικής μορφής η οποία θα είναι πραγματική όταν ο πίνακας είναι πραγματικός.

Το κίνητρο για την αναζήτηση της νέας κανονικής μορφής δημιουργείται αν παρατηρήσουμε ότι η Jordan κανονική μορφή ενός πραγματικού πίνακα A δεν είναι απαραίτητα πραγματική. Έτσι, αν ο A έχει κάποια ιδιοτιμή $\lambda \in \mathbb{C}-\mathbb{R}$, τότε σ' αυτήν αντιστοιχεί τουλάχιστον ένα Jordan μπλοκ, το οποίο (επειδή έχει στη διαγώνιο το $\lambda \in \mathbb{C}-\mathbb{R}$) δεν είναι πραγματικό. Κατά συνέπεια η Jordan κανονική μορφή του A δεν είναι πραγματική. Βεβαίως, αν όλες οι ιδιοτιμές είναι πραγματικές, η κανονική μορφή είναι πραγματική.

Έστω, λοιπόν, $\mathrm{A} \in M_n(\mathbb{R})$ και $\lambda \in \sigma(\mathrm{A}), \lambda \in \mathbb{C}-\mathbb{R}$. Είναι γνωστό ότι θα είναι $\overline{\lambda} \in \sigma(\mathrm{A})$. Ακόμη, παρατηρούμε ότι

$$rank(\mathrm{A}-\lambda I)^k = rank\left(\overline{\mathrm{A}-\lambda I}\right)^k = rank\left(\mathrm{A}-\overline{\lambda} I\right)^k,$$

για κάθε $\lambda \in \mathbb{C}$ και $k=1,2,\ldots$. Άρα, τα μπλοκ της Jordan κανονικής μορφής J του A που αντιστοιχούν στις ιδιοτιμές λ και $\overline{\lambda}$ έχουν το ίδιο πλήθος και είναι του ίδιου τύπου.

Ας υποθέσουμε ότι στον J υπάρχει το μπλοκ $J_r(\lambda) \in M_r$, οπότε θα υπάρχει και το μπλοκ $J_r\left(\overline{\lambda}\right) \in M_r$. Θεωρούμε, επιπλέον, χωρίς περιορισμό της γενικότητας, ότι τα $J_r(\lambda)$ και $J_r\left(\overline{\lambda}\right)$ είναι σε γειτονικές θέσεις στον J. Έτσι, στην διαγώνιο του J εμφανίζεται ο πίνακας

$$\begin{pmatrix} J_r(\lambda) & 0 \\ 0 & J_r(\overline{\lambda}) \end{pmatrix} = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 & & & & & \\ 0 & \lambda & 1 & \dots & 0 & & & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & & & & & \\ 0 & 0 & 0 & \cdots & \lambda & & & & & \\ & & & & & \overline{\lambda} & 1 & 0 & \dots & 0 \\ & & & 0 & & 0 & \overline{\lambda} & 1 & \dots & 0 \\ & & & & & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & & & 0 & 0 & 0 & \cdots & \overline{\lambda} \end{pmatrix}.$$

Με εναλλαγές γραμμών και στηλών ο παραπάνω πίνακας τίθεται στη μορφή

$$\begin{pmatrix} \lambda & 0 & 1 & 0 & & \cdots & & \\ 0 & \overline{\lambda} & 0 & 1 & & & & \\ & & \lambda & 0 & 1 & \cdots & 0 & \\ & & 0 & \overline{\lambda} & 0 & & & \\ \vdots & & \vdots & & & \ddots & \vdots & \\ & & & & & & 1 & 0 \\ & & & 0 & & \dots & 0 & 1 \\ & & & & & & \lambda & 0 \\ & & & & & \dots & 0 & \overline{\lambda} \end{pmatrix} = \begin{pmatrix} D(\lambda) & I & & \dots & \\ & D(\lambda) & I & \dots & 0 \\ & 0 & & \ddots & \vdots \\ & & \vdots & \dots & I \\ & & & \dots & D(\lambda) \end{pmatrix}$$

Όπου I ο 2×2 μοναδιαίος πίνακας και

$$D(\lambda) = \begin{pmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{pmatrix} \in M_2 .$$

Έστω, $\lambda = \alpha + i\beta, \alpha, \beta \in R$, και $U = \begin{pmatrix} -i & -i \\ 1 & -1 \end{pmatrix}$. Εύκολα επιβεβαιώνεται ότι

$$UD(\lambda)U^{-1} = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix},$$

δηλαδή ο $D(\lambda)$ είναι όμοιος με τον πίνακα

$$C(\alpha,\beta) = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}.$$

Κατά συνέπεια, κάθε ζεύγος από "συζυγή" Jordan μπλοκ $J_r(\lambda)$ και $J_r(\overline{\lambda})$ που αντιστοιχούν στις ιδιοτιμές $\lambda = a + i\beta$ και $\overline{\lambda} = \alpha - i\beta$, $\beta \neq 0$, είναι όμοιο με τον $2r \times 2r$ πίνακα

$$C_r(\alpha,\beta) = \begin{pmatrix} C(\alpha,\beta) & I & & \dots & \\ & C(\alpha,\beta) & I & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ & 0 & & \dots & I \\ & & & \dots & C(\alpha,\beta) \end{pmatrix} \in M_{2r}(\mathbb{R}) \qquad (1)$$

Λαμβάνοντας υπόψη τα παραπάνω διατυπώνουμε την εξής πρόταση:

Κάθε πραγματικός πίνακας A είναι όμοιος με έναν μπλοκ διαγώνιο πίνακα της μορφής

$$\begin{pmatrix} C_{n_1}(\alpha_1,\beta_1) & & & & & & \\ & C_{n_2}(\alpha_2,\beta_2) & & & & & \\ & & \ddots & & & 0 & \\ & & & C_{n_p}(\alpha_p,\beta_p) & & & \\ & & & & J_{n_r}(\lambda_r) & & \\ & 0 & & & & \ddots & \\ & & & & & & J_{n_s}(\lambda_s) \end{pmatrix}, \qquad (2)$$

όπου $\lambda_j = a_j + i\beta_j$, $j = 1,2,...,p$, είναι μη πραγματικές και $\lambda_r,...,\lambda_s$ πραγματικές ιδιοτιμές του A. Τα Jordan μπλοκ $J_{n_r}(\lambda_r),...,J_{n_s}(\lambda_s)$ είναι ακριβώς εκείνα που εμφανίζονται στην Jordan κανονική μορφή J του A και αντιστοιχούν στις πραγματικές ιδιοτιμές. Τα μπλοκ $C_{n_j}(\alpha_j,\beta_j)$, $j = 1,2,...,p$, αντιστοιχούν στις μη πραγματικές ιδιοτιμές του A. Για κάθε ζεύγος από Jordan μπλοκ $J_r(\lambda)$ και $J_r(\overline{\lambda})$ του J, τα οποία αντιστοιχούν σε συζυγείς μιγαδικές ιδιοτιμές $\lambda = a_j + i\beta_j$ και $\overline{\lambda} = \alpha_j - i\beta_j$, $\beta_j \neq 0$, και έχουν διάσταση r το καθένα, εμφανίζεται στον πίνακα (2) το μπλοκ $C_r(\alpha_j,\beta_j)$ διάστασης $2r$ που έχει τη μορφή (1).

Ο πίνακας (2) καλείται πραγματική Jordan κανονική μορφή ή απλά πραγματική Jordan μορφή του A.[1]

2.5 Πίνακας μετάβασης

Σε ένα διανυσματικό χώρο συχνά συναντάμε το πρόβλημα αλλαγής βάσης. Θέλουμε δηλαδή να βρούμε ποια σχέση συνδέει τις συντεταγμένες ενός διανύσματος σε δύο διαφορετικές βάσεις $\{\vec{e}_1,....,\vec{e}_\nu\}$ και $\{\vec{e}_1^{'},....,\vec{e}_\nu^{'}\}$ ενός χώρου E.

Για να τη βρούμε εργαζόμαστε ως εξής: Αναλύουμε καθένα από τα διανύσματα $\vec{e}_1^{'},....,\vec{e}_\nu^{'}$ στη βάση $\{\vec{e}_1,....,\vec{e}_\nu\}$.

$$\left\{\begin{array}{l}\vec{e}_1^{'}=a_{11}\vec{e}_1+a_{21}\vec{e}_2+...+a_{\nu 1}\vec{e}_\nu\\ \vec{e}_2^{'}=a_{12}\vec{e}_1+a_{22}\vec{e}_2+...+a_{\nu 2}\vec{e}_\nu\\ \vdots\\ \vec{e}_\nu^{'}=a_{1\nu}\vec{e}_1+a_{2\nu}\vec{e}_2+...+a_{\nu\nu}\vec{e}_\nu\end{array}\right\}$$

Συμβολίζουμε με P τον $\nu\times\nu$ πίνακα που έχει για 1^{η} στήλη του τις συντεταγμένες του $\vec{e}_1^{'}$, για 2^{η} στήλη του τις συντεταγμένες του $\vec{e}_1^{'}$ κ.λπ.

$$P=\begin{pmatrix}a_{11} & a_{12} & ... & a_{1\nu}\\ a_{21} & a_{22} & ... & a_{2\nu}\\ ... & ... & ... & ...\\ a_{\nu 1} & a_{\nu 2} & ... & a_{\nu\nu}\end{pmatrix}$$

Τον πίνακα αυτόν τον ονομάζουμε πίνακα μετάβασης από τη βάση $\{\vec{e}_i\}$ στην $\{\vec{e}_i^{'}\}$.

Ας είναι τώρα

$\vec{x}=x_1\vec{e}_1+,...,+x_\nu\vec{e}_\nu$ και $\vec{x}=x_1^{'}\vec{e}_1^{'}+,...,+x_\nu^{'}\vec{e}_\nu^{'}$ οι αναλύσεις ενός διανύσματος $\vec{x}$ κατά μήκος των βάσεων $(\vec{e}_i)$ και $(\vec{e}_i^{'})$ αντίστοιχα.

Θα έχουμε

$$\vec{x}=\sum_{k=1}^{\nu}x_k^{'}\vec{e}_k^{'}=\sum_{k=1}^{\nu}x_k^{'}\left(\sum_{j=1}^{\nu}a_{jk}\vec{e}_j\right)=\sum_{j=1}^{\nu}\left(\sum_{k=1}^{\nu}a_{jk}x_k^{'}\right)\vec{e}_j$$

έτσι

$$\sum_{j=1}^{\nu} x_j \vec{e}_j = \sum_{j=1}^{\nu} \left(\sum_{k=1}^{\nu} a_{jk} x_k^{'} \right) \vec{e}_j$$

Αφού κάθε διάνυσμα αναλύεται κατά μοναδικό τρόπο ως προς μια βάση, από την προηγούμενη ισότητα παίρνουμε

$$x_j = \sum_{k=1}^{\nu} a_{jk} x_k^{'} \ , \ j = 1, ..., \nu$$

Οι σχέσεις αυτές εκφράζουν την ακόλουθη ισότητα πινάκων

$$\begin{pmatrix} x_1 \\ \vdots \\ x_\nu \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1\nu} \\ \vdots & & \vdots \\ a_{\nu 1} & \cdots & a_{\nu\nu} \end{pmatrix} \begin{pmatrix} x_1^{'} \\ \vdots \\ x_\nu^{'} \end{pmatrix}$$

Η τελευταία σχέση μας λέει ότι αν πολλαπλασιάσουμε την στήλη των συντεταγμένων του $\vec{x}$ στη βάση $\left(\vec{e}_i^{\,'}\right)$ με τον πίνακα μετάβασης βρίσκουμε την στήλη των συντεταγμένων του $\vec{x}$ στη βάση $\left(\vec{e}_i\right)$. Αυτή είναι και η αρχικά ζητούμενη σχέση.

Ο πίνακας $P = \left(p_{ij}\right)$ μετάβασης από την βάση $\left\{\vec{e}_i\right\}$ στη βάση $\left\{\vec{e}_i^{\,'}\right\}$ είναι αντιστρέψιμος και μάλιστα ο αντίστροφός του είναι ο πίνακας μετάβασης από την βάση $\left\{\vec{e}_i^{\,'}\right\}$ στην $\left\{\vec{e}_i\right\}$. [3]

2.6 Πίνακες συνιστώσες

Θεώρημα :

Αν $A \in M_n$ με ελάχιστο πολυώνυμο

$$\psi(\lambda) = (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} ... (\lambda - \lambda_k)^{m_k}$$

και f συνάρτηση που ορίζεται στο $\sigma(A)$, τότε υπάρχουν πίνακες Z_{ij}, έτσι ώστε

$$f(A) = \sum_{i=1}^{k} \sum_{j=0}^{m_i - 1} f_i^{(j)} Z_{ij} \quad (2)$$

Οι πίνακες Z_{ij} δεν εξαρτώνται από την f και είναι γραμμικά ανεξάρτητοι.

Ορισμός :

Αν $A \in M_n$, οι πίνακες Z_{ij} του θεωρήματος 1 καλούνται (πίνακες-) συνιστώσες του A.

Από το παραπάνω θεώρημα προκύπτουν τα εξής επί μέρους συμπεράσματα για τις συνιστώσες Z_{ij} ενός πίνακα A:

Επειδή οι Z_{ij} είναι πίνακες γραμμικά ανεξάρτητοι, κανένας από αυτούς δεν μπορεί να είναι ο μηδενικός πίνακας. Ακόμη, επειδή οι Z_{ij} είναι πολυωνυμικές συναρτήσεις του A (αφού $Z_{ij} = \varphi_{ij}(A)$), άρα οποιοιδήποτε δυο από αυτούς αντιμετατίθενται μεταξύ τους. Για τον ίδιο λόγο οποιοσδήποτε από τους Z_{ij} αντιμετατίθεται με τον A. Τέλος, αν απαιτείται ο υπολογισμός δυο η περισσοτέρων συναρτήσεων του A, τότε η μέθοδος του θεωρήματος 1 είναι ιδιαίτερα χρήσιμη, αφού οι Z_{ij} υπολογίζονται μόνο μια φορά (ως εξαρτώμενοι μόνο από τον A και όχι από τις συναρτήσεις). Εξ' άλλου, η ανεξαρτησία των Z_{ij} από οποιαδήποτε συνάρτηση μας επιτρέπει να τους υπολογίσουμε από την (2), θεωρώντας την (2) ως σύστημα με αγνώστους τα Z_{ij} για κατάλληλες "απλές" συναρτήσεις f. [1]

2.7 Πίνακας προβολής

Έχουμε την εξίσωση $Ax = b$. Αποδεικνύεται ότι η εξίσωση έχει λύσεις μόνον όταν το διάνυσμα b ανήκει στο χώρο στηλών του πίνακα A. Συχνά όμως θέλουμε να βρούμε την καλύτερη δυνατή λύση της εξίσωσης, ακόμη και όταν το b δεν ανήκει στον $R(A)$. Αυτό συμβαίνει συχνά στην ανάλυση πειραματικών δεδομένων, όπου για να περιορίσουμε την πιθανότητα τυχαίου σφάλματος, παίρνουμε περισσότερες μετρήσεις. Το αποτέλεσμα είναι να έχουμε ένα σύστημα με αρκετά περισσότερες εξισώσεις παρά αγνώστους, όπου δεν περιμένουμε να υπάρχει ακριβής λύση.

Μπορούμε να βρούμε μια βέλτιστη λύση της εξίσωσης, εάν αντικαταστήσουμε το διάνυσμα b από το διάνυσμα του χώρου στηλών του A που είναι πλησιέστερο στο b από κάθε άλλο διάνυσμα του χώρου στηλών. Αυτό το διάνυσμα είναι η ορθογώνια προβολή του b στο χώρο στηλών.

Εάν συμβολίσουμε p την ορθογώνια προβολή του b στο χώρο στηλών, έχουμε μια νέα εξίσωση

$$A\hat{x} = p.$$

Οι λύσεις αυτής της εξίσωσης είναι βέλτιστες λύσεις της αρχικής εξίσωσης $Ax = b$.

