Μέθοδοι Πολυδιάστατης Ανάλυσης Δεδομένων με τη γλώσσα R

Ν. Κουτσουπιάς και Γ. Παπαδημητρίου
Πανεπιστήμιο Μακεδονίας

Περίληψη

Σύγχρονα εργαλεία ανοικτού κώδικα της "στατιστικής" γλώσσας R περιλαμβάνουν ποικίλες ρουτίνες μεθόδων της Πολυδιάστατης Ανάλυσης Δεδομένων (Π.Α.Δ.). Στο παρόν συνοπτικά παρουσιάζονται, ο τρόπος εγκατάστασης και οι βασικές λειτουργίες της R και του γραφικού περιβάλλοντος της, R-Studio, καθώς και η χρήση και παραμετροποίηση πρόσθετου πακέτου για την εφαρμογή μεθόδων της Π.Α.Δ.

Εισαγωγή

Η γλώσσα R[1] (Ihaka & Gentlemen 1996, Team 2000), διάδοχος της γλώσσας S (Chambers 1998), είναι μία υψηλού επιπέδου γλώσσα προγραμματισμού διαχείρισης, ανάλυσης και οπτικοποίησης δεδομένων. Έχει καταστεί ευρείας χρήσης και αναγνώρισης, τόσο στον ακαδημαϊκό (Jones 2009, Kelley *et al.* 2008) όσο και τον επιχειρηματικό (Level 2017) χώρο παγκοσμίως, λόγω της ευελιξίας, της ευχρηστίας, του ευρύτατου φάσματος εφαρμογών, αλλά και του ότι διατίθεται δωρεάν για τα σημαντικότερα λειτουργικά συστήματα (Windows, MacOS, Linux).

Κατά την τελευταία εικοσαετία έχουν αναπτυχθεί πληθώρα πρόσθετων πακέτων (packages) για το περιβάλλον της γλώσσας ώστε να καλυφθούν οι αυξανόμενες ανάγκες εμπειρικής έρευνας του συνόλου σχεδόν των επιστημονικών πεδίων[2]. Ειδικότερα, για την πολυδιάστατη ανάλυση δεδομένων (Παπαδημητρίου 2007) με χρήση της R, έχουν κατασκευαστεί πολυάριθμα πακέτα, όπως αυτά στο CRAN Task View: Multivariate statistics (Hewson 2015) και στο CRAN task view: Cluster analysis & finite mixture models (Leisch & Gruen, 2016), ενώ η σχετική

[1] Για πληροφορίες ανάκτησης και εγκατάστασης των λογισμικών R και R-Studio δείτε και στη σελίδα https://sites.google.com/view/gsda-soft/r-packages
[2] Κατά τη συγγραφή του παρόντος, τα διαθέσιμα πακέτα ήταν πάνω από 11,5 χιλ.

βιβλιογραφία είναι εκτενής (Murtagh 2005, Everitt & Hothorn 2011, Peng 2012, Pagès 2016, Husson *et.al* 2017).

Επικεντρώνοντας στις δυνατότητες επεξεργασίας και τους αλγόριθμους, καθώς και στα αποτελέσματα (πίνακες ερμηνείας και γραφικά) που παρέχουν τα πακέτα της R που εμφανίζουν κοινά χαρακτηριστικά με τα λογισμικά Praxitelis (Karakos 2012), S-Pro (Κουτσουπιάς 2002), MAD (Καραπιστόλης 2003) και Chic Analysis (Markos et al. 2010), επιλέχθηκε για παρουσίαση και περεταίρω διερεύνηση το πακέτο FactoMineR(Le et al. 2008). Φυσικά, για να καλυφθεί η ευρύτητα των εφαρμογών και το μεγάλο πλήθος των μεθόδων της Π.Α.Δ. ο ενδιαφερόμενος ερευνητής ενδείκνυται να μελετήσει και άλλα πακέτα, όπως τα CAvariants (Beh & Lombardo 2014), CAinterprTools (Alberti 2015), homals (De Leeuw & Mair 2009), FactoClass (Pardo & Del Campo 2007), CAvariants (Beh & Lombardo 2014), CAinterprTools (Alberti 2015), ca (Nenadic & Greenacre 2007), ade4 (Dray & Dufour 2007), hclust, βασισμένο σε κώδικα του Murtagh (1985), agnes (Struyf et al. 1997), clustrd (Markos, et al. 2017) και idm (Iodice D'Enza et. al. 2017). Επιπλέον, η ΑΚΣ διατίθεται στο βασικό (built-in) πακέτο stats της R, με τις συναρτήσεις prcomp() και princomp().

