

Πολυμεταβλητή Ανάλυση

Παραγοντική ανάλυση

Λουκία Μελιγκοτσίδου

Τμήμα Μαθηματικών, ΕΚΠΑ

1 Εισαγωγή

- Μείωση διαστάσεων προβλήματος
- Δημιουργία νέων μεταβλητών (μη μετρήσιμων εννοιών)
- Ερμηνεία συσχετίσεων στα δεδομένα

Παρατήρηση

Η παραγοντική ανάλυση προσπαθεί περισσότερο να ερμηνεύσει τη δομή παρά τη μεταβλητότητα

- Υπάρχει ένα δομημένο μοντέλο και κάποιες υποθέσεις, δεν είναι απλώς ένας μαθηματικός μετασχηματισμός.
- Το ενδιαφέρον στηρίζεται στο να εξηγηθεί η **συνδιακύμανση** των μεταβλητών και όχι η **διακύμανση** αυτών.

- Στηρίζεται σε ένα πλήθος υποθέσεων οι οποίες δεν είναι απαραίτητα ρεαλιστικές - δεν μπορούν να ελεγχθούν εύκολα.
- Δεν έχει μοναδική λύση
- Οι παράγοντες που προκύπτουν μπορούν να δεχτούν διαφορετικές ερμηνείες
- Ο αριθμός των παραγόντων δεν είναι προφανής - εξαρτάται από τη μέθοδο εκτίμησης.

Το βασικό μοντέλο της FA

Έστω $\mathbf{X} = (X_1, \dots, X_p)^\top$ τυχαίο διάνυσμα με

$$E(\mathbf{X}) = \boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top, \quad \text{Var}(\mathbf{X}) = \boldsymbol{\Sigma}_{p \times p}.$$

Θεωρούμε ότι υπάρχουν τυχαία διανύσματα (μη παρατηρήσιμα)

$$\mathbf{F}_{m \times 1} = (F_1, \dots, F_m)^\top, \quad \mathbf{u}_{p \times 1} = (u_1, \dots, u_p)^\top,$$

με $p > m$ και ένας πίνακας $\mathbf{L}_{p \times m} = \{\ell_{ij}\}$ τάξης m τέτοιος ώστε

$$X_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + \dots + \ell_{1m}F_m + u_1$$

$$X_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + \dots + \ell_{2m}F_m + u_2$$

$$\vdots = \vdots$$

$$X_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \dots + \ell_{pm}F_m + u_p$$

Το βασικό μοντέλο της FA (2)

$$X_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + \dots + \ell_{1m}F_m + u_1$$

$$X_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + \dots + \ell_{2m}F_m + u_2$$

$$\vdots = \vdots$$

$$X_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \dots + \ell_{pm}F_m + u_p$$

Έπεται ότι μοντέλο μπορεί να γραφεί ισοδύναμα σε μορφή πινάκων ως

$$\mathbf{X}_{p \times 1} - \boldsymbol{\mu}_{p \times 1} = \mathbf{L}_{p \times m} \mathbf{F}_{m \times 1} + \mathbf{u}_{p \times 1}$$

Παρατήρηση

- Θυμίζει ένα μοντέλο γραμμικής παλινδρόμησης με $\mathbf{X}$ εξαρτημένη και $\mathbf{u}$ σφάλματα
- Η διαφορά είναι ότι εδώ δεν γνωρίζουμε καμία από τις ποσότητες του δεξιού μέλους.

Στο μοντέλο

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \mathbf{u}$$

η X_i μοντελοποιείται ως

$$X_i - \mu_i = \ell_{i1}F_1 + \dots + \ell_{ij}F_j + \dots + \ell_{im}F_m + u_i$$

- Οι $F_1, F_2, \dots, F_m$ λέγονται κοινοί παράγοντες (common factors).
- Το u_i λέγεται ειδικός παράγοντας (specific factor) ή σφάλμα (error).
- Τα ℓ ονομάζονται φορτία (loading). $\ell_{ij} \rightarrow$ φορτίο της i μεταβλητής στον j παράγοντα.

Το βασικό μοντέλο της FA: Παρατηρήσεις

Στο μοντέλο

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \mathbf{u}$$

η X_i μοντελοποιείται ως

$$X_i - \mu_i = \ell_{i1}F_1 + \dots + \ell_{ij}F_j + \dots + \ell_{im}F_m + u_i$$

- Στην FA κάθε μεταβλητή γράφεται ως γραμμικός συνδυασμός $m < p$ μη παρατηρηθέντων κοινών παραγόντων συν ένα μη παρατηρηθέν ειδικό σφάλμα
- Το ειδικό σφάλμα μπορεί να περιέχει εκτός από τυχαία μεταβλητότητα και άλλους ειδικούς όρους που αφορούν μόνο στην συγκεκριμένη μεταβλητή
- Με αυτόν τον τρόπο ουσιαστικά μοντελοποιούμε τον πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}$.

Το βασικό μοντέλο της FA: Παραδοχές

Το βασικό μοντέλο FA κάνει τις παρακάτω παραδοχές

- 1 Οι κοινοί παράγοντες έχουν μηδενική μέση τιμή, είναι ασυσχέτιστοι μεταξύ τους και έχουν διακύμανση 1.

$$E(\mathbf{F}) = \mathbf{0}, \quad \text{Var}(\mathbf{F}) = E(\mathbf{F}\mathbf{F}^\top) = \mathbf{I}_m$$

- 2 Οι ειδικοί παράγοντες έχουν μηδενική μέση τιμή, είναι ασυσχέτιστοι μεταξύ τους, αλλά όχι αναγκαστικά με την ίδια διασπορά

$$E(\mathbf{u}) = \mathbf{0}, \quad \text{Var}(\mathbf{u}) = E(\mathbf{u}\mathbf{u}^\top) = \mathbf{\Psi} = \text{diag}(\psi_1, \dots, \psi_p)$$

- 3 Σφάλματα και παράγοντες ασυσχέιστα μεταξύ τους

$$\text{Cov}(\mathbf{u}, \mathbf{F}) = E(\mathbf{u}\mathbf{F}^\top) = \mathbf{0}_{p \times m}$$

Υπενθυμίζεται ότι $\text{Cov}(\mathbf{x}_1, \mathbf{x}_2) = E(\mathbf{x}_1\mathbf{x}_2^\top) - E(\mathbf{x}_1)E(\mathbf{x}_2)^\top$.

Οι παραπάνω παραδοχές και ειδικότερα και ειδικότερα η πρώτη ορίζουν το **μοντέλο ορθογώνιων παραγόντων**.

Το βασικό μοντέλο της FA: Πίνακας συνδιακύμανσης

Υπολογίζεται εύκολα

$$\begin{aligned}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top &= (\mathbf{LF} + \mathbf{u})(\mathbf{LF} + \mathbf{u})^\top \\&= (\mathbf{LF})(\mathbf{LF})^\top + (\mathbf{LF})\mathbf{u}^\top + \mathbf{u}(\mathbf{LF})^\top + \mathbf{uu}^\top \\&= \mathbf{L}(\mathbf{FF}^\top)\mathbf{L}^\top + \mathbf{L}(\mathbf{Fu}^\top) + (\mathbf{uF}^\top)\mathbf{L}^\top + \mathbf{uu}^\top\end{aligned}$$

και επειδή $\text{Var}(\mathbf{F}) = \mathbf{I}_m$, $\text{Var}(\mathbf{u}) = \boldsymbol{\Psi}$, και $\text{Cov}(\mathbf{u}, \mathbf{F}) = \mathbf{0}$,

$$\begin{aligned}\boldsymbol{\Sigma} &= E\left\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right\} \\&= \mathbf{LE}(\mathbf{FF}^\top)\mathbf{L}^\top + \mathbf{LE}(\mathbf{Fu}^\top) + E(\mathbf{uF}^\top)\mathbf{L}^\top + E(\mathbf{uu}^\top) \\&= \mathbf{LL}^\top + \boldsymbol{\Psi}.\end{aligned}$$

Το βασικό μοντέλο της FA: $\Sigma = \mathbf{L}\mathbf{L}^\top + \Psi$

Γράφοντας τον πίνακα $\mathbf{L}$ ως $\mathbf{L}_{p \times m} = \begin{pmatrix} \ell_1^\top \\ \vdots \\ \ell_p^\top \end{pmatrix}$, προκύπτει ότι

$$\begin{aligned} \mathbf{L}\mathbf{L}^\top &= \begin{pmatrix} \ell_1^\top \\ \vdots \\ \ell_p^\top \end{pmatrix} (\ell_1 \quad \dots \quad \ell_p) \\ &= \begin{pmatrix} \ell_1^\top \ell_1 & \ell_1^\top \ell_2 & \dots & \ell_1^\top \ell_p \\ \ell_2^\top \ell_1 & \ell_2^\top \ell_2 & \dots & \ell_2^\top \ell_p \\ \vdots & \vdots & \ddots & \vdots \\ \ell_p^\top \ell_1 & \ell_p^\top \ell_2 & \dots & \ell_p^\top \ell_p \end{pmatrix} \end{aligned}$$

