

Πολυμεταβλητή Ανάλυση Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA)

Λουκία Μελιγκοτσίδου

Τμήμα Μαθηματικών, ΕΚΠΑ

1 Εισαγωγή

Η κεντρική ιδέα Η ανάλυση κυρίων συνιστωσών (PCA) έχει ως στόχο την **ελάττωση** της διάστασης (αριθμός μεταβλητών) ενός συνόλου δεδομένων που αποτελείται από ένα μεγάλο σύνολο **συσχετισμένων** μεταβλητών

Η κεντρική ιδέα Η ανάλυση κυρίων συνιστωσών (PCA) έχει ως στόχο την **ελάττωση** της διάστασης (αριθμός μεταβλητών) ενός συνόλου δεδομένων που αποτελείται από ένα μεγάλο σύνολο **συσχετισμένων** μεταβλητών

Πώς επιτυγχάνεται αυτό; Μέσω ενός γραμμικού μετασχηματισμού των μεταβλητών προκύπτουν νέες μεταβλητές (οι κύριες συνιστώσες), οι οποίες

- 1 είναι ασυσχέτιστες
- 2 έχουν την **ίδια συνολική διασπορά** με τις αρχικές

και επιπλέον

Η κεντρική ιδέα Η ανάλυση κυρίων συνιστωσών (PCA) έχει ως στόχο την **ελάττωση** της διάστασης (αριθμός μεταβλητών) ενός συνόλου δεδομένων που αποτελείται από ένα μεγάλο σύνολο **συσχετισμένων** μεταβλητών

Πώς επιτυγχάνεται αυτό; Μέσω ενός γραμμικού μετασχηματισμού των μεταβλητών προκύπτουν νέες μεταβλητές (οι κύριες συνιστώσες), οι οποίες

- 1 είναι ασυσχέτιστες
- 2 έχουν την **ίδια συνολική διασπορά** με τις αρχικές

και επιπλέον

- Ελπίζουμε ότι ένας μικρός αριθμός εξ αυτών εξηγεί το μεγαλύτερο μέρος της διασποράς των αρχικών μεταβλητών.
- Εάν συμβαίνει αυτό → ίσως μπορούμε να αντικαταστήσουμε τα αρχικά δεδομένα με ένα μικρό αριθμό κυρίων συνιστωσών (χωρίς να χάσουμε πολλή πληροφορία).

- 1 Σε μία μελέτη για την αντίδραση των ασθενών με καρκίνο στη χημειοθεραπεία έχουμε μετρήσει 10 μεταβλητές. Επειδή είναι δύσκολο να δουλέψουμε και με τις 10 για να αξιολογήσουμε την επίδραση κάποιου φαρμάκου θα θέλαμε να συμπυκνώσουμε την πληροφορία από τις 10 μεταβλητές σε λιγότερες, η PCA μπορεί να βοηθήσει σε αυτό.

- 1 Σε μία μελέτη για την αντίδραση των ασθενών με καρκίνο στη χημειοθεραπεία έχουμε μετρήσει 10 μεταβλητές. Επειδή είναι δύσκολο να δουλέψουμε και με τις 10 για να αξιολογήσουμε την επίδραση κάποιου φαρμάκου θα θέλαμε να συμπυκνώσουμε την πληροφορία από τις 10 μεταβλητές σε λιγότερες, η PCA μπορεί να βοηθήσει σε αυτό.
- 2 Για να μετρήσουμε την ικανοποίηση των πελατών θα πρέπει να μετρήσουμε διαφορετικές οπτικές της ικανοποίησης και να τις συνθέσουμε σε έναν ενδεχομένως δείκτη.

- 1 Σε μία μελέτη για την αντίδραση των ασθενών με καρκίνο στη χημειοθεραπεία έχουμε μετρήσει 10 μεταβλητές. Επειδή είναι δύσκολο να δουλέψουμε και με τις 10 για να αξιολογήσουμε την επίδραση κάποιου φαρμάκου θα θέλαμε να συμπυκνώσουμε την πληροφορία από τις 10 μεταβλητές σε λιγότερες, η PCA μπορεί να βοηθήσει σε αυτό.
- 2 Για να μετρήσουμε την ικανοποίηση των πελατών θα πρέπει να μετρήσουμε διαφορετικές οπτικές της ικανοποίησης και να τις συνθέσουμε σε έναν ενδεχομένως δείκτη.
- 3 Συνήθως οι φωτογραφίες αποθηκεύονται σε RGB format και επομένως αν μπορούμε να τις συμπιέσουμε σε λιγότερες μεταβλητές θα ήταν μεγάλο κέρδος. Πόσο θα χάσουμε όμως σε ποιότητα;

Φασματική ανάλυση συμμετρικού πίνακα $\mathbf{A}_{p \times p}$:

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$$

όπου $\mathbf{\Lambda}_{p \times p}$ διαγώνιος πίνακας με τις ιδιοτιμές του $\mathbf{A}$

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

$\mathbf{E}_{p \times p}$ ορθογώνιος ($\mathbf{E}\mathbf{E}^T = \mathbf{I}_p$) από τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών. Προκύπτει ότι:

$$\mathbf{\Lambda} = \mathbf{E}^T \mathbf{A} \mathbf{E}$$

Αν ξεκινήσουμε από έναν τετραγωνικό πίνακα $\mathbf{A}$ μπορούμε να καταλήξουμε σε έναν διαγώνιο πίνακα $\mathbf{\Lambda}$.

Χρησιμότητα: Ανάλυση πίνακα διακύμανσης Σ

Η πρώτη κύρια συνιστώσα

Έστω $\mathbf{X} = (X_1, X_2, \dots, X_p)$ τυχαίο διάνυσμα με $\text{Var}(\mathbf{X}) = \Sigma_{p \times p}$ θετικά ορισμένο.

Πρώτη κύρια συνιστώσα

Γραμμικός συνδυασμός

$$Y_1 = \mathbf{a}^\top \mathbf{X} = a_1 X_1 + \dots + a_p X_p$$

με την **μέγιστη** δυνατή διασπορά.

Παρατηρήσεις:

- Η διασπορά αυτή είναι μη φραγμένη ποσότητα
- Περιοριζόμαστε σε γραμμικούς συνδυασμούς που ικανοποιούν κάποια συνθήκη
- Συνηθέστερη συνθήκη: $\mathbf{a}^\top \mathbf{a} = 1$

Η πρώτη κύρια συνιστώσα μιας κατανομής (συνέχεια)

Επειδή

$$\text{Var}(Y_1) = \text{Var}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \Sigma \mathbf{a},$$

ζητάμε να μεγιστοποιήσουμε την $\mathbf{a}^\top \Sigma \mathbf{a}$ ως προς $\mathbf{a}$ υπό τον περιορισμό $\|\mathbf{a}\| = \mathbf{a}^\top \mathbf{a} = 1$.

Μεγιστοποίηση τετραγωνικών μορφών

Αποδεικνύεται ότι (δείτε τις σημειώσεις της γραμμικής άλγεβρας) το μέγιστο αυτό ισούται με λ_1 , την μέγιστη ιδιοτιμή του Σ , και επιτυγχάνεται για $\mathbf{a} = \mathbf{e}_1$, το ιδιοδιάνυσμα που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή.

