

Παραγοντική ανάλυση

Λουκία Μελικοτσίδου

Εισαγωγή

Στην R, η ανάλυση παραγόντων υλοποιείται από τη συνάρτηση `factanal()` του βασικού πακέτου της R. Η συνάρτηση εκτελεί ανάλυση παραγόντων με τη μέθοδο μέγιστης πιθανοφάνειας σε έναν πίνακα συνδιακύμανσης/συσχέτισης ή έναν πίνακα δεδομένων. Ο αριθμός των παραγόντων καθορίζεται από το όρισμα `factors`. Επιπλέον, οι βαθμολογίες παραγόντων (`factor scores`) μπορούν να υπολογιστούν είτε χρησιμοποιώντας τον εκτιμητή Thompson είτε τη μέθοδο σταθμισμένων ελαχίστων τετραγώνων του Bartlett. Η συγκεκριμένη μέθοδος καθορίζεται από ένα πρόσθετο όρισμα `scores = "regression"` ή `scores = "Bartlett"`. Επιπλέον, με το όρισμα `rotation`, ο μετασχηματισμός των παραγόντων μπορεί να καθοριστεί είτε με `rotation = "varimax"` για ορθογώνια περιστροφή, `rotation = "Bartlett"` για μη ορθογώνια περιστροφή ή `rotation = "none"` για καμία περιστροφή.

Food dataset

Σαν παράδειγμα, θα χρησιμοποιήσουμε δεδομένα που αφορούν στην υφή προϊόντων ζαχαροπλαστικής με βάση το ζυμάρι (πίτες, κρουασάν και άλλα παρόμοια). Τα δεδομένα περιλαμβάνουν 50 προϊόντα και τις 5 ακόλουθες μεταβλητές:

- Oil: ποσοστό λαδιού στη ζύμη
- Density: η πυκνότητα του προϊόντος (όσο μεγαλύτερος είναι ο αριθμός, τόσο πιο πυκνό είναι το προϊόν)
- Crispy: μια μέτρηση τραγανότητας, σε κλίμακα από το 7 έως το 15, με το 15 να είναι πιο τραγανό
- Fracture: η γωνία, σε μοίρες, την οποία το προϊόν μπορεί να λυγίσει αργά πριν σπάσει
- Hardness: χρησιμοποιείται ένα αιχμηρό αντικείμενο για τη μέτρηση της ποσότητας της δύναμης που απαιτείται πριν από τη θραύση

Φορτώνουμε πρώτα τη βιβλιοθήκη που θα χρειαστεί

```
# Load Libraries  
library(psych)
```

Εισάγουμε τα δεδομένα `food` ως `data.frame` στην R.

```
# Load data
food <- read.csv("food.csv")
```

Έλεγχος συσχετίσεων

Πριν ξεκινήσουμε την παραγοντική ανάλυση είναι απαραίτητο να εκτελέσουμε μια διερευνητική ανάλυση στις συσχετίσεις μεταξύ των μεταβλητών. Αρχικά, ας ρίξουμε μια ματιά στις μέσες τιμές και τις τυπικές αποκλίσεις των μεταβλητών.

```
#####
### Descriptive characteristics ###
#####
round(cbind(colMeans(food), apply(food, 2, sd)), 2)

##           [,1] [,2]
## Oil         17.20  1.59
## Density    2857.60 124.50
## Crispy       11.52  1.78
## Fracture     20.86  5.47
## Hardness    128.18 31.13
```

Υπενθυμίζεται ότι η `colMeans` υπολογίζει το διανυσματικό μέσο ενός πίνακα/`data.frame`, ενώ η `apply(food, 2, sd)` εφαρμόζει τη συνάρτηση `sd` (τυπική απόκλιση) στις στήλες του `data.frame` `food`. Παρατηρήστε ότι οι μέσες τιμές διαφέρουν αρκετά, αλλά αυτό δεν μας αφορά καθόλου. Οι τυπικές αποκλίσεις επίσης διαφέρουν σημαντικά, το οποίο όμως δεν έχει σημασία για τη μέθοδο μέγιστης πιθανοφάνειας, αλλά είναι σημαντικό εάν χρησιμοποιήσουμε την μέθοδο κυρίων συνιστωσών για την εκτίμηση των παραγόντων. Επομένως για τη μέθοδο των κυρίων συνιστωσών, θα προτιμούσαμε μάλλον τον πίνακα συσχέτισης αντί του πίνακα συνδιακύμανσης.