Εξετάζουμε πρώτα την προβολή σε μία ευθεία. Θεωρούμε τα διανύσματα a και b στο επίπεδο. Γνωρίζουμε από την αναλυτική γεωμετρία την προβολή του b πάνω στην ευθεία που ορίζει το a. Το διάνυσμα προβολής p χαρακτηρίζεται από τις ακόλουθες ιδιότητες:

1. Το p είναι συγγραμμικό με το a, $p = a\hat{x}$ για $\hat{x} \in \mathbb{R}$.
2. Η διαφορά $b - p$ είναι ορθογώνια στο a, $a^T(b - p) = 0.$

Από αυτές τις ιδιότητες λαμβάνουμε την εξίσωση

$$a^T(b - a\hat{x}) = 0$$

την οποία μπορούμε να λύσουμε για να βρούμε το $\hat{x}$:

$$\hat{x} = \frac{a^T b}{a^T a}.$$

Συνεπώς το διάνυσμα προβολής είναι

$$p = a\hat{x} = a\frac{a^T b}{a^T a},$$

και εφαρμόζοντας την προσεταιριστική ιδιότητα

$$p\frac{1}{a^T a}aa^T b.$$

Παρατηρούμε ότι $a^T a$ είναι θετικός αριθμός, το τετράγωνο του μήκους του a, ενώ aa^T είναι τετραγωνικός πίνακας.

Τον πίνακα $P = \frac{1}{a^T a}aa^T$ τον ονομάζουμε πίνακα προβολής (projection matrix).

Ισχύει ότι: Ένας $m \times m$ πίνακας P είναι πίνακας προβολής σε ένα υποχώρο του $\mathbb{R}^m$ εάν και μόνον εάν P είναι συμμετρικός και $P^2 = P.$

Απόδειξη. Έστω V ένας υπόχωρος του $\mathbb{R}^m$, και A ο πίνακας με στήλες τα διανύσματα μίας βάσης του V. Τότε ο πίνακας προβολής στον υπόχωρο V είναι ο $P = A\left(A^T A\right)^{-1} A^T$. Εύκολα ελέγχουμε ότι $P^2 = P$,

$$\begin{aligned} P^2 &= A\left(A^T A\right)^{-1} A^T A\left(A^T A\right)^{-1} A^T \\ &= A\left(A^T A\right)^{-1} A^T \\ &= P. \end{aligned}$$

Ο ανάστροφος του P είναι ο πίνακας

$$\begin{aligned} P^T &= (A(A^T A)^{-1} A^T)^T \\ &= (A^T)^T \left((A^T A)^{-1}\right)^T A^T \\ &= A\left((A^T A)^T\right)^{-1} A^T \\ &= A(AA^T)^{-1} A^T \\ &= P \end{aligned}$$

Αντιστρόφως, εάν ο $m \times m$ πίνακας P ικανοποιεί τις σχέσεις $P^2 = P$ και $P = P^T$, θα δείξουμε ότι P είναι ο πίνακας προβολής στο χώρο στηλών του. Προφανώς, για κάθε $b \in \mathbb{R}^m$, Pb ανήκει στο χώρο στηλών του P. Για να δείξουμε ότι Pb είναι η προβολή του b στον υπόχωρο $V = R(P)$ αρκεί να δείξουμε ότι $b - Pb$ είναι ορθογώνιο στον V.

Έστω u διάνυσμα του V. Τότε u είναι γραμμικός συνδυασμός των στηλών του P, δηλαδή υπάρχει $c \in \mathbb{R}^m$ τέτοιο ώστε $u = Pc$, και έχουμε

$$\begin{aligned} \left(b - Pb\right)^T u &= \left(b - Pb\right)^T Pc \\ &= \left(b^T - b^T P^T\right) Pc \\ &= b^T \left(I - P^T\right) Pc \\ &= b^T \left(P - P^T P\right) c. \end{aligned}$$

Αλλά $P^T = P$ και $P^2 = P$, άρα $P - P^T P = P - P = 0$.

Τέλος, κάθε πίνακας προβολής P ικανοποιεί τις εξής δύο ιδιότητες:

1. $P^2 = P$ και

2. Ο P είναι συμμετρικός.

Επίσης, κάθε πίνακας που ικανοποιεί αυτές τις δύο ιδιότητες είναι ο πίνακας προβολής για κάποιο υπόχωρο του $\mathbb{R}^n$. [13]

3. Πολυσυγγραμικότητα

3.1 Πολυσυγγραμμικότητα

Μια σημαντική δυσκολία στην διαδικασία της πολλαπλής γραμμικής παλινδρόμησης είναι η πολυσυγγραμμικότητα δηλαδή η ύπαρξη δύο ή περισσοτέρων μεταβλητών $X_i, X_j, \ldots$ οι οποίες συνδέονται με μία σχέση γραμμική δηλαδή ισχύει

$X_j = a + bX_j (a, b \in \mathbb{R})$

Η ύπαρξη πολυσυγγραμμικότητας σ' ένα πρόβλημα πολλαπλής γραμμικής παλινδρόμησης μας δημιουργεί μία σειρά από προβλήματα όπως η ύπαρξη μη αντιστρέψιμων πινάκων με χαρακτηριστικότερο τον $X^T.X$, την πολύ μεγάλη διασπορά των συντελεστών b_i του μοντέλου καθώς και τις πολύ μεγάλες τιμές των S και SSE.

Έτσι αν οι μεταβλητές x_i, x_j συνδέονται με μία γραμμική σχέση τότε διαγράφουμε την μία από τις δύο (συνήθως αυτή με το μικρότερο R^2) και παίρνουμε ένα νέο μοντέλο με μία λιγότερη μεταβλητή.

Η διαδικασία μπορεί να συνεχιστεί μέχρι να απαλείψουμε αυτές τις μεταβλητές που είναι γραμμικά εξαρτημένες με άλλες μεταβλητές. [8]

3.2 Ο συντελεστής γραμμικής συσχέτισης του Pearson

Ο συντελεστής γραμμικής συσχέτισης του Pearson ορίζεται από τη σχέση

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Όπου $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ η συνδιακύμανση των x και y

s_x, s_y οι τυπικές αποκλίσεις των x και y (std. deviation)

- $-1 \le r_{xy} \le 1$

Όταν $r_{xy} \to 1$ τότε οι μεταβλητές x, y έχουν έντονη θετική γραμμική συσχέτιση

Όταν $r_{xy} \to -1$ τότε οι μεταβλητές x, y έχουν έντονη αρνητική γραμμική συσχέτιση

Όταν $r_{xy} \to 0$ τότε οι μεταβλητές x, y δεν έχουν γραμμική συσχέτιση

Ο r_{xy} είναι λοιπόν ένα μέτρο γραμμικής συμμεταβολής δύο τυχαίων μεταβλητών.

Έστω $x_1, x_2, ..., x_m$ μεταβλητές και X ο πίνακας n παρατηρήσεων των μεταβλητών αυτών δηλαδή

$$X = \begin{bmatrix} x_{11} & x_{21} & . & . & . & x_{m1} \\ x_{12} & x_{22} & . & . & . & x_{m2} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ x_{1n} & x_{2n} & . & . & . & x_{mn} \end{bmatrix} \ (n \times m \text{ πίνακας})$$

Εάν $\overline{x_j}, j = 1, 2, ..., m$ είναι οι μέσες τιμές των μεταβλητών x_j και s_j είναι οι τυπικές τους αποκλίσεις τότε συμβολίζω με X^* τον πίνακα

$$X^* = \begin{bmatrix} \frac{x_{11} - \overline{x_1}}{s_1\sqrt{n-1}} & \frac{x_{21} - \overline{x_2}}{s_2\sqrt{n-1}} & . & . & . & \frac{x_{m1} - \overline{x_m}}{s_m\sqrt{n-1}} \\ \frac{x_{12} - \overline{x_1}}{s_1\sqrt{n-1}} & \frac{x_{22} - \overline{x_2}}{s_2\sqrt{n-1}} & . & . & . & \frac{x_{m2} - \overline{x_m}}{s_m\sqrt{n-1}} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ \frac{x_{1n} - \overline{x_1}}{s_1\sqrt{n-1}} & \frac{x_{2n} - \overline{x_2}}{s_2\sqrt{n-1}} & . & . & . & \frac{x_{mn} - \overline{x_m}}{s_m\sqrt{n-1}} \end{bmatrix}$$

Από τον ορισμό γινομένου πινάκων έχω

$$X^{*T}\cdot X^{*}=\left[\frac{\sum_{k=1}^{n}(x_{jk}-\overline{x_j})(x_{ik}-\overline{x_i})}{(n-1)s_j\cdot s_i}\right]=\left[r_{ji}\right] \text{ με } j,i=1,2,...,m \text{ ή}$$

$$X^{*T}\cdot X^{*}=\begin{bmatrix} \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})^2}{(n-1)s_1^2} & \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{2k}-\overline{x_2})}{(n-1)s_1\cdot s_2} & \cdots & \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{2k}-\overline{x_2})}{(n-1)s_1\cdot s_2} \\ \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{2k}-\overline{x_2})}{(n-1)s_1\cdot s_2} & \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})^2}{(n-1)s_1^2} & \cdots & \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{2k}-\overline{x_2})}{(n-1)s_1\cdot s_2} \\ \vdots & \vdots & & \vdots \\ \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{2k}-\overline{x_2})}{(n-1)s_1\cdot s_2} & \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{2k}-\overline{x_2})}{(n-1)s_1\cdot s_2} & \cdots & \frac{\sum_{k=1}^{n}(x_{1k}-\overline{x_1})(x_{2k}-\overline{x_2})}{(n-1)s_1\cdot s_2} \end{bmatrix}$$

Συνεπώς ο $X^{*T}\cdot X^{*}$ είναι ο πίνακας των γραμμικών συσχετίσεων των μεταβλητών $x_i, i=1,2,...,m$ καθότι

$$r_{ji}=r(x_j,x_i)=\frac{\sum_{k=1}^{n}(x_{jk}-\overline{x_j})(x_{ik}-\overline{x_i})}{(n-1)s_j\cdot s_i}$$

Προφανώς ο πίνακας $X^{*T}\cdot X^{*}$ είναι συμμετρικός και τα στοιχεία της κύριας διαγωνίου του είναι 1 καθότι

$$r_{jj}=\frac{\sum_{k=1}^{n}(x_{jk}-\overline{x_j})^2}{(n-1)s_j^2}=\frac{(n-1)s_j^2}{(n-1)s_j^2}=1$$

άρα έχω ότι

$$X^{*T}\cdot X^{*}=R=\begin{bmatrix}1 & r_{12} & r_{13} & \dots & r_{1m}\\ r_{21} & 1 & r_{23} & \dots & r_{2m}\\ r_{31} & r_{32} & 1 & \dots & r_{3m}\\ \dots & \dots & \dots & \dots & \dots\\ r_{m1} & r_{m2} & r_{m3} & \dots & 1\end{bmatrix}$$

με $r_{ji}=r_{ij}$

Αν είναι $X^{*T}\cdot X^{*}=I_m$ τότε λέμε ότι οι μεταβλητές $x_1,x_2,...,x_m$ είναι ορθογώνιες και καμία σχέση συγγραμικότητας δεν υπάρχει μεταξύ των μεταβλητών.

Ο αντίστροφος πίνακας του R (correlation matrix) συμβολίζεται με S και τα διαγώνια στοιχεία του S παριστάνουν τα VIF (variance inflation factor) των αντίστοιχων μεταβλητών. Δηλαδή είναι

$VIF_j=s_{jj}$ όπου $s=\lfloor s_{ji}\rfloor \quad j,i=1,2,...,m$

Προφανώς και ο $S=R^{-1}$ είναι συμμετρικός σαν αντίστροφος συμμετρικού πίνακα.

Προφανώς $VIF_j\geq 1$

Αν $VIF_j=1$ τότε η μεταβλητή x_j είναι ορθογώνια προς τις υπόλοιπες μεταβλητές $x_1,x_2,...,x_{j-1},x_{j+1},...,x_m$

Αν $VIF_j\neq 1$ τότε η x_j έχει κάποια γραμμική εξάρτηση ως προς κάποια ή κάποιες από τις υπόλοιπες.

Όσο πιο μεγάλη είναι η τιμή του VIF_j τόσο μεγαλύτερη σχέση εξάρτησης έχει η x_j προς τις υπόλοιπες.

Οι μεταβλητές που έχουν μεγάλα VIF είναι μεταξύ τους γραμμικά εξαρτημένες.

Ένας τρίτος τρόπος για να ελέγξουμε την γραμμική εξάρτηση των μεταβλητών (συγγραμικότητα) είναι να βρούμε τις ιδιοτιμές του πίνακα R (correlation matrix).

Αν υπάρχει ιδιοτιμή 0 ή κοντά στο 0 τότε υπάρχουν γραμμικά εξαρτημένες μεταβλητές. Στην περίπτωση αυτή ορίζουμε την ποσότητα

$$\varphi = \frac{\lambda_{\max}}{\lambda_{\min}}$$

Όπου $\lambda_{\max}$ είναι η μεγαλύτερη ιδιοτιμή του R και $\lambda_{\min}$ η μικρότερη ιδιοτιμή του R και η φ ονομάζεται ο condition number του correlation matrix.

Αν το φ είναι μεγάλος αριθμός αυτό σημαίνει ότι υπάρχει «σοβαρή» πολυσυγγραμικότητα. Εάν το φ είναι εξαιρετικά μεγάλος αριθμός αυτό σημαίνει ότι οι συντελεστές του μοντέλου μου είναι ασταθείς.

Τέλος, ο συντελεστής συσχέτισης υπολογισμένος στην δεύτερη δύναμη (τετράγωνο), ορίζει τον «συντελεστή προσδιορισμού» (R squared ή R^2) και εξηγεί το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που οφείλεται στην επίδραση της ανεξάρτητης. Λαμβάνει τιμές μεταξύ μηδέν και ένα $\left(0 < R^2 \le 1\right)$ και όσο η τιμή του πλησιάζει τη μονάδα τόσο καλύτερη είναι η ερμηνευτική ικανότητα του μοντέλου. [8]

3.3 Κανονικοποιημένος πίνακας

Έστω $x_1, x_2, ..., x_m$ μεταβλητές και X ο πίνακας n παρατηρήσεων των μεταβλητών αυτών δηλαδή

$$X = \begin{bmatrix} x_{11} & x_{21} & . & . & . & x_{m1} \\ x_{12} & x_{22} & . & . & . & x_{m2} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ x_{1n} & x_{2n} & . & . & . & x_{mn} \end{bmatrix} \ (n \times m \text{ πίνακας})$$

Εάν $\overline{x_j}, j = 1, 2, ..., m$ είναι οι μέσες τιμές των μεταβλητών x_j και s_j είναι οι τυπικές τους αποκλίσεις τότε συμβολίζω με X^* τον πίνακα

$$X^* = \begin{bmatrix} \frac{x_{11}-\overline{x_1}}{s_1\sqrt{n-1}} & \frac{x_{21}-\overline{x_2}}{s_2\sqrt{n-1}} & \cdots & \frac{x_{m1}-\overline{x_m}}{s_m\sqrt{n-1}} \\ \frac{x_{12}-\overline{x_1}}{s_1\sqrt{n-1}} & \frac{x_{22}-\overline{x_2}}{s_2\sqrt{n-1}} & \cdots & \frac{x_{m2}-\overline{x_m}}{s_m\sqrt{n-1}} \\ \vdots & \vdots & & \vdots \\ \frac{x_{1n}-\overline{x_1}}{s_1\sqrt{n-1}} & \frac{x_{2n}-\overline{x_2}}{s_2\sqrt{n-1}} & \cdots & \frac{x_{mn}-\overline{x_m}}{s_m\sqrt{n-1}} \end{bmatrix}$$

Ο πίνακας αυτός είναι ο κανονικοποιημένος πίνακας X (Standardized Matrix), δηλαδή το αποτέλεσμα έχει μέση τιμή 0 και τυπική απόκλιση 1. Όμως θέλουμε τη μαθηματική κανονικοποίηση γι' αυτό πολλαπλασιάσαμε με $\frac{1}{\sqrt{n-1}}$. (Μετατρέψαμε τα διανύσματα σε μοναδιαία με μήκος 1).