Μεγιστοποίηση τετραγωνικών μορφών

Έστω $\mathbf{A}_{p \times p}$ ένας θετικά ορισμένος πίνακας με ιδιοτιμές

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$$

και τα αντίστοιχα (κανονικοποιημένα) ιδιοδιανύσματα $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$. Τότε ισχύουν τα εξής

$$\max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_1 \quad (\text{για } \mathbf{x} = \mathbf{e}_1)$$

$$\max_{\mathbf{x} \neq \mathbf{0}, \mathbf{x} \perp \mathbf{e}_1, \dots, \mathbf{e}_k} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \lambda_{k+1} \quad (\text{για } \mathbf{x} = \mathbf{e}_{k+1}), \quad k = 1, 2, \dots, p-1.$$

Η πρώτη κύρια συνιστώσα μιας κατανομής (ορισμός)

Ορισμός

Η πρώτη κύρια συνιστώσα μιας κατανομής είναι ο γραμμικός συνδυασμός $Y_1 = \mathbf{e}_1^T \mathbf{X}$ με διασπορά $Var(Y_1) = \lambda_1$, όπου λ_1 η μέγιστη ιδιοτιμή του $\mathbf{\Sigma}$ και $\mathbf{e}_1$ το (κανονικοποιημένο) ιδιοδιάνυσμα που αντιστοιχεί σε αυτήν την ιδιοτιμή.

Από φασματική ανάλυση του $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \sum_{i=1}^p \lambda_i \mathbf{e}_i \mathbf{e}_i^T$$

$$\begin{aligned} Var(Y_1) &= Var(\mathbf{e}_1^T \mathbf{X}) = \mathbf{e}_1^T \mathbf{\Sigma} \mathbf{e}_1 = \mathbf{e}_1^T \sum_{i=1}^p \lambda_i \mathbf{e}_i \mathbf{e}_i^T \mathbf{e}_1 \\ &= \sum_{i=1}^p \lambda_i (\mathbf{e}_1^T \mathbf{e}_i) (\mathbf{e}_i^T \mathbf{e}_1) = \lambda_1 (\mathbf{e}_1^T \mathbf{e}_1)^2 = \lambda_1 \end{aligned}$$

επειδή $\mathbf{e}_1 \perp \mathbf{e}_2, \dots, \mathbf{e}_p$ και $\mathbf{e}_1^T \mathbf{e}_1 = 1$.

Η δεύτερη κύρια συνιστώσα μιας κατανομής

Η δεύτερη κύρια συνιστώσα μιας κατανομής ορίζεται ως ο γραμμικός συνδυασμός

$$Y_2 = \mathbf{a}^\top \mathbf{X} = a_1 X_1 + \dots + a_p X_p$$

τέτοιος ώστε

- 1 $\text{Cov}(Y_1, Y_2) = 0$
- 2 $\text{Var}(Y_2)$ να γίνει μέγιστη
- 3 $\mathbf{a}^\top \mathbf{a} = 1$

Επειδή θέλουμε

$$0 = \text{Cov}(Y_1, Y_2) = \text{Cov}(\mathbf{e}_1^\top \mathbf{X}, \mathbf{a}^\top \mathbf{X}) = \mathbf{e}_1^\top \mathbf{\Sigma} \mathbf{a} = \mathbf{e}_1^\top \left(\sum_{i=1}^p \lambda_i \mathbf{e}_i \mathbf{e}_i^\top \right) \mathbf{a} = \lambda_1 \mathbf{e}_1^\top \mathbf{a}$$

$\Leftrightarrow \mathbf{e}_1 \perp \mathbf{a}$. Επομένως, από το θεώρημα μεγιστοποίησης τετραγωνικών μορφών, προκύπτει ότι η μέγιστη διασπορά για την Y_2 είναι ίση με λ_2 και επιτυγχάνεται για $\mathbf{a} = \mathbf{e}_2$.

Η δεύτερη κύρια συνιστώσα μιας κατανομής (ορισμός)

Ορισμός

Η δεύτερη κύρια συνιστώσα μιας κατανομής είναι ο γραμμικός συνδυασμός $Y_1 = \mathbf{e}_2^T \mathbf{X}$ με διασπορά $Var(Y_2) = \lambda_2$, όπου λ_2 η δεύτερη μεγαλύτερη ιδιοτιμή του $\mathbf{\Sigma}$ και $\mathbf{e}_2$ το (κανονικοποιημένο) ιδιοδιάνυσμα που αντιστοιχεί σε αυτήν την ιδιοτιμή.

Η k -οστή κύρια συνιστώσα μιας κατανομής

Συνεχίζοντας ανάλογα, ορίζουμε την k -οστή κύρια συνιστώσα μιας κατανομής ως τον γραμμικό συνδυασμό

$$Y_k = \mathbf{a}^\top \mathbf{X}$$

($\mathbf{a}^\top \mathbf{a} = 1$) ο οποίος είναι ασυσχέτιστος με τις $k - 1$ προηγούμενες κύριες συνιστώσες $Y_1, Y_2, \dots, Y_{k-1}$, και έχει την μέγιστη δυνατή διασπορά.

Μέσω του θεωρήματος μεγιστοποίησης τετραγωνικών μορφών παρομοίως αποδεικνύεται ότι η μέγιστη διασπορά είναι ίση με $\text{Var}(Y_k) = \lambda_k$ και επιτυγχάνεται για $\mathbf{a} = \mathbf{e}_k$.

Συνολική διασπορά (πληροφορία) κυρίων συνιστωσών

Συνολικά λοιπόν έχουμε p κύριες συνιστώσες. Επειδή

$$\sum_{i=1}^p \text{Var}(X_i) = \sum_{i=1}^p \sigma_{ii} = \text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \text{Var}(Y_i)$$

επομένως οι p κύριες συνιστώσες $Y_1, Y_2, \dots, Y_p$ διατηρούν ακριβώς την συνολική διασπορά (πληροφορία) των αρχικών μεταβλητών $X_1, X_2, \dots, X_p$.

Ορισμός

Έστω $\mathbf{X}$ τυχαίο διάνυσμα με $\text{Var}(\mathbf{X}) = \mathbf{\Sigma}$ θετικά ορισμένο και $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$ τα ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων του $\mathbf{\Sigma}$ με $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$. Τότε οι γραμμικοί συνδυασμοί

$$Y_1 = \mathbf{e}_1^\top \mathbf{X}$$

$$\vdots = \vdots$$

$$Y_k = \mathbf{e}_k^\top \mathbf{X}$$

$$\vdots = \vdots$$

$$Y_p = \mathbf{e}_p^\top \mathbf{X}$$

καλούνται οι κύριες συνιστώσες της κατανομής του $\mathbf{X}$. Η Y_k καλείται η k -οστή κύρια συνιστώσα και έχει διασπορά $\text{Var}(Y_k) = \lambda_k, k = 1, 2, \dots, p$. Οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και έχουν συνολική διασπορά ίση με την συνολική διασπορά των αρχικών μεταβλητών $X_1, X_2, \dots, X_p$.

