ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ
(PRINCIPAL COMPONENT ANALYSIS – PCA)

Κεντρική Ιδέα

- Η ανάλυση κυρίων συνιστωσών (PCA) έχει ως στόχο την ελάττωση της διάστασης ενός συνόλου δεδομένων που αποτελείται από έναν μεγάλο αριθμό συσχετισμένων μεταβλητών.
- **Πώς επιτυγχάνεται αυτό;** Μέσω ενός (γραμμικού) μετασχηματισμού των μεταβλητών προκύπτουν κάποιες νέες μεταβλητές (οι κύριες συνιστώσες), οι οποίες (i) είναι ασυσχέτιστες και (ii) έχουν την ίδια συνολική διασπορά με τις αρχικές μεταβλητές. Επιπλέον, ελπίζουμε ότι ένας μικρός αριθμός από αυτές εξηγεί το μεγαλύτερο μέρος της διασποράς των αρχικών μεταβλητών.
- Αν πράγματι συμβαίνει το τελευταίο, μπορούμε να αντικαταστήσουμε το αρχικό σύνολο δεδομένων με αυτόν το μικρό αριθμό των κυρίων συνιστωσών, χωρίς να χάσουμε σημαντική πληροφορία.

Κύριες Συνιστώσες μιας Κατανομής

- Έστω $\mathbf{X}_{p\times 1} = (X_1, \dots, X_p)'$ τυχαίο διάνυσμα με $\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma}_{p\times p} > 0.$ Ορίζουμε ως πρώτη κύρια συνιστώσα της κατανομής το γραμμικό συνδυασμό

$$Y_1 = \mathbf{a}'\mathbf{X} = a_1 X_1 + \cdots + a_p X_p,$$

με τη μέγιστη δυνατή διασπορά.

- Είναι φανερό ότι αυτή η μέγιστη δυνατή διασπορά είναι μια μη φραγμένη ποσότητα. Για το λόγο αυτό, περιοριζόμαστε σε γραμμικούς συνδυασμούς $\mathbf{a}'\mathbf{X}$ όπου το διάνυσμα $\boldsymbol{a} \in \mathbb{R}^p$ ικανοποιεί κάποια συνθήκη. Η συνηθέστερη συνθήκη είναι $\mathbf{a}\mathbf{a}' = \|\mathbf{a}\|^2 = 1.$
- Επειδή $\mathrm{Var}(Y_1) = \mathrm{Var}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\mathbf{\Sigma}\mathbf{a}$, απαιτούμε τη μεγιστοποίηση του $\mathbf{a}'\mathbf{\Sigma}\mathbf{a}$ ως προς $\mathbf{a}$ υπό τον περιορισμό $\|\mathbf{a}\|^2 = 1.$
- Αποδεικνύεται ότι το μέγιστο αυτό ισούται με λ_1, τη μεγαλύτερη ιδιοτιμή του $\mathbf{\Sigma}$ και επιτυγχάνεται για $\mathbf{a} = \boldsymbol{\epsilon}_1$, το ιδιοδιάνυσμα που αντιστοιχεί σε αυτή την ιδιοτιμή.

Κύριες Συνιστώσες μιας Κατανομής

Πρόταση (Μεγιστοποίηση τετραγωνικών μορφών): Έστω $\mathbf{A}_{p\times p}$ ένας θετικά ορισμένος πίνακας με ιδιοτιμές $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ και αντίστοιχα (κανονικοποιημένα) ιδιοδιανύσματα $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_p$. Τότε:

$$\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_1 \ (\text{για } \mathbf{x} = \boldsymbol{\epsilon}_1),$$

$$\max_{\mathbf{x}\neq 0, \mathbf{x}\perp \boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_\kappa} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_{\kappa+1} \ (\text{για } \mathbf{x} = \boldsymbol{\epsilon}_{\kappa+1}),\ \kappa = 1,2,\ldots,p-1.$$

Ορισμός: Η πρώτη κύρια συνιστώσα της κατανομής είναι ο γραμμικός συνδυασμός $Y_1 = \boldsymbol{\epsilon}_1'\mathbf{X}$ με διασπορά $\mathrm{Var}(Y_1) = \lambda_1$, όπου λ_1 είναι η μέγιστη ιδιοτιμή του $\boldsymbol{\Sigma}$ και $\boldsymbol{\epsilon}_1$ το ιδιοδιάνυσμα που αντιστοιχεί σε αυτή την ιδιοτιμή.

Κύριες Συνιστώσες μιας Κατανομής

- Η δεύτερη κύρια συνιστώσα της κατανομής ορίζεται ως ο γραμμικός συνδυασμός

$$Y_2 = \mathbf{a}'\mathbf{X} = a_1 X_1 + \cdots + a_p X_p,$$

ο οποίος είναι ασυσχέτιστος με την Y_1 και έχει τη μέγιστη δυνατή διασπορά. Όπως προηγουμένως, θέτουμε πάλι τον περιορισμό $\|\mathbf{a}\|^2 = 1$.

- Πριν συνεχίσουμε, ας θεωρήσουμε τη Φασματική Ανάλυση του $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i'.$$

Επειδή

$$\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(\boldsymbol{\epsilon}_1'\mathbf{X}, \mathbf{a}'\mathbf{X}) = \boldsymbol{\epsilon}_1'\boldsymbol{\Sigma}\mathbf{a} = \boldsymbol{\epsilon}_1'\left(\sum_{i=1}^{p} \lambda_i \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i'\right)\mathbf{a} = \lambda_1 \boldsymbol{\epsilon}_1'\mathbf{a} = 0,$$

το οποίο ισχύει αν και μόνο αν $\mathbf{a} \perp \boldsymbol{\epsilon}_1$, προκύπτει ότι η μέγιστη διασπορά για το Y_2 είναι λ_2 και επιτυγχάνεται για $\mathbf{a} = \boldsymbol{\epsilon}_2$.

Ορισμός: Η δεύτερη κύρια συνιστώσα της κατανομής είναι ο γραμμικός συνδυασμός $Y_2 = \boldsymbol{\epsilon}_2'\mathbf{X}$ με διασπορά $\mathrm{Var}(Y_2) = \lambda_2$, όπου λ_2 είναι η δεύτερη μεγαλύτερη ιδιοτιμή του $\boldsymbol{\Sigma}$ και $\boldsymbol{\epsilon}_2$ το ιδιοδιάνυσμα που αντιστοιχεί σε αυτή την ιδιοτιμή.