Ο κανονικοποιημένος αυτός πίνακας έχει τις εξής ιδιότητες:

- Οι στήλες του είναι οι αρχικές μεταβλητές με αλλαγμένες τιμές, οι στήλες του έχουν μέση τιμή 0 και μήκος 1.
- Όλα τα στοιχεία του X^* είναι μεταξύ -1 και 1
- Είναι καθαροί αριθμοί, απαλλαγήκαμε από τις μονάδες μέτρησης των αριθμών
- Το άθροισμα των τετραγώνων των στοιχείων είναι μονάδα (έγιναν μοναδιαία διανύσματα οι μεταβλητές). [8]

4. Παραγοντική ανάλυση

4.1 Εισαγωγή

Η τεχνική της παραγοντικής ανάλυσης (factor analysis) επιτυγχάνει τη μείωση ενός μεγάλου αριθμού μεταβλητών σ' ένα μικρότερο αριθμό σημαντικών παραγόντων. Κριτήριο της τεχνικής αυτής είναι ο υπολογισμός των συσχετίσεων που παρατηρούνται μεταξύ των μεταβλητών που εξετάζονται. Ουσιαστικά, πρόκειται για μια στατιστική μέθοδο που «κατασκευάζει» μια ομάδα κοινών παραγόντων ανάμεσα σε μια ομάδα πολυάριθμων παραγόντων. Σκοπός αυτής της μεθόδου είναι η μείωση

των διαστάσεων του προβλήματος, διατηρώντας τις πληροφορίες των αρχικών μεταβλητών. Επιπλέον, μπορούμε να δημιουργήσουμε μεταβλητές που στην πραγματικότητα είναι μη μετρήσιμες άμεσα, όπως είναι για παράδειγμα η ευφυΐα, η σχιζοφρένεια και πολλές άλλες πτυχές της ανθρώπινης συμπεριφοράς. Τέλος, αναλύουμε τις συσχετίσεις μεταξύ των μεταβλητών που οφείλονται στην ύπαρξη των κοινών παραγόντων που δημιούργησαν τα δεδομένα. Η ανάλυση παραγόντων χρησιμοποιείται για πολλούς λόγους. Αναφέρουμε ενδεικτικά μερικούς:

- Καθορίζει τους κύριους παράγοντες που επηρεάζουν τις προτιμήσεις καταναλωτών.
- Ορίζει ποιες ερωτήσεις αξιολογούν την ίδια έννοια σε ένα ερωτηματολόγιο.
- Προσδιορίζει τις διαστάσεις μια κλίμακας.
- Ερευνά ποια χαρακτηριστικά είναι τα πιο σημαντικά στην ομαδοποίηση των καταναλωτικών ιδεών και στάσεων.
- Διαπιστώνει αν οι διαστάσεις μιας κλίμακας επαληθεύονται στα δεδομένα μιας έρευνας.

Η παραγοντική ανάλυση είναι μια τεχνική που έχει πολλές ομοιότητες με την τεχνική της ανάλυσης σε κύριες συνιστώσες. Η διαφορά της μεθόδου της παραγοντικής ανάλυσης από αυτήν των κύριων συνιστωσών είναι ότι, ενώ η τελευταία κατασκευάζει έναν ορθογώνιο μετασχηματισμό των μεταβλητών ο οποίος δεν εξαρτάται από το μοντέλο, η παραγοντική ανάλυση βασίζεται σε ένα κατάλληλο στατιστικό μοντέλο.

Επίσης, η παραγοντική ανάλυση εστιάζεται περισσότερο στην εξήγηση της δομής της συνδιακύμανσης των μεταβλητών παρά με την εξήγηση της διακύμανσης, κάτι που συμβαίνει στην ανάλυση σε κύριες συνιστώσες. Οποιαδήποτε διακύμανση δεν εξηγείται από τους κοινούς παράγοντες, θεωρείται ότι προέρχεται από τους όρους των καταλοίπων. [11], [12]

4.2 Το ορθογώνιο μοντέλο της παραγοντικής ανάλυσης

Στο ορθογώνιο μοντέλο της παραγοντικής ανάλυσης, το οποίο είναι και το πιο διαδεδομένο, υποθέτουμε ότι οι όποιες συσχετίσεις μεταξύ των μεταβλητών οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων τους οποίους δεν ξέρουμε και θέλουμε να εκτιμήσουμε.

Σύμφωνα με το παραγοντικό μοντέλο, θεωρούμε ένα τυχαίο X το οποίο μπορεί να παρατηρηθεί με p συνιστώσες, με μέσο μ και πίνακα συνδιακύμανσης Σ. Το X εξαρτάται γραμμικά από μια σειρά τυχαίων μεταβλητών $F_1, F_2, ..., F_k$ οι οποίες δεν είναι δυνατόν να παρατηρηθούν. Οι μεταβλητές αυτές ονομάζονται κοινοί παράγοντες (common factors). Το X εξαρτάται επίσης από p πρόσθετες πηγές μεταβλητότητας $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ που ονομάζονται σφάλματα (errors).

Ειδικότερα το μοντέλο της παραγοντικής ανάλυσης γράφεται ως εξής:

$$\mathrm{X} - \mu = LF + \varepsilon$$

Όπου

X είναι το διάνυσμα των αρχικών μεταβλητών μεγέθους $p \times 1$,

μ είναι το διάνυσμα των μέσων μεγέθους $p \times 1$,

L είναι ένας πίνακας $p \times k$ όπου το L_{ij} είναι η επιβάρυνση (loading) του παράγοντα L_j στη μεταβλητή X_i. Ονομάζεται πίνακας των επιβαρύνσεων (matrix of factor loadings),

F είναι ένα $k \times 1$ διάνυσμα με τους (κοινούς) παράγοντες (common factors) και

ε είναι το σφάλμα (error) ή ειδικός παράγοντας (specific factor). Το σφάλμα ε_i είναι ο μοναδικός παράγοντας της i μεταβλητής και είναι το μέρος της μεταβλητής το οποίο δεν μπορεί να εξηγηθεί από τους παράγοντες.

Μπορούμε να υποθέσουμε ότι όλες οι μεταβλητές έχουν μέσο μηδέν $(\mu = 0)$, οπότε διάνυσμα μ δεν χρειάζεται στο παραπάνω μοντέλο. Επίσης, είναι προφανές ότι $k < p$, δηλαδή ο αριθμός των παραγόντων πρέπει να είναι μικρότερος του αριθμού των μεταβλητών, γιατί αλλιώς θα ήταν χωρίς νόημα να γίνει παραγοντική ανάλυση. Σύμφωνα με τα παραπάνω υποθέτουμε ότι κάθε μεταβλητή μπορούμε να τη γράψουμε με τη μορφή:

$$\begin{aligned}
\mathrm{X}_1 &= L_{11}F_1 + L_{12}F_2 + ... + L_{1k}F_k + \varepsilon_1 \\
\mathrm{X}_1 &= L_{21}F_1 + L_{22}F_2 + ... + L_{2k}F_k + \varepsilon_2 \\
&... \\
\mathrm{X}_p &= L_{p1}F_1 + L_{p2}F_2 + ... + L_{pk}F_k + \varepsilon_p
\end{aligned}$$

Είναι σημαντικό να σημειωθεί ότι:

- Το παραπάνω μοντέλο, αν και μοιάζει, με ένα γραμμικό μοντέλο παλινδρόμησης , έχει μερικές διαφορές. Πρώτον, τα χ δεν είναι παρατηρήσεις αλλά μεταβλητές. Δεύτερον, το δεξί μέλος της εξίσωσης δεν είναι παρατηρήσιμο και έτσι πρέπει να εκτιμηθεί.
- Οι παράγοντες φ μπορούν να γραφούν και αυτοί σαν γραμμικός συνδυασμός των μεταβλητών. Αυτό είναι χρήσιμο να γίνεται, όταν θέλουμε αν δημιουργήσουμε νέες μεταβλητές. Αυτοί οι συντελεστές όμως διαφέρουν από τις επιβαρύνσεις. Οι συντελεστές κάθε μεταβλητής όταν εκφράζουμε κάθε παράγοντα σαν γραμμικό συνδυασμό των μεταβλητών καλούνται συντελεστές των σκορ (factor score coefficients).
- Οι παράγοντες έχουν την ίδια διακύμανση. Συνεπώς οι παράγοντες που προκύπτουν δεν είναι απαραιτήτως σε μια σειρά.
- Το μοντέλο αυτό προσπαθεί να εκφράσει τις μεταβλητές ως γραμμικό συνδυασμό των παραγόντων. [11]

4.3 Βήματα παραγοντικής ανάλυσης

Η παραγοντική ανάλυση ακολουθεί μια διαδικασία με διαδοχικά βήματα, τα οποία είναι τα εξής:

- Έλεγχος ύπαρξης ικανοποιητικού βαθμού συσχετίσεων για να πραγματοποιηθεί η παραγοντική ανάλυση.
- Εύρεση του αριθμού των παραγόντων και εκτίμηση των παραμέτρων του μοντέλου.
- Περιστροφή του μοντέλου με σκοπό να αυξήσουμε την ερμηνευτική του ικανότητα.
- Εκτίμηση των συντελεστών των παραγοντικών σκορ (factor score coefficients) για περαιτέρω στατιστική ανάλυση.

Για να προχωρήσουμε σε παραγοντική ανάλυση θα πρέπει να είμαστε σίγουροι ότι τα δεδομένα μας συσχετίζονται σε βαθμό τέτοιο έτσι ώστε να είναι δυνατή η όρυξη κοινών παραγόντων. Σε αντίθετη περίπτωση δεν είναι καθόλου φρόνιμο να προχωρήσουμε σε παραγοντική ανάλυση καθώς δεν θα είναι εφικτή η εξαγωγή συμπερασμάτων.

Στην στατιστική μελέτη συνηθίζουμε να ξεκινάμε την μελέτη μας από την εξέταση των δεδομένων μας περιγραφικά. Έτσι θα ξεκινήσουμε με την εκτίμηση των συσχετίσεων των μεταβλητών. Είναι επιθυμητό να έχουμε μεγάλες συσχετίσεις διότι αν είναι ασυσχέτιστα τα δεδομένα ή με μικρές συσχετίσεις τότε δεν θα βρούμε κοινούς παράγοντες που να μπορούμε να δουλέψουμε με αυτούς. Όταν αναφερόμαστε σε μεγάλες συσχετίσεις δεν εννοούμε να είναι στατιστικά σημαντικές οι συσχετίσεις (δηλαδή διάφορες του μηδενός), αλλά να πρόκειται για τιμές μεγαλύτερες σε απόλυτη τιμή από 0,40. Χρειάζονται μεγάλες συσχετίσεις τουλάχιστον σε μεγάλο ποσοστό του πίνακα συσχετίσεων. Σε οποιαδήποτε περίπτωση που δεν υπάρχουν τόσο μεγάλες συσχετίσεις δεν υπάρχει λόγος να συνεχίσουμε την παραγοντική ανάλυση, επιπλέον αν έχουμε κάποιες μεταβλητές με μικρές συσχετίσεις θα πρέπει να τις παραλείψουμε από την ανάλυση, αφού και στην πορεία θα αναγκαστούμε να τις αφαιρέσουμε γιατί θα αποτελούν ένα ξεχωριστό παράγοντα.

Για να ελέγξουμε την στατιστική σημαντικότητα ενός δειγματικού συντελεστή συσχέτισης, χρειάζεται να ελεγχθεί η μηδενική υπόθεση H_0 έναντι της εναλλακτικής H_α, ως εξής:

$$\mathrm{H}_0 : \rho = 0$$
$$\mathrm{H}_\alpha : \rho \neq 0$$

Υπολογίζοντας την ελεγχοσυνάρτηση $t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$

η οποία ακολουθεί την t κατανομή με $n-2$ βαθμούς ελευθερίας.

Στη συνέχεια ένα πολύ βασικό ερώτημα στην παραγοντική ανάλυση είναι ο καθορισμός του αριθμού των παραγόντων που θα χρησιμοποιηθούν. Ο αριθμός, όμως, δεν είναι γνωστός και υπάρχουν διάφοροι μέθοδοι για να εκτιμηθεί. Μία από τις βασικές μέθοδοι εκτίμησης των παραγόντων είναι η μέθοδος των κύριων συνιστωσών, η οποία περιγράφεται στη συνέχεια. [12]

4.4 Η αδράνεια

Ο όρος «Αδράνεια» προέρχεται από τη Μηχανική. Είναι γνωστό ότι κάθε φυσικό αντικείμενο έχει ένα κέντρο βάρους. Αν θεωρήσουμε ότι κάθε τμήμα του

αντικειμένου έχει μάζα m και απέχει απόσταση d από το κέντρο βάρους του, τότε η αδράνεια του αντικειμένου είναι ίση με το άθροισμα των ποσοτήτων md^2 για κάθε τμήμα του. Επομένως,

$$\text{Αδράνεια} = \sum md^2 .$$

Τα σημεία που αντιστοιχούν στα προφίλ των γραμμών (στηλών) και αποτελούν το νέφος των γραμμών (στηλών), θεωρούνται ως υλικά σημεία εφοδιασμένα με μάζα. Κατά συνέπεια, τα σημεία-προφίλ των γραμμών (στηλών) είναι εφοδιασμένα με μάζες, έχουν οριστεί οι μεταξύ τους αποστάσεις $\left(x^2\right)$ και ένα κέντρο βάρους (μέσο προφίλ). Συνεκδοχικά, κάθε σημείο γραμμής ή στήλης συνεισφέρει στην ολική αδράνεια του αντιστοίχου νέφους σημείων στο οποίο ανήκει, ανάλογα με τη μάζα του και την απόστασή του από το κέντρο βάρους του. Ειδικότερα, αν με g_r και g_c συμβολίσουμε το μέσο προφίλ γραμμών και στηλών αντίστοιχα, τότε τα τετράγωνα των αποστάσεων της i γραμμής και της j στήλης από τα αντίστοιχα κέντρα βάρους τους δίνονται από τις σχέσεις:

$$d^2_{x^2}\left(i, g_r\right) = \sum_{j=1}^{l} \frac{n}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{+j}}{n} \right)^2 = \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{n_{ij}}{n_{i+}} - c_j \right)^2$$

$$d^2_{x^2}\left(j, g_c\right) = \sum_{i=1}^{k} \frac{n}{n_{i+}} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{i+}}{n} \right)^2 = \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{n_{ij}}{n_{+j}} - r_i \right)^2$$

Σύμφωνα με τον ορισμό της Αδράνειας στο πλαίσιο της Μηχανικής, η αδράνεια I_r του νέφους των σημείων γραμμών θα δίνεται από τη σχέση:

$$I_r = \sum_i (\mu\acute{\alpha}\zeta\alpha \ i \ \gamma\rho\alpha\mu\mu\acute{\eta}\varsigma) \times d^2_{x^2}\left(i, g_r\right) = \sum_{i=1}^{k} r_i \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{n_{ij}}{n_{i+}} - c_j \right)^2 = \sum_{i=1}^{k} r_i \sum_{j=1}^{l} \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - c_j \right)^2$$

ενώ η αδράνεια I_c του νέφους των σημείων στηλών από τη σχέση:

$$I_c = \sum_j (\mu\acute{\alpha}\zeta\alpha \ j \ \gamma\rho\alpha\mu\mu\acute{\eta}\varsigma) \times d^2_{x^2}\left(j, g_c\right) = \sum_{j=1}^{l} c_j \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{n_{ij}}{n_{+j}} - r_i \right)^2 = \sum_{j=1}^{l} c_j \sum_{i=1}^{k} \frac{1}{r_i} \left(\frac{p_{ij}}{c_j} - r_i \right)^2$$

Από στατιστική σκοπιά, η ολική αδράνεια του νέφους των σημείων γραμμών ή στηλών μπορεί να οριστεί ως μια γενικευμένη διασπορά και πιο συγκεκριμένα, ως ο σταθμισμένος μέσος όρος των τετραγώνων των x^2 αποστάσεων των προφίλ γραμμών, ή ισοδύναμα των προφίλ στηλών, από το κέντρο τους. Γεωμετρικά, η αδράνεια μπορεί να θεωρηθεί ως ένα μέτρο της διασποράς των σημείων-προφίλ στον πολυδιάστατο χώρο στον οποίο ανήκουν. Όσο μεγαλύτερη είναι η αδράνεια, τόσο μεγαλύτερη είναι και η διασπορά των σημείων στο χώρο.