Το βασικό μοντέλο της FA: $\Sigma = \mathbf{L}\mathbf{L}^\top + \Psi$

Βάσει των προηγούμενων, στο μοντέλο ορθογώνιων παραγόντων ισχύουν τα εξής:

$$\sigma_i^2 = \text{Var}(X_i) = \sum_{k=1}^m \ell_{ik}^2 + \psi_i$$

$$\sigma_{ij} = \text{Cov}(X_i, X_j) = \sum_{k=1}^m \ell_{ik} \ell_{jk}$$

$$\text{Cov}(X_i, F_j) = \ell_{ij} \quad \text{αφού} \quad \text{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}.$$

Η διασπορά της i μεταβλητής αναλύεται ως εξής:

$$\underbrace{\sigma_i^2}_{\text{Var}(X_i)} = \underbrace{\ell_{i1}^2 + \dots + \ell_{ij}^2 + \dots + \ell_{im}^2}_{\substack{\text{communality} \\ (\text{εταιρικότητα})}} + \underbrace{\psi_i}_{\substack{\text{specific variance} \\ (\text{ειδική διασπορά})}}$$

Θέτουμε την i εταιρικότητα

$$h_i^2 = \ell_{i1}^2 + \dots + \ell_{ij}^2 + \dots + \ell_{im}^2$$

οπότε

$$\sigma_i^2 = h_i^2 + \psi_i.$$

Η παραπάνω ορολογία δικαιολογείται από το γεγονός ότι

- h_i είναι το μέρος της διασποράς που οφείλεται στους κοινούς παράγοντες
- ψ_i είναι το μέρος της διασποράς που οφείλεται στον ειδικό παράγοντα.

- Με βάση τα παραπάνω αυτό που χρειαζόμαστε είναι να εκτιμήσουμε τους πίνακες L και Ψ .
- Η παραπάνω σχέση σημαίνει ότι αν έχουμε διαθέσιμους τους πίνακες L και Ψ ή έστω μια εκτίμηση τους μπορούμε να βρούμε τον πίνακα διακύμανσης και άρα να μετρήσουμε πόσο καλό είναι το μοντέλο που υποθέσαμε.
- Προσέξτε επίσης ότι δεν χρειαζόμαστε όλα τα δεδομένα αλλά μόνο τον πίνακα διακύμανσης!

- Οι παράγοντες F_i μπορούν να γραφτούν και αυτοί σαν γραμμικός συνδυασμός των μεταβλητών - χρήσιμο όταν θέλουμε να δημιουργήσουμε νέες μεταβλητές.
- Θα πρέπει όμως να γίνει σαφές ότι οι συντελεστές αυτοί διαφέρουν από τις επιβαρύνσεις και δεν πρέπει να γίνεται σύγχυση
- Οι συντελεστές κάθε παράγοντα όταν εκφράζουμε τις μεταβλητές ως γραμμικό συνδυασμό των παραγόντων καλούνται επιβαρύνσεις ενώ αντίστοιχα οι συντελεστές κάθε μεταβλητής όταν εκφράζουμε κάθε παράγοντα ως γραμμικό συνδυασμό των μεταβλητών καλούνται συντελεστές των σκορ (factor scores coefficients).

- Έλεγχος για ικανοποιητικές συσχετίσεις
- Εύρεση αριθμού παραγόντων και εκτίμηση παραμέτρων μοντέλου
- Περιστροφή μοντέλου με σκοπό την αύξηση της ερμηνευτικής του ικανότητας
- Εκτίμηση των σκορ των παραγόντων για περαιτέρω στατιστική χρήση

- Μας ενδιαφέρει να υπάρχουν μεγάλες συσχετίσεις τουλάχιστον σε μεγάλο ποσοστό του R (πχ τιμές > 0.40)
- Για να ελέγξουμε τη στατιστική σημαντικότητα ενός δειγματικού συντελεστή συσχέτισης:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0,$$

- Ελεγχοςυνάρτηση

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$$

- **Παρατήρηση:** Δεν μπορεί να χρησιμοποιηθεί για υποθέσεις διάφορες από τη μορφή $\rho = 0$ (η ελεγχοςυνάρτηση δεν ακολουθεί t)

Έλεγχος σφαιρικότητας του Bartlett

$$H_0 : \mathbf{P} = \mathbf{I}_p \quad \text{έναντι της εναλλακτικής}$$

$$H_1 : \mathbf{P} \neq \mathbf{I}_p$$

Υποθέτοντας κανονικότητα, μπορεί να χρησιμοποιηθεί η ελεγχοσυνάρτηση

$$L = - \left[n - \frac{1}{6(2p+5)} \log |\mathbf{R}| \sim \chi_{p(p-1)/2}^2 \right].$$

Υπολογίζει τη συσχέτιση αφού αφαιρέσει την επίδραση λοιπών μεταβλητών. Στο πλαίσιο της παραγοντικής ανάλυσης:

- Μερικοί συντελεστές συσχέτισης $\rightarrow$ εκτιμήσεις των συσχετίσεων μεταξύ των μοναδικών παραγόντων ***u***
- Θα πρέπει ≈ 0 όταν οι υποθέσεις του παραγοντικού μοντέλου ισχύουν

- Αν οι μεταβλητές μοιράζονται κοινούς παράγοντες, μερικός συντελεστής συσχέτισης ανάμεσα σε δύο μεταβλητές μικρός, αφού η ακύρωση της επίδρασης των υπολοίπων μεταβλητών ακυρώνει σε μεγάλο βαθμό την επίδραση των κοινών παραγόντων.
- Επομένως οι μερικοί συντελεστές συσχέτισης είναι εκτιμήσεις των συσχετίσεων μεταξύ των μοναδικών παραγόντων και θα πρέπει να είναι κοντά στο 0 όταν οι υποθέσεις του παραγοντικού μοντέλου ισχύουν.
- Θυμηθείτε ότι από τις υποθέσεις του μοντέλου οι μοναδικοί παράγοντες είναι ασυσχέτιστοι.

Μέτρο σύγκρισης συντελεστών συσχέτισης με μερικούς συντελεστές συσχέτισης:

$$KMO = \frac{\sum_{i,j=1:i \neq j}^p r_{ij}^2}{\sum_{i,j=1:i \neq j}^p r_{ij}^2 + \sum_{i,j=1:i \neq j}^p a_{ij}^2}$$

$$a_{ij} = \text{Corr}(X_i, X_j | \mathbf{X} \setminus \{X_i, X_j\})$$

όπου δηλαδή a_{ij} είναι ο μερικός συντελεστής συσχέτισης. Αν οι μερικοί συντελεστές συσχέτισης είναι μικροί (ιδανικά 0), θα έχουμε $KMO = 1$. Στην πράξη δεν εφαρμόζουμε παραγοντική ανάλυση εάν $KMO < 0.5$. Τιμές γύρω στο 0.8 θεωρούνται εξαιρετικές.

Μέτρο δειγματικής καταλληλότητας (measure of sampling adequacy) για την i μεταβλητή:

$$KMO_i = \frac{\sum_{j=1:j \neq i}^p r_{ij}^2}{\sum_{j=1:j \neq i}^p r_{ij}^2 + \sum_{j=1:j \neq i}^p a_{ij}^2}$$

Τιμές κοντά στο 1 είναι ενδείξεις ότι η μεταβλητή είναι πολύ καλή για να χρησιμοποιηθεί στην ανάλυση.

- Η οποιαδήποτε λύση για τον πίνακα L δεν είναι μοναδική!
- Έστω $Q_{m \times m}$ ένας ορθογώνιος πίνακας ($QQ^T = I_m$) και

$$L_{p \times m}^* = L_{p \times m} Q_{m \times m}$$

Τότε

$$L^* L^{*\top} + \Psi = (LQ)(LQ)^T + \Psi = LQQ^T L + \Psi = LL^T + \Psi = \Sigma,$$

δηλαδή ο πίνακας συνδιακύμανσης αναπαράγεται χρησιμοποιώντας είτε τον L είτε τον L^* .

- Επιπλέον, ισχύει ότι

$$X - \mu = LF + u = L(QQ^T)F + u = L^*F^* + u$$

όπου $F^* = Q^T F$.