Συνολική διασπορά μεταβλητών: $\sum_{i=1}^p \lambda_i$

- Η Y_1 εξηγεί το $100 \times \frac{\lambda_1}{\sum_{i=1}^p \lambda_i} \%$ της συνολικής διακύμανσης
- Οι Y_1, Y_2 εξηγούν το $100 \times \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^p \lambda_i} \%$ της συνολικής διακύμανσης
- $\vdots$
- Οι $Y_1, Y_2, \dots, Y_k$ εξηγούν το $100 \times \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i} \%$ της συνολικής διακύμανσης

Θέλουμε ένας μικρός αριθμός κυρίων συνιστωσών να ερμηνεύει μεγάλο ποσοστό της συνολικής μεταβλητότητας (π.χ. $> 80\%$)

Πρέπει οι αρχικές μεταβλητές να είναι συσχετισμένες!

Έστω ότι ο πίνακας συνδιακύμανσης του $\mathbf{X} = (X_1, X_2, X_3)^\top$ είναι

$$\mathbf{\Sigma} = \begin{pmatrix} 100 & -5 & 1 \\ -5 & 1 & 0.5 \\ 1 & 0.5 & 1 \end{pmatrix}$$

	λ_i	% πληρ.	αθρ. % πληρ.	Ιδιοδιανύσματα
1	100.2	0.98	0.98	$\mathbf{e}_1^\top = (0.9987, -0.0503, 0.0098)$
2	1.4	0.01	0.99	$\mathbf{e}_2^\top = (0.0238, 0.6260, 0.7795)$
3	0.3	0.00	1.00	$\mathbf{e}_3^\top = (-0.0453, -0.7782, 0.6264)$

	λ_i	% πληρ.	αθρ. % πληρ.	Ιδιοδιανύσματα
1	100.2	0.98	0.98	$\mathbf{e}_1^T = (0.9987, -0.0503, 0.0098)$
2	1.4	0.01	0.99	$\mathbf{e}_2^T = (0.0238, 0.6260, 0.7795)$
3	0.3	0.00	1.00	$\mathbf{e}_3^T = (-0.0453, -0.7782, 0.6264)$

Άρα, η πρώτη κύρια συνιστώσα είναι $Y_1 = 0.9987X_1 - 0.0503X_2 + 0.0098X_3$.

	λ_i	% πληρ.	αθρ. % πληρ.	Ιδιοδιανύσματα
1	100.2	0.98	0.98	$\mathbf{e}_1^T = (0.9987, -0.0503, 0.0098)$
2	1.4	0.01	0.99	$\mathbf{e}_2^T = (0.0238, 0.6260, 0.7795)$
3	0.3	0.00	1.00	$\mathbf{e}_3^T = (-0.0453, -0.7782, 0.6264)$

Άρα, η πρώτη κύρια συνιστώσα είναι $Y_1 = 0.9987X_1 - 0.0503X_2 + 0.0098X_3$.

- Η πρώτη κύρια συνιστώσα εξηγεί το 98% της συνολικής διασποράς (ουσιαστικά ταυτίζεται με την X_1).
- Αυτό συμβαίνει διότι η διασπορά της X_1 είναι πολύ μεγαλύτερη από τις υπόλοιπες.
- Δεν είναι καλή περίπτωση - θέλουμε > 1 μεταβλητές να συνεισφέρουν στις κύριες συνιστώσες.

- $Var(Y_j) = \lambda_j$
- Πίνακας διακύμανσης $\mathbf{\Lambda}$ κυρίων συνιστωσών: Διαγώνιος με στοιχεία τις ιδιοτιμές λ_j
- $tr(\mathbf{\Sigma}) = \sum_{i=1}^p \lambda_i = tr(\mathbf{\Lambda})$, η συνολική διακύμανση διατηρείται
- $|\mathbf{\Sigma}| = \prod_{i=1}^p \lambda_i = |\mathbf{\Lambda}|$ αφού $\mathbf{\Lambda}$ διαγώνιος πίνακας.
- % συνολικής διακύμανσης που εξηγεί η j συνιστώσα: $\frac{\lambda_j}{\sum_{i=1}^p \lambda_i}$
- **Ασυσχέτιστες αρχικές μεταβλητές θα ταυτιστούν με κάποια συνιστώσα** (επόμενο παράδειγμα)
- Οι κύριες συνιστώσες δεν είναι μοναδικές (μπορώ να αλλάζω τα πρόσημα, οι τιμές των συντελεστών δεν αλλάζουν σε απόλυτη τιμή)

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Ιδιοτιμές: $\lambda_1 = 5.83, \lambda_2 = 2, \lambda_3 = 0.17$

Ιδιοδιανύσματα:

$$\mathbf{e}_1^T = (-0.383, 0.924, 0)$$

$$\mathbf{e}_2^T = (0, 0, 1)$$

$$\mathbf{e}_3^T = (0.924, 0.383, 0)$$

Κύριες συνιστώσες:

$$Y_1 = -0.383X_1 + 0.924X_2$$

$$Y_2 = X_3$$

$$Y_3 = 0.924X_1 + 0.383X_2$$

Μειονέκτημα PCA βάσει Σ

- Αν αλλάξει η κλίμακα μέτρησης, αλλάζουν και οι κύριες συνιστώσες και η ερμηνεία τους.
- Αν κάποια μεταβλητή έχει πολύ μεγαλύτερη διακύμανση από τις υπόλοιπες, αυτή θα τείνει να ταυτιστεί με την πρώτη κύρια συνιστώσα

Έστω $\Sigma = \begin{pmatrix} 10 & 5 \\ 5 & 3 \end{pmatrix}$ για ηλικία X_1 σε χρόνια και βάρος X_2 σε κιλά:

- Ιδιοτιμές: (12.6033, 0.3967)
- Αν αντί X_2 σε κιλά έχουμε X'_2 σε τόνους

$$\Sigma' = \begin{pmatrix} 10 & 0.005 \\ 0.005 & 0.0000003 \end{pmatrix}$$

Ιδιοτιμές: (10.00000250, 0.00000050) $\rightarrow$ η πρώτη κύρια συνιστώσα θα ταυτίζεται με την X_1 !

Γενικά, διαφορά στις διακυμάνσεις ενδέχεται να περιέχει πληροφορία πολύτιμη για το θέμα που εξετάζουμε. Στην πράξη

- Αποφεύγουμε Σ όταν υπάρχουν μεταβλητές με πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες.
- Αν οι διακυμάνσεις είναι συγκρίσιμες καλό είναι να χρησιμοποιούμε Σ .
- Εναλλακτικά μετασχηματίζουμε δεδομένα ώστε διακυμάνσεις συγκρίσιμες

$$\Sigma = \begin{pmatrix} 50 & -1 & 0.1 \\ -1 & 1 & -0.5 \\ 0.1 & -0.5 & 0.3 \end{pmatrix}$$

Ιδιοτιμές: 50.0206, 1.2425 και 0.0368

$$Y_1 = 0.9997X_1 - 0.0190X_2 + 0.0075X_3$$

$$Y_2 = -0.0204X_1 - 0.8840X_2 + 0.4669X_3$$

$$Y_3 = 0.0022X_1 + 0.4669X_2 + 0.8842X_3$$

Αντιμετώπιση: Χρησιμοποίηση **πίνακα συσχετίσεων**.

Οι συσχετίσεις

- δεν αλλάζουν όταν αλλάξουν οι μονάδες μέτρησης ή η κλίμακα
- δίνουν ίδιο βάρος σε όλες τις μεταβλητές

Ο πίνακας συσχετίσεων είναι ο πίνακας συνδιακύμανσης των τυποποιημένων μεταβλητών!