Αν συμβολίσουμε με:

D_r και D_c τους διαγώνιους πίνακες που έχουν ως στοιχεία τις μάζες των γραμμών και στηλών αντίστοιχα, δηλαδή $D_r = diag(r)$ και $D_c = diag(c)$, $\tilde{a}_i$ το προφίλ της i γραμμής, $\tilde{b}_j$ το προφίλ της j στήλης ($i = 1,...,k$ και $j = 1,...,l$) και **1**: το διάνυσμα (κατάλληλων ανά περίπτωση διαστάσεων) με στοιχεία $[1,1,...,1]^T$, τότε τα διανύσματα με στοιχεία τις μάζες γραμμών και στηλών μπορούν να δοθούν από τις σχέσεις:

$$r = P1 \text{ και } c = P^T 1,$$

Ενώ οι πίνακες των προφίλ Α και Β, μπορούν να γραφούν και ως εξής:

$$A = D_r^{-1}P = \begin{bmatrix} \tilde{a}_1^T \\ \tilde{a}_2^T \\ \vdots \\ \tilde{a}_k^T \end{bmatrix} \text{ και } B = D_c^{-1}P^T = \begin{bmatrix} \tilde{b}_1^T \\ \tilde{b}_2^T \\ \vdots \\ \tilde{b}_1^T \end{bmatrix}$$

Επίσης, μπορεί να δειχθεί ότι τα μέσα προφίλ (κέντρα βάρους) των προφίλ γραμμών και στηλών δίνονται αντίστοιχα από τις σχέσεις:

$$c = A^T r \text{ και } r = B^T c\,.$$

Με βάση τους παραπάνω συμβολισμούς η ολική αδράνεια του νέφους των σημείων γραμμών μπορεί να υπολογιστεί από τη σχέση:

$$I_r = \sum_i r_i \left(\tilde{a}_i - c\right)^T D_c^{-1} \left(\tilde{a}_i - c\right) = \sum_i r_i \sum_j \frac{\left(\frac{p_{ij}}{r_i} - c_j\right)^2}{c_j}, \qquad (1)$$

Ενώ η ολική αδράνεια του νέφους των σημείων στηλών από τη σχέση:

$$I_r = \sum_i r_i \left(\tilde{a}_i - c\right)^T D_c^{-1} \left(\tilde{a}_i - c\right) = \sum_i r_i \sum_j \frac{\left(\frac{p_{ij}}{r_i} - c_j\right)^2}{c_j}. \qquad (2)$$

Με τη βοήθεια των δύο αυτών σχέσεων αποδεικνύεται ότι η αδράνεια του νέφους των σημείων γραμμών είναι ίση με την αδράνεια του νέφους των σημείων στηλών, δηλαδή $I_r = I_c$.

Έτσι, μπορούμε να μιλάμε πλέον για «Ολική Αδράνεια» του πίνακα **N** η οποία δίνεται από τη σχέση:

$$\text{Ολική Αδράνεια } I = I_r = I_c = \sum_i \sum_j \frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} = \sum_i \sum_j \frac{\left(\frac{n_{ij}}{n} - \frac{n_{i+}}{n}\frac{n_{+j}}{n}\right)^2}{\frac{n_{i+}}{n}\frac{n_{+j}}{n}} \qquad (3)$$

Η Ολική Αδράνεια του πίνακα **N** συνδέεται υπολογιστικά με το γνωστό έλεγχο ανεξαρτησίας x^2 του Pearson και πιο συγκεκριμένα ισχύει:

$$I = \frac{Q}{n} = \varphi^2, \qquad (4)$$

όπου Q η τιμή του στατιστικού x^2 που υπολογίζεται κατά τον έλεγχο ανεξαρτησίας (ή ομοιογένειας) των δύο μεταβλητών και φ^2 ο συντελεστής συνάφειας μέσου τετραγώνου (mean square contingency coefficient) του Pearson.

Αν συμβολίσουμε με s_{ij} την ποσότητα $\frac{\left(p_{ij} - r_i c_j\right)}{\sqrt{r_i c_j}}$, τότε από τις σχέσεις (3) και (4) συνεπάγεται ότι:

$$I = \frac{Q}{n} = \varphi^2 = \sum_i \sum_j s_{ij}^2 \qquad (5)$$

Έστω **S** ο $k \times l$ πίνακας με στοιχεία τις ποσότητες s_{ij}. Ο **S** ονομάζεται πίνακας των «Σχετικών Τυποποιημένων Υπολοίπων». Κάθε στοιχείο του είναι η διαφορά της παρατηρούμενης σχετικής συχνότητας, που αντιστοιχεί στο κελί (i, j)

του πίνακα **N**, από την αντίστοιχη αναμενόμενη (θεωρητική) σχετική συχνότητα κάτω από την ισχύ της μηδενικής υπόθεσης της ανεξαρτησίας των δύο μεταβλητών, διαιρεμένη με την τετραγωνική ρίζα της θεωρητικής αυτής συχνότητας. Από την (5) είναι φανερό ότι ο πίνακας **S** σχετίζεται άμεσα με την ολική αδράνεια του **N** και γενικότερα με τη διασπορά των στοιχείων του. Αν τα στοιχεία του **S** έχουν μικρή απόλυτη τιμή, τότε η υπόθεση της ανεξαρτησίας είναι μάλλον ισχυρή και επομένως η αδράνεια του πίνακα αναμένεται να είναι μικρή. Στην περίπτωση της πλήρους ανεξαρτησίας των δύο μεταβλητών τα στοιχεία s_{ij} είναι ίσα με μηδέν και συνεπώς $Q = I = 0$. [9]

4.5 Παραγοντικοί άξονες

Με την παραγοντική ανάλυση προσδιορίζονται οι κύριοι άξονες της αδράνειας και για κάθε άξονα υπολογίζεται η αντίστοιχη ιδιοτιμή (eigenvalue), η οποία είναι ίση με την αδράνεια του νέφους προς την κατεύθυνση του αντίστοιχου άξονα. Ο πρώτος παραγοντικός άξονας είναι η ευθεία προς την κατεύθυνση της οποίας η αδράνεια του νέφους είναι μέγιστη. Ο δεύτερος παραγοντικός άξονας είναι η αμέσως επόμενη ευθεία, κάθετη στον πρώτο άξονα, για την οποία η αδράνεια του νέφους είναι επίσης μέγιστη. Το ίδιο συμβαίνει με τους υπόλοιπους άξονες. Η «μερική» αδράνεια (μάζα σημείου × απόσταση σημείου από την αρχή του άξονα εις το τετράγωνο) όλων των σημείων γραμμών (ή στηλών) κατά μήκος του άξονα ισοδυναμεί με την αδράνεια του άξονα.

Η αδράνεια ενός παραγοντικού άξονα είναι ο σταθμισμένος μέσος όρος των τετραγώνων των x^2 αποστάσεων των προβολών των σημείων γραμμών (ή στηλών) επί του άξονα, από το κέντρο βάρους τους. Με άλλα λόγια, είναι το μέτρο της διασποράς των σημείων γραμμών (ή στηλών) προς την κατεύθυνση του άξονα. Η αδράνεια ενός άξονα μπορεί να διασπαστεί στις μερικές αδράνειες κάθε σημείου επί του άξονα. Σημεία γραμμών (ή στηλών) με υψηλή συνεισφορά στην αδράνεια ενός παραγοντικού άξονα καθορίζουν σε μεγάλο βαθμό τον προσανατολισμό και την ταυτότητά του (δηλαδή τη φυσική του ερμηνεία).

Τα συνημίτονα των γωνιών που σχηματίζουν τα διανύσματα θέσης των σημείων γραμμών (ή στηλών) με τους παραγοντικούς άξονες εκφράζουν το βαθμό συσχέτισής τους με τους αντίστοιχους άξονες. Αποτελούν δείκτες ποιότητας της απεικόνισης των σημείων στον υποχώρο που προβάλλονται και εκφράζουν το πόσο

κοντά στην πραγματική τους θέση βρίσκεται η απεικόνισή τους στον επιλεγμένο υποχώρο.

Τέλος η στατιστική σημαντικότητα των παραγοντικών αξόνων έγκειται στον έλεγχο σημαντικότητας των αντίστοιχων ιδιοτιμών (αδρανειών).

Έστω λ_s η αδράνεια που αντιστοιχεί στον παραγοντικό άξονα $s\left(s=1,...,p\right)$, με $p=\min\left(k-1,l-1\right)$ και $k,\left(l\right)$ τις γραμμές (στήλες) του απλού πίνακα συμπτώσεων δύο μεταβλητών. Η αδράνεια αυτού του πίνακα μπορεί να υπολογιστεί με την παρακάτω σχέση:

$$I_N=\sum_{s=1}^{p}\lambda_s=\frac{Q}{n} \text{ με } \lambda_1\geq\lambda_2\geq...\geq\lambda_p,$$

όπου η ποσότητα Q αντιστοιχεί στο στατιστικό χ^2 που υπολογίζεται κάτω από την υπόθεση ανεξαρτησίας των δύο μεταβλητών και n είναι το μέγεθος του δείγματος.

Έτσι από τη προηγούμενη σχέση έχουμε:

$$Q=n\sum_{s}\lambda_s=n\lambda_1+n\lambda_2+...+n\lambda_p.$$

Τελικά η αδράνεια του συνολικού συστήματος ισούται με το άθροισμα των ιδιοτιμών και σε κάθε παραγοντικό άξονα i αντιστοιχεί αδράνεια ίση με $\frac{\lambda_i}{\sum\lambda_j}$ ή $100\frac{\lambda_i}{\sum\lambda_j}$. [9], [10]

4.5.1 Προβολή στους παραγοντικούς άξονες

Ο χώρος στον οποίον βρίσκεται το νέφος σημείων παρατηρήσεων εξαρτάται από το πλήθος των μεταβλητών που το χαρακτηρίζουν και που ορίζουν τη διάστασή του. Όπως είναι φυσικό, πολύ δύσκολα μπορούμε να φανταστούμε τη μορφή αυτού του νέφους. Αν δε θελήσουμε να το παρατηρήσουμε διαπιστώνουμε ότι αυτό είναι σχεδόν αδύνατο. Ο μόνος τρόπος που μπορούμε να έχουμε μια εικόνα του είναι να το προβάλλουμε σε κάποιο επίπεδο. Ποιο είναι το επίπεδο επί του οποίου έχουμε την αντιπροσωπευτικότερη προβολή;

Οι άξονες αδράνειας, που όπως είδαμε στην ανάλυση δεδομένων ονομάζονται παραγοντικοί άξονες, ορίζουν ανά δύο, επίπεδα που μας παρέχουν την καλύτερη δυνατή εικόνα του νέφους.

Επανερχόμαστε στην απλούστερη περίπτωση του ελλειψοειδούς του χώρου των τριών διαστάσεων όπου οι άξονες αδράνειας (παραγοντικοί άξονες) είναι ο μεγάλος, ο μεσαίος και ο μικρός άξονας του ελλειψοειδούς.

Οι δύο πρώτοι παραγοντικοί άξονες ορίζουν ένα επίπεδο, επί του οποίου η προβολή του νέφους είναι η πιο ενδιαφέρουσα που μπορούμε να έχουμε. Η ποιότητα αυτής της προβολής είναι δυνατό να μετρηθεί αν συγκρίνουμε την αδράνεια $\lambda_1 + \lambda_2$ που ερμηνεύει αυτό το πρώτο παραγοντικό επίπεδο, με τη συνολική αδράνεια $\lambda_1 + \lambda_2 + \lambda_3$ (όπου $\lambda_1, \lambda_2, \lambda_3$ οι αντίστοιχες ιδιοτιμές).

Η προβολή του νέφους στο πρώτο παραγοντικό επίπεδο μπορεί να θεωρηθεί σαν η καλύτερη φωτογραφία του νέφους των σημείων και ανταποκρίνεται τόσο περισσότερο στην πραγματικότητα όσο πιο μικρή είναι η τρίτη χαρακτηριστική τιμή (αδράνεια) λ_3 ως προς το άθροισμα $\lambda_1 + \lambda_2$ των δύο πρώτων.

Ακόμη και στις περιπτώσεις που η διάσταση του χώρου, στον οποίο βρίσκεται το νέφος των σημείων, είναι πολύ μεγαλύτερη του 3, συχνά οι δύο πρώτοι παραγοντικοί άξονες φέρουν επάνω τους το 80% ή και το 90% της ολικής αδράνειας.

Σ’ αυτές τις περιπτώσεις η προβολή του νέφους των σημείων των παρατηρήσεων στο πρώτο παραγοντικό επίπεδο παρέχει μια εξαιρετική αναπαράστασή του και δεν χρειάζεται να προσφύγουμε και σε άλλες. Αν όμως αυτό δεν συμβαίνει, ή υπάρχουν άλλοι ιδιαίτεροι λόγοι, μπορούμε να έχουμε προβολές και σ’ άλλα παραγοντικά επίπεδα που δημιουργούνται από τους παραγοντικούς άξονες αν τους πάρουμε ανά δύο με τη φθίνουσα σειρά τους. [2]

4.6 Ανάλυση σε Κύριες Συνιστώσες (ACP)

4.6.1 Μία πρώτη προσέγγιση

Η Ανάλυση σε Κύριες Συνιστώσες (ACP, Analyse en Composantes Principales) είναι μια τεχνική ανάλυσης δεδομένων με σκοπό τη δημιουργία καινούργιων μεταβλητών, οι οποίες είναι συνδυασμοί των αρχικών μεταβλητών, έτσι

ώστε να είναι ασυσχέτιστες μεταξύ τους και να περιέχουν όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών.

Οι νέες μεταβλητές που παράγονται ονομάζονται Κύριες Συνιστώσες. Το τι επιτυγχάνεται από τη μέθοδο αυτή είναι ότι από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών, το οποίο είναι χρήσιμο για αρκετές στατιστικές μεθόδους. Επίσης οι κύριες συνιστώσες που προκύπτουν μπορούν να ερμηνεύσουν το μεγαλύτερο ποσοστό της διακύμανσης, που σημαίνει πως καταλήγουμε σε ένα πιο μικρό ποσοστό της συνολικής μεταβλητότητας. Αυτό είναι πολύ σημαντικό ιδιαίτερα στις περιπτώσεις που έχουμε λίγες παρατηρήσεις και πολλές μεταβλητές. Συνεπώς, αν σε μια τέτοια περίπτωση θέλαμε να εφαρμόσουμε ένα (γενικευμένο) γραμμικό μοντέλο, η υπερπαραμετροποίηση του μοντέλου μπορεί να ξεπεραστεί χρησιμοποιώντας την παραπάνω μέθοδο. Η πρώτη παρουσίαση της μεθόδου έγινε από τον Pearson (1901), ενώ ο Hotelling (1933) την εξέλιξε σε σημαντικό βαθμό.