Στην συνέχεια εξετάζουμε αρχικά τις συσχετίσεις μέσω ενός γραφήματος και του πίνακα συσχετίσεων.

```
# Correlation plot
pairs(food)
```

```
# Correlation matrix
print(cor(food), 2)

##           Oil Density Crispy Fracture Hardness
## Oil         1.000   -0.75   0.59   -0.53   -0.096
## Density    -0.750    1.00  -0.67    0.57    0.108
## Crispy       0.593   -0.67   1.00   -0.84    0.411
## Fracture    -0.534    0.57  -0.84    1.00   -0.373
## Hardness    -0.096    0.11   0.41   -0.37    1.000
```

Οι συσχετίσεις είναι γενικά ικανοποιητικές (>0.40), επομένως έχει νόημα να συνεχίσουμε με παραγοντική ανάλυση. Παρατηρήστε ότι η Hardness έχει τη μικρότερη συσχέτιση με τις υπόλοιπες. Επίσης, με βάση το γράφημα, βλέπουμε ότι οι συσχετίσεις μεταξύ των μεταβλητών είναι γραμμικές (το οποίο είναι καλό). Οι κατανομές των μεταβλητών, λαμβάνοντας υπόψιν το μικρό μέγεθος δείγματος ($n = 50$) δεν αποκλίνουν ιδιαίτερα από την κανονική κατανομή (σημαντικό εάν χρησιμοποιήσουμε την μέθοδο μέγιστης πιθανοφάνειας).

Ένα καλό περιγραφικό μέτρο καταλληλότητας των συσχετίσεων είναι το κριτήριο των Kaiser-Meyer-Olkin (KMO), το οποίο συγκρίνει τις ανά δύο συσχετίσεις με τους συντελεστές μερικής συσχέτισης. Το στατιστικό KMO παίρνει τιμές στο 0-1 και δείχνει πόσο υψηλές είναι οι συσχετίσεις συνολικά.

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

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

το οποίο υπολογίζεται και για κάθε μεταβλητή χωριστά ως

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

Επίσης, θα θέλαμε α απορρίψουμε την μηδενική υπόθεση $H_0: \mathbf{P} = \mathbf{I}_5$ (Έλεγχος σφαιρικότητας του Bartlett).

Οι δύο αυτοί έλεγχοι μπορούν να υπολογιστούν αυτόματα χρησιμοποιώντας τις συναρτήσεις `KMO` και `cortest.bartlett` του πακέτου `psych`.

```
# KMO
KMO(food)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = food)
## Overall MSA = 0.71
## MSA for each item =
##      Oil Density   Crispy Fracture Hardness
##      0.82      0.71      0.67      0.79      0.43

# Testing correlations
cortest.bartlett(food)

## $chisq
## [1] 154.9936
##
## $p.value
## [1] 3.492964e-28
##
## $df
## [1] 10
```

Η τιμή του KMO είναι ίση με 0.71, το οποίο δεν είναι ιδιαίτερα υψηλό, αλλά τουλάχιστον δεν είναι αποθαρρυντικό (ιδανικά θέλουμε >0.80). Όπως αναμενόταν, απορρίπτουμε την υπόθεση ότι ο πίνακας συσχέτισης είναι μοναδιαίος.

Παραγοντική ανάλυση με την `factanal()`

Εκτός από το σύνολο δεδομένων, η συνάρτηση `factanal()` απαιτεί μια εκτίμηση του αριθμού των παραγόντων (`factanal(data, factors = n)`). Αυτή είναι μια δύσκολη πτυχή της παραγοντικής ανάλυσης. Εάν έχουμε μια υπόθεση για τους παράγοντες, μπορούμε να ξεκινήσουμε με μια τεκμηριωμένη εικασία. Εάν δεν έχουμε καμία ιδέα για τον αριθμό των παραγόντων και ο αριθμός των μεταβλητών στο σύνολο δεδομένων δεν είναι πολύ μεγάλος, μπορούμε απλώς να δοκιμάσουμε πολλές τιμές για την προετοιμασία του μοντέλου. Μια άλλη, πιο περίπλοκη προσέγγιση είναι η χρήση της ανάλυσης κύριων συνιστωσών για να ληφθεί μια καλή αρχική εκτίμηση του αριθμού των παραγόντων.

Σε αυτό το απλό παράδειγμα κάνουμε απλώς μια εικασία και ορίζουμε τον αριθμό του παράγοντα σε 2. Επιπλέον, διατηρούμε τις προεπιλογές για τις βαθμολογίες (`score = "none"`) και την περιστροφή (`rotation = "varimax"`).