- Επομένως έχει νόημα να χρησιμοποιήσουμε τις κύριες συνιστώσες του πίνακα συσχετίσεων.

Εάν $E(\mathbf{X}) = \boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$, έστω

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \dots, Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

οι τυποποιημένες μεταβλητές. Αν θέσουμε $\mathbf{V} = \text{diag}(\sigma_{11}, \dots, \sigma_{pp})$, τότε $\mathbf{Z} = (Z_1, \dots, Z_p) = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ και κατά συνέπεια

$$\text{Var}(\mathbf{Z}) = \text{Var}\left\{\mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})\right\} = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2} = \mathbf{P}$$

όπου $\mathbf{P}$ είναι ο πίνακας συσχετίσεων (με ij στοιχείο ρ_{ij} , $i \neq j$ και άσσους στην κύρια διαγώνιο)

- Ο πίνακας $\mathbf{P}$ είναι θετικά ορισμένος εφόσον ο πίνακας $\mathbf{\Sigma}$ είναι θετικά ορισμένος
- Η ανάλυση σε κύριες συνιστώσες μπορεί να εφαρμοστεί μέσω του $\mathbf{P}$.
- Σε αυτήν την περίπτωση, η συνολική διασπορά είναι ίση με

$$\text{tr}(\mathbf{P}) = p.$$

αφού στην διαγώνιο υπάρχουν άσσοι.

$$\mathbf{P} = \begin{pmatrix} 1 & -0.5 & 0.1 \\ -0.5 & 1 & 0.5 \\ 0.1 & 0.5 & 1 \end{pmatrix} \text{ από } \mathbf{\Sigma} = \begin{pmatrix} 100 & -5 & 1 \\ -5 & 1 & 0.5 \\ 1 & 0.5 & 1 \end{pmatrix}$$

	λ_i	% πληρ.	αθρ. % πληρ.	Ιδιοδιανύσματα
1	1.66	0.55	0.55	$\mathbf{e}_1^T = (-0.482, 0.732, 0.482)$
2	1.10	0.37	0.92	$\mathbf{e}_2^T = (0.707, 0.000, 0.707)$
3	0.24	0.08	1.00	$\mathbf{e}_3^T = (-0.517, -0.682, 0.517)$

Βλέπουμε ότι τώρα η πρώτη κύρια συνιστώσα εξηγεί το 55% της τυποποιημένης διασποράς, ενώ οι δύο πρώτες κύριες συνιστώσες μαζί εξηγούν το 92%

Παράδειγμα (συνέχεια)

	λ_i	% πληρ.	αθρ. % πληρ.	Ιδιοδιανύσματα
1	1.66	0.55	0.55	$\mathbf{e}_1^T = (-0.482, 0.732, 0.482)$
2	1.10	0.37	0.92	$\mathbf{e}_2^T = (0.707, 0.000, 0.707)$
3	0.24	0.08	1.00	$\mathbf{e}_3^T = (-0.517, -0.682, 0.517)$

Παρατηρήστε ότι η πρώτη κύρια συνιστώσα είναι ίση με

$$\begin{aligned}Y_1 &= \mathbf{e}_1^T \mathbf{Z} = -0.482Z_1 + 0.732Z_2 + 0.482Z_3 \\&= -0.482 \frac{X_1 - \mu_1}{\sqrt{100}} + 0.732 \frac{X_2 - \mu_2}{\sqrt{1}} + 0.482 \frac{X_3 - \mu_3}{\sqrt{1}} \\&= -0.0482X_1 + 0.732X_2 + 0.482X_3\end{aligned}$$

ενώ βάσει του πίνακα συνδιακύμανσης ήταν

$$Y_1 = 0.9987X_1 - 0.0503X_2 + 0.0098X_3.$$

Ταιριάζουν τα αποτελέσματα;

- Γενικά, ανάλυση με τους δύο πίνακες → διαφορετικά αποτελέσματα
- Το ένα δεν είναι συνάρτηση του άλλου

Γιατί καλύτερα στον πίνακα συσχέτισης;

- Αποτελέσματα σε διαφορετικά σύνολα μεταβλητών μπορούν να συγκριθούν ευκολότερα μεταξύ τους
- Ανάλυση με πίνακα συνδιακύμανσης είναι πολύ ευαίσθητα στις μονάδες μέτρησης

Γιατί καλύτερα στον πίνακα συνδιακύμανσης;

- Ευκολότερη στατιστική συμπερασματολογία στις κύριες συνιστώσες
- Ευκολότερη ερμηνεία ως 'νέες' μεταβλητές.

Ποια από τις δύο αναλύσεις προτιμούμε;

- Δεν υπάρχει αυστηρός κανόνας
- Περιγραφικοί σκοποί: → **συνήθως πίνακα συσχέτισης.**
- Κοινές μονάδες μέτρησης για όλες τις μεταβλητές: → **συνήθως πίνακα συνδιακύμανσης.**

Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα από κάποια p -διάστατη κατανομή με $\text{Var}(X) = \Sigma$ θετικά ορισμένο και

$$\mathbf{S} = (s_{ij}) = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

ο δειγματικός πίνακας συνδιακύμανσης και

$$\mathbf{R} = (r_{ij}) = \mathbf{V}^{-1/2} \mathbf{S} \mathbf{V}^{-1/2}$$

ο δειγματικός πίνακας συσχέτισης, όπου $\mathbf{V} = \text{diag}(s_{11}, \dots, s_{pp})$.

Διαδικασία

Εφαρμόζουμε την ίδια διαδικασία που περιγραφηκε προηγουμένως είτε στον $\mathbf{S}$ είτε στον $\mathbf{R}$, αφού είναι θετικά ορισμένοι με πιθανότητα 1.

Απαραίτητη προϋπόθεση: Οι μεταβλητές πρέπει να είναι συσχετισμένες, οπωσδήποτε να απορρίπτεται η υπόθεση $\mathbf{P} = \mathbf{I}_p$!

- Σημαντικό κομμάτι της ανάλυσης,
 - Δεν έχει εύκολη και κοινώς αποδεκτή απάντηση.
- Επιλέγοντας λιγότερες κύριες συνιστώσες από όσες μεταβλητές είχαμε αρχικά, χάνουμε αναγκαστικά πληροφορία
- Ενδιαφερόμαστε για κάποιον μικρότερο αριθμό συνιστωσών.
- Πόσες όμως;

Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες

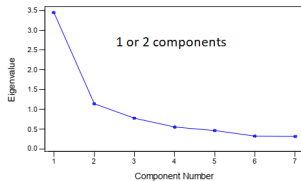
- Βάζουμε κάποιο όριο (π.χ. 80%) και διαλέγουμε τόσες συνιστώσες ώστε αθροιστικά να εξηγούν μεγαλύτερο ποσοστό από το στόχο που βάλαμε.
- Απλό και εύκολο, αλλά όχι πάντα καλά αποτελέσματα.
- Δεν είναι ξεκάθαρο ποιο ποσοστό της διακύμανσης πρέπει να βάλουμε σαν στόχο.