Η ACP λοιπόν, παρέχει τη δυνατότητα ανάλυσης και μελέτης πινάκων (αντικείμενα x μεταβλητές) των οποίων όλες οι μεταβλητές είναι ποσοτικές. Η ACP περιγράφει τα αντικείμενα του πίνακα που αναλύει, λαμβάνοντας υπόψη τις προσεγγίσεις αυτών μεταξύ τους, στο χώρο διάστασης που ορίζει το πλήθος των μεταβλητών, καθώς και τις μεταβλητές, λαμβάνοντας υπόψη τις μεταξύ τους συσχετίσεις, σε χώρο διάστασης όσο το σύνολο των αντικειμένων. Η ACP διά των αποτελεσμάτων της παρέχει την εποπτική εικόνα των αντικειμένων, των μεταβλητών καθώς και των σχέσεων σύνδεσης αυτών μεταξύ τους. Η εικόνα αυτή επιτυγχάνεται με την ταυτόχρονη παρουσίαση, ως σημείων (διανύσματα) των αντικειμένων και των μεταβλητών, σε επίπεδα (παραγοντικά επίπεδα – χώροι δύο διαστάσεων) που ορίζουν ανά δύο νέοι άξονες (κύριοι άξονες).

Το βασικότερο πλεονέκτημα της μεθόδου, όπως και των υπολοίπων της Ανάλυσης Δεδομένων, είναι ότι εφαρμόζεται χωρίς καμία a priori υπόθεση, για το ποιες μεταβλητές ή ποια αποτελέσματα παίζουν το σημαντικότερο ρόλο στο φαινόμενο που περιγράφει ο πίνακας που αναλύεται. Με την ACP, θεωρώντας ότι η αλληλεξάρτηση των μεταβλητών παίζει σημαντικό ρόλο στη στατιστική ανάλυση, αντικαθίστανται ομάδες παρατηρούμενων μεταβλητών με νέες (η κάθε ομάδα από μία μεταβλητή) τους παράγοντες.

Η συνολική συμπεριφορά των παραγόντων είναι ίδια με αυτή των αρχικών μεταβλητών, δηλαδή, η εικόνα του πίνακα των αρχικών δεδομένων (αντικείμενα *x* μεταβλητές) παραμένει αναλλοίωτη και στον πίνακα των παραγόντων (αντικείμενα *x* παράγοντες).

Οι παράγοντες είναι ανεξάρτητοι ανά δύο μεταξύ τους, αναδεικνύουν τη δομή του πίνακα δεδομένων και φανερώνουν την οργανική συνοχή των μεταβλητών, προσδιορίζοντας τις αντιπροσωπευτικότερες με τη μεγαλύτερη σημαντικότητα. Η ACP μπορεί να εφαρμοστεί και σαν είδος ενός ανιχνευτή που προειδοποιεί τόσο για το μη πραγματικό όσο και για το ασήμαντο που πρέπει να αποφύγουμε, αναδεικνύοντας με τις πρώτες αναλύσεις τις πραγματικές δομές που κρύβονται μέσα στην ποικιλία των αλληλοεπιδρώμενων μεταξύ τους μεταβλητών. Η ACP, βασιζόμενη κυρίως στις έννοιες της αλληλοσυσχέτισης μεταξύ των μεταβλητών και στην αντικατάστασή τους από νέες, αντικαθιστά τον πίνακα των συντελεστών συσχέτισης του Pearson με τον πίνακα των παραγόντων (συντεταγμένες των αντικειμένων στους κύριους άξονες – παράγοντες). Συνέπεια αυτού είναι αφενός μεν η πλήρης περιγραφή του φαινομένου με πίνακα πολύ λιγότερων στοιχείων (παράγοντες στη θέση των μεταβλητών), αφετέρου δε καθόσον οι κύριοι άξονες είναι ανεξάρτητοι – ασυσχέτιστοι – ορθογώνιοι μεταξύ τους, έχουν δικό τους όνομα – ταυτότητα και εκφράζουν ο καθένας μια ομάδα μεταβλητών με κοινά χαρακτηριστικά.

Οι κύριες συνιστώσες είναι διανύσματα που σχηματίζονται σαν γραμμικοί συνδυασμοί των μεταβλητών του συνόλου των δεδομένων και κατασκευάζονται έτσι ώστε να είναι κάθετα μεταξύ τους (στον πολυδιάστατο χώρο που ορίζουν ανάλογα με το πλήθος τους) και να αντιπροσωπεύουν κατά φθίνουσα τάξη ποσοστά της αρχικής μεταβλητότητας των δεδομένων. Το πλήθος τους είναι ίσο με το πλήθος των αρχικών μεταβλητών. Όμως, από κάποιο σημείο και μετά, πολλά από αυτά δεν είναι χρήσιμα γιατί δεν επεξηγούν κάποιο ποσοστό από την αρχική μεταβλητότητα. Αυτό συμβαίνει στην περίπτωση που οι μεταβλητές είναι έντονα συσχετισμένες μεταξύ τους.

Τα αποτελέσματα της τεχνικής των κυρίων συνιστωσών προκύπτουν μετά από ανάλυση του πίνακα συνδιακύμανσης ή του πίνακα συσχέτισης των δεδομένων. Με αυτόν τον τρόπο δεν χρησιμοποιούνται τα ίδια τα δεδομένα για την εξαγωγή συμπερασμάτων αλλά η "εσωτερική δομή συσχέτισής" τους. Τις περισσότερες φορές χρησιμοποιείται ο πίνακας συσχέτισης στον οποίο οι μεταβλητές με μεγάλη διασπορά

δεν έχουν βαρύτητα μεγαλύτερη από τις υπόλοιπες. Αυτό συμβαίνει γιατί η συσχέτιση είναι ουσιαστικά ένα τυποποιημένο μέτρο συνδιακύμανσης. Έτσι, η επιλογή των μονάδων μέτρησης των μεταβλητών δεν παίζει κανένα ρόλο όταν τα δεδομένα δεν περιέχουν ομοειδείς μεταβλητές. Όταν όλες οι μεταβλητές που περιέχονται στα δεδομένα εκφράζονται με την ίδια μονάδα μέτρησης, τότε συνήθως χρησιμοποιείται ο πίνακας συνδιακύμανσης. Όσον αφορά την επιλογή του πίνακα συνδιακύμανσης ή του πίνακα συσχέτισης για την ανάλυση του προβλήματος, η δεύτερη περίπτωση αντιστοιχεί στην παραδοχή ότι οι μεταβλητές είναι της ίδιας βαρύτητας. Αντίθετα, η χρησιμοποίηση του πίνακα συνδιακύμανσης οδηγεί στη μεροληπτική σε σχέση και με το μέγεθος της διασποράς της κάθε μεταβλητής (πέρα από τα μεγέθη συνδιακύμανσης) επιλογή του βάρους που έχει κάθε μεταβλητή στον σχηματισμό της κάθε κύριας συνιστώσας. [2],[11],[16]

4.6.2 Εκτίμηση με τη μέθοδο των Κύριων Συνιστωσών

Η εκτίμηση με τη μέθοδο των κύριων συνιστωσών βασίζεται στην φασματική ανάλυση του πίνακα διακύμανσης (συσχέτισης) Σ. Ο πίνακας Σ αποτελείται από ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων (λ_i, e_i) με $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Οπότε

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_1 e_2 e_2' + \ldots + \lambda_p e_p e_p' = \left[\sqrt{\lambda_1} e_1 \vdots \sqrt{\lambda_2} e_2 \vdots \cdots \vdots \sqrt{\lambda_p} e_p\right] \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix} \qquad (1)$$

Η παραπάνω αναπαράσταση αφορά το μοντέλο της παραγοντικής ανάλυσης, το οποίο έχει τόσους παράγοντες όσες και μεταβλητές $(m = p)$ και ειδική διακύμανση $\psi_i = 0$ για όλα τα i. Τα φορτία του j κοινού παράγοντα ισούνται με τους συντελεστές της j κύριας συνιστώσας πολλαπλασιασμένους με $\sqrt{\lambda_j} e_j$. Συνεπώς, μπορούμε να γράψουμε ότι

$$\underset{(p\times p)}{\Sigma} = \underset{(p\times p)}{L}\underset{(p\times p)}{L'} + \underset{(p\times p)}{0} = LL' \qquad (2)$$

Όμως, η αναπαράσταση του πίνακα Σ στην παραπάνω σχέση δεν είναι ιδιαίτερα χρήσιμη γιατί δημιουργεί τόσους κοινούς παράγοντες όσες είναι οι μεταβλητές και δεν επιτρέπει καμία μεταβολή των ειδικών παραγόντων ε της σχέσης

$X-\mu=LF+\varepsilon$. Προτιμούνται μοντέλα τα οποία εξηγούν τη διακύμανση χρησιμοποιώντας λίγους παράγοντες. Μία προσέγγιση είναι, όταν οι $p-m$ τελευταίες ιδιοτιμές είναι μικρές, να αγνοούμε τη συνεισφορά των $\lambda_1 e_1 e_1^{'}+\lambda_1 e_2 e_2^{'}+...+\lambda_p e_p e_p^{'}$ στον Σ στη σχέση (1). Αγνοώντας αυτή τη συνεισφορά, καταλήγουμε στον πίνακα

$$\Sigma=\underset{(p\times m)}{L}\underset{(p\times m)}{L^{'}} \tag{3}$$

Η αναπαράσταση (3) υποθέτει ότι οι ειδικοί παράγοντες ε στη σχέση $X-\mu=LF+\varepsilon$ είναι αμελητέας σημαντικότητας και μπορούν, επίσης, να αγνοηθούν στην παραγοντοποίηση του πίνακα Σ. Αν συμπεριληφθούν οι ειδικοί παράγοντες στο μοντέλο, οι διακυμάνσεις τους θα θεωρηθούν ότι είναι διαγώνια στοιχεία του $\Sigma=LL^{'}$. Θεωρούμε τότε την προσέγγιση

$$\Sigma\approx LL^{'}+\Psi$$

$$=\left[\sqrt{\lambda_1}e_1\vdots\sqrt{\lambda_2}e_2\vdots\cdots\vdots\sqrt{\lambda_m}e_m\right]\begin{bmatrix}\sqrt{\lambda_1}e_1^{'}\\ \sqrt{\lambda_2}e_2^{'}\\ \vdots\\ \sqrt{\lambda_m}e_m^{'}\end{bmatrix}+\begin{bmatrix}\psi_1 & 0 & \cdots & 0\\ 0 & \psi_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \psi_p\end{bmatrix} \tag{4}$$

Όπου $\psi_i=\sigma_{ii}-\sum_{j=1}^{m}l_{ij}^2$ για i από $1,2,...,p$.

Η αναπαράσταση της σχέσης (4), όταν εφαρμόζεται στον δειγματικό πίνακα διακύμανσης ή στον δειγματικό πίνακα συσχέτισης, είναι γνωστή ως η λύση των κύριων συνιστωσών. Συνεπώς, η παραγοντική ανάλυση με τη μέθοδο των κύριων συνιστωσών για τον δειγματικό πίνακα διακύμανσης είναι η εξής:

Έστω $\left(\hat{\lambda}_1,\hat{e}_1\right),\left(\hat{\lambda}_2,\hat{e}_2\right),...,\left(\hat{\lambda}_p,\hat{e}_p\right)$, όπου $\hat{\lambda}_1\geq\hat{\lambda}_2\geq...\geq\hat{\lambda}_p$. Τότε στο μοντέλο με $m>p$ αριθμό των κοινών παραγόντων, εκτιμούμε τον πίνακα των εκτιμώμενων φορτίων των παραγόντων $\left\{\tilde{l}_{ij}\right\}$, ο οποίος δίνεται από

$$\hat{L}=\left[\sqrt{\hat{\lambda}_1}\hat{e}_1\vdots\sqrt{\hat{\lambda}_2}\hat{e}_2\vdots\cdots\vdots\sqrt{\hat{\lambda}_m}\hat{e}_m\right]$$

Και τις ειδικές διασπορές $\tilde{\psi}_{ii} = s_{ii} - \sum_{j=1}^{m} \tilde{l}_{ij}^2$

Οι εταιρικότητες υπολογίζονται ως εξής:

$\tilde{h}_i^2 = \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + ... + \tilde{l}_{im}^2$. [11]

4.7 Στροφή συστήματος συντεταγμένων στο χώρο, με πίνακα ορθογώνιο

4.7.1 Εισαγωγή

Γενικά, φυσικά μεγέθη που περιγράφονται από διατεταγμένες τριάδες αριθμών και έχουν την ιδιότητα να μετασχηματίζονται κάτω από στροφές του συστήματος συντεταγμένων όπως και οι συντεταγμένες της θέσης, λέγονται διανύσματα. Φυσικά μεγέθη που περιγράφονται από έναν μοναδικό αριθμό και παραμένουν αναλλοίωτα κάτω από στροφές λέγονται μονόμετρα ή βαθμωτά μεγέθη. Για παράδειγμα στην περίπτωση του νόμου του Νεύτωνα η δύναμη και η επιτάχυνση είναι διανύσματα επειδή μετασχηματίζονται κάτω από στροφές όπως και οι μετατοπίσεις, ενώ η μάζα παραμένει αναλλοίωτη. Ο διανυσματικός χαρακτήρας βασικών φυσικών ποσοτήτων είναι πειραματικό αποτέλεσμα.

Μια και η στροφή ενός συστήματος συντεταγμένων παίζει καθοριστικό ρόλο στην κατάταξη ενός φυσικού μεγέθους σαν διάνυσμα ή βαθμωτό, ας ξεκινήσουμε με την μαθηματική περιγραφή και τις ιδιότητες των μετασχηματισμών στροφής. Αυτό θα μας βοηθήσει να εισάγουμε και τους τανυστές που αποτελούν ένα γενικότερο/ευρύτερο μαθηματικό αντικείμενο το οποίο περιλαμβάνει τα βαθμωτά, διανυσματικά και άλλα πιο περίπλοκα φυσικά μεγέθη που χρειάζονται περισσότερες συνιστώσες για την περιγραφή τους.

Ένας μετασχηματισμός στροφής του συστήματος συντεταγμένων αναπαρίσταται μαθηματικά από τον πίνακα στροφής. Αυτός ο πίνακας είναι ορθογώνιος και όταν πολλαπλασιάσει ένα διάνυσμα δίνει ένα άλλο διάνυσμα με διαφορετική κατεύθυνση αλλά το ίδιο μέτρο με το αρχικό, όπως ακριβώς ο αντίστοιχος γεωμετρικός μετασχηματισμός στροφής έχει σαν αποτέλεσμα την αλλαγή της κατεύθυνσης του διανύσματος αλλά όχι και του μέτρου του. [17]

4.7.2 Στροφή συστήματος συντεταγμένων στον χώρο

Από εδώ και στο εξής για απλότητα στον συμβολισμό και στις πράξεις θα συμβολίζουμε με $a_i, i=1,2,3$ τις συνιστώσες ενός διανύσματος $\vec{A}$ σε ένα καρτεσιανό σύστημα συντεταγμένων xyz στον χώρο (δηλαδή $a_1 = a_x, a_2 = a_y$ και $a_3 = a_z$) και με $\hat{e}_i, i=1,2,3$ τα αντίστοιχα μοναδιαία διανύσματα κατά τους άξονες Ox, Oy και Oz (δηλαδή $\hat{e}_1 = \hat{i}, \hat{e}_2 = \hat{j}$ και $\hat{e}_3 = \hat{k}$). Επομένως το διάνυσμα $\vec{A} = a_x\hat{i} + a_y\hat{j} + a_z\hat{k}$ θα γράφεται σαν

$$\vec{A} = a_1\hat{e}_1 + a_2\hat{e}_2 + a_3\hat{e}_3 = \sum_{i=1}^{3} a_i\hat{e}_i \, .$$

Ανάλογα, με a_i' και $\hat{e}_i', i=1,2,3$ θα συμβολίσουμε τα αντίστοιχα μεγέθη ως προς ένα σύστημα συντεταγμένων $x'y'z'$, που προκύπτει από το αρχικό xyz με στροφή γύρω από κάποιον άξονα που περνάει από την κοινή αρχή O των δύο συστημάτων συντεταγμένων.