Ερμηνεία των αποτελεσμάτων

Πριν ερμηνεύσουμε τα αποτελέσματα της παραγοντικής ανάλυσης, θυμηθείτε τη βασική ιδέα πίσω από αυτήν. Η παραγοντική ανάλυση δημιουργεί γραμμικούς συνδυασμούς παραγόντων για να δώσει μια δομή στην ομοιότητα μεταξύ των μεταβλητών. Στο βαθμό που οι μεταβλητές έχουν μια υποκείμενη ομοιότητα, λιγότεροι παράγοντες καταγράφουν το μεγαλύτερο μέρος της διακύμανσης στο σύνολο δεδομένων. Αυτό μας επιτρέπει να συγκεντρώσουμε μεγάλο αριθμό παρατηρήσιμων μεταβλητών σε ένα μοντέλο για να αναπαραστήσουμε μια υποκείμενη έννοια, καθιστώντας ευκολότερη την κατανόηση των δεδομένων. Η μεταβλητότητα στα δεδομένα μας, \mathbf{X} , δίνεται από το $\mathbf{\Sigma}$, και η εκτίμησή του $\hat{\mathbf{\Sigma}}$ αποτελείται από τη μεταβλητότητα που εξηγείται από τους παράγοντες (**communality**) και της μεταβλητότητας η οποία δεν μπορεί να εξηγηθεί από ένα γραμμικό συνδυασμό των παραγόντων (**uniqueness**).

$$\hat{\mathbf{\Sigma}} = \underbrace{\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T}_{\text{communality}} + \underbrace{\mathbf{\Psi}}_{\text{uniqueness}}.$$

```
#####
### Fitting MLE model (default options) ###
#####
food.fa <- factanal(food, factors = 2)
food.fa

##
## Call:
## factanal(x = food, factors = 2)
##
```

```
## Uniquenesses:
##      Oil   Density   Crispy Fracture Hardness
##      0.334   0.156    0.042   0.256    0.407
##
## Loadings:
##           Factor1 Factor2
## Oil        -0.816
## Density     0.919
## Crispy     -0.745   0.635
## Fracture    0.645  -0.573
## Hardness           0.764
##
##           Factor1 Factor2
## SS loadings      2.490   1.316
## Proportion Var   0.498   0.263
## Cumulative Var   0.498   0.761
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 0.27 on 1 degree of freedom.
## The p-value is 0.603
```

- Το output του μοντέλου με την κλήση της συνάρτησης για να μας υπενθυμίζει τι ζητήσαμε.
- Το πρώτο κομμάτι παρέχει τις ειδικότητες (*uniquenesses*), οι οποίες κυμαίνονται από 0 έως 1. Η ειδικότητα, που μερικές φορές αναφέρεται ως θόρυβος-σφάλμα, αντιστοιχεί στο ποσοστό της μεταβλητότητας η οποία δεν μπορεί να εξηγηθεί από έναν γραμμικό συνδυασμό των παραγόντων. Αυτό είναι το Ψ στην παραπάνω εξίσωση. Μια **υψηλή ειδικότητα** για μια μεταβλητή υποδηλώνει ότι **οι παράγοντες δεν εξηγούν καλά τη διακύμανσή της**.

```
# uniquenesses
```

```
food.fa$uniquenesses
```

```
##      Oil   Density   Crispy Fracture Hardness
## 0.3338599 0.1555255 0.0422238 0.2560235 0.4069459
```

- Η επόμενη ενότητα είναι τα φορτία, τα οποία κυμαίνονται από -1 έως 1. Αυτό είναι το Λ στην παραπάνω εξίσωση. Τα φορτία είναι η συμβολή κάθε παράγοντα στην κάθε μεταβλητή. Οι μεταβλητές με υψηλά φορτία εξηγούνται καλά από τους παράγοντες. Σημειώστε ότι δεν υπάρχει καταχώρηση για ορισμένες μεταβλητές. Αυτό συμβαίνει επειδή η R δεν δείχνει φορτία μικρότερα από 0.1. Αυτό έχει σκοπό να μας βοηθήσει να εντοπίσουμε ομάδες μεταβλητών. Πληκτρολογήστε `help(loadings)` στην κονσόλα σας για περισσότερες λεπτομέρειες.
- Τετραγωνίζοντας το φορτίο υπολογίζουμε το ποσοστό της συνολικής διακύμανσης της μεταβλητής που εξηγείται από τον παράγοντα. Αυτή η