Σε πρόσφατη συνεδρίαση των μελών της Ελληνικής Εταιρίας Ανάλυσης Δεδομένων (Ε.Ε.Α.Δ.)[3], αποφασίστηκε η διεύρυνση του φάσματος επιλογών λογισμικού Π.Α.Δ., τόσο για ερευνητικούς όσο και για διδακτικούς σκοπούς. Η κύρια κατεύθυνση είναι η ανάπτυξη εξελληνισμένων πακέτων και διεπαφών για υπάρχοντα πακέτα αξιοποίησης των μεθόδων με την προσέγγιση της Γαλλικής Σχολής μεθόδων της Π.Α.Δ (βλ. και Μενεξές 2013). Με τη συγκεκριμένη οπτική, το παρόν αποτελεί μια πρώτη εισαγωγή στο κύριο διαθέσιμο αγγλόφωνο πακέτο των μεθόδων αυτών.

Μέθοδοι της Π.Α.Δ. με το FactoMineR

Το συγκεκριμένο πακέτο της R, αναπτύχθηκε[4] στη μονάδα Παιδαγωγικής και Εφαρμοσμένων Μαθηματικών στο AgroCampus Ouest στο Rennes Cedex της Γαλλίας. Περιλαμβάνει ρουτίνες τόσο για τις κλασσικές μεθόδους της Π.Α.Δ., όσο και για μεθόδους περισσότερο εξελιγμένες, όπως η MFA (Pagès, 2015) και άλλες.

[3] Πραγματοποιήθηκε το Σάββατο 31/9/17 στις εγκαταστάσεις του Α.Π.Θ. στο πλαίσιο του 9ου Πανελλήνιου Συνεδρίου της Ε.Ε.Α.Δ. με Διεθνή Συμμετοχή.
[4] Για περισσότερες λεπτομέρειες δείτε και στη σελίδα http://factominer.free.fr/index.html

Αναφορικά με τις κλασσικές μεθόδους και όπως θα δούμε παρακάτω, το συγκεκριμένο πακέτο εμφανίζει αξιοσημείωτη λειτουργικότητα και συμβατά εξαγόμενα, μεταξύ άλλων, για την Ανάλυση σε Κύριες Συνιστώσες (ΑΚΣ), την Πολλαπλή Παραγοντική Ανάλυση των Αντιστοιχιών (ΠΠΑ) και την Αυτόματη Ιεραρχική Ταξινόμηση (ΑΙΤ). Έχει επιλεγεί ως η προσφορότερη λύση για τη διδασκαλία των συγκεκριμένων μεθόδων σε προπτυχιακό και μεταπτυχιακό επίπεδο σε οικείο Τμήμα ΑΕΙ (Τμήμα Διεθνών & Ευρωπαϊκών Σπουδών/ΠΑΜΑΚ) με πολύ ενθαρρυντικά αποτελέσματα, ειδικά όταν συνδυάζεται με τη χρήση του συμπληρωματικού πακέτου Factoshiny (Vaissie *et al.* 2016).

Για εποπτικούς λόγους και πριν από τη συνοπτική παρουσίαση των λειτουργιών του FactoMiner, θα παρουσιάσουμε παρακάτω τις βασικές λειτουργίες εισαγωγής δεδομένων στο περιβάλλον του R-Studio (Gandrud 2013, Studio 2013) χρησιμοποιώντας ποσοτικά δεδομένα δεικτών ανάπτυξης Ευρωπαϊκών χωρών (Πίνακας 1α)[5] που αξιοποιήθηκαν σε πρόσφατη έρευνα (Koutsoupias & Boutsiouki, 2017) και ποιοτικά δεδομένα 10 περιστατικών μαθησιακών διαταραχών[6] (Πίνακας 1b), αντίστοιχης έρευνας (Hatzilias & Koutsoupias 2017).