Ας θεωρήσουμε το σύστημα $x'y'z'$ που προκύπτει από μια στροφή του xyz γύρω από τον άξονα Oz κατά γωνία θ. Τότε ο αντίστοιχος μετασχηματισμός στροφής είναι ο $U = U_z(\theta)$, που δίνεται από τον πίνακα

$$U_z(\theta) = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Αν $A = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$ και $A' = \begin{pmatrix} a_1' \\ a_2' \\ a_3' \end{pmatrix}$ είναι τα διανύσματα στήλες με τις αντίστοιχες συντεταγμένες ενός διανύσματος $\vec{A} = a_1\hat{e}_1 + a_2\hat{e}_2 + a_3\hat{e}_3 = a_1'\hat{e}_1' + a_2'\hat{e}_2' + a_3'\hat{e}_3'$ ως προς τα δύο συστήματα συντεταγμένων, τότε αυτά συνδέονται μεταξύ τους με τον μετασχηματισμό στροφής:

$$A' = UA \Leftrightarrow a_i' = \sum_{j=1}^{3} U_{ij}a_j, \text{ για κάθε } i = 1,2,3$$

Στην γενική περίπτωση που έχουμε στροφή του συστήματος συντεταγμένων ως προς τυχαίο άξονα (όχι γύρω από τον άξονα Oz), η στροφή προσδιορίζεται πλήρως από τις τρεις γωνίες Euler (α,β,γ) και ο αντίστοιχος πίνακας του μετασχηματισμού στροφής είναι ο

$$U(\alpha,\beta,\gamma)=\begin{pmatrix}\cos\alpha\cos\gamma-\sin\alpha\cos\beta\sin\gamma & -\sin\alpha\cos\gamma-\cos\alpha\cos\beta\sin\gamma & \sin\beta\sin\gamma\\ \cos\alpha\sin\gamma+\sin\alpha\cos\beta\cos\gamma & -\sin\alpha\sin\gamma+\cos\alpha\cos\beta\cos\gamma & -\sin\beta\cos\gamma\\ \sin\alpha\sin\beta & \cos\alpha\sin\beta & \cos\beta\end{pmatrix}$$ [17]

4.7.3 Ιδιότητες μετασχηματισμών στροφής

Είναι εύκολο να διαπιστώσει κανείς ότι ο πίνακας $U=U_z(\theta)$ που δίνεται στην προηγούμενη παράγραφο είναι ορθογώνιος, δηλαδή $U_z^T(\theta)U_z(\theta)=U_z(\theta)U_z^T(\theta)=I$, όπου $U_z^T(\theta)$ είναι ο ανάστροφος του $U_z(\theta)$ και I ο ταυτοτικός πίνακας. Οι ορθογώνιοι μετασχηματισμοί αφήνουν αναλλοίωτη την απόσταση μεταξύ δύο σημείων κι επομένως διατηρούν σταθερό το μέτρο ενός διανύσματος. Αυτές είναι ιδιότητες που έχει κάθε μετασχηματισμός στροφής. Εν γένει, οποιοσδήποτε μετασχηματισμός στροφής περιγράφεται από έναν ορθογώνιο πίνακα. (Το αντίστροφο δεν ισχύει, δηλαδή όλοι οι ορθογώνιοι πίνακες δεν αντιστοιχούν σε μετασχηματισμούς στροφής).

Αν λοιπόν ο πίνακας

$$U=\begin{pmatrix}u_{11} & u_{12} & u_{13}\\ u_{21} & u_{22} & u_{23}\\ u_{31} & u_{32} & u_{33}\end{pmatrix}$$

περιγράφει κάποιον μετασχηματισμό στροφής θα είναι υποχρεωτικά ορθογώνιος και θα ισχύει

$$U^T=U^{-1}\Leftrightarrow U^TU=UU^T=I,$$

όπου U^T ο ανάστροφος του U (που προκύπτει από την εναλλαγή των γραμμών με τις στήλες) και U^{-1} ο αντίστροφος του U (που το γινόμενό τους δίνει τον ταυτοτικό πίνακα).

Οι ορθογώνιοι πίνακες (κι επομένως οι μετασχηματισμοί στροφής) έχουν την ιδιότητα τα διανύσματα που έχουν σαν συνιστώσες οποιαδήποτε από τις στήλες τους να είναι ορθοκανονικά. Το ίδιο ισχύει και για τα διανύσματα που έχουν σαν συνιστώσες τις γραμμές τους. Αυτό εκφράζεται από τις σχέσεις:

$$\sum_{i=1}^{3} U_{ij} U_{ik} = \delta_{jk} \text{ για κάθε } j,k = 1,2,3$$

$$\sum_{i=1}^{3} U_{ji} U_{ki} = \delta_{jk} \text{ για κάθε } j,k = 1,2,3$$

όπου δ_{jk} είναι το σύμβολο δέλτα του Kronecker που ισούται με 1 αν $j = k$ και -1 αν $j \neq k$:

$$\delta_{jk} = \begin{cases} 0, j \neq k \\ 1, j = k \end{cases}$$

Η πρώτη από τις πιο πάνω σχέσεις εκφράζει την ορθοκανονικότητα των στηλών και η δεύτερη των γραμμών του ορθογώνιου πίνακα στροφής U. Λόγω αυτών των ιδιοτήτων, ένας ορθογώνιος πίνακας έχει μόνο τρία ανεξάρτητα στοιχεία.

Επειδή ο U είναι ορθογώνιος ισχύει $U^T U = I \Rightarrow \det\left(U^T U\right) = 1$ και επειδή είναι $\det\left(U^T\right) = \det\left(U\right)$ και $\det\left(AB\right) = \det\left(A\right)\det\left(B\right)$ προκύπτει ότι η ορίζουσα κάθε ορθογώνιου πίνακα ισούται με ± 1. Οι μετασχηματισμοί στροφής δίνονται μονάχα από ορθογώνιους πίνακες με ορίζουσα $\det\left(U\right) = +1.$ (Οι ορθογώνιοι μετασχηματισμοί με ορίζουσα -1 αντιστοιχούν σε ανακλάσεις και όχι σε στροφές). [17]

5. Πολυδιάστατη στατιστική ανάλυση δεδομένων

Υποθέτουμε ότι έχουμε στη διάθεσή μας να μελετήσουμε, ένα σύνολο δεδομένων n παρατηρήσεων, που η κάθε μια χαρακτηρίζεται από τις τιμές που λαμβάνει σε P αριθμητικές μεταβλητές $X_1, X_2, ..., X_p$.

Η τιμή που εμφανίζεται στον πίνακα δεδομένων για τη μεταβλητή X_j στην παρατήρηση i θα συμβολίζεται με x_{ij}.

Η ανάλυση σε κύριες συνιστώσες είναι μια στατιστική μέθοδος που έχει σαν κύριο σκοπό να περιγράψει τον πίνακα δεδομένων $X\left(n,p\right)=x_{ij}$.

Θεωρούμε την κάθε γραμμή i του πίνακα X, δηλαδή την καθεμία από τις n παρατηρήσεις, σαν ένα σημείο στο χώρο των P διαστάσεων. Οι τιμές που παίρνει η παρατήρηση i στις P μεταβλητές $\left(x_i : i=1,2,...,P\right)$ θεωρούνται ως οι συντεταγμένες του αντίστοιχου σημείου στους P άξονες αυτού του χώρου. Το σύνολο λοιπόν, των πληροφοριών που μας παρέχει ο πίνακας των δεδομένων X, μπορεί να παρομοιαστεί με ένα νέφος n σημείων στο χώρο R^p. Σκοπός μας είναι να κατορθώσουμε να παρατηρήσουμε το νέφος των σημείων σ' έναν χώρο πολύ λιγότερων διαστάσεων από τις αρχικές, έστω $m\left(m<P\right)$.

Για να το επιτύχουμε αυτό, προσπαθούμε να ορίσουμε ένα γραμμικό συνδυασμό m διαστάσεων που να διέρχεται όσο το δυνατό πλησιέστερα από το κέντρο μάζας του αρχικού νέφους των σημείων των δεδομένων με την παρακάτω έννοια.

«Η αδράνεια του νέφους των σημείων ως προς αυτόν τον γραμμικό συνδυασμό να είναι ελάχιστη».

Δηλαδή, ο μέσος όρος των τετραγώνων των αποστάσεων των σημείων x_i του νέφους από τον γραμμικό αυτό συνδυασμό να είναι ελάχιστος και το κάθε σημείο x_i να απεικονίζεται σ' αυτόν τον γραμμικό συνδυασμό από την προβολή του.

Αν θεωρήσουμε $m=0$ τη διάσταση, αυτό σημαίνει ότι αναζητούμε ένα σημείο που να βρίσκεται όσο το δυνατό πλησιέστερα στο κέντρο των σημείων, που δεν είναι άλλο από το κέντρο μάζας G (ή κέντρο βάρους) του νέφους.

$$G=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Αν θεωρήσουμε $m=1$, αυτό σημαίνει ότι αναζητούμε μια ευθεία που να διέρχεται όσο το δυνατό πλησιέστερα από το κέντρο μάζας G του νέφους των σημείων. Αυτή η ευθεία Δ_1 θα διέρχεται από το κέντρο μάζας του νέφους G και ονομάζεται πρώτος κύριος άξονας του νέφους. Αν λάβουμε ως αρχή αυτού του άξονα το G, έχουμε δημιουργήσει την πρώτη κύρια συνιστώσα Y_1.

Αν στη συνέχεια θεωρήσουμε $m = 2$, τη διάσταση του επιθυμητού γραμμικού συνδυασμού, προσπαθούμε να ορίσουμε ένα επίπεδο που να διέρχεται όσο το δυνατόν πλησιέστερα στο κέντρο μάζας G του νέφους. Αυτό το επίπεδο διέρχεται από το κέντρο μάζας G και περιέχει τον πρώτο κύριο άξονα Δ_1. Για να ορίσουμε λοιπόν, πλήρως το επίπεδο αρκεί να ορίσουμε μια δεύτερη ευθεία Δ_2, διερχόμενη από το κέντρο μάζας και ορθογώνια στον πρώτο κύριο άξονα Δ_1. Αυτή η δεύτερη ευθεία Δ_2 είναι ο δεύτερος κύριος άξονας. Οι συντεταγμένες των προβολών των σημείων (παρατηρήσεων) x_i επί του δεύτερου κύριου άξονα Δ_2, λαμβάνονται πάλι με αρχή το κέντρο μάζας G, που συμπίπτει πλέον με την αρχή των αξόνων και δημιουργούν τη δεύτερη κύρια συνιστώσα Y_2.

Σαν συμπέρασμα του ορισμού των δύο πρώτων κύριων αξόνων Δ_1 και Δ_2 μπορούμε να αναφέρουμε ότι:

Ο καλύτερος τρόπος για να παρουσιάσουμε εποπτικά τις πληροφορίες του πίνακα των δεδομένων X σε ένα επίπεδο, είναι να λάβουμε σαν επίπεδο αυτό που σχηματίζουν οι δύο πρώτοι κύριοι άξονες Δ_1 και Δ_2 και να θεωρήσουμε τις παρατηρήσεις x_i στις θέσεις των προβολών τους σ' αυτό.

Επειδή οι κύριες συνιστώσες δεν είναι συσχετισμένες μεταξύ τους, μας επιτρέπουν τη μελέτη της θέσης των προβολών των σημείων x_i πάνω σ' αυτές, καθώς και των συσχετισμών μεταξύ των αρχικών μεταβλητών x_i και των κύριων συνιστωσών Y_i.

Με τον τρόπο αυτόν, πετυχαίνουμε να περιορίσουμε το σοβαρό πρόβλημα που δημιουργείται, όταν οι διαστάσεις του πίνακα των δεδομένων που μελετάμε είναι μεγάλες.

Τα δεδομένα μας λοιπόν που θα μελετήσουμε βρίσκονται στον παρακάτω πίνακα X. Το κυρίαρχο στοιχείο του πίνακα X, είναι ότι τα στοιχεία του δεν είναι αριθμοί που εκφράζουν συχνότητα αλλά ποσότητες τελείως ανομοιογενείς (μπορεί να έχουμε για παράδειγμα εισόδημα σε χιλιάδες ευρώ, απόσταση σε χιλιόμετρα και έναν δείκτη). Έχουμε δηλαδή 3 μεταβλητές, κάθε στήλη αντιπροσωπεύει και μια μεταβλητή, αλλά με διαφορετικές μονάδες η κάθε μία. Το άθροισμα των στοιχείων

μιας γραμμής δεν έχει κανένα νόημα, η δε σύγκριση των αθροισμάτων των τριών στηλών δεν μπορεί να μας δείξει τίποτε, πέρα από τις κλασικές μέσες τιμές των τριών μεταβλητών.

Έτσι ο πίνακας X των δεδομένων είναι:

$$X = \begin{bmatrix} 20 & 75 & 184 \\ 350 & 120 & 170 \\ 120 & 110 & 180 \\ 170 & -20 & -180 \\ 370 & 10 & 132 \end{bmatrix}$$

Όπου $m = 3$ είναι ο αριθμός των στηλών – μεταβλητών και $n = 5$ είναι ο αριθμός των γραμμών – σημείων.

Γίνεται φανερό ότι για να μπορέσουμε να μελετήσουμε αυτόν τον πίνακα δεδομένων, θα πρέπει να προσαρμόσουμε την κάθε μεταβλητή στην τυπική κατανομή $N(0,1)$.

Πρώτα λοιπόν κανονικοποιούμε τις τιμές της κάθε μεταβλητής, αφαιρώντας από αυτές την αντίστοιχη μέση τιμής τους και στη συνέχεια τις προσαρμόζουμε στη $N(0,1)$ διαιρώντας τη διαφορά $X_i - \overline{X}$ με την τυπική απόκλιση s_x.

Κάναμε δηλαδή κανονικοποίηση του πίνακα (στατιστική κανονικοποίηση) με την εντολή standardize, ώστε το αποτέλεσμα να έχει μέση τιμή 0 και τυπική απόκλιση 1. Εμείς όμως θέλουμε τη μαθηματική κανονικοποίηση γι'αυτό πολλαπλασιάζουμε με $\frac{1}{\sqrt{n-1}}$.(Μετατρέπει τα διανύσματα σε μοναδιαία με μήκος 1).

Έτσι παίρνουμε τον κανονικοποιημένο πίνακα X_0:

$$X_0 = \begin{bmatrix} -0.61686 & 0.12969 & 0.27764 \\ 0.47757 & 0.49445 & 0.23286 \\ -0.28521 & 0.41339 & 0.26485 \\ -0.11939 & -0.64035 & -0.88666 \\ 0.54389 & -0.39718 & 0.11131 \end{bmatrix}$$

Ο οποίος, όπως έχουμε αναφέρει έχει τις εξής ιδιότητες:

- Οι στήλες του είναι οι αρχικές μεταβλητές με αλλαγμένες τιμές, οι στήλες έχουν μέση τιμή 0 και μήκος 1, δηλαδή για οποιαδήποτε από τις μεταβλητές – στήλες το άθροισμα των στοιχείων της είναι μηδέν.
- Όλα τα στοιχεία του X_0 είναι μεταξύ -1 και 1
- Είναι καθαροί αριθμοί, απαλλαγήκαμε από τις μονάδες μέτρησης των αριθμών
- Το άθροισμα των τετραγώνων όλων των στοιχείων του πίνακα είναι μονάδα (έγιναν μοναδιαία διανύσματα οι μεταβλητές)

Μπορούμε να λέμε ότι:

Η ανάλυση σε κύριες συνιστώσες συνίσταται κυρίως, στη μετατροπή του πίνακα δεδομένων σε πίνακα με στήλες που ακολουθούν την τυπική κανονική κατανομή $N(0,1)$ και στην ανάλυση των δύο νεφών.

Γενικά, στην ανάλυση σε κύριες συνιστώσες όλες οι μεταβλητές j και όλες οι παρατηρήσεις i θεωρούνται ότι έχουν την ίδια μάζα (βαρύτητα). Γι' αυτόν το λόγο ο πίνακας δεδομένων X μετατρέπεται σε πίνακα συντεταγμένων. Καθώς ο πίνακας των δεδομένων έχει περισσότερες γραμμές από στήλες, ο πίνακας της αδράνειας V θα έχει διαστάσεις που συμπίπτουν με το πλήθος των μεταβλητών.