- Παίρνουμε τόσες ιδιοτιμές όσες είναι μεγαλύτερες από $\bar{\lambda} = \frac{\sum_{j=1}^p \lambda_j}{p}$.
- Πίνακας συσχέτισης: $\bar{\lambda} = 1$, άρα διαλέγουμε όσες ιδιοτιμές είναι μεγαλύτερες της μονάδας.
- Ασυσχέτιστες μεταβλητές $\Rightarrow \mathbf{R} = \mathbf{I}_p$ (μοναδιαίος) $\Rightarrow$ οι ιδιοτιμές είναι ίσες με 1. Επομένως, κάθε ιδιοτιμή > 1 προσφέρει κάτι
- Απλοϊκή υπόθεση
- Λόγω τυχαίας μεταβλητότητας του δείγματος, κάποιες από τις ιδιοτιμές > 1 .
- Το κριτήριο συνήθως υπερεκτιμά τον αριθμό των συνιστωσών που χρειάζονται.

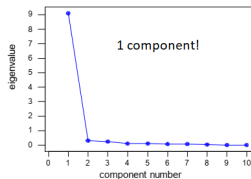
Το scree plot είναι ένα γράφημα που έχει στον οριζόντιο άξονα των x τη σειρά και στον κάθετο άξονα των y την τιμή της κάθε ιδιοτιμής. Το κριτήριο αυτό προτείνει να πάρουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου επίπεδο, στην ουσία μέχρι να διαπιστώσουμε ότι αρχίζει να αλλάζει η κλίση.

Scree plot

Scree Plot

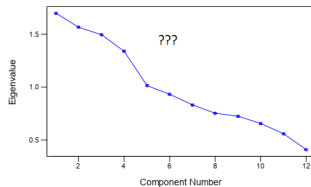


α



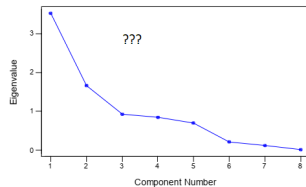
β

Scree Plot



γ

Scree Plot



δ

- Δεν εμπιστευόμαστε το κριτήριο του Καισερ γιατί κάποιες ιδιοτιμές λόγω τυχαίου σφάλματος μπορεί > 1 .
- Υποθέτουμε ότι $X_1, \dots, X_p$ ακολουθούν πολυμεταβλητή κανονική κατανομή.
- Αν το 95% διάστημα εμπιστοσύνης για την i ιδιοτιμή δεν περιέχει το 1 και είναι μεγαλύτερο από αυτή την τιμή κρατάμε την αντίστοιχη κύρια συνιστώσα (υποθέτουμε πως δουλεύουμε με τον πίνακα συσχετίσεων).

Εάν $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p)^\top$ οι ιδιοτιμές του $\mathbf{S}$ και $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p)^\top$ και $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ τότε

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \approx N(\mathbf{0}, 2\boldsymbol{\Lambda}^2)$$

$\Rightarrow \hat{\lambda}_i \approx N(\lambda_i, 2\lambda_i^2/n)$, και συνεπώς ένα $100(1 - \alpha)\%$ ΔΕ για την ιδιοτιμή λ_i δίνεται από

$$\left[\frac{\hat{\lambda}_i}{1 + z_{\alpha/2} \sqrt{2/n}}, \frac{\hat{\lambda}_i}{1 - z_{\alpha/2} \sqrt{2/n}} \right]$$

- Επαναληπτική δειγματοληψία με επανάθεση να εκτιμήσουμε ποσότητες του πληθυσμού και κυρίως τα τυπικά σφάλματα των εκτιμητριών τους.
- ΔΕ για τις ιδιοτιμές χωρίς να βασιζόμαστε σε **ασυμπτωτικά** αποτελέσματα από την **Κανονική** κατανομή.

Βήματα:

- Παίρνουμε ένα δείγμα ίσου μεγέθους με το πραγματικό κάνοντας δειγματοληψία με επανάθεση (κάποια παρατήρηση μπορεί να εμφανιστεί παραπάνω από μια φορά)
- Για αυτό το δείγμα υπολογίζουμε τον πίνακα διακύμανσης (συσχέτισης) και βρίσκουμε τις ιδιοτιμές.
- Αν επαναλάβουμε τη διαδικασία πολλές φορές, έχουμε σχηματίσει μια σειρά από τιμές της κατανομής των ιδιοτιμών. Ένα 95% bootstrap ΔΕ είναι κάτω και άνω όριο το 2.5% και 97.5% ποσοστημόρια της βοοστραπ κατανομής.

Αν μπορούμε να υποθέσουμε κανονικότητα του πληθυσμού ο Bartlett περιέγραψε έναν έλεγχο υπόθεσης για να ελέγξουμε αν οι τελευταίες π-κ ιδιοτιμές είναι ίσες (και επομένως δεν πρέπει να τις χρησιμοποιήσουμε)

H_0 : οι τελευταίες $p - k$ ιδιοτιμές είναι ίσες

H_1 : δεν είναι ίσες

Ελεγχουσυνάρτηση:

$$\chi^2 = -(n-1) \sum_{j=k+1}^p \log(\hat{\lambda}_j) + (n-1)(p-k) \log \left(\frac{\sum_{j=k+1}^p \hat{\lambda}_j}{p-k} \right)$$

η οποία ακολουθεί την κατανομή χ^2 με $\frac{1}{2}(p-k-1)(p-k+2)$.

- Αν δεν υπάρχουν συσχετίσεις είναι άσκοπο να συνεχίσουμε.
- Μεταβλητές ασυσχέτιστες: τις διώχνουμε από την ανάλυση.
- 'Μεγάλη συσχέτιση' $\rightarrow 0.4$ ή και μεγαλύτερη σε απόλυτη τιμή.

Μέτρο συνολικής συσχέτισης

$$\phi = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2 - p}{p(p-1)}}$$

- $\phi \rightarrow 1$ εάν υπάρχουν μεγάλες συσχετίσεις
- $\phi \rightarrow 0$ αν δεν υπάρχουν συσχετίσεις
- Στην πράξη τιμές πάνω από 0.4 θεωρούνται ικανοποιητικές

- Μεταβλητή ασυσχέτιστη με τις υπόλοιπες καλό είναι να την αφαιρέσουμε από την ανάλυση
- Ιδies ιδιοτιμές $\rightarrow$ Έλεγχος για πρόβλημα σε δεδομένα (π.χ. στήλες που επαναλαμβάνονται)
- Μηδενικές ιδιοτιμές $\rightarrow$ Κάποιες μεταβλητές είναι γραμμικά εξαρτημένες
- Σε διαφορετικά σεί δεδομένων μπορεί να πάρουμε ίδια ιδιοδιανύσματα με διαφορετικές ιδιοτιμές
- Αν $\mathbf{R}$ έχει μόνο θετικά στοιχεία $\rightarrow Y_i$ σταθμικός μέσος όρος των X_i
- Αν έχουμε τα σκορ των συνιστωσών μπορούμε εύκολα να βρούμε τα αρχικά δεδομένα

Παράδειγμα: Αποτελέσματα Επτάθλου (Ολυμπιακοί αγώνες, Λος Άντζελες 1984)