Κύριες Συνιστώσες μιας Κατανομής

- Συνεχίζοντας έτσι, ορίζουμε την κ-οστη κύρια συνιστώσα της κατανομής ως το γραμμικό συνδυασμό $Y_\kappa = \mathbf{a}'\mathbf{X}$ (με $\|\mathbf{a}\| = 1$) ο οποίος είναι ασυσχέτιστος με τις $\kappa - 1$ προηγούμενες κύριες συνιστώσες $Y_1, \dots, Y_{\kappa-1}$ και έχει τη μέγιστη δυνατή διασπορά.
- Αυτή η διασπορά θα ισούται με $\mathrm{Var}(Y_\kappa) = \lambda_\kappa$ και θα επιτυγχάνεται για $\mathbf{a} = \boldsymbol{\epsilon}_\kappa$. Συνολικά έχουμε p κύριες συνιστώσες και επειδή

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

οι p κύριες συνιστώσες $Y_1, \dots, Y_p$ έχουν συνολική διασπορά ίση με τη συνολική διασπορά των $X_1, \dots, X_p$.

Κύριες Συνιστώσες μιας Κατανομής

Ορισμός: Έστω $\mathbf{X}$ τυχαίο διάνυσμα με $\mathrm{Cov}(\mathbf{X}) = \mathbf{\Sigma} > 0$ και $(\lambda_1, \boldsymbol{\epsilon}_1), \dots, (\lambda_p, \boldsymbol{\epsilon}_p)$ τα ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων του $\mathbf{\Sigma}$ με $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Τότε, οι γραμμικοί συνδυασμοί

$$Y_1 = \boldsymbol{\epsilon}_1'\mathbf{X}, \cdots, Y_\kappa = \boldsymbol{\epsilon}_\kappa'\mathbf{X}, \cdots, Y_p = \boldsymbol{\epsilon}_p'\mathbf{X},$$

καλούνται κύριες συνιστώσες της κατανομής του $\mathbf{X}$. Η Y_κ καλείται κ-οστη κύρια συνιστώσα και έχει διασπορά $\mathrm{Var}(Y_\kappa) = \lambda_\kappa$, $\kappa = 1,2, \dots, p$. Οι p κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και έχουν συνολική διασπορά ίση με τη συνολική διασπορά των $X_1, \dots, X_p$.

Κύριες Συνιστώσες μιας Κατανομής

- Η συνολική διασπορά των συντεταγμένων του $\mathbf{X}$ (και του $\mathbf{Y} = (Y_1, \dots, Y_p)'$) είναι όπως είπαμε $\sum_{i=1}^{p} \lambda_i$. Έτσι:
  - Η Y_1 εξηγεί το $100 \times \frac{\lambda_1}{\sum_{i=1}^{p} \lambda_i}\%$ της συνολικής διασποράς.
  - Οι Y_1, Y_2 εξηγούν το $100 \times \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i}\%$ της συνολικής διασποράς.
  - ...
  - Οι $Y_1, \dots, Y_\kappa$ εξηγούν το $100 \times \frac{\sum_{i=1}^{\kappa} \lambda_i}{\sum_{i=1}^{p} \lambda_i}\%$ της συνολικής διασποράς.
- Επειδή οι ιδιοτιμές δεν είναι όλες ίσες μεταξύ τους (εκτός αν $\mathbf{\Sigma} = \lambda \mathbf{I}_p$), ελπίζουμε ότι ένας μικρός αριθμός από αυτές (οι κ πρώτες, για κ μικρό) θα εξηγεί ένα μεγάλο ποσοστό της συνολικής διασποράς, π.χ. το 80% ή και περισσότερο. Εκμεταλλευόμενοι αυτό το γεγονός, μπορούμε αντί του p-διάστατου $\mathbf{X}$ να θεωρήσουμε το κ-διάστατο $(Y_1, \dots, Y_\kappa)'$, χωρίς μεγάλη απώλεια πληροφορίας.

Κύριες Συνιστώσες μιας Κατανομής

- Η συνεισφορά της συντεταγμένης X_j του $\mathbf{X}$ στην κύρια συνιστώσα Y_κ μπορεί να μετρηθεί (αξιολογηθεί) είτε με το μέγεθος του $e_{\kappa j}$ στο γραμμικό συνδυασμό

$$Y_\kappa = \boldsymbol{\epsilon}_\kappa' \mathbf{X} = e_{\kappa 1} X_1 + \cdots + e_{\kappa j} X_j + \cdots + e_{\kappa p} X_p,$$

είτε με το συντελεστή συσχέτισης των X_j, Y_κ.

Πρόταση: $\mathrm{Corr}(X_j, Y_\kappa) = \dfrac{e_{\kappa j}\sqrt{\lambda_\kappa}}{\sqrt{\sigma_{jj}}}$.

- Κλείνοντας, να σχολιάσουμε ότι στην περίπτωση κατά την οποία ο πίνακας συνδιακύμανσης $\boldsymbol{\Sigma}$ είναι διαγώνιος, δηλ. οι συντεταγμένες του $\mathbf{X}$ είναι ασυσχέτιστες, οι ιδιοτιμές είναι οι ίδιες οι διασπορές και συνεπώς οι κύριες συνιστώσες της κατανομής είναι οι ίδιες οι συντεταγμένες του X.
- Είναι λοιπόν φανερό ότι η εφαρμογή της PCA έχει έννοια μόνο στην περίπτωση που οι συντεταγμένες είναι ασυσχέτιστες. Για το λόγο αυτό, πριν ξεκινήσουμε μια τέτοια ανάλυση, καλό είναι να «βεβαιωθούμε» γι'αυτό (μέσω ελέγχων, γραφικών παραστάσεων κλπ.)

Κύριες Συνιστώσες Από τον Πίνακα Συσχετίσεων

- Ας θεωρήσουμε το εξής παράδειγμα:

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

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

Στον παρακάτω πίνακα, δίνονται οι ιδιοτιμές και τα ιδιοδιανύσματα του $\mathbf{\Sigma}$.

	λ_i	$\lambda_i/\mathrm{tr}(\mathbf{\Sigma})$	Αθροιστικό %	Ιδιοδιανύσματα
1	100.2614	0.983	98.3%	$\boldsymbol{\epsilon}_1' = (0.9987, -0.0503, 0.0098)$
2	1.4321	0.014	99.7%	$\boldsymbol{\epsilon}_2' = (0.0238, 0.6260, 0.7795)$
3	0.3064	0.003	100.0%	$\boldsymbol{\epsilon}_3' = (-0.0453, -0.7782, 0.6264)$

Σύμφωνα με τα προηγούμενα, η πρώτη κύρια συνιστώσα είναι

$$Y_1 = \boldsymbol{\epsilon}_1'\mathbf{X} = 0.9987X_1 - 0.0503X_2 + 0.0098X_3,$$

εξηγεί το 98% της συνολικής διασποράς, συνεπώς θα μπορούσε να πάρει μόνη της τη θέση του **X**. Παρατηρούμε ότι στην Y_1 συνεισφέρει σχεδόν αποκλειστικά η X_1. Αυτό συμβαίνει διότι αυτή η συντεταγμένη έχει εξαιρετικά μεγαλύτερη διασπορά από τις υπόλοιπες ($\sigma_{11} = 100$ έναντι $\sigma_{22} = \sigma_{33} = 1$) και συνεπώς είναι η κυρίως υπεύθυνη για τη συνολική διασπορά.