αναλογία της μεταβλητότητας ονομάζεται ως εταιρικότητα (communality). Ένας άλλος τρόπος υπολογισμού της εταιρικότητας είναι να αφαιρέσετε τις ειδικότητες από το 1. Ένα κατάλληλο μοντέλο παραγόντων έχει ως αποτέλεσμα χαμηλές τιμές για τη ειδικότητα και υψηλές τιμές για την εταιρικότητα.

```
# communality
round(rowSums(food.fa$loadings^2),3)

##      Oil   Density   Crispy Fracture Hardness
##      0.666    0.844    0.958    0.744    0.593

round(1 - rowSums(food.fa$loadings^2),3) # uniqueness

##      Oil   Density   Crispy Fracture Hardness
##      0.334    0.156    0.042    0.256    0.407
```

Ο πίνακας κάτω από τα φορτία δείχνει την αναλογία διακύμανσης που εξηγείται από κάθε παράγοντα. Η σειρά Cumulative Var δίνει την αθροιστική αναλογία διακύμανσης που εξηγείται. Αυτοί οι αριθμοί κυμαίνονται από το 0 έως το 1. Η γραμμή Proportion Var δίνει την αναλογία διακύμανσης που εξηγείται από κάθε παράγοντα και οι τα SS loadings δίνουν το άθροισμα των τετραγωνικών φορτίων. Αυτό μερικές φορές χρησιμοποιείται για τον προσδιορισμό της αξίας ενός συγκεκριμένου παράγοντα. Ένας παράγοντας αξίζει να κρατηθεί εάν το φορτίο του είναι μεγαλύτερο από 1 (ανάλογο με τον κανόνα Kaiser για ιδιοτιμές).

Το τελευταίο τμήμα του output δείχνει τα αποτελέσματα ενός ελέγχου. Η μηδενική υπόθεση, H_0 , είναι ότι ο αριθμός των παραγόντων στο μοντέλο, στο παράδειγμά μας 2 παράγοντες, είναι επαρκής για να εξηγήσει την δομή συνδιακύμανσης του συνόλου δεδομένων. Συμβατικά, απορρίπτουμε την H_0 εάν η τιμή p είναι μικρότερη από 0.05. Ένα τέτοιο αποτέλεσμα δείχνει ότι ο αριθμός των παραγόντων είναι πολύ μικρός. Αντίθετα, δεν απορρίπτουμε την H_0 εάν η τιμή p υπερβαίνει το 0.05. Ένα τέτοιο αποτέλεσμα δείχνει ότι υπάρχουν πιθανώς αρκετοί (ή περισσότεροι από αρκετοί) παράγοντες που αποτυπώνουν την πλήρη πολυπλοκότητα του συνόλου δεδομένων. Η υψηλή τιμή p στο παραπάνω παράδειγμά μας μας οδηγεί να μην απορρίψουμε την H_0 και υποδεικνύει ότι χρησιμοποιήσαμε ένα κατάλληλο μοντέλο. Αυτός ο έλεγχος υποθέσεων είναι διαθέσιμος χάρη στη μέθοδο μέγιστης πιθανοφάνειας. Σημειώστε ότι εάν δώσετε έναν πίνακα συνδιακύμανσης στη συνάρτηση `factanal()` και όχι ένα σύνολο δεδομένων, ο έλεγχος υπόθεσης δεν είναι διαθέσιμος εάν δεν παρέχουμε ρητά τον αριθμό των παρατηρήσεων (`n.obs`) ως πρόσθετο όρισμα στην κλήση της συνάρτησης.

Ο πίνακας καταλοίπων

Η συνάρτηση `factanal` ουσιαστικά δουλεύει με τον πίνακα συσχέτισης και όχι με τον πίνακα συνδιακύμανσης. Αυτό όμως δεν έχει επίδραση στα αποτελέσματα αφού χρησιμοποιούμε την μέθοδο μέγιστης πιθανοφάνειας.

$$\hat{\mathbf{P}} = \underbrace{\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T}_{\text{communality}} + \underbrace{\hat{\mathbf{\Psi}}}_{\text{uniqueness}}.$$

όπου \mathbf{P} ο πίνακας συσχέτισης των δεδομένων.

Χρησιμοποιώντας το παραγοντικό μας μοντέλο `food.fa` μπορούμε να υπολογίσουμε το $\hat{\mathbf{P}}$ και να το συγκρίνουμε με τον παρατηρούμενο πίνακα συσχέτισης, \mathbf{R} , με απλή άλγεβρα πινάκων. Ο τελεστής `%%` εκτελεί πολλαπλασιασμό πινάκων. Η συνάρτηση `t()` δημιουργεί τον ανάστροφο ενός πίνακα. Η συνάρτηση `diag()` παίρνει ένα διάνυσμα k αριθμών και δημιουργεί έναν πίνακα $k \times k$ με τους αριθμούς στη διαγώνιο και 0 αλλού.

```
# The residual matrix
Lambda <- food.fa$loadings
Psi <- diag(food.fa$uniquenesses)
R <- food.fa$correlation
Rhat <- Lambda %*% t(Lambda) + Psi
```