Θα περιοριστούμε στη διαδικασία εισαγωγής δεδομένων από αρχεία τύπου CSV, καθώς είναι ο πλέον διαδεδομένος τύπος για την μεταφορά τους, τόσο μεταξύ διαφορετικών εφαρμογών, όσο και μεταξύ λειτουργικών συστημάτων.

ΧΩΡΑ	IM	EX	HD	ED	CO	UN
CY	0,59	0,91	0,83	2,7	37	9,6
FR	0,64	0,63	0,853	16,2	32	6,1
DE	0,49	0,71	0,873	6,4	18	8,5
GR	0,43	0,44	0,813	5,9	52	6,5
IT	0,37	1,99	0,895	20,5	13	5,7
MT	0,45	0,66	0,86	3,3	90	12,9
PT	0,41	0,4	0,877	16	56	8,3
SI	0,33	0,78	0,89	17,8	23	10
ES	0,35	0,35	0,853	7,4	48	5,9
UK	1,02	2,42	0,91	15,7	7	5,4

ΑΑ	ΦΥΛΟ	ΔΙΑΓΝΩΣΗ	ΗΛΙΚΙΑ
1	Α	ΑΥΤ	ΠΑΙΔ
2	Κ	ΔΙΑ	ΕΦ
3	Α	ΔΙΑ	ΠΑΙΔ
4	Κ	ΑΥΤ	ΠΑΙΔ
5	Κ	ΔΙΑ	ΠΑΙΔ
6	Α	ΔΙΑ	ΕΦ
7	Α	ΔΙΑ	ΕΦ
8	Α	ΔΙΑ	ΕΦ
9	Α	ΑΥΤ	ΠΑΙΔ
10	Κ	ΑΥΤ	ΠΑΙΔ

Πίνακας 1a/b: Ποσοτικά & Ποιοτικά Δεδομένα

Στα αρχεία CSV (comma separated value), οι στήλες διαχωρίζονται με κόμμα (“,” ή “;”) και η εντολή από την κονσόλα του R-Studio για την εισαγωγή στο περιβάλλον

[5] Αφορά το έτος 2015 (πηγές: Eurostat, World Economic Forum, ΟΗΕ & Παγκόσμια Τράπεζα).
[6] ΑΥΤ = Αυτισμός / ΔΙΑ = Διαταραχή Προσοχής

της R των δεδομένων του αρχείου mydata.csv που βρίσκεται αποθηκευμένο στον κατάλογο c:\data είναι η ακόλουθη:

```
mydata <-  read.csv2(file="c:\\data\\mydata.csv", row.names=1)
```

Με την εντολή αυτή ορίζεται η μεταβλητή[7] mydata στην οποία αποθηκεύονται τα δεδομένα του αρχείου mydata.csv που περιέχει τα ονόματα των γραμμών στην 1η γραμμή (row.names=1)

Στο σημείο αυτό θα πρέπει να αναφερθεί ότι συνήθως, μετά από παρόμοιες εντολές και για λόγους ελέγχου των δεδομένων, χρησιμοποιούνται και δυο ακόμη εντολές, η View() και η str(), ώστε να εμφανιστεί στο περιβάλλον του R-Studio ένα μέρος των περιεχομένων και η δομή τους στη νέα μεταβλητή (mydata). Γράφουμε, λοιπόν:

```
> View(mydata) #εμφανίζει την αρχή των περιεχομένων της μεταβλητής mydata[8]
> str(mydata) #εμφανίζει τη δομή στα δεδομένα
```