- Δεδομένα 26 αθλητριών του επτάθλου στα 7 αγωνίσματα κατά τη διάρκεια της Ολυμπιάδας του Los Angeles το 1984.
- Δρόμοι (100 μέτρα με εμπόδια, 200 μέτρα και 800 μέτρα), ρίψεις (σφαιροβολία και ακοντισμό) καθώς και άλματα (άλμα εις μήκος και άλμα εις ύψος)
- Η σειρά των αγωνισμάτων είναι όπως στον παρακάτω πίνακα

Πίνακας: Τα δεδομένα: Οι επιδόσεις των 26 αθλητριών που έλαβαν μέρος στο έπταθλο στην Ολυμπιάδα του Λος Άντζελες το 1984

Αθλήτρια	100μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιροβολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντισμός (m)	800μ (sec)
1	13.87	1.70	13.11	25.44	6.23	45.42	134.31
2	14.03	1.79	13.05	24.39	6.22	45.18	130.90
3	14.79	1.55	10.71	25.66	5.76	0.00	141.59

Πίνακας συσχέτισης δεδομένων επτάθλου

	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιρο- βολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντι- σμός (m)	800 μέτρα (sec)
100 μ με εμπόδια (sec)	1						
Άλμα εις ύψος (m)	-0.493	1					
Σφαιροβολία (m)	-0.500	0.441	1				
200 μέτρα (sec)	0.629	-0.374	-0.433	1			
Άλμα εις μήκος (m)	-0.527	0.461	0.391	-0.621	1		
Ακοντισμός (m)	-0.252	0.331	0.531	-0.170	0.135	1	
800 μέτρα (sec)	0.368	-0.442	-0.384	0.328	-0.445	-0.036	1

- Μέτριες συσχετίσεις γενικά
- Κάποιες ισχυρές συσχετίσεις (π.χ. τα 100 μέτρα με εμπόδια και τα 200 μέτρα) και κάποιες χαμηλές (π.χ. ο ακοντισμός έχει υψηλή συσχέτιση μόνο με τη σφαιροβολία)
- $\phi = 0.42$ (μέτρια συσχέτιση αλλά αξίζει να διερευνηθεί).

Πίνακας συνδιακύμανσης δεδομένων επτάθλου

	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιροβο- λία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντισμ- ός (m)	800 μέτρα (sec)
100 μ με εμπόδια (sec)	0.5085						
Άλμα εις ύψος (m)	-0.0337	0.0091					
Σφαιροβολία (m)	-0.3739	0.0443	1.0980				
200 μέτρα (sec)	0.3493	-0.0278	-0.3532	0.6073			
Άλμα εις μήκος (m)	-0.1166	0.0137	0.1271	-0.1501	0.0963		
Ακοντισμός (m)	-1.7400	0.3065	5.3859	-1.2833	0.4050	93.6593	
800 μέτρα (sec)	1.7624	-0.2845	-2.7019	1.7199	-0.9275	-2.3498	45.1422

Πίνακα συνδιακύμανσης ή πίνακα συσχέτισης;

- Πίνακα διακύμανσης ή συσχέτισης;
- Θα προχωρήσουμε στην ανάλυση και με τους δύο πίνακες για να συγκρίνουμε τα αποτελέσματά τους
- Οι μονάδες είναι διαφορετικές (κάποιες μεταβλητές μετριοούνται σε μέτρα ενώ κάποιες σε δευτερόλεπτα)
- Το μέγεθος των διακυμάνσεων διαφέρει πάρα πολύ. Ο ακοντισμός έχει διακύμανση 93 περίπου ενώ στο άλμα σε ύψος η διακύμανση είναι μόλις 0.009.
- Πίνακα διακυμάνσεων; όχι καλή ιδέα

Πίνακας συνδιακύμανσης: Ιδιοτιμές

Πίνακας Διακύμανσης		
Ιδιοτιμή	Ποσοστό της διακύμανσης	Αθροιστικό Ποσοστό της διακύμανσης
94.160	0.667	0.667
45.308	0.321	0.988
0.959	0.007	0.995
0.438	0.003	0.998
0.200	0.001	1.000
0.050	0.000	1.000
0.005	0.000	1.000

- 2 πρώτες ιδιοτιμές είναι πολύ μεγάλες σε σχέση με τις υπόλοιπες
- διαλέγοντας τις δύο πρώτες συνιστώσες → 99% της συνολικής διακύμανσης. Αν διαλέξουμε μόνο μια συνιστώσα εξηγούμε το 66,7

Πίνακας συνδιακύμανσης: Κύριες συνιστώσες

Κύρια Συνιστώσα	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιροβολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντισμός (m)	800 μέτρα (sec)
1	0.019802	-0.003443	-0.059340	0.014941	-0.004937	-0.996535	0.052393
2	0.037744	-0.005976	-0.054564	0.037455	-0.020285	0.057006	0.995238
3	-0.491883	0.024709	0.616033	-0.588679	0.148971	-0.051861	0.080738
4	-0.319610	0.011340	-0.780255	-0.524139	0.114575	0.031103	-0.010309
5	-0.805739	0.026062	-0.071453	0.583657	-0.065764	-0.002578	0.003639
6	-0.061431	-0.066489	0.008981	-0.190797	-0.977351	-0.000086	-0.010312
7	-0.033084	-0.997052	0.004458	0.007154	0.068589	0.002106	-0.003479

- Οι δύο πρώτες συνιστώσες σχεδόν ταυτίζονται με τον ακοντισμό και τα 800 μέτρα αντίστοιχα (τα πρόσημα δεν παίζουν κάποιο ρόλο).
- Αυτές ήταν οι δύο μεταβλητές με τη μεγαλύτερη διακύμανση
- Τα αποτελέσματα δεν έχουν κάποιο ενδιαφέρον και αυτό οφείλεται στις μεγάλες διαφορές των μεταβλητών ως προς τις διακυμάνσεις τους.

Πίνακας συνδιακύμανσης: Πρώτη κύρια συνιστώσα

$$\begin{aligned} Y_1 &= 0.019802(100\mu) - 0.003443(\text{ύψος}) - 0.059340(\text{σφαιροβολία}) \\ &+ 0.014941(200\mu) - 0.004937(\text{μήκος}) - 0.996535(\text{ακοντισμός}) \\ &+ 0.052393(800\mu) \end{aligned}$$

- Παρατηρείστε πως η κύρια συνιστώσα δεν έχει μονάδες.
- Ακόμα και αν έχουμε χρησιμοποιήσει τον πίνακα διακυμάνσεων όπου υπήρχαν μονάδες το αποτέλεσμα πρέπει να το βλέπουμε σαν αριθμό χωρίς μονάδες (δεν έχει νόημα να προσθέσουμε δευτερόλεπτα με μέτρα)!
- Σκίαση τους συντελεστές που είναι μεγαλύτεροι σε απόλυτη τιμή από 0.30
- Μικροί συντελεστές μπορεί απλά να οφείλονται στην τυχαιότητα του δείγματος
- Το άθροισμα τετραγώνων των συντελεστών κάθε συνιστώσας είναι 1.