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \mathbf{u} = \mathbf{L}(\mathbf{Q}\mathbf{Q}^\top)\mathbf{F} + \mathbf{u} = \mathbf{L}^*\mathbf{F}^* + \mathbf{u}$$

- $\mathbf{F}^* = \mathbf{Q}^\top \mathbf{F}$.
- Εάν χρησιμοποιήσουμε τον πίνακα $\mathbf{L}^*$ και το μετασχηματισμένο διάνυσμα παραγόντων $\mathbf{F}^* \rightarrow$ ίδιο πίνακα συνδιακύμανσης
- Όταν πολλαπλασιάζουμε ένα διάνυσμα με έναν ορθογώνιο πίνακα, το περιστρέφουμε στο χώρο.
- Για αυτό τον λόγο, η χρησιμοποίηση του $\mathbf{L}^*$ αντί του $\mathbf{L}$ αναφέρεται ως **περιστροφή των παραγόντων (factor rotation)**.
- Περιστροφή γίνεται για να ερμηνευτούν καλύτερα οι παράγοντες (περισσότερα σε λίγο)

- **Μέθοδος κυρίων συνιστωσών:** δεν αφορά την PCA που είδαμε απλά χρησιμοποιεί τη φασματική ανάλυση του πίνακα διακύμανσης. Δεν πρέπει να συγχέεται με την PCA.
- **Μέθοδος μεγίστης πιθανοφάνειας:** βασίζεται στην υπόθεση πολυμεταβλητής κανονικότητας των δεδομένων

Εκτίμησης φορτιών: Μέθοδος κυρίων συνιστωσών

Έστω $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ οι ιδιοτιμές και $\epsilon_1, \epsilon_2, \dots, \epsilon_p$ τα αντίστοιχα ιδιοδιανύσματα του Σ . Τότε

$$\Sigma = \sum_{j=1}^p \lambda_j \epsilon_j \epsilon_j^\top$$

Θεωρήστε τον πίνακα

$$L = (\mathbf{q}_1 \quad \mathbf{q}_2 \quad \dots \quad \mathbf{q}_p)$$

όπου $\mathbf{q}_j = \sqrt{\lambda_j} \epsilon_j, j = 1, 2, \dots, p$. Επομένως, έχουμε ότι

$$LL^\top = (\mathbf{q}_1 \quad \mathbf{q}_2 \quad \dots \quad \mathbf{q}_p) \begin{pmatrix} \mathbf{q}_1^\top \\ \mathbf{q}_2^\top \\ \vdots \\ \mathbf{q}_p^\top \end{pmatrix} = \sum_{j=1}^p \mathbf{q}_j \mathbf{q}_j^\top = \sum_{j=1}^p \lambda_j \epsilon_j \epsilon_j^\top = \Sigma$$

$$\mathbf{L}\mathbf{L}^\top = (\mathbf{q}_1 \quad \mathbf{q}_2 \quad \dots \quad \mathbf{q}_p) \begin{pmatrix} \mathbf{q}_1^\top \\ \mathbf{q}_2^\top \\ \vdots \\ \mathbf{q}_p^\top \end{pmatrix} = \sum_{j=1}^p \mathbf{q}_j \mathbf{q}_j^\top = \sum_{j=1}^p \lambda_j \boldsymbol{\epsilon}_j \boldsymbol{\epsilon}_j^\top = \boldsymbol{\Sigma}$$

Άρα έχουμε γράψει ότι $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top + \boldsymbol{\Psi}$, όπου $\boldsymbol{\Psi} = \mathbf{0}_{p \times p}$.

- $\mathbf{L} = \{\ell_{ij}\} = \sqrt{\lambda_j} \boldsymbol{\epsilon}_{j,i}$
- Τα φορτία δηλαδή του j κοινού παράγοντα ισούνται με τους συντελεστές της j κύριας συνιστώσας πολλαπλασιασμένους με $\sqrt{\lambda_j}$.
- Η παραπάνω αναπαράσταση δεν είναι χρήσιμη γιατί οδηγεί σε p παράγοντες (όσος δηλαδή ο αριθμός των μεταβλητών).

Εκτίμησης φορτιών: Μέθοδος κυρίων συνιστώσων (2)

Εάν όμως θεωρήσουμε ότι οι τελευταίες $p - m$ ιδιοτιμές είναι μικρές, μπορούν να αγνοηθούν, άρα και να αγνοήσουμε την συνεισφορά του $\sum_{j=m+1}^p \lambda_j \epsilon_j \epsilon_j^T$ στον Σ . Άρα θέτουμε

$$\mathbf{L}_m = (\mathbf{q}_1 \quad \mathbf{q}_2 \quad \dots \quad \mathbf{q}_m)_{p \times m}$$

και

$$\mathbf{L}_m \mathbf{L}_m^T = \lambda_1 \epsilon_1 \epsilon_1^T + \dots + \lambda_m \epsilon_m \epsilon_m^T = \sum_{j=1}^m \lambda_j \epsilon_j \epsilon_j^T$$

Επομένως

$$\Sigma = \sum_{j=1}^p \lambda_j \epsilon_j \epsilon_j^T = \mathbf{L}_m \mathbf{L}_m^T + \sum_{j=m+1}^p \lambda_j \epsilon_j \epsilon_j^T$$

Ο πίνακας $\sum_{j=m+1}^p \lambda_j \epsilon_j \epsilon_j^T$ δεν είναι διαγώνιος (όπως θα έπρεπε στην παραγοντική ανάλυση).

Εκτίμησης φορτιών: Μέθοδος κυρίων συνιστώσων (3)

Ιδανικά θα θέλαμε έναν διαγώνιο πίνακα $\Psi_{p \times p}$ ώστε

$$\Sigma = L_m L_m^T + \Psi.$$

Αυτό όμως δεν είναι πάντα εφικτό. Σαν προσέγγιση παίρνουμε

$$\Psi_m = \text{diag}(\psi_1, \psi_2, \dots, \psi_p)$$

όπου $\psi_i = \sigma_i^2 - \sum_{j=1}^m \ell_{ij}^2, i = 1, \dots, p$. Δηλαδή έχουμε ορίσει τον πίνακα Ψ_m ώστε τα διαγώνια στοιχεία του να είναι με τα αντίστοιχα διαγώνια στοιχεία του $\Sigma - L_m L_m^T$. Άρα υποθέτουμε

$$\Sigma \approx L_m L_m^T + \Psi_m.$$

Ο πίνακας Ψ_m είναι 'καλή' προσέγγιση εάν ο πίνακας $D = \Sigma - L_m L_m^T + \Psi_m$ είναι περίπου $\mathbf{0}_{p \times p}$ (αν δηλαδή τα μη διαγώνια στοιχεία είναι 0 γιατί τα διαγώνια είναι 0 εξ ορισμού).

Μέτρο καλής εφαρμογής μεθόδου κυρίων συνιστωσών

Ένα μέτρο του πόσο ‘κοντά’ στο 0 είναι ο $\mathbf{D}$ είναι το άθροισμα των τετραγωνικών αποκλίσεων

$$\text{tr}(\mathbf{D}^\top \mathbf{D}) = \sum_{i=1}^p \sum_{j=1}^p d_{ij}$$

Η προσέγγιση είναι καλή εάν $\text{tr}(\mathbf{D}^\top \mathbf{D}) \approx 0$.

Παρατηρήσεις

- Εάν χρησιμοποιήσουμε p κοινούς παράγοντες: $\Rightarrow \Psi_p = \mathbf{0} \Rightarrow \mathbf{D} = \mathbf{0}$.
- Γενικά, όσο μεγαλώνει το m , τόσο πιο κοντά στο $\mathbf{0}$ θα βρίσκεται ο πίνακας $\mathbf{\Sigma} - \mathbf{L}_m \mathbf{L}_m^\top = \sum_{j=m+1}^p \lambda_j \epsilon_j \epsilon_j^\top$ αφού οι ιδιοτιμές $\lambda_{m+1}, \lambda_{m+2}, \dots$ θα γίνονται όλο και μικρότερες.
 - Καλύτερη προσέγγιση με κόστος όμως την αύξηση του αριθμού των παραγόντων.
- Ένα πλεονέκτημα αυτής της μεθόδου εκτίμησης παραμέτρων είναι ότι αν προσθέσουμε επιπλέον παράγοντες, τα φορτία των προηγούμενων παραγόντων παραμένουν αναλοίωτα.

Το ενδιαφέρον στην παραγοντική ανάλυση συνήθως επικεντρώνεται στην εκτίμηση των L και Ψ . Μερικές φορές όμως χρειάζεται να εκτιμήσουμε και τις τιμές των παραγόντων.

Αποτέλεσμα

Εάν χρησιμοποιήσουμε την μέθοδο των κυρίων συνιστωσών τότε οι κοινοί παράγοντες είναι οι τυποποιημένες κύριες συνιστώσες !