Με την εντολή View() εμφανίζονται στην περιοχή δεδομένων και κώδικα οι πρώτες δέκα γραμμές αρχείου (Πίνακας 1β) που εισάγαμε και με την εντολή str(), η δομή τους:

```
'data.frame':        10 obs. of  3 variables:
 $ ΦΥΛΟ    : Factor w/ 2 levels "Α","Κ": 1 2 1 2 2 1 1 1 1 2
 $ ΔΙΑΓΝΩΣΗ: Factor w/ 2 levels "ΑΥΤ","ΔΙΑ": 2 1 1 2 1 1 1 1 2 2
 $ ΗΛΙΚΙΑ  : Factor w/ 2 levels "ΕΦ","ΠΑΙΔ": 2 1 2 2 2 1 1 1 2 2
```

Πριν από τη χρήση οποιουδήποτε πακέτου στο περιβάλλον του R-Studio είναι επιβεβλημένη η φόρτωση και η ενεργοποίησή του, με τις εντολές install.packages() και library() αντίστοιχα. Έτσι για την περίπτωση του FactoMineR η φόρτωση πραγματοποιείται με την εντολή install.packages("FactoMineR") και η ενεργοποίηση με την library(FactoMiner). Οι διαδικασία φόρτωσης ενός πακέτου απαιτείται μία φορά για κάθε εγκατάσταση του R-Studio, ενώ η ενεργοποίησή του μια φορά ανά εκτέλεσή του R-Studio.

Βοήθεια και περισσότερες λεπτομέρειες για κάθε εγκατεστημένο πακέτο εμφανίζεται στην καρτέλα Packages, επιλέγοντας το εν λόγω πακέτο. Επιπλέον στην ίδια περιοχή και σε παρακείμενη καρτέλα υπάρχει βοήθεια (Help) με δυνατότητα

[7] Η μεταβλητή mydata ορίζεται από την ρουτίνα read.csv2() και είναι τύπου Dataframe - ένα είδος πίνακα (array) με δυνατότητα αποθήκευσης μεταβλητών διαφορετικών τύπων.

[8] Στο περιβάλλον της R ο χαρακτήρας # και ότι τον ακολουθεί εκλαμβάνονται ως σχόλια και δεν εκτελούνται.

αναζήτησης, για οτιδήποτε αφορά πακέτα, αλλά και λειτουργίες του περιβάλλοντος χρήσης.

Η Ανάλυση σε Κύριες Συνιστώσες

Αξιοποιώντας τα παραπάνω δεδομένα (Πίνακας 1α) στη μεταβλητή mydata και αφού έχει φορτωθεί και ενεργοποιηθεί το FactoMineR, η εντολή για την εκτέλεση της ΑΚΣ έχει ως εξής:

```
mydata.pca <- PCA(mydata)
```

Οι πίνακες ερμηνείας αποθηκεύονται στη λίστα με μεταβλητές-πίνακες mydata.pca, ενώ τα γραφήματα εμφανίζονται στην καρτέλα plots του R-Studio.

Μεταβλητή	**Περιγραφή**	**Εξαγόμενα (Παράρτημα)**
mydata.pca$call	Κάποια Περιγραφικά	E1
mydata.pca$eig	Ιδιοτιμές	E2
mydata.pca$ind	Πίνακες Ερμηνείας Αντικειμένων	E3
mydata.pca$var	Πίνακες Ερμηνείας Μεταβλητών	E4
mydata.pca$svd	Ανάλυση Ιδιοτιμών	E5