Αφαιρούμε τώρα τον εκτιμημένο πίνακα συσχέτισης, `Rhat`, από τον παρατηρούμενο πίνακα συσχέτισης, `R`. Στρογγυλοποιούμε επίσης το αποτέλεσμα σε 6 ψηφία.

```
round(R - Rhat, 6)

##           Oil   Density   Crispy   Fracture   Hardness
## Oil      0.000000  0.000001 -0.002613 -0.018220 -0.000776
## Density  0.000001  0.000000 -0.001081 -0.007539 -0.000320
## Crispy   -0.002613 -0.001081  0.000000  0.000000  0.000005
## Fracture -0.018220 -0.007539  0.000000  0.000000  0.000033
## Hardness -0.000776 -0.000320  0.000005  0.000033  0.000000
```

Ο πίνακας που προκύπτει ονομάζεται πίνακας καταλοίπων (**residual matrix**). Αριθμοί κοντά στο 0 υποδεικνύουν ότι το παραγοντικό μοντέλο μας είναι μια καλή αναπαράσταση των δεδομένων. Δεν ασχολούμαστε με την κύρια διαγώνιο διότι είναι πάντα 0 εξ ορισμού. Κοιτάμε μόνο τα μη διαγώνια στοιχεία, εάν δηλαδή η δομή συσχέτισης που προτείνει το μοντέλο απέχουν πολύ από τον παρατηρούμενο πίνακα συσχέτισης που δεν κάνει καμία υπόθεση για τις συσχετίσεις μεταξύ των μεταβλητών.

Ερμηνεία των παραγόντων

Ο σκοπός μιας περιστροφής είναι να δημιουργήσει παράγοντες με συνδυασμό υψηλών και χαμηλών φορτίων και λίγα φορτία μετρίου μεγέθους. Η ιδέα είναι να δοθεί κάποιο νόημα στους παράγοντες, κάτι που βοηθά στην ερμηνεία τους. Από μαθηματική άποψη, δεν υπάρχει διαφορά μεταξύ ενός περιστραμμένου και μη περιστραμμένου πίνακα. Το προσαρμοσμένο μοντέλο είναι το ίδιο, οι ειδικότητες είναι ίδιες και η αναλογία διακύμανσης που εξηγείται είναι η ίδια.

Ας προσαρμόσουμε τρία μοντέλα παραγόντων, ένα χωρίς περιστροφή, ένα με περιστροφή varimax και ένα με περιστροφή promax, και να κάνουμε ένα διάγραμμα σημείων (scatterplot) των φορτίων του πρώτου και δεύτερου παράγοντα.