Εκτίμηση παραγόντων με τη μέθοδο των κυρίων συνιστωσών

Μοντέλο FA: $\mathbf{X}^* = \begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_p - \mu_p \end{pmatrix} = \mathbf{L}\mathbf{F} + \mathbf{u}$. Γνωρίζουμε ότι οι κύριες

συνιστώσες για τις αρχικές μεταβλητές $\mathbf{X}^* = (X_1^*, \dots, X_p^*)$ υπολογίζονται ως $Z_j = \boldsymbol{\epsilon}_j^\top \mathbf{X}^*, j = 1, 2, \dots, p$. Σε μορφή πινάκων, $\mathbf{Z} = \mathbf{E}^\top \mathbf{X}^*$, όπου

$\mathbf{Z} = (Z_1, \dots, Z_p)^\top$ και $\mathbf{E}^\top = \begin{pmatrix} \boldsymbol{\epsilon}_1^\top \\ \vdots \\ \boldsymbol{\epsilon}_p^\top \end{pmatrix}$. Επειδή $\mathbf{E}$ ορθογώνιος

$$(\mathbf{E}\mathbf{E}^\top)\mathbf{X}^* = \mathbf{E}\mathbf{Z} \Rightarrow \mathbf{X}^* = \mathbf{E}\mathbf{Z}.$$

Επιπλέον, γνωρίζουμε ότι $E(Z_j) = 0$ και $\text{Var}(Z_j) = \lambda_j$. Ορίζουμε τις τυποποιημένες κύριες συνιστώσες

$$\tilde{Z}_j = \frac{Z_j}{\sqrt{\lambda_j}} \Rightarrow \mathbf{Z} = \mathbf{\Lambda}^{1/2} \tilde{\mathbf{Z}}$$

όπου $\mathbf{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$ και $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_p)$.

Με βάση τα προηγούμενα $\mathbf{X}^* = \mathbf{E}\mathbf{\Lambda}^{1/2}\tilde{\mathbf{Z}}$. Θέτοντας $F_j = \tilde{Z}_j, j = 1, \dots, p$, $\mathbf{F} = (F_1, \dots, F_p)^\top$ και $\mathbf{L}_p = \mathbf{E}\mathbf{\Lambda}^{1/2} = (\sqrt{\lambda_1}\epsilon_1, \dots, \sqrt{\lambda_p}\epsilon_p)$, έχουμε

$$\mathbf{X}^* = \mathbf{L}_p\mathbf{F}$$

το οποίο αντιστοιχεί σε $\mathbf{u} = \mathbf{0}$. Δηλαδή έχουμε p κοινούς παράγοντες καθένας από τους οποίους αντιστοιχεί σε μια (τυποποιημένη) κύρια συνιστώσα και όλοι οι ειδικοί παράγοντες είναι 0.

Εκτίμηση παραγόντων με τη μέθοδο των κυρίων συνιστωσών

Εάν πάρουμε τις m πρώτες κύριες συνιστώσες, $\tilde{z}_j, j = 1, 2, \dots, m$ ως κοινούς παράγοντες

$$\begin{aligned} \mathbf{X}^* &= \mathbf{E}\mathbf{\Lambda}^{1/2}\tilde{\mathbf{Z}} = (\mathbf{q}_1 \quad \dots \quad \mathbf{q}_p) \tilde{\mathbf{Z}} \\ &= (\mathbf{q}_1 \quad \dots \quad \mathbf{q}_m \quad \mathbf{q}_{m+1} \quad \dots \quad \mathbf{q}_p) \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_m \\ \tilde{z}_{m+1} \\ \vdots \\ \tilde{z}_p \end{pmatrix} \\ &= (\mathbf{q}_1 \quad \dots \quad \mathbf{q}_m) \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_m \end{pmatrix} + (\mathbf{q}_{m+1} \quad \dots \quad \mathbf{q}_p) \begin{pmatrix} \tilde{z}_{m+1} \\ \vdots \\ \tilde{z}_p \end{pmatrix} \end{aligned}$$

$$\mathbf{X}^* = (\mathbf{q}_1 \quad \dots \quad \mathbf{q}_m) \begin{pmatrix} \tilde{z}_1 \\ \vdots \\ \tilde{z}_m \end{pmatrix} + (\mathbf{q}_{m+1} \quad \dots \quad \mathbf{q}_p) \begin{pmatrix} \tilde{z}_{m+1} \\ \vdots \\ \tilde{z}_p \end{pmatrix}.$$

Θέτουμε $F_j = \tilde{z}_j, j = 1, \dots, m$ και $\mathbf{L}_m = (\mathbf{q}_1 \quad \dots \quad \mathbf{q}_m)$, οπότε $\mathbf{X}^* = \mathbf{L}_m \mathbf{F} + \mathbf{u}$,

όπου $\mathbf{u} = (\mathbf{q}_{m+1} \quad \dots \quad \mathbf{q}_p) \begin{pmatrix} \tilde{z}_{m+1} \\ \vdots \\ \tilde{z}_p \end{pmatrix} = \sum_{j=m+1}^p \sqrt{\lambda_j} \epsilon_j \tilde{z}_j$ και θυμηθείτε ότι οι

κύριες συνιστώσες είναι ανεξάρτητες μεταξύ τους. Άρα

$$\mathbf{\Sigma} = \text{Var}(\mathbf{L}_m \mathbf{F} + \mathbf{u}) = \mathbf{L}_m \mathbf{I}_m \mathbf{L}_m^\top + \text{Var}(\mathbf{u}) = \mathbf{L}_m \mathbf{L}_m^\top + \text{Var}(\mathbf{u})$$

όπου $\text{Var}(\mathbf{u}) = \sum_{j=m+1}^p \lambda_j \epsilon_j \epsilon_j^\top$, όπως είπαμε προηγουμένως. Επομένως, η μέθοδος των κυρίων συνιστωσών $\rightarrow$ μια προσεγγιστική λύση για την εύρεση των κοινών παραγόντων.

Ποσοστό συνολικής διακύμανσης

- Η communality της τμ $\mathbf{X}_i$: $h_i^2 = \sum_{j=1}^m \ell_{ij}^2 = \sum_{j=1}^m \lambda_j \epsilon_{j,i}^2$
- Συνολική διακύμανση:
 $\sum_{i=1}^p h_i^2 = \sum_{i=1}^p \sum_{j=1}^m \lambda_j \epsilon_{j,i}^2 = \sum_{j=1}^m \lambda_j \sum_{i=1}^p \epsilon_{j,i}^2 = \sum_{j=1}^m \lambda_j$, αφού έχουμε μοναδιαία ιδιοδιανύσματα.
- Άρα,

$$\frac{\sum_{j=1}^p h_j^2}{\sum_{j=1}^p \sigma_j^2} = \frac{\sum_{j=1}^m \lambda_j}{\text{tr}(\mathbf{\Sigma})} = \frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j}$$

- Το ποσοστό της συνολικής διασποράς που ερμηνεύεται από τους m κοινούς παράγοντες, εάν χρησιμοποιήσουμε τη μέθοδο των κυρίων συνιστωσών είναι ίσο με

$$\frac{\sum_{j=1}^m \lambda_j}{\sum_{j=1}^p \lambda_j}$$

FA με την μέθοδο των κυρίων συνιστωσών στον πίνακα συσχέτισης

- Η μέθοδος των κυρίων συνιστωσών μπορεί επίσης να χρησιμοποιηθεί παρομοίως στον πίνακα συσχέτισης
- Πώς αιτιολογείται αυτό;
- Δουλεύουμε με τις τυποποιημένες παρατηρήσεις

$$\tilde{\mathbf{X}} = \left\{ \begin{array}{c} \frac{X_1 - \mu_1}{\sqrt{\text{Var}(X_1)}} \\ \vdots \\ \frac{X_p - \mu_p}{\sqrt{\text{Var}(X_p)}} \end{array} \right\} = \tilde{\mathbf{L}}_m \tilde{\mathbf{F}} + \tilde{\mathbf{u}}$$

και ο πίνακας συνδιακύμανσης των τυποποιημένων μεταβλητών είναι ο πίνακας συσχέτισης των αρχικών μεταβλητών.

FA με την μέθοδο των κυρίων συνιστωσών στο δείγμα

Στην πράξη η μέθοδος εφαρμόζεται στον δειγματικό πίνακα συνδιακύμανσης $\mathbf{S} = \frac{1}{n-1} \sum_i^n (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}})$ είτε στον δειγματικό πίνακα συσχέτισης $\mathbf{R}$.

FA με την μέθοδο των κύριων συνιστωσών

Έστω $(l_1, \mathbf{e}_1), \dots, (l_p, \mathbf{e}_p)$ τα ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων του $\mathbf{S}$ (ή του $\mathbf{R}$), με $l_1 \geq \dots \geq l_p > 0$. Τότε στο μοντέλο με $m < p$ παράγοντες εκτιμούμε τον πίνακα των φορτίων με

$$\hat{\mathbf{L}}_{p \times m} = (\hat{l}_{ij}) = (\sqrt{l_1} \mathbf{e}_1 \quad \sqrt{l_2} \mathbf{e}_2 \quad \dots \quad \sqrt{l_p} \mathbf{e}_p)$$

και τις ειδικές διασπορές με $\hat{\psi}_i = s_i^2 - \hat{h}_i^2 = s_i^2 - \sum_{j=1}^m \hat{l}_{ij}^2$.

Παρατηρήσεις

- Ένα πλεονέκτημα αυτής της μεθόδου είναι ότι εάν προσθέσουμε επιπλέον παράγοντες, τα φορτία των προηγούμενων παραγόντων παραμένουν αναλλοίωτα.
- Γενικά, όπως και στην PCA, οι πίνακες $\mathbf{S}$ και $\mathbf{R}$ δεν δίνουν τα ίδια αποτελέσματα.