Κύριες Συνιστώσες Από τον Πίνακα Συσχετίσεων

- Μια συνήθης τακτική στην PCA είναι η τυποποίηση των μεταβλητών πριν την εφαρμογή της μεθόδου, ειδικότερα όταν οι μεταβλητές μετρούνται σε διαφορετικές μονάδες μέτρησης (και συνεπώς οι μεταβλητότητές τους είναι μη συγκρίσιμες).
- $E(\mathbf{X}) = \boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$ και 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 $\mathbf{V}_{p\times p} = \text{diag}(\sigma_{11}, \dots, \sigma_{pp})$, τότε

$$\mathbf{Z} = (Z_1, \dots, Z_p)' = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}),$$

και κατά συνέπεια

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

όπου **P** είναι ο πίνακας συσχετίσεων, με το (i, j) στοιχείο να είναι $\rho_{ij} = \text{Corr}(X_i, X_j)$, $i \neq j$ και μονάδες στη διαγώνιο.

- Ο πίνακας **P** είναι θετικά ορισμένος (αφού ο **Σ** είναι θετικά ορισμένος) και συνεπώς η PCA μπορεί να εφαρμοσθεί με τον ίδιο τρόπο σε αυτόν. Μπορεί κανείς βέβαια να δει αμέσως ότι στην περίπτωση αυτή, η συνολική διασπορά των κύριων συνιστωσών είναι $\mathrm{tr}(\mathbf{P}) = p$ (αφού στη διαγώνιο είναι p άσσοι).

Παράδειγμα (συνέχεια): Ο πίνακας συσχετίσεων του $\mathbf{X} = (X_1, X_2, X_3)'$ (και του $\mathbf{Z} = (Z_1, Z_2, Z_3)'$) είναι

$$\mathbf{P} = \begin{pmatrix} 1 & -0.5 & 0.1 \\ -0.5 & 1 & 0.5 \\ 0.1 & 0.5 & 1 \end{pmatrix}.$$

Στον παρακάτω πίνακα, δίνονται οι ιδιοτιμές και τα ιδιοδιανύσματα του **P**.

	λ_i	$\lambda_i/\mathrm{tr}(\mathbf{\Sigma})$	Αθροιστικό %	Ιδιοδιανύσματα
1	1.6589	0.553	55.3%	$\boldsymbol{\epsilon}_1' = (0.4820, 0.7316, 0.4820)$
2	1.1000	0.367	92.0%	$\boldsymbol{\epsilon}_2' = (0.7071, 0.0000, 0.7071)$
3	0.2411	0.080	100.0%	$\boldsymbol{\epsilon}_3' = (-0.5173, -0.6817, 0.5173)$

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

- Αξίζει επίσης να παρατηρήσουμε ότι η πρώτη κύρια συνιστώσα είναι τώρα

$$Y_1 = \epsilon_1' Z = 0.4820Z_1 - 0.7316Z_2 + 0.4820Z_3 = 0.4820\frac{X_1-\mu_1}{\sqrt{100}} - 0.7316\frac{X_2-\mu_2}{\sqrt{1}} +$$

$$0.4820\frac{X_3-\mu_3}{\sqrt{1}} = 0.0482(X_1-\mu_1) - 0.7316(X_2-\mu_2) + 0.4820(X_3-\mu_3),$$

ενώ προηγουμένως ήταν

$$Y_1 = \boldsymbol{\epsilon}_1' \mathbf{X} = 0.9987X_1 - 0.0503X_2 + 0.0098X_3.$$

Ερώτηση: PCA στον πίνακα συνδιακύμανσης ή στον πίνακα συσχέτισης;

Εάν προσπαθήσουμε σε γενικές γραμμές να κάνουμε μια σύγκριση της PCA στον πίνακα συνδιακύμανσης και της PCA στον πίνακα συσχέτισης, μπορούμε να παρατηρήσουμε τα εξής:

- **Ταιριάζουν τα αποτελέσματα;** Γενικά, η ανάλυση στους δύο πίνακες δίνει διαφορετικά αποτελέσματα (όπως είδαμε και προηγουμένως). Επίσης, γενικά το ένα αποτέλεσμα δεν είναι συνάρτηση του άλλου.

- **Γιατί καλύτερα στον πίνακα συσχέτισης;** Γιατί τα αποτελέσματα της ανάλυσης σε διαφορετικά σύνολα μεταβλητών μπορούν να συγκριθούν ευκολότερα μεταξύ τους απ' ότι αν έχουν γίνει στους πίνακες συνδιακύμανσης. Επίσης, γιατί η ανάλυση στον πίνακα συνδιακύμανσης είναι πολύ ευαίσθητη στις μονάδες μέτρησης των μεταβλητών και συνεπώς τυποποιώντας παρακάμπτουμε αυτό το πρόβλημα.
- **Γιατί καλύτερα στον πίνακα συνδιακύμανσης;** Γιατί μπορεί να γίνει στατιστική συμπερασματολογία για τις κύριες συνιστώσες ευκολότερα. Επίσης, γιατί οι ίδιες κύριες συνιστώσες μπορούν να ερμηνευθούν ευκολότερα (ως νέες «μεταβλητές»).

Ποια από τις δύο αναλύσεις να προτιμούμε;

- Δεν υπάρχει αυστηρός κανόνας.
- Πάντως, αν θέλουμε να πραγματοποιήσουμε την ανάλυση μόνο για περιγραφικούς λόγους και όχι για συμπερασματολογία, καλύτερα να προτιμούμε τον πίνακα συσχέτισης.
- Τον πίνακα συνδιακύμανσης να τον προτιμούμε μάλλον όταν όλες οι μεταβλητές μετρούνται στις ίδιες μονάδες (και σε αυτή την περίπτωση, η διαφορετική μεταβλητότητα είναι σημαντική πληροφορία).

PCA στο Δείγμα μας

- Έστω $\mathbf{X}_1, \dots, \mathbf{X}_n$ τυχαίο δείγμα από κάποια p-διάστατη κατανομή με $\mathrm{Cov}(\mathbf{X}_1) = \mathbf{\Sigma}$ θετικά ορισμένο και

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

ο δειγματικός πίνακας συνδιακύμανσης και

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

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

- Η διαδικασία που περιγράψαμε προηγουμένως εφαρμόζεται είτε στον **S** είτε στον **R**, αφού είναι θετικά ορισμένοι με πιθανότητα 1. Θα συμβολίζουμε στο εξής με $l_1, \dots, l_p$ τις ιδιοτιμές είτε του **S** είτε του **R** και με $\mathbf{e}_1, \dots, \mathbf{e}_p$ τα αντίστοιχα ιδιοδιανύσματα.
- Απαραίτητη προϋπόθεση: Οι μεταβλητές πρέπει να είναι συσχετισμένες. Συνεπώς μία PCA πρέπει πάντοτε να αρχίζει με «επιβεβαίωση» αυτού (πώς το ελέγχω αυτό;).