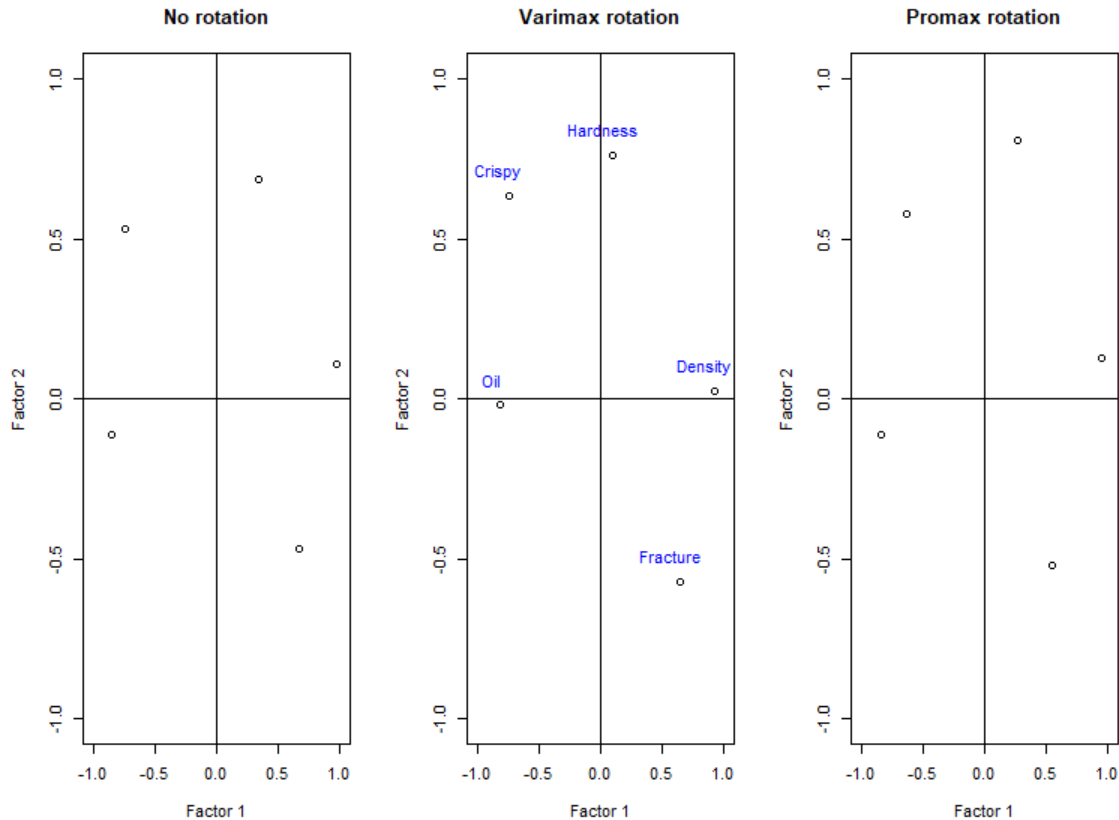
```
# Interpretation of the factors
food.fa.none <- factanal(food, factors = 2, rotation = "none")
food.fa.varimax <- factanal(food, factors = 2, rotation = "varimax")
food.fa.promax <- factanal(food, factors = 2, rotation = "promax")

par(mfrow = c(1,3))
plot(food.fa.none$loadings[,1],
      food.fa.none$loadings[,2],
      xlab = "Factor 1",
      ylab = "Factor 2",
      ylim = c(-1,1),
      xlim = c(-1,1),
      main = "No rotation")
abline(h = 0, v = 0)

plot(food.fa.varimax$loadings[,1],
      food.fa.varimax$loadings[,2],
      xlab = "Factor 1",
      ylab = "Factor 2",
      ylim = c(-1,1),
      xlim = c(-1,1),
      main = "Varimax rotation")

text(food.fa.varimax$loadings[,1]-0.08,
      food.fa.varimax$loadings[,2]+0.08,
      colnames(food),
      col="blue")
abline(h = 0, v = 0)

plot(food.fa.promax$loadings[,1],
      food.fa.promax$loadings[,2],
      xlab = "Factor 1",
      ylab = "Factor 2",
      ylim = c(-1,1),
      xlim = c(-1,1),
      main = "Promax rotation")
abline(h = 0, v = 0)
```

Τώρα έρχεται η δύσκολη πτυχή της παραγοντικής ανάλυσης: Η ερμηνεία των ίδιων των παραγόντων. Εάν δύο μεταβλητές έχουν και οι δύο μεγάλα φορτία για τον ίδιο παράγοντα, τότε ξέρουμε ότι έχουν κάτι κοινό. Ως ερευνητές πρέπει να κατανοήσουμε τα δεδομένα και τη σημασία τους για να δώσουμε ένα όνομα σε αυτήν την ομοιότητα. Ρίχνοντας μια ματιά στα παραπάνω σχήματα φαίνεται ότι ο παράγοντας 1 είναι υπεύθυνος για αρτοσκευάσματα/σφολιατοειδή τα οποία είναι πυκνά και μπορούν να λυγίσουν πολύ πριν σπάσουν. Ενώ ο παράγοντας 2 ευθύνεται για αρτοσκευάσματα που είναι τραγανά και δύσκολο να διαρραγούν. Έτσι, αν χρειαστεί να ονομάσουμε αυτούς τους παράγοντες, θα τους ονομάζαμε πιθανώς μαλακότητα/ευπλασία προϊόντος (παράγοντας 1) και σκληρότητα/τραγανότητα προϊόντος (παράγοντας 2).