Εκτίμηση με τη μέθοδο μέγιστης πιθανοφάνειας

- Χρειάζεται να κάνουμε συγκεκριμένες υποθέσεις για τις κατανομές των μεταβλητών.
- **Υπόθεση κανονικότητας:** $F_j \sim N(0, 1), j = 1, \dots, m$ ανεξάρτητες τμ και $u_j \sim N(0, \psi_j), j = 1, \dots, p$ ανεξάρτητες τμ. Σε μορφή πινάκων $\mathbf{F} \sim N(\mathbf{0}, \mathbf{I}_m)$ και $\mathbf{u} \sim N(\mathbf{0}, \mathbf{\Psi})$ ($\mathbf{F}$ και $\mathbf{u}$ ανεξάρτητα). Αυτό συνεπάγεται ότι

$$\mathbf{X}^* = \begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_p - \mu_p \end{pmatrix} = \mathbf{L}\mathbf{F} + \mathbf{u} \sim N(\mathbf{0}, \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi})$$

- Επειδή δεν υπάρχει μοναδική λύση για τον $\mathbf{L}$, πρέπει να εισάγουμε περιορισμούς!
- Συνήθης περιορισμός στα στατιστικά πακέτα: $\mathbf{L}^\top \mathbf{\Psi}^{-1} \mathbf{L} \rightarrow \text{διαγώνιος}$.

Έχουμε ότι $\mathbf{x}_i^* = \begin{pmatrix} x_{i1} - \mu_1 \\ \vdots \\ x_{ip} - \mu_p \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma})$, όπου $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}$. Η συνάρτηση πιθανοφάνειας

$$\begin{aligned} L(\mathbf{L}, \mathbf{\Psi}) &= \prod_{i=1}^n f(\mathbf{x}_i^* | \mathbf{L}, \mathbf{\Psi}) = \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\mathbf{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} \mathbf{x}_i^{*\top} \mathbf{\Sigma}^{-1} \mathbf{x}_i^* \right) \Rightarrow \\ \ell(\mathbf{L}, \mathbf{\Psi}) &= \log \{L(\mathbf{L}, \mathbf{\Psi})\} = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\mathbf{\Sigma}| \\ &\quad - \frac{1}{2} \text{tr} \left(\mathbf{\Sigma}^{-1} \sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*\top} \right) \Rightarrow \\ \ell(\mathbf{L}, \mathbf{\Psi}) &= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}| - \frac{n-1}{2} \text{tr} \left\{ (\mathbf{L}\mathbf{L}^\top + \mathbf{\Psi})^{-1} \mathbf{s} \right\} \end{aligned}$$

$$\ell(\mathbf{L}, \mathbf{\Psi}) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}| - \frac{n-1}{2} \text{tr} \left\{ (\mathbf{L}\mathbf{L}^\top + \mathbf{\Psi})^{-1} \mathbf{s} \right\}$$

όπου $\mathbf{s} = (n-1)^{-1} \sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*\top}$. Η μεγιστοποίηση της $\ell(\mathbf{L}, \mathbf{\Psi})$ ως προς $\mathbf{L}$ και $\mathbf{\Psi}$ υπό τον περιορισμό $\mathbf{L}^\top \mathbf{\Psi}^{-1} \mathbf{L}$ διαγώνιος γίνεται με αριθμητικές μεθόδους.

Εκτίμηση μέγιστης πιθανοφάνειας: Αριθμός παραγόντων

- Στην παραγοντική ανάλυση με m κοινούς παράγοντες, ο πίνακας $\mathbf{L}$ έχει pm στοιχεία ενώ ο πίνακας $\mathbf{\Psi}$ p μη μηδενικά στοιχεία. Άρα έχουμε $p(m+1)$ παραμέτρους ένας πίνακας συνδιακύμανσης (χωρίς δομή) έχει $\frac{p(p+1)}{2}$ μοναδικά στοιχεία.
- Μπορούμε να εκτιμήσουμε το πολύ $\frac{p(p+1)}{2}$ παραμέτρους, υπάρχει δηλαδή περιορισμός στον αριθμό των παραγόντων που μπορεί να μπει στο μοντέλο.
- Επομένως, μπορούμε να έχουμε το πολύ m_0 παράγοντες όπου m_0 είναι η λύση της εξίσωσης

$$p(m_0 + 1) = \frac{p}{2}(p + 1) \Rightarrow m_0 = \left\lfloor \frac{p-1}{2} \right\rfloor,$$

όπου $\lfloor a \rfloor$ συμβολίζει το ακέραιο μέρος του a .

Μέθοδος πιθανοφάνειας: Έλεγχος καλής εφαρμογής

Ένα πλεονέκτημα της μεθόδου μέγιστης πιθανοφάνειας είναι ότι μπορούμε να εκτελέσουμε έναν έλεγχο καλής εφαρμογής. Θεωρήστε τον έλεγχο

$$H_0 : \quad \mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$$

$$H_1 : \quad \text{όχι δομή στον } \mathbf{\Sigma}$$

Η ελεγχοσυνάρτηση βασίζεται στον λόγο των λογαρίθμων των πιθανοφανειών κάτω από την H_0 και H_1

$$LR = n \{ \text{tr}(\mathbf{A}) - \log(|\mathbf{A}|) - p \}, \quad \text{όπου } \mathbf{A} = (\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}})^{-1}\mathbf{s}$$

Ισχύει ότι $LR \sim \chi_s^2$ όπου $s = \{(p - m)^2 - (p + m)\}/2$. Απορρίπτουμε την H_0 , σε επίπεδο στατιστικής σημαντικότητας α , εάν $LR > \chi_{s,\alpha}^2$.

Παρατήρηση

Εάν το μοντέλο έχει καλή εφαρμογή, $\mathbf{A} \approx \mathbf{I} \Rightarrow |\mathbf{A}| \approx 1 \Rightarrow \log |\mathbf{A}| \approx 0$ και $\text{tr}(\mathbf{A}) \approx p$. Άρα τιμές του $LR \approx 0 \rightarrow$ καλή εφαρμογή, ενώ μεγάλες τιμές του $LR \rightarrow$ απόκλιση από την H_0 .

Έλεγχος καλής εφαρμογής: βαθμοί ελευθερίας

Η ελεγχουσυνάρτηση βασίζεται στον λόγο των λογαρίθμων των πιθανοφανειών κάτω από την H_0 και H_1

$$LR = n \{ \text{tr}(\mathbf{A}) - \log(|\mathbf{A}|) - p \}, \quad \text{όπου } \mathbf{A} = (\widehat{\mathbf{L}}\widehat{\mathbf{L}}^\top + \widehat{\Psi})^{-1} \mathbf{s}$$

Ισχύει ότι $LR \sim \chi_s^2$ όπου $s = \{(p - m)^2 - (p + m)\}/2$. Θυμηθείτε ότι έχουμε επιβάλει τον περιορισμό ο $\mathbf{L}^\top \Psi^{-1} \mathbf{L}$ να είναι διαγώνιος $\rightarrow \frac{m(m-1)}{2}$ μη διαγώνια στοιχεία να είναι διάφορα του 0. **Πώς προκύπτουν οι βαθμοί ελευθερίας;**

$$H_1 : p \text{ (μέσες τιμές)} + \frac{p(p+1)}{2} (\Sigma)$$

$$H_1 : p \text{ (μέσες τιμές)} + p(m+1) (\mathbf{L}, \Psi) - \frac{m(m-1)}{2} \text{ (περιορισμοί)}$$

Διαφορά στον αριθμό (ελεύθερων) παραμέτρων

$$\frac{p(p+1)}{2} - p(m+1) + \frac{m(m-1)}{2} = \frac{(p-m)^2 - (p+m)}{2}.$$

Παρατήρηση

Μια ενδιαφέρουσα ιδιότητα της μεθόδου μέγιστης πιθανοφάνειας στην FA είναι ότι δίνει τα ίδια αποτελέσματα οποιονδήποτε από τους **S** και **R** και να χρησιμοποιήσουμε.

Ποιά από τις δύο μεθόδους να χρησιμοποιούμε ;

Μέθοδος κυρίων συνιστωσών

- Αν η ανάλυση γίνεται κυρίως για περιγραφικούς σκοπούς
- Αν δεν μπορούμε να υποθέσουμε κανονικότητα
- Αν δεν θέλουμε η προσθήκη ή η απαλοιφή ενός παράγοντα να επηρεάζει τα προηγούμενα αποτελέσματα

Μέθοδος μέγιστης πιθανοφάνειας

- Αν η ανάλυση γίνεται για επιβεβαίωση και συμπερασματολογία
- Αν μπορούμε να υποθέσουμε κανονικότητα
- Αν είμαστε αναποφάσιστοι σχετικά με ποιον πίνακα να χρησιμοποιήσουμε

Παρατήρηση

Εάν το μοντέλο FA είναι κατάλληλο και το δείγμα αρκετά μεγάλο, οι δύο μέθοδοι θα πρέπει να δίνουν παρεμφερή αποτελέσματα. Εάν δεν συμβαίνει αυτό, τότε μάλλον η FA δεν είναι η κατάλληλη μέθοδος.