Κριτήρια Επιλογής του Αριθμού των Κυρίων Συνιστωσών

Δεν υπάρχουν αυστηρά κριτήρια επιλογής. Διάφορες μέθοδοι είναι οι παρακάτω:

- Επιλέγουμε όσες κύριες συνιστώσες χρειάζονται για να ερμηνευθεί ένα συγκεκριμένο ποσοστό ολικής μεταβλητότητας π.χ. 80% ή 90% (τυποποιημένης ή μη).
- Επιλέγουμε όσες κύριες συνιστώσες αντιστοιχούν σε ιδιοτιμές μεγαλύτερες του $\bar{\lambda} = \sum_{i=1}^{p} \lambda_i / p$ (= 1 για PCA σε πίνακα συσχέτισης)

Κριτήριο Kaizer, επιλέγουμε αυτές που είναι > 1, αν κάνουμε PCA σε πίνακα συσχέτισης ή αυτές που είναι $> \bar{\lambda}$, αν κάνουμε PCA σε πίνακα συνδιακύμανσης.

- Επιλέγουμε βάσει του Scree plot.
- Υποθέτοντας ότι η κατανομή των δεδομένων είναι (πολυμεταβλητή) κανονική, πραγματοποιούμε κατάλληλους ελέγχους για τις q μικρότερες ιδιοτιμές (PCA σε πίνακα συνδιακύμανσης) με σκοπό να διαπιστώσουμε αν υπάρχουν σημαντικές διαφορές με τις υπόλοιπες.
- Εφαρμόζουμε τη μέθοδο bootstrap σε συνδυασμό με κάποια από τα παραπάνω κριτήρια.

Γιατί να εφαρμόσουμε PCA

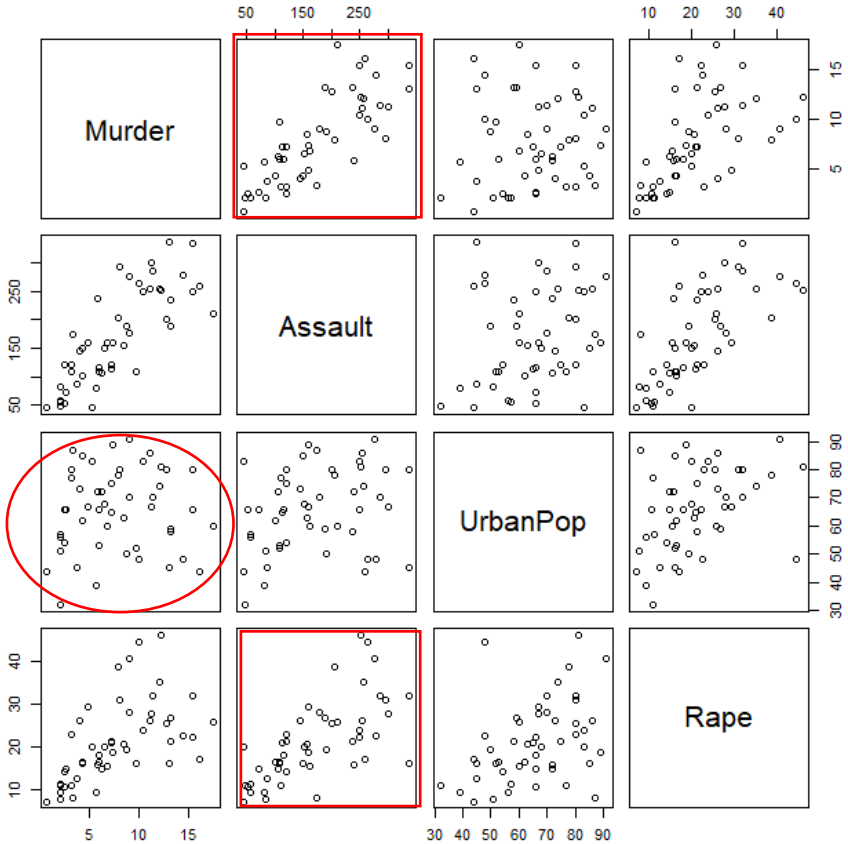
Μέσω της PCA επιτυγχάνουμε:

1. Μείωση του αριθμού των μεταβλητών και κατά συνέπεια της διάστασης των δεδομένων. Έτσι:
    - Μπορούμε να αναπαραστήσουμε γραφικά τα δεδομένα ευκολότερα, να τα ομαδοποιήσουμε και να κάνουμε ανίχνευση «περίεργων» παρατηρήσεων.
    - Μπορούμε να αποθηκεύσουμε τα δεδομένα ευκολότερα (αποθηκεύοντας αντί p μεταβλητές, κ κύριες συνιστώσες.
    - Μπορεί οι κύριες συνιστώσες να έχουν λογική ερμηνεία ως μεταβλητές και με αυτόν τον τρόπο είναι δυνατόν να ποσοτικοποιηθούν ακόμη και μη ποσοτικές μεταβλητές (περίπτωση Ανάλυσης Παραγόντων).
2. Μετάβαση από συσχετισμένες σε ασυσχέτιστες μεταβλητές. Έτσι:
    - Μπορούμε να ερμηνεύσουμε κάθε νέα μεταβλητή (κύρια συνιστώσα) χωρίς να επηρεαζόμαστε από τις υπόλοιπες.
    - Μπορούμε να ξεπεράσουμε το πρόβλημα της *πολυσυγγραμμικότητας* στην πολλαπλή γραμμική παλινδρόμηση.