Πίνακας συσχέτισης: Ιδιοτιμές

Α/Α	Πίνακας Συσχετίσεων		
	Ιδιοτιμή	Ποσοστό της διακύμανσης	Αθροιστικό Ποσοστό της διακύμανσης
1	3.4392	0.491	0.491
2	1.1408	0.163	0.654
3	0.7811	0.112	0.766
4	0.5461	0.078	0.844
5	0.4580	0.065	0.909
6	0.3248	0.046	0.956
7	0.3100	0.044	1.000

- Οι 2 πρώτες συνιστώσες εξηγούν μόλις το 65% της συνολικής διακύμανσης, ενώ οι 3 ανεβαίνουν στο 76

Πίνακας συσχέτισης: Κύριες συνιστώσες

Κύρια Συνιστώσα	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιροβολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντισμός (m)	800 μέτρα (sec)
1	0.429654	-0.391738	-0.401327	0.407340	-0.408511	-0.240791	0.331796
2	-0.082156	-0.066687	-0.377340	-0.230331	0.316381	-0.770490	-0.315616
3	-0.292121	-0.381523	-0.073663	-0.511351	0.159719	0.036265	0.689474
4	-0.100602	0.759488	-0.491387	0.113055	0.068449	0.019332	0.392117
5	-0.655432	0.063855	0.105759	0.030354	-0.693881	-0.267612	-0.033737
6	0.193855	0.226798	0.661771	0.162150	0.128028	-0.522687	0.396430
7	-0.496006	-0.253373	-0.023201	0.692493	0.452729	0.043106	0.053611

- Κάθε συνιστώσα εξαρτάται από πολλές μεταβλητές
- Η πρώτη συνιστώσα έχει θετικό πρόσημο στους δρόμους και αρνητικό στα υπόλοιπα αγωνίσματα.

Πίνακας συσχέτισης: Κύριες συνιστώσες

Κύρια Συνιστώσα	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιροβολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντισμός (m)	800 μέτρα (sec)
1	0.429654	-0.391738	-0.401327	0.407340	-0.408511	-0.240791	0.331796
2	-0.082156	-0.066687	-0.377340	-0.230331	0.316381	-0.770490	-0.315616
3	-0.292121	-0.381523	-0.073663	-0.511351	0.159719	0.036265	0.689474
4	-0.100602	0.759488	-0.491387	0.113055	0.068449	0.019332	0.392117
5	-0.655432	0.063855	0.105759	0.030354	-0.693881	-0.267612	-0.033737
6	0.193855	0.226798	0.661771	0.162150	0.128028	-0.522687	0.396430
7	-0.496006	-0.253373	-0.023201	0.692493	0.452729	0.043106	0.053611

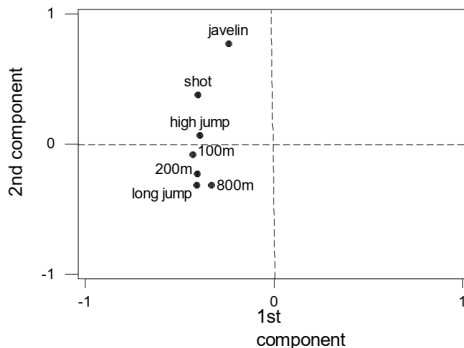
- Πρώτη συνιστώσα: σύγκριση ανάμεσα στους δρόμους και τα λοιπά αγωνίσματα;
 - Λάθος ερμηνεία διότι στους δρόμους οι καλές επιδόσεις είναι οι μικρές επιδόσεις (μικρότερος χρόνος τερματισμού λοιπόν) ενώ στις ρίψεις και τα άλματα οι μεγάλες επιδόσεις.
- Συνεπώς αυτή η διαφορά που βλέπουμε στα πρόσημα δεν σημαίνει σύγκριση αλλά είναι ένας σταθμικός μέσος όρος των επιδόσεων των αθλητριών.

Κύριες συνιστώσες/μετασχηματισμένα δεδομένα

Κύρια Συνιστώσα	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιροβολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντισμός (m)	800 μέτρα (sec)
1	-0.429654	-0.391738	-0.401327	-0.407340	-0.408511	-0.240791	-0.331796
2	-0.082156	0.066686	0.377340	-0.230331	-0.316381	0.770490	-0.315615
3	-0.292121	0.381523	0.073663	-0.511351	-0.159720	-0.036265	0.689474
4	0.100602	0.759488	-0.491387	-0.113056	0.068450	0.019332	-0.392117
5	0.655432	0.063855	0.105760	-0.030354	-0.693881	-0.267612	0.033736
6	-0.193857	0.226798	0.661771	-0.162148	0.128028	-0.522687	-0.396429
7	-0.496004	0.253373	0.023199	0.692493	-0.452730	-0.043105	0.053611

- Αφού αλλάξουμε πρόσημο στους δρόμους
- 1 συνιστώσα → γενικής αθλητικής ικανότητας
 - Μικρότερος συντελεστής στον ακοντισμό: δύσκολο αγώνισμα, συνήθως δεν τα πάνε καλά.
- Προτιμούμε την απεικόνιση των δύο πρώτων κύριων συνιστωσών σε γράφημα.

Απεικόνιση συντελεστών



- Οι δρόμοι (100μ, 200μ, 800μ) είναι πολύ κοντά μεταξύ τους, μαζί τους είναι και το άλμα εις μήκος - αυτά τα αθλήματα έχουν πολλά κοινά στοιχεία
- Αυτές οι μεταβλητές έχουν αρνητικά πρόσημα στη 2η κύρια συνιστώσα, για την οποία μπορούμε να δώσουμε μια ερμηνεία ως μια σύγκριση των δρόμων με τις ρίψεις

$$Y_1 = -0.429(100\mu) - 0.392(\text{ύψος}) - 0.401(\text{σφαιροβολία}) - 0.407(200 \mu.) \\ - 0.408(\text{μήκος}) - 0.240(\text{ακοντισμός}) - 0.332(800\mu)$$

	100 μ με εμπόδια (sec)	Άλμα εις ύψος (m)	Σφαιρο- βολία (m)	200 μέτρα (sec)	Άλμα εις μήκος (m)	Ακοντι- σμός (m)	800 μέτρα (sec)
Τιμή	-13.87	1.70	13.11	-25.44	6.23	45.42	-134.31
Τυποποιη- μένη τιμή	-0.074	-0.710	-0.441	-6.102	0.304	0.247	0.252

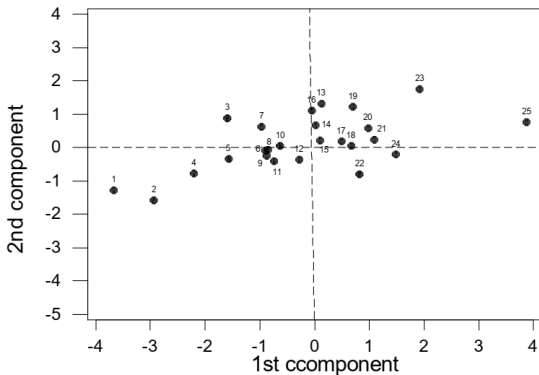
Επομένως η τιμή της αθλήτριας στην πρώτη κύρια συνιστώσα είναι

$$Y_1 = 0.429(-0.074) - 0.392(-0.710) - 0.401(-0.441) - 0.407(-1.102) \\ - 0.408(0.304) - -0.240(0.247) - 0.332(0.252) = 0.66845$$