Στην περίπτωσή μας θα είναι ένας 3×3 πίνακας.

Έτσι έχουμε τον πίνακα V, δηλαδή τον πίνακα αδράνειας που είναι ίσος με το γινόμενο του ανάστροφου X_0 επί του X_0 δηλαδή $V = X_0T.X_0$

$$V = X_0T.X_0 = \begin{pmatrix} 1 & -0.10135 & 0.03081 \\ -0.10135 & 1 & 0.78419 \\ 0.03081 & 0.78419 & 1 \end{pmatrix}$$

Για τον πίνακας V ισχύει:

- Είναι συμμετρικός δηλαδή τα διαγώνια στοιχεία του είναι μονάδες και τα στοιχεία του που είναι συμμετρικά ως προς την κύρια διαγώνιο είναι ίσα.
- Οι ιδιοτιμές είναι πραγματικοί αριθμοί και τα ιδιοδιανύσματα του.
- Ο πίνακας μετάβασης του είναι ορθογώνιος, δηλαδή οι στήλες του αποτελούν ορθοκανονική βάση του χώρου που ανήκει ο πίνακας. (Τα ιδιοδιανύσματα του

είναι μοναδιαία και κάθετα μεταξύ τους, αποτελούν δηλαδή ορθοκανονικό σύστημα).

- Η Jordan μορφή είναι διαγώνια, η αλγεβρική και γεωμετρική πολλαπλότητα των ιδιοδιανυσμάτων και ιδιοτιμών είναι ίσες.

Ο πίνακας των γωνιών Φ_{ij}° (όπου Φ_{ij} είναι η γωνία των μεταβλητών $Var[i]$, $Var[j]$) στο χώρο των 5 διαστάσεων είναι:

$$\begin{pmatrix} \Phi_{ij}^{\circ} & Var[1] & Var[2] & Var[3] \\ Var[1] & 0.00 & 95.82 & 88.23 \\ Var[2] & 95.82 & 0.00 & 38.35 \\ Var[3] & 88.23 & 38.35 & 0.00 \end{pmatrix}$$

- Όταν οι μεταβλητές είναι κάθετες μεταξύ τους σημαίνει είναι γραμμικά ανεξάρτητες.
- Όταν σχηματίζουν γωνία 0° ως 20° είναι έντονα θετικά συσχετισμένες.
- Όταν σχηματίζουν γωνία 180° είναι έντονα αρνητικά συσχετισμένες.

Έπειτα υπολογίζουμε ιδιοτιμές και ιδιοδιανύσματα (μοναδιαία και κάθετα μεταξύ τους) του V.

Οι ιδιοτιμές είναι:

$$L_1 = 1.78737$$
$$L_2 = 1.00782$$
$$L_3 = 0.204805$$

ενώ τα ιδιοδιανύσματα είναι :

$$v_1 = \{0.0636691, -0.708344, -0.70299\}$$
$$v_2 = \{0.991096, -0.0376598, 0.127709\}$$
$$v_3 = \{-0.116937, -0.704862, 0.699639\}$$

Ο πίνακας V αντιστοιχεί στην ανάλυση των $n(5)$ παρατηρήσεων στο χώρο των $k(3)$ μεταβλητών. Έχουμε, λοιπόν, για το παράδειγμά μας, 3 ορθογώνιους μεταξύ τους άξονες:

Τον άξονα της μεταβλητής $Var[1]$, τον άξονα της μεταβλητής $Var[2]$ και τον άξονα της μεταβλητής $Var[3]$. Με το να μετατρέψουμε τον αρχικό πίνακα των δεδομένων X στον κανονικοποιημένο X_0, ουσιαστικά μεταφέρουμε την αρχή των συντεταγμένων 0 στο κέντρο μάζας G του νέφους.

Καθώς στον πίνακα X_0 το άθροισμα των 5 συντεταγμένων της κάθε μεταβλητής είναι ίσο με το 0, μπορούμε να συμπεράνουμε ότι στο χώρο των 5 διαστάσεων το κάθε σημείο – στήλη θα βρίσκεται επί ενός υποχώρου, ορθογωνίου στο διάνυσμα με συντεταγμένες $(1,1,1,1,1)$.

Επίσης, καθώς το άθροισμα των τετραγώνων των 5 συντεταγμένων ισούται με 1, το σημείο στήλη (μεταβλητή) θα βρίσκεται στην επιφάνεια μιας «σφαίρας» με κέντρο την αρχή $G=0$ και με ακτίνα ίση με 1. Δηλαδή, όταν έχουμε, τρεις μόνο μεταβλητές, τότε αυτές βρίσκονται επί μιας περιφέρειας, που είναι η τομή της σφαίρας με κέντρο $0=G$ και ακτίνα 1, με επίπεδο κάθετο στο διάνυσμα $(1,1,1)$.

Στη συνέχεια παίρνουμε τη Jordan μορφή του V

$$Jordan[V]=\begin{pmatrix} 1.78737 & 0 & 0 \\ 0 & 1.00782 & 0 \\ 0 & 0 & 0.204805 \end{pmatrix}$$

Η Jordan μορφή έχει στη διαγώνιο της τις ιδιοτιμές.

Μπορώ να χρησιμοποιήσω αντί για τον V την Jordan μορφή του, ότι ιδιοτιμές έχει ο V έχει και η Jordan μορφή του. (Απλούστερη δυνατή μορφή του πίνακα V). Κάθε πίνακας έχει συγκεκριμένη αδράνεια που είναι το άθροισμα των ιδιοτιμών του, για τον $Jordan[V]$ προσθέτω διαγώνια τα στοιχεία του και βρίσκω αδράνεια ίση με 3.

Για κάθε παραγοντικό άξονα έχουμε βρει τη χαρακτηριστική τιμή (ιδιοτιμή) λ_i που τον ορίζει και που συγχρόνως δίνει την αδράνεια του νέφους κατά μήκος του, καθώς και το ποσοστό αδράνειας του άξονα αυτού ως προς τη συνολική αδράνεια το οποίο μετρά την ποιότητα της απεικόνισης των προβολών των σημείων του νέφους επ' αυτού.

Έτσι για τον πρώτο παραγοντικό άξονα, και αφού η μέγιστη ιδιοτιμή λ_i είναι $\lambda_1 = 1.78737$, έχουμε ότι :

$$\frac{\lambda_1}{\sum \lambda_j} = \frac{1.78737}{3} = 0.59579$$

Για τον δεύτερο παραγοντικό άξονα, αφού η ιδιοτιμή λ_2 είναι $\lambda_2 = 1.00782$, έχουμε ότι : $\frac{\lambda_2}{\sum \lambda_j} = \frac{1.00782}{3} = 0.33594$

Και τέλος για τον τρίτο παραγοντικό άξονα, αφού η ιδιοτιμή λ_3 είναι $\lambda_3 = 0.204805$ έχουμε ότι: $\frac{\lambda_3}{\sum \lambda_j} = \frac{0.204805}{3} = 0.0682683$.

Πιο αναλυτικά:

Βλέπουμε δηλαδή ότι ο πρώτος άξονας ερμηνεύει το 59.5792% της συνολικής πληροφορίας, ο δεύτερος το 33.594% και τέλος ο τρίτος το 6.82683%.

Ο πρώτος άξονας με τον δεύτερο σχηματίζουν το πρώτο παραγοντικό επίπεδο, επί του οποίου απεικονίζονται οι επικρατέστερες απομακρύνσεις από τη μέση κατάσταση που είναι και το κέντρο μάζας G το οποίο συμπίπτει με την αρχή των παραγοντικών αξόνων. Αν κρατήσουμε τους δύο πρώτους άξονες το ποσοστό τις πληροφορίας που χάνουμε είναι πάρα πολύ μικρό (6.82683%).

Δηλαδή το επίπεδο που δημιουργείται από τους δύο πρώτους άξονες αδράνειας ερμηνεύει το 93.1732% της ολικής αδράνειας

$$59.5792 + 33.594 = 93.1732\%$$

Αυτό το 93.1732% της ολικής αδράνειας μας επιτρέπει να συμπεράνουμε ότι το νέφος των σημείων στο χώρο των τριών διαστάσεων (που αντιστοιχούν στις τρεις μεταβλητές) είναι πάρα πολύ πεπλατυσμένο και, με σφάλμα 6.82683%, μπορούμε να το θεωρήσουμε επίπεδο. Η απεικόνισή του λοιπόν στο επίπεδο των δύο πρώτων κύριων αξόνων (αξόνων αδράνειας) είναι πάρα πολύ ικανοποιητική.

Στη συνέχεια έχουμε τον πίνακα μετάβασης U

$$U = \begin{pmatrix} 0.06367 & 0.99110 & -0.11694 \\ -0.70834 & -0.03766 & -0.70486 \\ -0.70299 & 0.12771 & 0.69964 \end{pmatrix}$$

Για τον οποίο ισχύει ότι :

- Οι στήλες του είναι τα ιδιοδιανύσματα του V .
- Είναι ορθογώνιος πίνακας, δηλαδή τα ιδιοδιανύσματα του είναι μοναδιαία και κάθετα ανά δύο μεταξύ τους.
- Όταν πολλαπλασιάζω οτιδήποτε με αυτόν τον πίνακα βρίσκω τις συντεταγμένες καινούριας βάσης.

Με τον πίνακα U αλλάζουμε βάση πάμε σε μια καινούρια για να έχουμε καλύτερη ορατότητα, καλύτερη οπτική γωνία, χωρίς να αλλάζει τίποτα.

Στη συνέχεια έχω τον πίνακα UX_0 που περιέχει τις καινούριες συντεταγμένες του X_0 ως προς την καινούρια βάση.

$$UX_0 = \begin{bmatrix} -0.32632 & -0.58079 & 0.17497 \\ -0.48353 & 0.48443 & -0.24145 \\ -0.49717 & -0.26442 & -0.07274 \\ 1.06930 & -0.20745 & -0.15502 \\ 0.23772 & 0.56822 & 0.29424 \end{bmatrix}$$

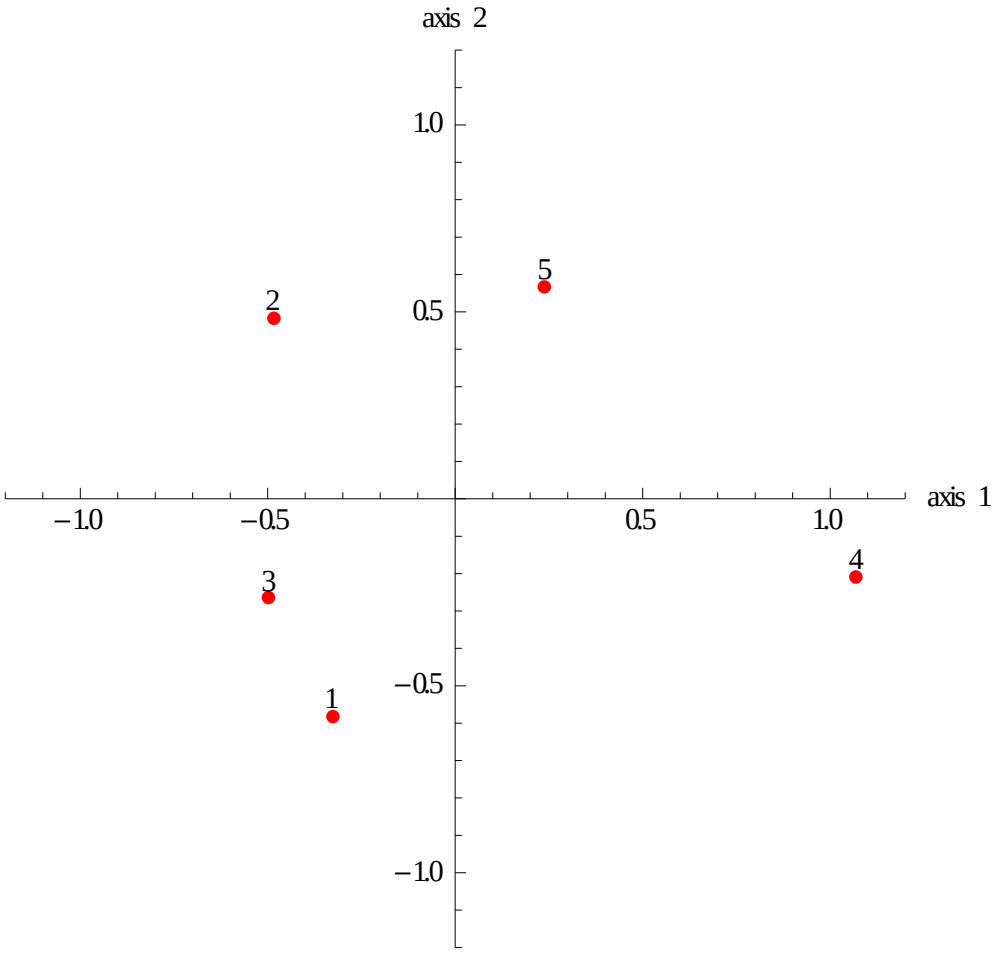
Από αυτόν τον πίνακα κρατάω τις δύο πρώτες στήλες, δηλαδή τις δύο πρώτες συντεταγμένες, αυτές που αντιστοιχούν στους δύο πρώτους άξονες. (Προβάλουμε τις μεταβλητές στο πρώτο παραγοντικό επίπεδο).

Ακολουθεί η χαρτογράφηση των σημείων στο πρώτο παραγοντικό επίπεδο.