ΕΦΑΡΜΟΓΗ PCA

- Από McNeil, D. R. (1977) *Interactive Data Analysis*. New York: Wiley, θα χρησιμοποιήσουμε το αρχείο δεδομένων USArrests (υπάρχει έτοιμο στην R) για να εφαρμόσουμε τη PCA.
- Τα δεδομένα περιέχουν στατιστικά στοιχεία σχετικά με τις συλλήψεις ανά 100000 πολίτες για τις κατηγορίες βίαιης επίθεσης (assault), φόνου (murder) και βιασμού (rape) σε καθεμία από τις 50 πολιτείες των ΗΠΑ το 1973. Είναι επίσης γνωστό το ποσοστό του πληθυσμού που ζει στα αστικά κέντρα (UrbanPop).
- Δεν είναι δύσκολο να διαπιστώσουμε (δείτε παρακάτω) ότι υπάρχει σημαντική διαφοροποίηση στις διακυμάνσεις κάθε μεταβλητής

```
> require(stats)
> require(graphics)
> summary(USArrests)
     Murder          Assault         UrbanPop          Rape
 Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
 1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
 Median : 7.250   Median :159.0   Median :66.00   Median :20.10
 Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
 3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
 Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00

> pairs(USArrests)
```


Murder
Assault
UrbanPop
Rape
50 150 250
10 20 30 40
5 10 15
50 150 250
30 40 50 60 70 80 90
10 20 30 40
5 10 15

```
>
c(var(USArrests$Murder),var(USArrests$Assault),var(USArrests$UrbanPop),
var(USArrests$Rape))
[1]    18.97047 6945.16571  209.51878   87.72916

> cov(USArrests)
             Murder   Assault   UrbanPop       Rape
Murder     18.970465  291.0624   4.386204  22.99141
Assault   291.062367 6945.1657 312.275102 519.26906
UrbanPop    4.386204  312.2751 209.518776  55.76808
Rape       22.991412  519.2691  55.768082  87.72916

> cov2cor(cov(USArrests))
             Murder   Assault   UrbanPop       Rape
Murder    1.00000000 0.8018733 0.06957262 0.5635788
Assault   0.80187331 1.0000000 0.25887170 0.6652412
UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
Rape      0.56357883 0.6652412 0.41134124 1.0000000

> Rm<-cov2cor(cov(USArrests))
> phiindex<-sqrt((sum(Rm^2)-4)/(4*(4-1)))
> phiindex
[1] 0.5234858
```

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

```
> Sm<-cov(USArrests)
> eigen(Sm)$values
[1] 7011.114851  201.992366   42.112651    6.164246

> valSm<-eigen(Sm)$values
> (cumsum(valSm)/sum(valSm))*100
[1]  96.55342  99.33516  99.91511 100.00000

> USArrests.pc.cov<-princomp(USArrests,cor=FALSE)
> USArrests.pc.cov
Call:
princomp(x = USArrests, cor = FALSE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4
82.890847 14.069560  6.424204  2.457837

 4  variables and  50 observations.

> summary(USArrests.pc.cov)
Importance of components:
                             Comp.1       Comp.2       Comp.3       Comp.4
Standard deviation     82.8908472 14.06956001 6.424204055 2.4578367034
Proportion of Variance  0.9655342  0.02781734 0.005799535 0.0008489079
Cumulative Proportion   0.9655342  0.99335156 0.999151092 1.0000000000
```

```
> loadings(USArrests.pc.cov)

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4
Murder                         0.995
Assault  -0.995
UrbanPop         -0.977 -0.201
Rape             -0.201  0.974
```

$Y_1 = -0.995X_2,\ Y_2 = -0.977X_3 - 0.201X_4,\ Y_3 = -0.201X_3 + 0.974X_4,\ Y_4 = 0.995X_1$

```
                Comp.1 Comp.2 Comp.3 Comp.4
SS loadings       1.00   1.00   1.00   1.00 # sum of squares of loadings
Proportion Var    0.25   0.25   0.25   0.25
Cumulative Var    0.25   0.50   0.75   1.00
```

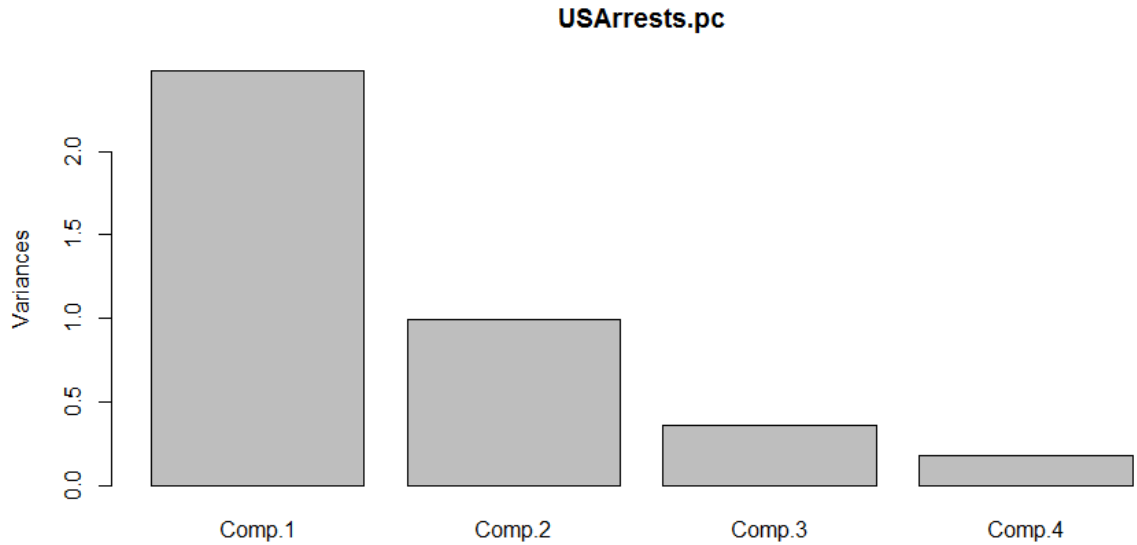
```
> valRm<-eigen(Rm)$values
> valRm
[1] 2.4802416 0.9897652 0.3565632 0.1734301
> (cumsum(valRm)/sum(valRm))*100
[1]  62.00604  86.75017  95.66425 100.00000
```

```
> USArrests.pc<-princomp(USArrests,cor=TRUE)
> USArrests.pc
Call:
princomp(x = USArrests, cor = TRUE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4
1.5748783 0.9948694 0.5971291 0.4164494

 4  variables and  50 observations.

> prcomp(USArrests, scale=TRUE)
Standard deviations:
[1] 1.5748783 0.9948694 0.5971291 0.4164494

Rotation:
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432

> summary(USArrests.pc)
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     1.5748783 0.9948694 0.5971291 0.41644938
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion  0.6200604 0.8675017 0.9566425 1.00000000
```