- Ανεξάρτητη της μεθόδου εκτίμησης
- Σκοπός: Αύξηση της ερμηνευτικής ικανότητας του μοντέλου.
- Γενικά: αν $\mathbf{L}$ πίνακας με επιβαρύνσεις και $\mathbf{Q}$ ορθογώνιος πίνακας τότε ισχύει

$$\mathbf{LQ}(\mathbf{LQ})^T = \mathbf{LQQL} = \mathbf{LL}^T.$$

Βασικές μέθοδοι περιστροφής:

- Varimax: Ελαχιστοποίηση αριθμού μεταβλητών με μεγάλες επιβαρύνσεις για κάθε παράγοντα
- Quartimax: Ελαχιστοποίηση αριθμού παραγόντων που εξηγούν μια μεταβλητή
- Equimax: Συνδυασμός varimax - quartimax
- Oblique: Μη ορθογώνια περιστροφή

Εφαρμογή σε πραγματικά δεδομένα

- Δεδομένα από 406 αυτοκίνητα και τα χαρακτηριστικά τους συμπεριλαμβανομένων τεχνικών χαρακτηριστικών αλλά και ήπειρο προέλευσης.

Αγγλική ονομασία	Ελληνική μετάφραση	Ονομασία που θα χρησιμοποιούμε
Miles per gallon	Κατανάλωση σε μίλα ανά γαλόνι βενζίνης	Κατανάλωση
Engine Displacement (cub. Inches)	Μέγεθος μηχανής σε κυβικές ίντσες (ίντσα $\approx$ 2.5 εκατ)	Μέγεθος μηχανής
Horsepower	Ιπποδύναμη	Ιπποδύναμη
Vehicle Weight (lbs)	Βάρος οχήματος σε λίβρες	Βάρος οχήματος
Time to accelerate from 0 to 60 miles per hour	Χρόνος επιτάχυνσης από τα 0 στα 60 μίλια την ώρα	Επιτάχυνση
Model Year (year –1900)	Χρονιά κατασκευής (κρατώντας μόνο τα 2 τελευταία ψηφία)	Χρονιά κατασκευής
Number of cylinders	Αριθμός κυλίνδρων	Αριθμός κυλίνδρων

- Να δούμε αν και κατά πόσο τα παραπάνω δεδομένα μπορούν να ερμηνευτούν με τη χρήση του ορθογώνιου παραγοντικού μοντέλου και κατά πόσο μπορούμε να ερμηνεύσουμε τους παράγοντες που προκύπτουν

$$\begin{aligned}\text{καταναλωση} &= L_{11}F_1 + L_{12}F_2 + \dots + L_{1m}F_m + u_1 \\ \text{μέγεθος μηχανης} &= L_{21}F_1 + L_{22}F_2 + \dots + L_{2m}F_m + u_2 \\ &\vdots = \vdots \\ \text{κυλινδροι} &= L_{71}F_1 + L_{72}F_2 + \dots + L_{7m}F_m + u_7\end{aligned}$$

Βασικά περιγραφικά χαρακτηριστικά

- Δεν ασχολούμαστε τόσο με τις μέσες τιμές (δεν είναι άμεσα συγκρίσιμες)
- Οι διακυμάνσεις διαφέρουν αρκετά → ο πίνακας συσχετίσεων μοιάζει πιο λογικός
- Έχει σημασία μόνο για μέθοδο κυρίων συνιστωσών αφού η μέθοδος μεγίστης πιθανοφάνειας θα δώσει τα ίδια αποτελέσματα.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Miles per Gallon	23.48	7.78	391
Engine Displacement (cu. inches)	194.13	104.63	391
Horsepower	104.24	38.28	391
Vehicle Weight (lbs.)	2973.10	845.83	391
Time to Accelerate from 0 to 60 mph (sec)	15.53	2.76	391
Model Year (modulo 100)	75.99	3.68	391
Number of Cylinders	5.47	1.70	391

Πίνακας συσχετίσεων

- Οι συσχετίσεις είναι ικανοποιητικές (> 0.4) - δεν ασχολούμαστε με στατιστική σημαντικότητα (επηρεάζεται πολύ από μέγεθος δείγματος).
- Η μεταβλητή που έχει τις μικρότερες συσχετίσεις με τις υπόλοιπες είναι η χρονιά κατασκευής.

Correlation Matrix^a

	Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)	Model Year (modulo 100)	Number of Cylinders
Miles per Gallon	1.000	-.805 .000	-.776 .000	-.831 .000	.431 .000	.577 .000	-.773 .000
Engine Displacement	-.805 .000	1.000	.898 .000	.934 .000	-.548 .000	-.367 .000	.951 .000
Horsepower	-.776 .000	.898 .000	1.000	.863 .000	-.701 .000	-.411 .000	.842 .000
Vehicle Weight (lbs.)	-.831 .000	.934 .000	.863 .000	1.000	-.425 .000	-.303 .000	.897 .000
Time to Accelerate	.431 .000	-.548 .000	-.701 .000	-.425 .000	1.000	.296 .000	-.511 .000
Model Year (modulo 100)	.577 .000	-.367 .000	-.411 .000	-.303 .000	.296 .000	1.000	-.342 .000
Number of Cylinders	-.776 .000	.951 .000	.842 .000	.897 .000	-.511 .000	-.342 .000	1.000

a. Determinant = 1.306E-04

- $KMO = 0.813$ υψηλό $\rightarrow$ ικανοποιητικές συσχετίσεις
- Ο έλεγχος σφαιρικότητας του Bartlett απορρίπτει τη μηδενική υπόθεση πως ο πίνακας συσχέτισης είναι ο μοναδιαίος (τιμή της ελεγκοσυνάρτησης 3459.6, βαθμοί ελευθερίας 21, $p < 0.001$).

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.813
Bartlett's Test of Sphericity	Approx. Chi-Square	3459.634
	df	21
	Sig.	.000

Μερικές συσχετίσεις και KMO_j

- Διαγώνια στοιχεία: KMO_j για κάθε μεταβλητή
- Μη διαγώνια στοιχεία: μερικές συσχετίσεις (θέλουμε να είναι ≈ 0)
- Ο πίνακας αυτός ονομάζεται Anti-Image.
- Χρονιά κατασκευής: η μικρότερη τιμή (0.631), εξηγείται και από τον πίνακα συσχετίσεων και δηλώνει πως αυτή η μεταβλητή είναι λιγότερο 'σχετική' με τις υπόλοιπες. Γενικά, οι τιμές κρίνονται ικανοποιητικές και δεν υπάρχει λόγος να διώξουμε κάποια μεταβλητή.

Anti-image Matrices

	Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)	Model Year (modulo 100)	Number of Cylinders
Miles per Gallon	.844 ^a	-.057	.007	.460	-.037	-.590	.053
Engine Displacement (cu. inches)	-.057	.837 ^a	-.264	-.415	.079	.098	-.674
Horsepower	.007	-.264	.839 ^a	-.379	.647	.167	.119
Vehicle Weight (lbs.)	.460	-.415	-.379	.795 ^a	-.434	-.392	-.079
Time to Accelerate from 0 to 60 mph (sec)	-.037	.079	.647	-.434	.707 ^a	.077	.063
Model Year (modulo 100)	-.590	.098	.167	-.392	.077	.631 ^a	-.023
Number of Cylinders	.053	-.674	.119	-.079	.063	-.023	.876 ^a

a. Measures of Sampling Adequacy(MSA)

- Δυναμική διαδικασία - προϋποθέτει επαναληπτικά την εκτίμηση και αξιολόγηση του μοντέλου εκτός εάν υπάρχουν ισχυρά στοιχεία ότι ο αριθμός των παραγόντων είναι συγκεκριμένος.

Κριτήρια βασισμένα στις ιδιοτιμές του πίνακα συνδιακύμανσης (ή συσχέτισης)

- κανόνα του Kaiser,
- το ποσοστό της διακύμανσης που εξηγείται
- το scree plot
- αλλά και όλα τα υπόλοιπα κριτήρια που αναφέραμε στην ανάλυση σε κύριες συνιστώσες

Κριτήρια βασισμένα στο μοντέλο (για μέθοδο μέγιστης πιθανοφάνειας)

Βάσει των ιδιοτιμών του R

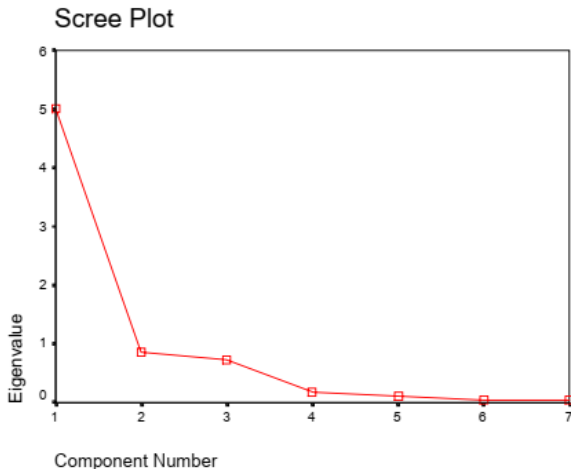
- Με αυτή την προσέγγιση, 2 παράγοντες εξηγούν πάνω από το 80% της διακύμανσης
- Αυτό ισχύει εάν χρησιμοποιήσουμε τις κύριες συνιστώσες σαν μέθοδο εκτίμησης των παραγόντων!