Σύγκριση πραγματικής βαθμολογίας με κύριες συνιστώσες

Αθλήτρια	Τελική επίδοση στο αγώνισμα (βαθμοί)	Κατάταξη με βάση την επίδοση	Σκορ της αθλήτριας στην 1η συνιστώσα	Σκορ της αθλήτριας στην 2η συνιστώσα	Κατάταξη αθλήτριας με την πρώτη κύρια συνιστώσα
1	6030	18	0.66845	0.06047	18
2	6251	12	-0.27815	-0.35951	12
3	4530	26	4.97808	-3.5323	26
4	6434	6	-0.90751	-0.0929	7
5	6845	2	-2.93584	-1.58375	2
6	5897	21	1.0869	0.24001	22
7	6649	3	-1.59017	0.88324	4
8	6256	11	-0.74466	-0.40209	10
9	5278	25	3.88302	0.76698	25
10	6388	7	-0.96898	0.62067	6
11	5994	19	0.7007	1.2179	19
12	6464	5	-1.56705	-0.3311	5
13	6300	9	-0.87144	-0.24021	8
14	7044	1	-3.668	-1.27114	1
15	6095	17	0.49016	0.1956	17
16	6263	10	-0.62701	0.05633	11
17	6141	15	0.09514	0.22018	15
18	5749	24	1.48899	-0.20147	23
19	6619	4	-2.20334	-0.78184	3
20	6142	14	0.01656	0.66866	14
21	5993	20	0.98095	0.58074	21
22	5869	22	0.82542	-0.80352	20
23	5847	23	1.92205	1.73972	24
24	6152	13	0.12688	1.31927	16
25	6333	8	-0.84747	-0.06985	9
26	6123	16	-0.05369	1.0999	13

Συγκριση πραγματικής βαθμολογίας με κύριες συνιστώσες

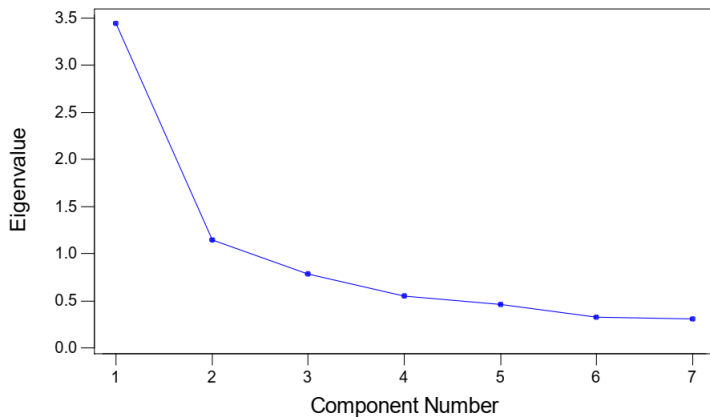


- Ετικέτες → η τελική τους κατάταξη με βάση τα αποτελέσματα της Ολυμπιάδας
- Ομαδοποίηση αθλητριών (μαύρη φυλή vs υπόλοιποι)
- Μικρότερη μεταβλητότητα στη 2η συνιστώσα
- Σχεδόν σε σειρά από δεξιά προς τα αριστερά

Eigenvalues	Proportion	Cumulative Proportion
3.4392	0.491	0.491
1.1408	0.163	0.654
0.7811	0.112	0.766
0.5461	0.078	0.844
0.458	0.065	0.909
0.3248	0.046	0.956
0.3100	0.044	1

- Κριτήριο Kaiser: έχουμε 2 ιδιοτιμές πάνω από 1 επομένως διαλέγουμε 2 συνιστώσες (τυχαία μεταβλητότητα);;;
- Συνολική διακύμανση $> 80\%$: 4 συνιστώσες ενώ αν ρίξουμε το ποσοστό αυτό στο 70% χρειαζόμαστε 3 συνιστώσες

Scree Plot for heptathlon data



- 1 ή 2 κύριες συνιστώσες
- Υποκειμενικό κριτήριο.

Eigenvalue	95% CI based on normal approximation		Bootstrap 95% CI		Τυπική απόκλιση
3.439	2.228	7.536	2.269	4.610	0.597
1.141	0.739	2.500	0.699	1.582	0.225
0.781	0.506	1.712	0.401	1.161	0.194
0.546	0.354	1.196	0.264	0.828	0.144
0.458	0.297	1.004	0.174	0.742	0.145
0.325	0.210	0.712	0.183	0.466	0.072
0.310	0.201	0.679	0.131	0.489	0.091

Πίνακας 7.9. Διαστήματα εμπιστοσύνης για τις ιδιοτιμές βασισμένα σε προσέγγιση με την κανονική κατανομή και στη μέθοδο bootstrap

- Διαλέγουμε τόσες συνιστώσες ώστε οι αντίστοιχες ιδιοτιμές να έχουν 95% ΔΕ μεγαλύτερο του 1
- ΔΕ βασισμένα στην κανονική προσέγγιση είναι πολύ μεγαλύτερα από αυτά της μεθόδου bootstrap
- Αυτό συμβαίνει και για τις δύο μεθόδους μόνο για μία ιδιοτιμή και επομένως διαλέγουμε να διατηρήσουμε μόνο μία συνιστώσα

Ιδέα

Αν κάποιες συνιστώσες δεν έχουν στατιστική σημαντικότητα (δηλαδή δεν εξηγούν σημαντικό κομμάτι ή αλλιώς δεν οφείλονται σε κάποια υπάρχουσα δομή των δεδομένων αλλά απλά οφείλονται στην τυχαία δειγματοληψία), τότε οι ιδιοτιμές τους θα είναι θεωρητικά ίσες στον πληθυσμό

- Επομένως ο έλεγχος ελέγχει τη μηδενική υπόθεση ότι οι τελευταίες ιδιοτιμές είναι ίσες έναντι της εναλλακτικής ότι δεν είναι.
- Ξεκινάμε από τον έλεγχο ότι οι 2 τελευταίες ιδιοτιμές είναι ίσες και αν δεν απορρίψουμε, συνεχίζουμε ελέγχοντας για τις 3 τελευταίες, και ούτω καθεξής μέχρι να απορρίψουμε για πρώτη φορά τη μηδενική υπόθεση.

Έλεγχος Bartlett

k	Test statistic	df	p-value
7	64.04026	27	0.01
6	16.65579	20	0.68
5	7.506863	14	0.92
4	2.81215	9	0.95
3	1.176038	5	0.95
2	0.01374	2	0.99

- Δεν μπορούμε να απορρίψουμε την υπόθεση πως οι 6 τελευταίες συνιστώσες είναι ίσες, αλλά μόνο την υπόθεση πως όλες οι συνιστώσες είναι ίσες
- Διαλέγουμε μόνο την πρώτη
- Επειδή ο έλεγχος γίνεται ακολουθιακά → πρόβλημα με το πραγματικό επίπεδο σημαντικότητας.

- Η απάντηση στο ερώτημα δεν υπάρχει με σιγουριά
- Όταν υπάρχει μια πραγματική δομή στον πληθυσμό, όλες οι μέθοδοι δίνουν σχετικά όμοια αποτελέσματα
- Αν επομένως πρέπει να προτείνουμε κάποιο κριτήριο, αυτό θα ήταν με τη χρήση της μεθόδου bootstrap
 - μειονέκτημα υπολογιστικού χρόνου