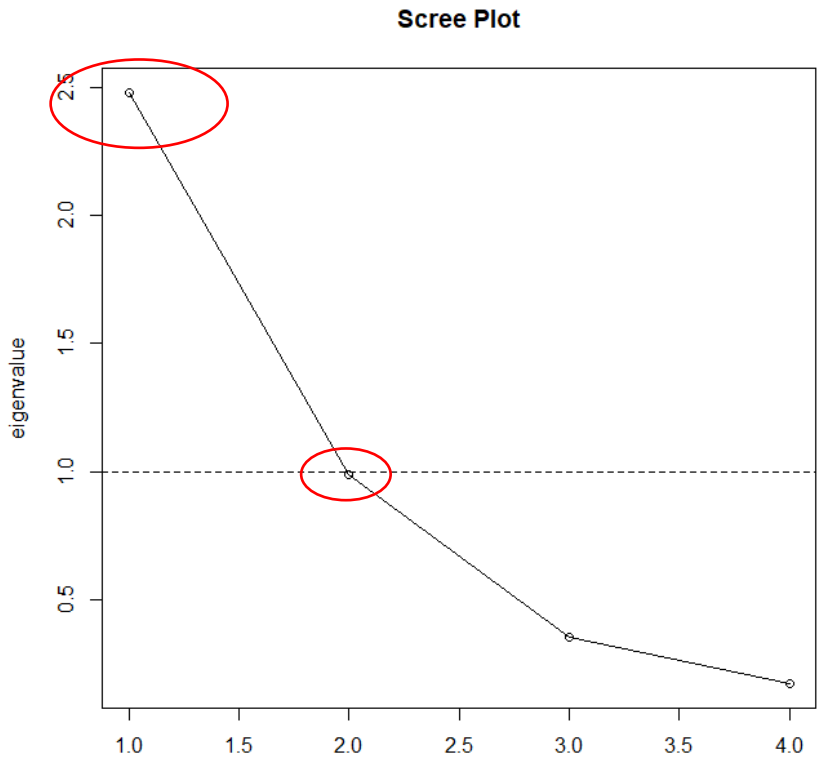
```
> loadings(USArrests.pc)

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4
Murder   -0.536  0.418 -0.341  0.649
Assault  -0.583  0.188 -0.268 -0.743
UrbanPop -0.278 -0.873 -0.378  0.134
Rape     -0.543 -0.167  0.818
```

$$
\begin{aligned}
Y_1 &= -0.536X_1 - 0.583X_2 - 0.278X_3 - 0.543X_4, \\
Y_2 &= 0.418X_1 + 0.188X_2 - 0.873X_3 - 0.167X_4, \\
Y_3 &= -0.341X_1 - 0.268X_2 - 0.378X_3 + 0.818X_4, \\
Y_4 &= 0.649X_1 - 0.743X_2 + 0.134X_3
\end{aligned}
$$

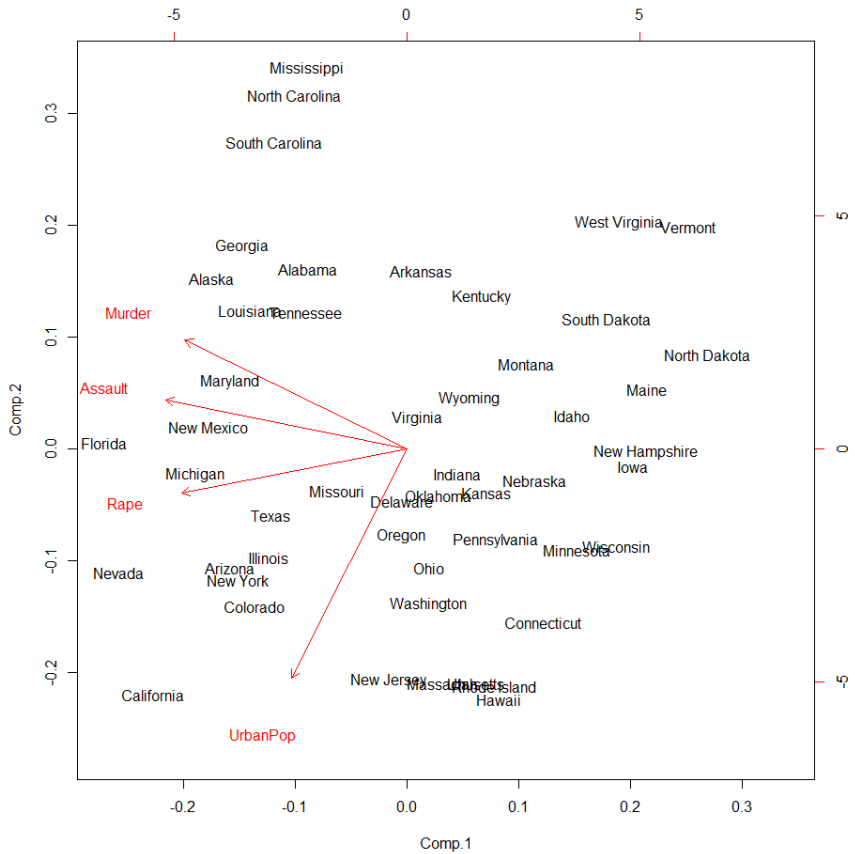
```
                Comp.1 Comp.2 Comp.3 Comp.4
SS loadings       1.00   1.00   1.00   1.00
Proportion Var    0.25   0.25   0.25   0.25
Cumulative Var    0.25   0.50   0.75   1.00

> plot(pc.cr) # shows a screeplot.
```




```
# Constructing Scree Plot manualy
> plot(valRm,type="o",main="Scree Plot",xlab="",ylab="eigenvalue")
> abline(h=1,lty=2)
```


Scree Plot
eigenvalue
2.5
2.0
1.5
1.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

```
> biplot(pc.cr)
```


-5
0
5
Mississippi
North Carolina
South Carolina
Georgia
Alaska
Alabama
Arkansas
Kentucky
Murder
Louisiana
Tennessee
West Virginia
Vermont
South Dakota
North Dakota
Montana
Maine
Assault
Maryland
Wyoming
Virginia
Idaho
New Mexico
Florida
New Hampshire
Iowa
Michigan
Indiana
Nebraska
Missouri
Oklahoma
Kansas
Delaware
Rape
Texas
Oregon
Pennsylvania
Minnesota
Wisconsin
Illinois
Nevada
Arizona
New York
Ohio
Colorado
Washington
Connecticut
New Jersey
Rhode Island
Hawaii
California
UrbanPop
0.3
0.2
0.1
0.0
-0.1
-0.2
Comp.2
5
0
-5
-0.2
-0.1
0.0
0.1
0.2
0.3
Comp.1