Total Variance Explained

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.015	71.638	71.638
2	.870	12.434	84.072
3	.721	10.305	94.377
4	.185	2.647	97.023
5	.120	1.710	98.733
6	.054	.773	99.506
7	.035	.494	100.000

Scree plot

- Όχι απόλυτα ξεκάθαρο, αλλά προς το παρόν, ας κρατήσουμε 2 κοινούς παράγοντες



Εκτίμηση παραμέτρων με PCA

Ο πίνακας φορτίων είναι $\hat{L} = (\sqrt{l_1}\mathbf{e}_1 \quad \sqrt{l_2}\mathbf{e}_2)$ όπου $(l_i, \mathbf{e}_i)$ είναι τα ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων του εκτιμημένου πίνακα συσχέτισης $\mathbf{R}$.

Component Matrix^a

	Component	
	1	2
Miles per Gallon	-.890	.188
Engine Displacement (cu. inches)	.961	.168
Horsepower	.947	.078
Vehicle Weight (lbs.)	.926	.216
Time to Accelerate from 0 to 60 mph (sec)	-.646	.015
Model Year (modulo 100)	-.510	.848
Number of Cylinders	.931	.187

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

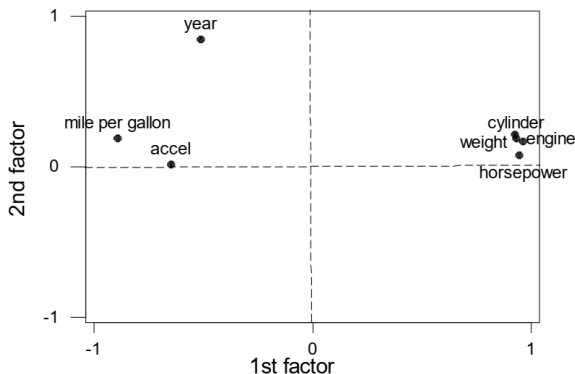
Εκτίμηση παραμέτρων με PCA

$$\begin{aligned}\text{Κατανάλωση} &= -0.890F_1 + 0.188F_2 \\ \text{Μέγεθος μηχανής} &= +0.961F_1 + 0.168F_2 \\ \text{Ιπποδύναμη} &= +0.947F_1 + 0.078F_2 \\ \text{Βάρος οχήματος} &= +0.926F_1 + 0.216F_2 \\ \text{Επιτάχυνση} &= -0.646F_1 + 0.015F_2 \\ \text{Χροιά κατασκευής} &= -0.510F_1 + 0.848F_2 \\ \text{Αριθμός κυλίνδρων} &= +0.931F_1 + 0.187F_2\end{aligned}$$

- Κατανάλωση, Επιτάχυνση και Χροιά κατασκευής έχουν αρνητικά πρόσημα ενώ οι υπόλοιπες θετικά πρόσημα για τον πρώτο παράγοντα
- Πρώτος παράγοντας: σύγκριση μεγέθους/όγκου και επιδόσεων (επιτάχυνση/κατανάλωση);

Οι επιβαρύνσεις των παραγόντων για κάθε μια μεταβλητή

- Σε κάθε άξονα αντιστοιχούμε έναν παράγοντα και για κάθε μεταβλητή την αναπαριστούμε με συντεταγμένες τις επιβαρύνσεις κάθε παράγοντα
- Παρατηρήστε πως οι μεταβλητές βάρος, ιπποδύναμη, μέγεθος μηχανής και αριθμός κυλίνδρων είναι πολύ κοντά η μία στην άλλη και σε αντιδιαστολή με τις υπόλοιπες τρεις



Οι εταιρικότητες των μεταβλητών

- Εταιρικότητες: → το ποσοστό της διακύμανσης κάθε μεταβλητής που εξηγείται από τον αριθμό των παραγόντων που προσαρμόσαμε
- ερμηνεύουμε το 83% της κατανάλωσης: $(-0.89)^2 + (0.188)^2 = 0.828$.
- 1-communalities → εκτιμήσεις ειδικοτήτων ψ_i

Communalities

	Initial	Extraction
Miles per Gallon	1.000	.828
Engine Displacement (cu. inches)	1.000	.952
Horsepower	1.000	.902
Vehicle Weight (lbs.)	1.000	.905
Time to Accelerate from 0 to 60 mph (sec)	1.000	.418
Model Year (modulo 100)	1.000	.978
Number of Cylinders	1.000	.902

Extraction Method: Principal Component Analysis.

Μέθοδος μέγιστης πιθανοφάνειας

- Μπορούμε να εκτιμήσουμε το πολύ 3 παράγοντες
- Επιλέγουμε 2 για λόγους συγκρισιμότητας

Factor Matrix^a

	Factor	
	1	2
Miles per Gallon	-.454	-.697
Engine Displacement (cu. inches)	.574	.802
Horsepower	.721	.608
Vehicle Weight (lbs.)	.453	.844
Time to Accelerate from 0 to 60 mph (sec)	-.999	3.173E-02
Model Year (modulo 100)	-.305	-.249
Number of Cylinders	.536	.789

Extraction Method: Maximum Likelihood.

a. 2 factors extracted. 7 iterations required.

Κατανάλωση	=	$-0.454F_1 - 0.697F_2$
Μέγεθος μηχανής	=	$+0.574F_1 + 0.802F_2$
Ιπποδύναμη	=	$+0.721F_1 + 0.608F_2$
Βάρος οχήματος	=	$+0.453F_1 + 0.844F_2$
Επιτάχυνση	=	$-0.999F_1 + 0.032F_2$
Χροιά κατασκευής	=	$-0.305F_1 - 0.249F_2$
Αριθμός κυλίνδρων	=	$+0.536F_1 + 0.789F_2$

- Σημαντικές διαφορές με την μέθοδο κυρίων συνιστωσών → Ισχυρή ένδειξη πως το μοντέλο δεν είναι σωστό.

Εταιρικότητες με τη μέθοδο μεγίστης πιθανοφάνειας

- Διακύμανση κάθε μεταβλητής που καταφέραμε να εξηγήσουμε με 2 παράγοντες (2η στήλη).
- Με τη μέθοδο μεγίστης πιθανοφάνειας εξηγώ μόλις το 15% της χρονιάς κατασκευής ενώ με την προηγούμενη μέθοδο είχα εξηγήσει το 98%.

Communalities^a

	Initial	Extraction
Miles per Gallon	.808	.692
Engine Displacement (cu. inches)	.951	.973
Horsepower	.897	.889
Vehicle Weight (lbs.)	.926	.917
Time to Accelerate from 0 to 60 mph (sec)	.632	.999
Model Year (modulo 100)	.474	.155
Number of Cylinders	.906	.910

Extraction Method: Maximum Likelihood.

- a. One or more communality estimates greater than 1.0 were encountered during iterations. The resulting solution should be interpreted with caution.

Το ποσοστό της διακύμανσης που εξηγούν οι 2 παράγοντες

- Η πρώτη στήλη είναι για σύγκριση με την PCA μέθοδο.

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.015	71.638	71.638	2.639	37.705	37.705
2	.870	12.434	84.072	2.896	41.377	79.082
3	.721	10.305	94.377			
4	.185	2.647	97.023			
5	.120	1.710	98.733			
6	.054	.773	99.506			
7	.035	.494	100.000			

Extraction Method: Maximum Likelihood.