Μέθοδος κυρίων συνιστωσών

Εφαρμόζουμε το μοντέλο με την συνάρτηση `principal` της βιβλιοθήκης `psych` η οποία εκτελεί παραγοντική ανάλυση βάσει της μεθόδου των κυρίων συνιστωσών. Η σύνταξη της εντολής είναι παρόμοια. Δίνουμε αρχικά το `data.frame` `food`, τον αριθμό των παραγόντων (`nfactors = 2`), το `covar = F` δηλώνει ότι έχουμε δώσει έναν πίνακα δεδομένων ως όρισμα και όχι έναν πίνακα συνδιακύμανσης, ενώ με το `rotate = "none"` δηλώνουμε ότι δεν θέλουμε περιστροφή.

```
# Factor analysis using principal component method
fitPCA <- principal(food, nfactors = 2, covar = F, rotate = "none")
fitPCA

## Principal Components Analysis
## Call: principal(r = food, nfactors = 2, rotate = "none", covar = F)
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PC1  PC2  h2  u2 com
## Oil      0.80 -0.42 0.81 0.19 1.5
## Density -0.83  0.41 0.86 0.14 1.4
## Crispy   0.93  0.22 0.91 0.09 1.1
## Fracture -0.88 -0.25 0.83 0.17 1.2
## Hardness 0.27  0.92 0.91 0.09 1.2
##
##          PC1  PC2
## SS loadings      3.03 1.30
## Proportion Var    0.61 0.26
## Cumulative Var    0.61 0.87
## Proportion Explained 0.70 0.30
## Cumulative Proportion 0.70 1.00
##
## Mean item complexity = 1.3
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is 0.06
## with the empirical chi square 3.55 with prob < 0.06
##
## Fit based upon off diagonal values = 0.99
```

Παρόμοια, για να βρούμε τα communalities αθροίζουμε τα τετράγωνα του πίνακα φορτίων ανά μεταβλητή.

```
round(rowSums(fitPCA$loadings[,1:2]^2),2)

##      Oil  Density  Crispy Fracture Hardness
##      0.81    0.86    0.91    0.83    0.91
```

Για εκπαιδευτικούς λόγους, επιβεβαιώνουμε τα αποτελέσματα εκτελώντας φασματική ανάλυση του πίνακα συσχέτισης. Θυμηθείτε ότι χρησιμοποιώντας τη μέθοδο των κυρίων συνιστωσών, οι κοινοί παράγοντες είναι οι τυποποιημένες κύριες συνιστώσες!

```
# Eigenvalue-vector decomposition of the correlation matrix
eigCor <- eigen(cor(food))
round(cumsum(eigCor$values/sum(eigCor$values))*100,2)

## [1] 60.62 86.54 92.74 97.58 100.00

round(t(sqrt(eigCor$values)*t(eigCor$vectors)),2)[,1:2]
```

```
##      [,1] [,2]
## [1,]  0.80  0.42
## [2,] -0.83 -0.41
## [3,]  0.93 -0.22
## [4,] -0.88  0.25
## [5,]  0.27 -0.92
```

Στην συνέχεια θέλουμε να συγκρίνουμε τα αποτελέσματα της μεθόδου κύριων συνιστωσών με αυτά της μέγιστης πιθανοφάνειας. Επομένως, εφαρμόζουμε πρώτα *varimax* περιστροφή πρώτα

```
# Factor analysis using principal component method
fitPCA <- principal(food, nfactors = 2, covar = F, rotate = "varimax")
fitPCA

## Principal Components Analysis
## Call: principal(r = food, nfactors = 2, rotate = "varimax", covar = F)
## Standardized loadings (pattern matrix) based upon correlation matrix
##          RC1  RC2  h2  u2 com
## Oil      -0.90 -0.08 0.81 0.19 1.0
## Density   0.93  0.05 0.86 0.14 1.0
## Crispy    -0.77  0.57 0.91 0.09 1.8
## Fracture   0.71 -0.57 0.83 0.17 1.9
## Hardness   0.11  0.95 0.91 0.09 1.0
##
##          RC1  RC2
## SS loadings      2.77 1.56
## Proportion Var    0.55 0.31
## Cumulative Var    0.55 0.87
## Proportion Explained 0.64 0.36
## Cumulative Proportion 0.64 1.00
##
## Mean item complexity = 1.4
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is 0.06
## with the empirical chi square 3.55 with prob < 0.06
##
## Fit based upon off diagonal values = 0.99
```