Πίνακας 2 : Αποτελέσματα και Πίνακες Ερμηνείας της ΑΚΣ

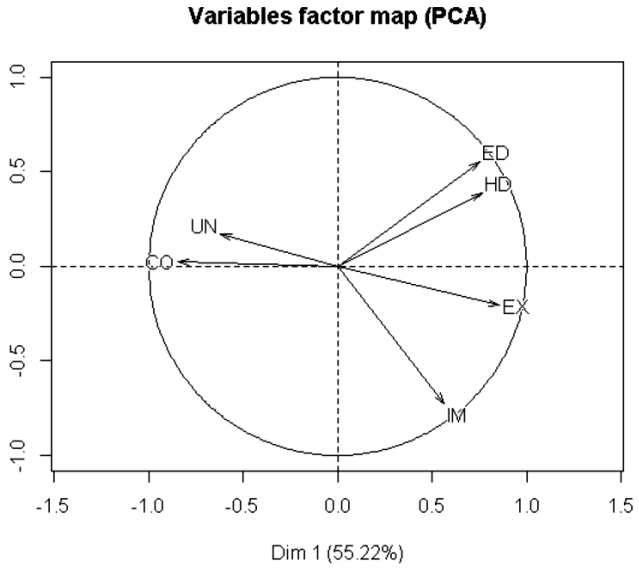
Στη συνέχεια, για να εμφανίσουμε τα αποτελέσματα και τους πίνακες ερμηνείας (π.χ. τις ιδιοτιμές) που αποθηκεύτηκαν στην μεταβλητή mydata.pca, χρησιμοποιούμε το όνομα της αντίστοιχης μεταβλητής ως εντολή στην κονσόλα του R-Studio (π.χ. mydata.pca$eig). Επιπλέον, είναι εφικτό να αποθηκεύσουμε σε αρχείο κειμένου τα περιεχόμενα της μεταβλητής που μας ενδιαφέρει. Για παράδειγμα, χρησιμοποιώντας την εντολή:

```
write.csv2(file="PCA-EIG.TXT",mydata.pca$eig)
```

αποθηκεύονται τα περιεχόμενα της μεταβλητής mydata.pca$eig σε αρχείο κειμένου με όνομα PCA-EIG.TXT (δες Εξαγόμενα E2). Για να ορίσουμε τον φάκελο εισαγόμενων και εξαγόμενων σε κάθε εκκίνηση της εφαρμογής εκτελούμε την εντολή setwd(), για παράδειγμα η εντολή setwd("c:\\data") ορίζει τη θέση c:\data ως την τρέχουσα θέση αρχείων, ενώ με την εντολή getwd() εμφανίζεται η θέση αυτή στην κονσόλα.

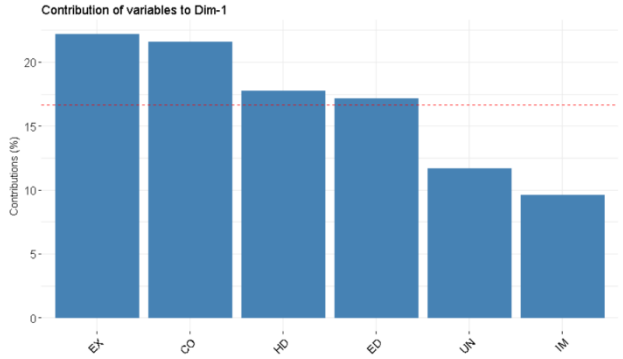
Τα γραφικά που δημιουργούνται είναι δυνατό να εξαχθούν ξεχωριστά, σε μορφή αρχείου γραφικών (JPEG, PNG, TIFF, BMP, Metafile, SVG ή EPS) ή ακόμη και ως

αρχείο PDF μέσα από το περιβάλλον του R-Studio (Διαγράμματα 1 & 3) Εμφανίζονται σε χωριστά παράθυρα ή στην καρτέλα Plots όπου η πλοήγηση μεταξύ διαδοχικών γραφημάτων επιτυγχάνεται με τη χρήση των αντίστοιχων δεικτών (γαλάζια βέλη) κάτω από τη λέξη Plots.



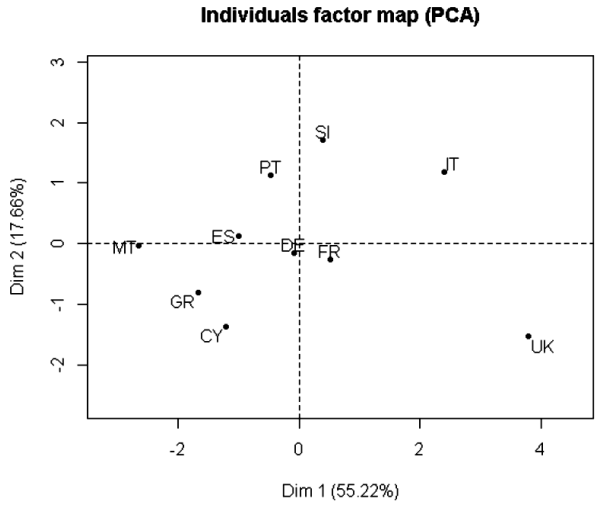
Διάγραμμα 1: 1ο Παραγοντικό Επίπεδο της PCA (μεταβλητές)