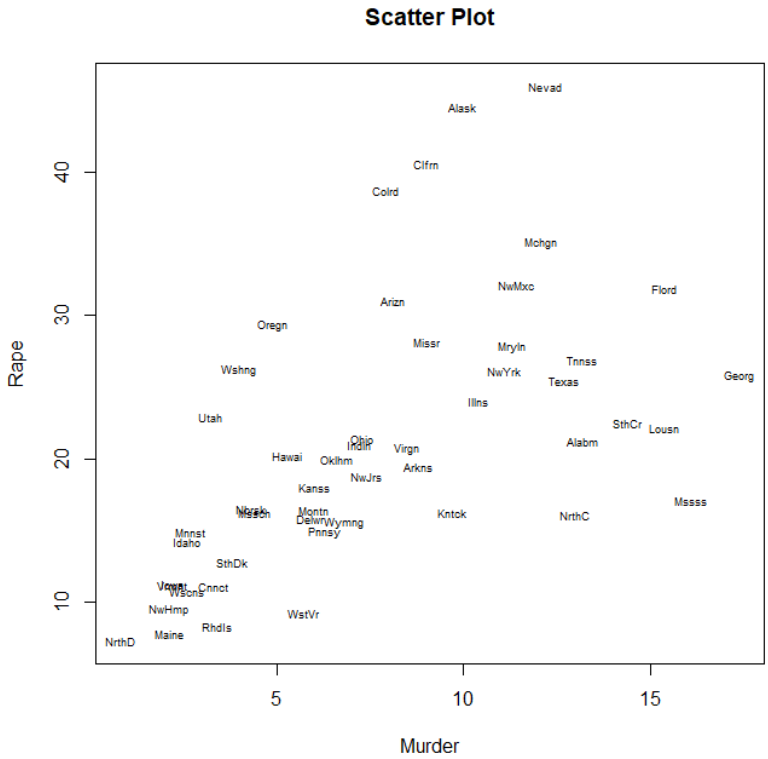
Ερμηνεία biplot

- Γενικά, υποθέτουμε οι 2 πρώτες κύριες συνιστώσες ερμηνεύουν επαρκές ποσοστό της συνολικής μεταβλητότητας ώστε να έχει νόημα να προχωρήσουμε σε γραφική αναπαράσταση της δομής των περιπτώσεων που εξετάζονται (cases, εδώ είναι οι πολιτείες των ΗΠΑ) και των μεταβλητών. Εδώ, το ποσοστό αυτό είναι 86.75%.
- Στο γράφημα φαίνονται:

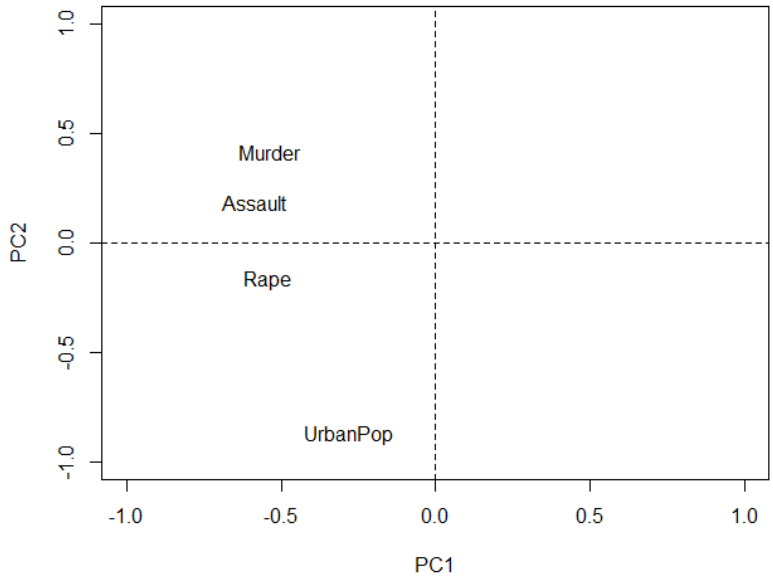
  (α) Το σκορ σε κάθε περίπτωση (πολιτεία των ΗΠΑ) σε καθεμία από τις 2 πρώτες κύριες συνιστώσες

  (β) Η τιμή του loading για κάθε μεταβλητή (Murder, Assault, Rape, UrbanPop) σε καθεμία από τις 2 πρώτες κύριες συνιστώσες.
- Ο κάτω και ο αριστερός άξονας δείχνουν τα (κανονικοποιημένα) σκορς των κυρίων συνιστωσών ενώ ο πάνω και ο δεξιός άξονας δείχνουν τα loadings. Κατά συνέπεια, ο πάνω και ο δεξιός άξονας χρησιμοποιούνται για να ερμηνεύσουμε τα κόκκινα βέλη (οι μεταβλητές του συνόλου των δεδομένων)

```
> statename<-(abbreviate(rownames(USArrests),5))
> plot(X1,X4,xlab="Murder",ylab="Rape",main="Scatter
Plot",type="n")
> text(X1,X4,labels=statename,cex=0.6)
```



```
> vecname<-c("Murder","Assault","UrbanPop","Rape")
> plot(c(-0.5358995,-0.5831836,-0.2781909,-
0.5434321),c(0.4181809,0.1879856,-0.8728062,-
0.1673186),xlab="PC1",ylab="PC2",xlim=c(-1,1),ylim=c(-
1,1),main="",type="n")
> text(c(-0.5358995,-0.5831836,-0.2781909,-
0.5434321),c(0.4181809,0.1879856,-0.8728062,-
0.1673186),labels=vecname)
> abline(h=0,lty=2)
> abline(v=0,lty=2)
```



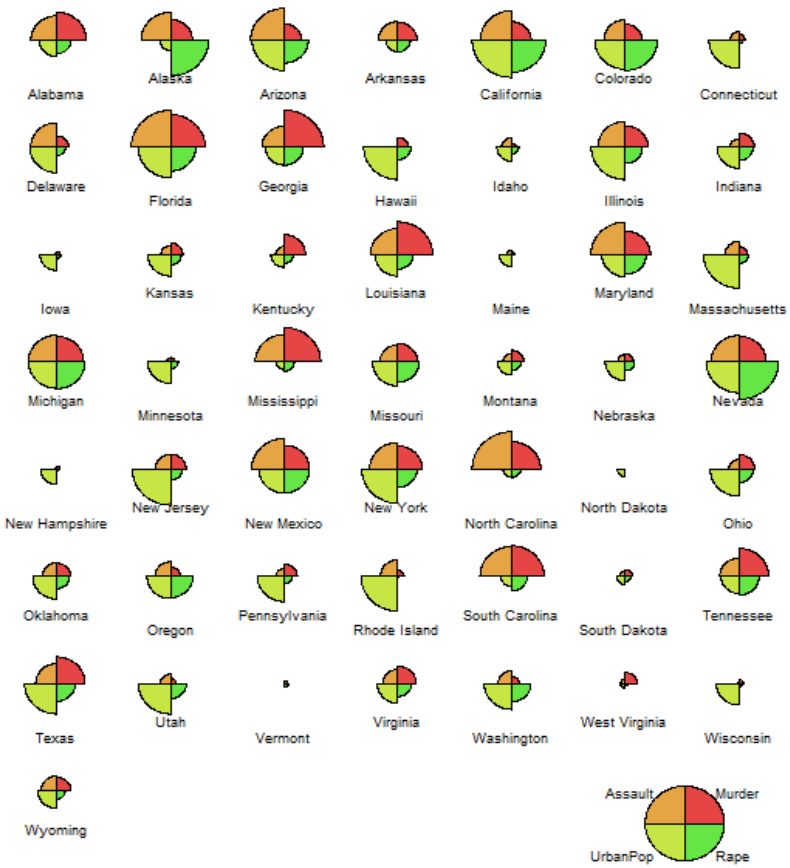
Κατασκευή Star Plot

```
> palette(rainbow(10,s=0.7,v=0.9))
> stars(USArrests,len=0.8,cex=0.55,key.loc=c(15,1.6),main="star
plot",draw.segments=TRUE)
```