Συγκρίνουμε τα αποτελέσματα με αυτά από την μέθοδο μέγιστης πιθανοφάνειας.

```
# Compare results with the MLE
par(mfrow = c(1,2))
plot(food.fa.varimax$loadings[,1],
      food.fa.varimax$loadings[,2],
      xlab = "Factor 1",
      ylab = "Factor 2",
      ylim = c(-1,1),
      xlim = c(-1,1),
```

```

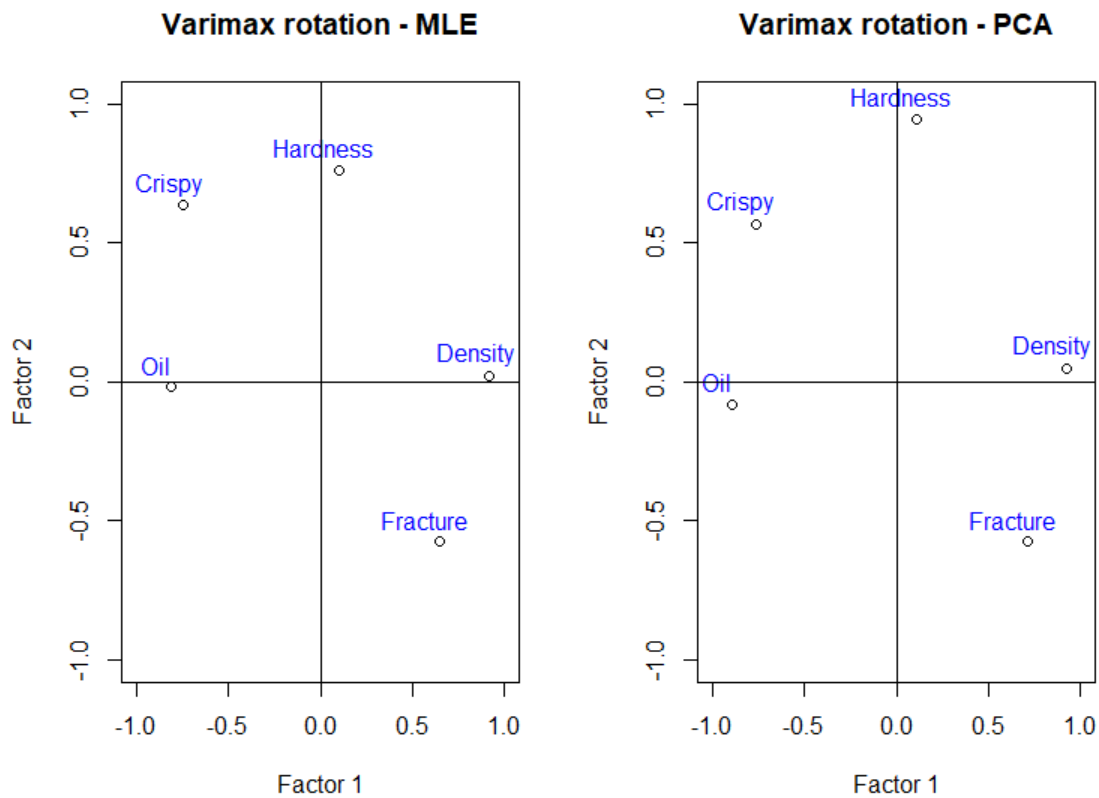
    main = "Varimax rotation - MLE")

text(food.fa.varimax$loadings[,1]-0.08,
     food.fa.varimax$loadings[,2]+0.08,
     colnames(food),
     col="blue")
abline(h = 0, v = 0)

plot(fitPCA$loadings[,1],
     fitPCA$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax rotation - PCA")

text(fitPCA$loadings[,1]-0.08,
     fitPCA$loadings[,2]+0.08,
     colnames(food),
     col="blue")
abline(h = 0, v = 0)

```



Υπολογίζουμε επίσης τα σκορ των δύο κοινών παραγόντων με τη μέθοδο της παλινδρόμησης (προεπιλογή της `principal`). Θυμηθείτε επίσης ότι χρειάζεται πρώτα να τυποποιήσουμε τα δεδομένα ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Αυτό γίνεται αυτόματα στην R μέσω της εντολής `scale`.

$$\tilde{\mathbf{X}}\mathbf{R}^{-1}\hat{\mathbf{\Lambda}}$$

```
scale(food)[1:5,] %*% ( solve(cor(food)) %*% fitPCA$loadings[,1:2] )

##              RC1              RC2
## [1,]  0.5208946 -0.8099158
## [2,] -1.3585163  0.9095213
## [3,]  0.1680757  0.7504091
## [4,]  0.6371111 -1.3568066
## [5,]  0.8974277  0.2544005

fitPCA$scores[1:5,]

##              RC1              RC2
## [1,]  0.5208946 -0.8099158
## [2,] -1.3585163  0.9095213
## [3,]  0.1680757  0.7504091
## [4,]  0.6371111 -1.3568066
## [5,]  0.8974277  0.2544005
```

Παρατηρήστε ότι τα σκορ των παραγόντων έχουν μέση τιμή 0, διακύμανση 1 και είναι ασυσχέτιστα μεταξύ τους. Αυτό συμβαίνει διότι έχουμε επιλέξει την *ορθογώνια περιστροφή* `varimax`.

```
# Factor scores
colMeans(fitPCA$scores)

##              RC1              RC2
## -2.137179e-16 -2.065015e-16

cov(fitPCA$scores)

##              RC1              RC2
## RC1 1.000000e+00 4.624172e-16
## RC2 4.624172e-16 1.000000e+00
```