Η βιβλιοθήκη πακέτων της γλώσσας παρέχει επιπλέον εργαλεία για τον εμπλουτισμό ή/και για παραλλαγές στις απεικονίσεις των σχετικών γραφημάτων, όπως π.χ. η απεικόνιση των παραγοντικών αξόνων χωριστά, η επανατοποθέτηση των περιγραφών (labels) των σημείων με πιο αναγνώσιμη διάταξη ή/και η απεικόνιση των σημαντικότερων συνεισφορών μεταβλητών (Διάγραμμα 2) ή/και αντικειμένων.



Διάγραμμα 2: Σημαντικότερες συνεισφορές μεταβλητών ΑΚΣ

Σε κάθε παραγοντικό επίπεδο, τόσο για την ΑΚΣ όσο και για την ΠΠΑ, το πακέτο εμφανίζει οριζόντια και κάθετα τα αντίστοιχα ποσοστά αδράνειας ανά άξονα (Διαγράμματα 1, 3 & 4).



Διάγραμμα 3: 1ο Παραγοντικό Επίπεδο της ΑΚΣ (αντικείμενα)

Η Πολλαπλή Παραγοντική Ανάλυση των Αντιστοιχιών

Για να λάβουμε αποτελέσματα από την ΠΠΑ, χρησιμοποιούμε την εντολή MCA:

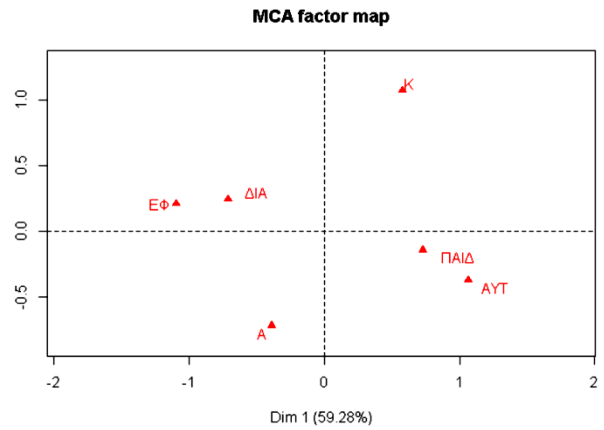
mydata.mca <- MCA(mydata)

Οι πίνακες ερμηνείας αποθηκεύονται στη μεταβλητή mydata.mca. Ακολούθως, εμφανίζονται (Πίνακας 3) τα αποτελέσματα για την ερμηνεία της συνάρτησης MCA:

Μεταβλητή	**Περιγραφή**	**Πίνακας (Παράρτημα)**
mydata.mca$call	Κάποια Περιγραφικά	Ε6
mydata.mca$eig	Ιδιοτιμές	Ε7
mydata.mca$ind	Πίνακες Ερμηνείας Αντικειμένων	Ε8
mydata.mca$var	Πίνακες Ερμηνείας Μεταβλητών	Ε9
mydata.mca$svd	Ανάλυση Ιδιοτιμών	Ε10

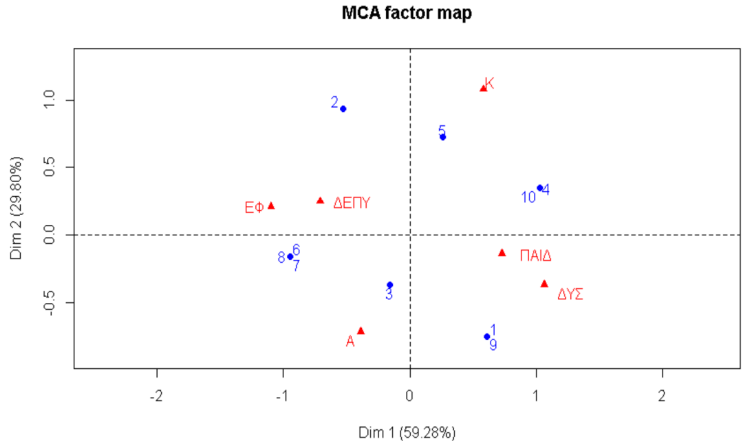
Πίνακας 3: Αποτελέσματα και Πίνακες Ερμηνείας της ΠΠΑ