Ερώτηση: Ποια η διαφορά μεταξύ Michigan και Ohio;

Απάντηση: Παρατηρούμε ότι ο πληθυσμός στις 2 αυτές πολιτείες είναι (σχεδόν) ίδιος αλλά το Michigan έχει πολύ μεγαλύτερο αριθμό συλλήψεων και για τις 3 κατηγορίες (Murder, Assault, Rape) σε σύγκριση με το Ohio

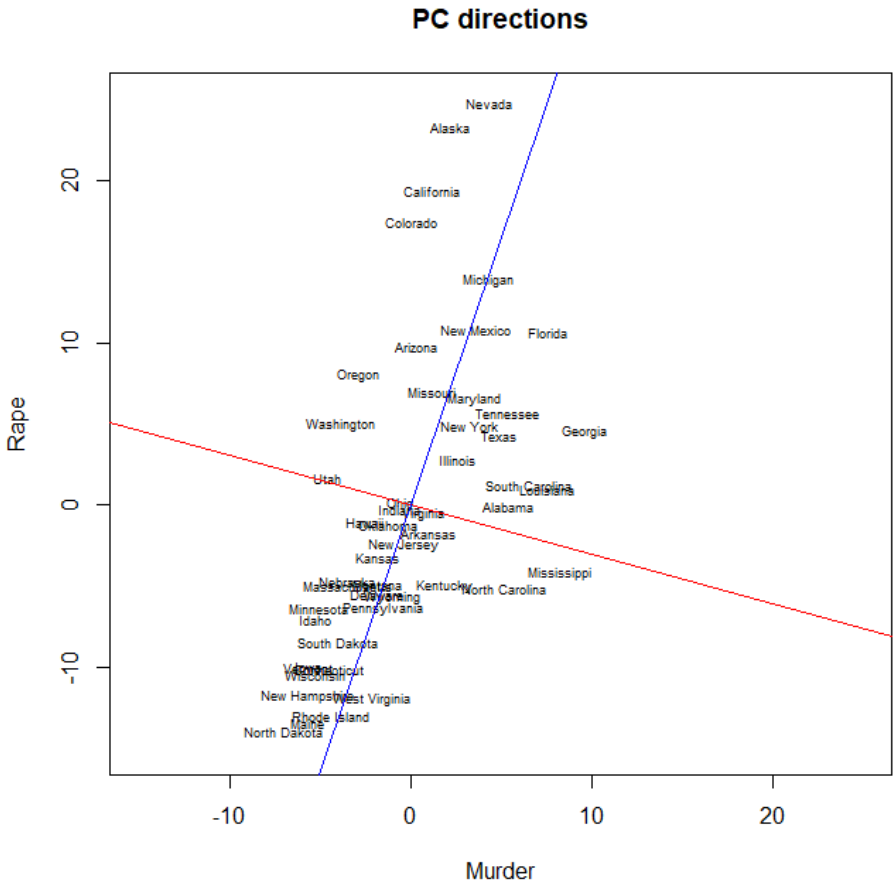
star plot


Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
Assault
Murder
UrbanPop
Rape

- Παρακάτω, χρησιμοποιούμε τις μεταβλητές Murder (=X1) και Rape (=X4) και κάνουμε PCA μόνο σε αυτές τις δύο (με χρήση της συνάρτησης `prcomp()`).
- Αρχικά, έχουμε κεντροποιήσει τις τιμές τους (αφαιρούμε τη μέση τιμή τους) και απεικονίζουμε σε ένα διάγραμμα διασποράς τα ζεύγη που προκύπτουν (V1, V2).
- Κατόπιν, βάζουμε και 2 άξονες στις διευθύνσεις των κυρίων συνιστωσών (μπλε για PC1, κόκκινο για PC2).
- Τέλος, αξίζει αν σημειωθεί πώς για να φαίνονται οι δύο διευθύνσεις ότι είναι ορθογώνιες, έχουμε «πειράξει» τα όρια κάθε άξονα (με χρήση των ορισμάτων xlim, ylim).
- Στη συνέχεια, μπορούμε να βγάλουμε συμπεράσματα σχετικά τις συλλήψεις για φόνο και βιασμό στις διάφορες αμερικανικές πολιτείες.

Κώδικας στην R

```
> V1<-X1-mean(X1)
> V4<-X1-mean(X4)
> Vdat<-cbind(V1,V4)
> pc14<-prcomp(Vdat)
> pc14.direction<-pc14$rotation
> a1<-pc14.direction[,1]
> a2<-pc14.direction[,2]
> plot(V1,V4,type="p",xlim=c(-15,25),ylim=c(-
15,25),xlab="Murder",ylab="Rape",main="PC directions")
> abline(0,a1[2]/a1[1],col="blue")
> abline(0,a2[2]/a2[1],col="red")
> rownames(Vdat)<-rownames(USArrests)
> plot(V1,V4,type="n",xlim=c(-15,25),ylim=c(-
15,25),xlab="Murder",ylab="Rape",main="PC directions")
> text(V1,V4,labels=rownames(Vdat),cex=0.6)
> abline(0,a2[2]/a2[1],col="red")
> abline(0,a1[2]/a1[1],col="blue")
```


PC directions
Rape
Murder
-10
0
10
20
Nevada
Alaska
California
Colorado
Michigan
New Mexico
Florida
Arizona
Oregon
Maryland
Tennessee
Washington
New York
Texas
Georgia
Illinois
Utah
South Carolina
Louisiana
Alabama
Arkansas
New Jersey
Kansas
Mississippi
Kentucky
North Carolina
Pennsylvania
Idaho
South Dakota
New Hampshire
West Virginia
Rhode Island
North Dakota

Manual Construction of the biplot

```
> X1<-USArrests$Murder
> X2<-USArrests$Assault
> X3<-USArrests$UrbanPop
> X4<-USArrests$Rape
> Z1<-(X1-mean(X1))/sd(X1)
> Z2<-(X2-mean(X2))/sd(X2)
> Z3<-(X3-mean(X3))/sd(X3)
> Z4<-(X4-mean(X4))/sd(X4)
> PC1Z<--0.536*Z1-0.583*Z2-
0.278*Z3-0.543*Z4
> PC2Z<-0.418*Z1+0.188*Z2-
0.873*Z3-0.167*Z4
> plot(PC1Z,PC2Z)
```