Έτσι οι συντεταγμένες των σημείων στο πρώτο παραγοντικό επίπεδο είναι:

i	$x[1]$	$x[2]$
1	−0.32632	−0.58079
2	−0.48353	0.48443
3	−0.49717	−0.26442
4	1.06930	−0.20745
5	0.23772	0.56822

Διάγραμμα των σημείων χωρίς την κανονικοποίηση



Ακολουθεί η χαρτογράφηση των μεταβλητών στο πρώτο παραγοντικό επίπεδο

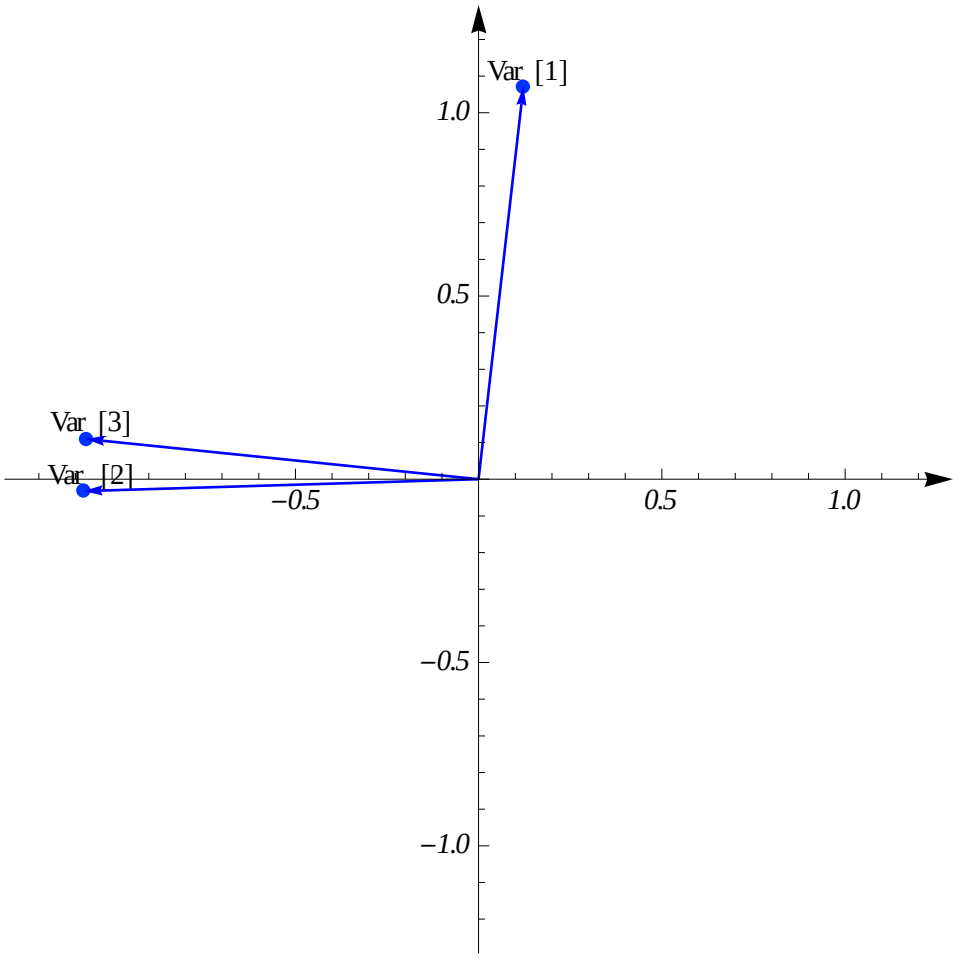
Οι συντεταγμένες των μεταβλητών στο πρώτο παραγοντικό επίπεδο είναι:

i	$x[1]$	$x[2]$
1	0.11380	0.99885
2	−1.26608	−0.03795
3	−1.25651	0.12871

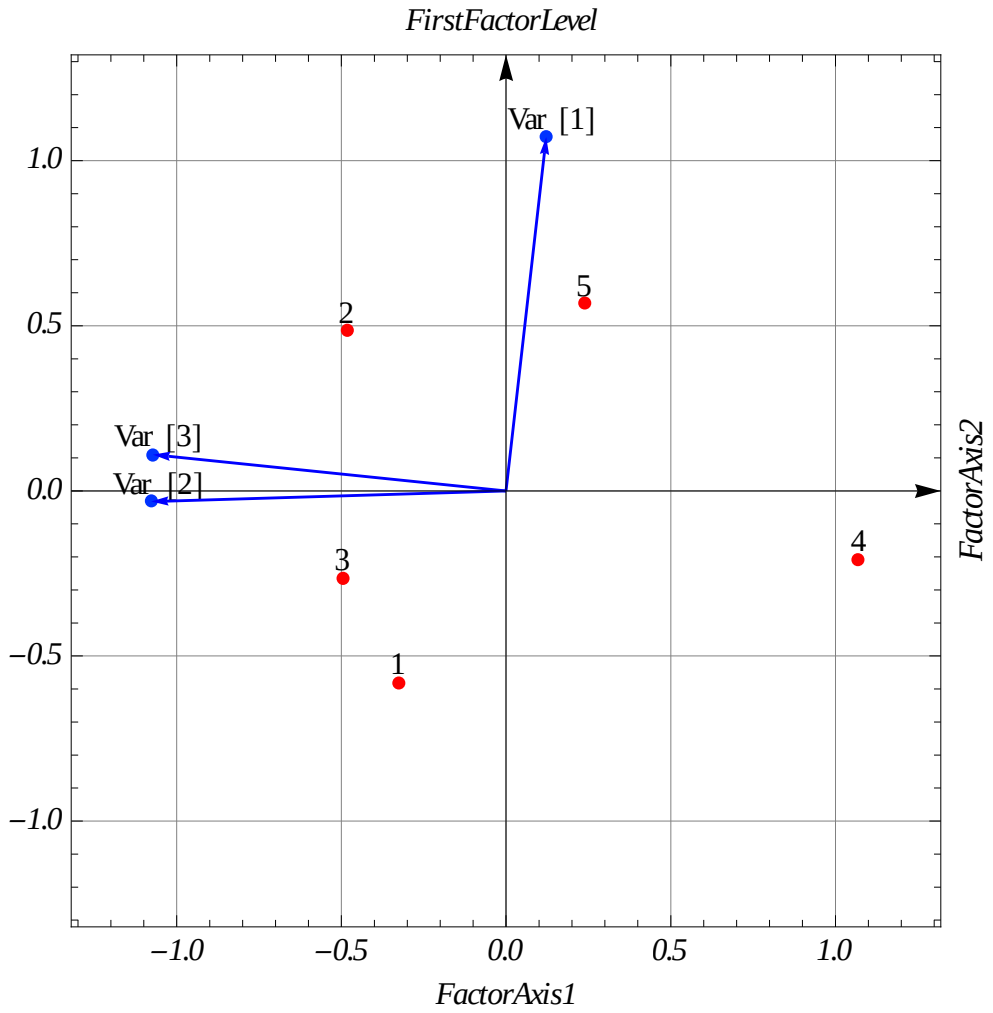
Ο πίνακας των γωνιών f_{ij}° των μεταβλητών ανά δύο, στο πρώτο παραγοντικό επίπεδο:

$$\begin{pmatrix} f_{ij}^{\circ} & Var[1] & Var[2] & Var[3] \\ Var[1] & 0.00 & 98.22 & 90.65 \\ Var[2] & 98.22 & 0.00 & 7.57 \\ Var[3] & 90.65 & 7.57 & 0.00 \end{pmatrix}$$

Διάγραμμα των μεταβλητών με κανονικοποίηση:



Χαρτογράφηση των σημείων και των μεταβλητών στο πρώτο παραγοντικό επίπεδο



Δύο μεταβλητές που βρίσκονται κοντά η μία στην άλλη είναι έντονα συσχετισμένες μεταξύ τους. Στο παράδειγμά μας, όπως διαπιστώνουμε από το παραπάνω σχήμα, η μεταβλητή $Var[3]$ είναι έντονα θετικά συσχετισμένη με τη μεταβλητή $Var[2]$. Αντίθετα δύο μεταβλητές που σχηματίζουν μεταξύ τους γωνία περίπου 90° είναι γραμμικά ανεξάρτητες μεταξύ τους. Στην περίπτωσή μας η $Var[1]$ είναι γραμμικά ανεξάρτητη της $Var[2]$ και της $Var[3]$.

Η συσχέτιση των μεταβλητών με τους κύριους άξονες παρουσιάζει ιδιαίτερο ενδιαφέρον, καθόσον μας επιτρέπει να εξηγήσουμε τους άξονες, δηλαδή να προσδιορίσουμε την ερμηνεία τους.

Επειδή οι δύο μεταβλητές $Var[2]$ και $Var[3]$ είναι έντονα συσχετισμένες με τον δεύτερο κύριο άξονα $FactorAxis2$, μπορούμε να τον θεωρήσουμε σαν μια μεταβλητή που εκφράζει και τις δύο συγχρόνως.

Η μεταβλητή $Var[1]$, όπως διαπιστώσαμε, είναι γραμμικά ανεξάρτητη των άλλων δύο. Επίσης είναι έντονα συσχετισμένη με τον πρώτο κύριο άξονα $FactorAxis1$ που θεωρούμε ότι εκφράζει και επ' αυτού μπορούμε να μελετήσουμε τις διαφορές που παρουσιάζουν ως προς αυτόν οι παρατηρήσεις.

Όταν οι μεταβλητές είναι μόνο τρεις, τότε είμαστε σχεδόν βέβαιοι ότι η απεικόνιση επί του πρώτου κύριου επιπέδου είναι πολύ κοντά στην πραγματικότητα.

Όταν όμως οι μεταβλητές είναι περισσότερες από τρεις, τότε χρειάζεται προσοχή, γιατί είναι δυνατό, δύο μεταβλητές να έχουν επί του επιπέδου προβολές γειτονικές, ενώ η γωνία που σχηματίζουν μεταξύ τους να είναι αρκετά μεγάλη.

Στην ανάλυση σε κύριες συνιστώσες δεν πρέπει να λαμβάνουμε υπόψη την απόσταση του σημείου στήλη (μεταβλητή) από την αρχή, αλλά τη διεύθυνση που ορίζει το διάνυσμα της θέσης του. Αυτή η διεύθυνση βοηθά στην ερμηνεία της θέσης των σημείων γραμμών (παρατηρήσεις) ως προς τις μεταβλητές.

Αν στο παραπάνω σχήμα τα κόκκινα σημεία (παρατηρήσεις) είναι πάνω σε μια μεταβλητή, τότε αυτά χαρακτηρίζονται από την μεταβλητή και την χαρακτηρίζουν. Τέλος τα κόκκινα σημεία είναι κοντά μεταξύ τους, άρα δεν έχουμε διαφοροποιήσεις.

Πίνακας αριθμητικών αποτελεσμάτων για τις μεταβλητές

i	$x[1]$	$x[2]$	$\lvert Var_i \rvert$	$Cos\Phi_i$	Φ_i°	$\Phi_i rad$
1	0.113801	0.998847	1.005594	0.999761	1.364681	0.023818
2	−1.266076	−0.037954	1.274845	0.993568	6.501930	0.113480
3	−1.256507	0.128708	1.271183	0.993627	6.472215	0.112961

Όπου:

$x[1], x[2]:$ είναι οι συντεταγμένες των μεταβλητών (Var) στο πρώτο παραγοντικό επίπεδο.

$\lvert Var_i \rvert:$είναι οι νόρμες των μεταβλητών (Var) στον τρισδιάστατο χώρο

$\Phi_i:$είναι οι γωνίες των μεταβλητών (Var) με το πρώτο παραγοντικό επίπεδο.

Για να διαπιστώσουμε αν οι μεταβλητές είναι κοντά κοιτάμε τα συνημίτονα των γωνιών και τα μήκη τους. Έτσι από τον παραπάνω πίνακα βλέπουμε ότι και οι τρεις

μεταβλητές είναι κοντά άρα είναι γραμμικά εξαρτημένες, δηλαδή τη μία μπορούμε να τη διαγράψουμε.

Πίνακας των γωνιών Φ_{ij}° των μεταβλητών ανά δύο στον χώρο των 5 διαστάσεων.

$$\begin{pmatrix} \Phi_{ij}^{\circ} & Var[1] & Var[2] & Var[3] \\ Var[1] & 0.00 & 95.82 & 88.23 \\ Var[2] & 95.82 & 0.00 & 38.35 \\ Var[3] & 88.23 & 38.35 & 0.00 \end{pmatrix}$$

Πίνακας των γωνιών f_{ij}° των μεταβλητών ανά δύο, στο πρώτο παραγοντικό επίπεδο:

$$\begin{pmatrix} f_{ij}^{\circ} & Var[1] & Var[2] & Var[3] \\ Var[1] & 0.00 & 98.22 & 90.65 \\ Var[2] & 98.22 & 0.00 & 7.57 \\ Var[3] & 90.65 & 7.57 & 0.00 \end{pmatrix}$$

Αριθμητικά αποτελέσματα των σημείων

$$\begin{pmatrix} i & x[1] & x[2] & \left|Point_i\right| & Cos\theta_i & \theta_i^{\circ} & \theta_i rad \\ 1 & -0.326319 & -0.580790 & 0.688777 & 0.967198 & 14.715746 & 0.256838 \\ 2 & -0.483532 & 0.484431 & 0.725791 & 0.943045 & 19.430707 & 0.339130 \\ 3 & -0.497167 & -0.264418 & 0.567788 & 0.991761 & 7.360143 & 0.128459 \\ 4 & 1.069300 & -0.207447 & 1.100212 & 0.990024 & 8.099807 & 0.141368 \\ 5 & 0.237719 & 0.568225 & 0.682615 & 0.902332 & 25.533632 & 0.445646 \end{pmatrix}$$

Όπου:

$x[1], x[2]:$ είναι οι συντεταγμένες των σημείων στο πρώτο παραγοντικό επίπεδο.

$\left|Point_i\right|:$ είναι οι νόρμες των σημείων στον τρισδιάστατο χώρο

$\theta_i:$ είναι οι γωνίες των σημείων με το πρώτο παραγοντικό επίπεδο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Π.Χ.Γ Βασιλείου, Γ. Τσακλίδης, Θεσσαλονίκη 2003, Εφαρμοσμένη Θεωρία Πινάκων, Εκδόσεις ΖΗΤΗ

2. Γιάννης Παπαδημητρίου, Ιανουάριος 2007, Η Ανάλυση Δεδομένων, Τυπωθήτω Γιώργος Δαρδανός

3. Σ. Μ. Μποζαπαλίδη, Θεσσαλονίκη 1997, Εισαγωγή στην Γραμμική Άλγεβρα

4. Ν. Κ. Στεφανίδη, Θεσσαλονίκη 1993, Εισαγωγή στη Γεωμετρία

5. Γρηγόρης Καμβύσας, Μαρία Χατζηνικολάου, Πάτρα 2005, Γενικά Μαθηματικά ΙΙ, Ελληνικό Ανοικτό Πανεπιστήμιο

6. Ευκλείδεια Γεωμετρία Α΄ και Β΄ Γενικού Λυκείου, Οργανισμός Εκδόσεων Διδακτικών Βιβλίων, Αθήνα

7. Επιστημονικό περιοδικό αειχώρος, Ειδικό τεύχος – Αφιέρωμα Γεωπληροφορική, τόμος 37, τεύχος 3, σελίδες 157-158

8. Ανδρέας Λ. Πετράκης, Κοζάνη 2008, Γραμμικά Στοχαστικά Υποδείγματα, Θεωρία και Εφαρμογές, Δεύτερη έκδοση

9. Άγγελος Ι. Μάρκος, Βοήθεια στην Ερμηνεία των Αποτελεσμάτων της Παραγοντικής Ανάλυσης των Αντιστοιχιών & Αλγόριθμοι Κατασκευής και Ανάλυσης Ειδικών Πινάκων Εισόδου, Η Περίπτωση του Λογισμικού CHIC Analysis, Διδακτορική Διατριβή, Πανεπιστήμιο Μακεδονίας, Θεσσαλονίκη 2006

10. Γεώργιος Χ. Μενεξές, Πειραματικοί Σχεδιασμοί στην Ανάλυση Δεδομένων, Διδακτορική Διατριβή, Πανεπιστήμιο Μακεδονίας, Θεσσαλονίκη 2006

11. Ναταλία Δ. Μαύρου, Μέθοδοι Πολυμεταβλητής Στατιστικής Ανάλυσης και Εφαρμογές, Διπλωματική Εργασία, Αθήνα, Απρίλιος 2012

12. Ελένη Κούπα, Στατιστικές Μέθοδοι σε Ψυχομετρικά Δεδομένα, Μεταπτυχιακή Διπλωματική Εργασία, Αθήνα, Ιανουάριος 2008

13. Χρήστος Κουρουνιώτης, Σημειώσεις μαθήματος Εισαγωγή στη Γραμμική Άλγεβρα, Τμήμα Μαθηματικών, Πανεπιστήμιο Κρήτης, 2006

14. Χρήστος Κουρουνιώτης, Σημειώσεις μαθήματος Επίπεδο και Χώρος, Τμήμα Μαθηματικών, Πανεπιστήμιο Κρήτης, 2009

15. http://el.wikipedia.org/wiki/%CE%95%CF%85%CE%BA%CE%BB%CE%B5%CE%AF%CE%B4%CE%B5%CE%B9%CE%BF_%CE%B4%CE%B9%CE%AC%CE%BD%CF%85%CF%83%CE%BC%CE%B1#.CE.9C.CE.AD.CF.84.CF.81.CE.BF_.CE.B4.CE.B9.CE.B1.CE.BD.CF.8D.CF.83.CE.BC.CE.B1.CF.84.CE.BF.CF.82 [20 Αυγούστου 2013]

16. Κωνσταντίνος Φωκιανός & Χαράλαμπος Χαραλάμπους, Εισαγωγή στην R, πρόχειρες σημειώσεις, Τμήμα Μαθηματικών & Στατιστικής, Πανεπιστήμιο Κύπρου, Ιανουάριος 2010

17. http://www.arnos.gr/dmdocuments/aei/panepistimio.patras/episthmhs.ylikwn/elastikotita/shmeiwseis.elastikothta.episthmhs.ylikvn.pan.patras.pdf
[20 Αυγούστου 2013]