Οι επιβαρύνσεις των παραγόντων για διαφορετικό αριθμό παραγόντων

- Κάθε φορά που προσθέτουμε έναν παράγοντα οι επιβαρύνσεις των προηγούμενων αλλάζουν

Μεταβλητή	Αριθμός Παραγόντων					
	3			2		1
	Παράγοντας			Παράγοντας		Παράγοντας
	1	2	3	1	2	1
Κατανάλωση	-0.856	-0.515	0.033	-0.454	-0.697	-0.823
Μέγεθος μηχανής	0.818	0.235	0.509	0.574	0.802	0.991
Ιπποδύναμη	0.884	0.059	0.319	0.721	0.608	0.907
Βάρος οχήματος	0.762	0.374	0.441	0.453	0.844	0.943
Επιτάχυνση	-0.834	0.550	0.011	-0.999	0.032	-0.553
Χρονιά κατασκευής	-0.516	-0.248	0.243	-0.305	-0.249	-0.378
Αριθμός κυλίνδρων	0.779	0.243	0.497	0.536	0.789	0.956

Αξιολόγηση του μοντέλου

- Η χρήση της μεθόδου μεγίστης πιθανοφάνειας μας επιτρέπει τη στατιστική αξιολόγηση του μοντέλου.
- Συγκεκριμένα μπορούμε να ελέγξουμε τη μηδενική υπόθεση

$$H_0 : \quad \Sigma = LL^T + \Psi$$

$$H_1 : \quad \text{Δεν υπάρχει δομή στον } \Sigma$$

- Στην πράξη, προσθέτουμε παράγοντες μέχρι το μοντέλο να μην μπορεί να απορριφθεί.

Αριθμός παραγόντων	Τιμή ελεγχ.	Βαθμοί ελευθερίας	<i>p</i> – value
1	499.618	14	0.000
2	282.514	8	0.000
3	108.140	3	0.000

- Το μοντέλο με 2 ή 3 παράγοντες δεν φαίνεται να περιγράφει καλά τα δεδομένα

Αποτυχία μοντέλου

Το μοντέλο με 2 ή 3 παράγοντες δεν φαίνεται να περιγράφει καλά τα δεδομένα. Τι μπορεί να φταιει για αυτό;

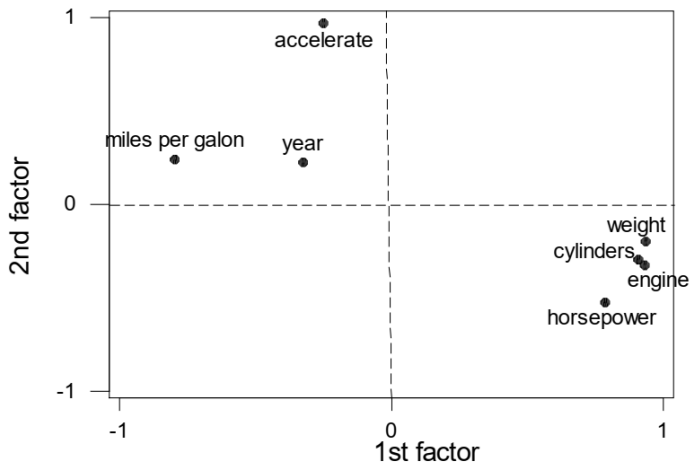
- 1 Μη κανονικότητα δεδομένων
- 2 Μη γραμμικότητα της σχέσης παραγόντων και μεταβλητών
- 3 Ύπαρξη συσχέτισης μεταξύ των παραγόντων (το μοντέλο δεν είναι ορθογώνιο)
- 4 Παραβίαση υπόθεσης πως η συσχέτιση ανάμεσα σε δύο μεταβλητές οφείλεται στην ύπαρξη κάποιων κοινών παραγόντων

Παρατήρηση

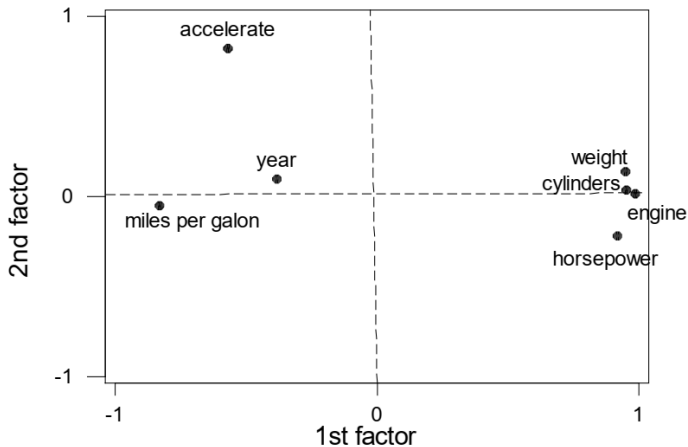
Έχειδειχτεί πως η ΕΜΠ είναι ισοδύναμη με τη μεγιστοποίηση της ορίζουσας του πίνακα των μερικών συσχετίσεων και επομένως η υπόθεση της κανονικότητας χρειάζεται περισσότερο για σκοπούς **στατιστικής συμπερασματολογίας** παρά για την **εκτίμηση των φορτίων**.

- Varimax: προσπαθεί να απομακρύνει όσο γίνεται τις μεταβλητές ώστε σε κάθε παράγοντα μόνο λίγες μεταβλητές να έχουν μεγάλες επιβαρύνσεις.
 - Αν σκοπός μας είναι να δούμε ποιες μεταβλητές επιδρούν σε κάθε παράγοντα για να αναγνωρίσουμε τον παράγοντα
- Quartimax: προσπαθεί να 'μαζέψει' τις μεταβλητές κοντά σε μια από τις δύο γραμμές που δείχνουν τους παράγοντες.
 - Σε ποιο παράγοντα είναι κάθε μεταβλητή πιο σημαντική, κι ενδεχομένως να δούμε οι μεταβλητές πως τείνουν να είναι μαζί στους παράγοντες,

Varimax Rotation



Quartimax Rotation



Δημιουργία των factor scores

$$F_1 = a_{11}X_1^* + a_{12}X_2^* + \dots a_{1p}X_p^*$$

$$F_2 = a_{21}X_1^* + a_{22}X_2^* + \dots a_{2p}X_p^*$$

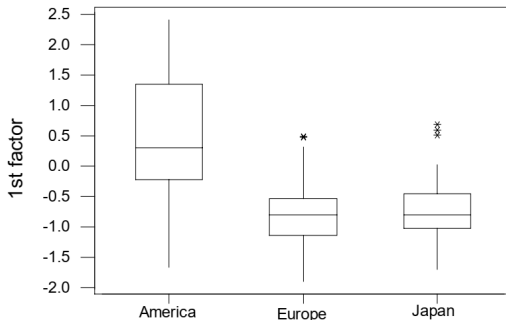
$$\vdots = \vdots$$

$$F_m = a_{m1}X_1^* + a_{m2}X_2^* + \dots a_{mp}X_p^*$$

$\mathbf{F} = \mathbf{A}\mathbf{X}^*$, $\mathbf{A} \rightarrow$ factor score coefficient matrix.

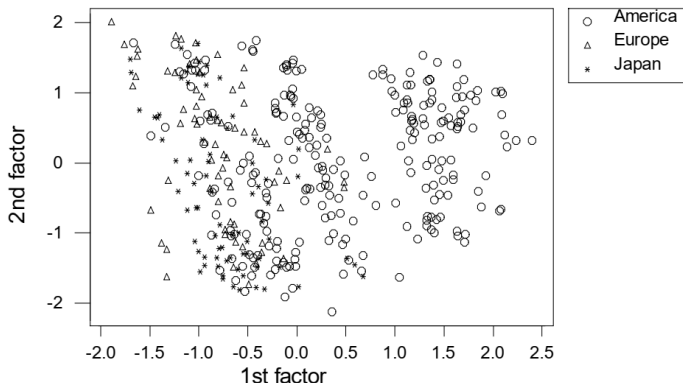
- Αποθηκεύουμε τους παράγοντες σαν νέες 'μεταβλητές'.
- Το αν θα είναι όμως ασυσχέτιστες ή όχι εξαρτάται από τη μέθοδο υπολογισμού που χρησιμοποιήσαμε αλλά και την περιστροφή που επιλέξαμε.
- Τα αυτοκίνητα είναι ταξινομημένα σε Αμερικάνικα, Ευρωπαϊκά και Ιαπωνικά.

Παράγοντας 1 ανά προέλευση αυτοκινήτου



- Πρώτος παράγοντας: σύγκριση ανάμεσα στο μέγεθος του αυτοκινήτου και τις επιδόσεις του.
- Λογικό αποτέλεσμα, αφού τα αμερικάνικα είναι συνήθως ογκώδη και για αυτό ξεχωρίζουν από τα υπόλοιπα.

Χρήση και των δύο παραγόντων



- Χρησιμοποιώντας την παραγοντική ανάλυση η πληροφορία που συμπυκνώνουν οι 2 πρώτοι παράγοντες είναι αρκετή για να αποκτήσουμε την ικανότητα να ξεχωρίζουμε τα αυτοκίνητα

- Το παραγοντικό μοντέλο απέτυχε στο συγκεκριμένο σετ δεδομένων.
- Χρησιμοποιώντας το απλά για περιγραφικούς σκοπούς κάποιος μπορεί να αποκτήσει μια εικόνα σχετικά με τον τρόπο που συσχετίζονται οι μεταβλητές ακόμα και αν το μοντέλο δεν είναι καλό
- Πολλές φορές το παραγοντικό μοντέλο χρησιμοποιείται απλά για περιγραφικούς σκοπούς αγνοώντας κάθε μορφή ελέγχου της καταλληλότητας του μοντέλου