Τα γραφήματα (Διάγραμμα 4 & 5) εμφανίζονται στην καρτέλα των γραφημάτων (Plots) του R-Studio που βρίσκεται στην ίδια περιοχή με αυτήν των καρτελών των Πακέτων (Packages) και της βοήθειας (Help) .



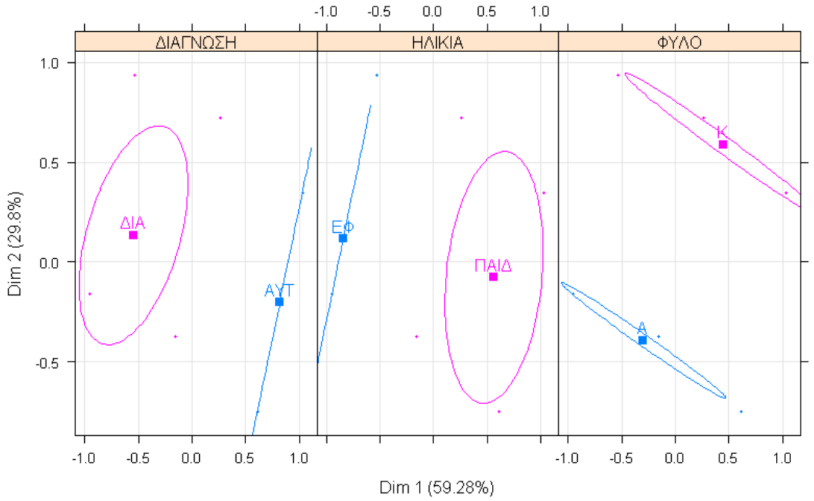
Διάγραμμα 4: 1ο Παραγοντικό Διάγραμμα της ΠΠΑ (μεταβλητές)

Με την εντολή plot.MCA(mydata.mca) λαμβάνουμε το παρακάτω διάγραμμα:



Διάγραμμα 5: 1ο Παραγοντικό Διάγραμμα της ΠΠΑ (κατηγορίες-αντικείμενα)

Επιπλέον, με κατάλληλη παραμετροποίηση τόσο για την ΑΚΣ, όσο και για τη ΠΠΑ, είναι εφικτή η επιλογή κύριων και συμπληρωματικών αντικειμένων και μεταβλητών στις αναλύσεις. Πρόσθετες ρουτίνες στη συγκεκριμένη βιβλιοθήκη παράγουν διαγράμματα ελλείψεων εμπιστοσύνης (Le Roux, 2010), όπως απεικονίζεται παρακάτω (Διάγραμμα 6Διάγραμμα 6).



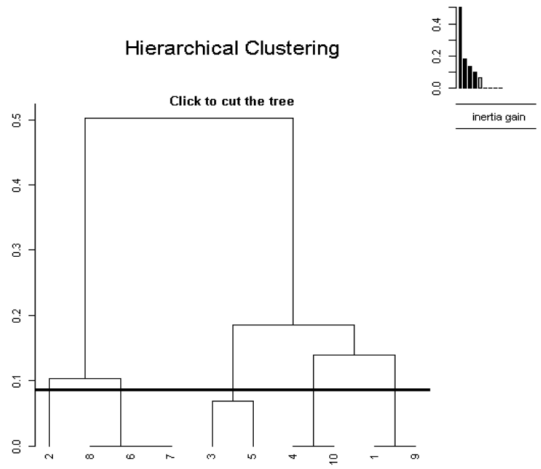
Διάγραμμα 6: Ελλείψεις εμπιστοσύνης ανά μεταβλητή

Η Αυτόματη Ιεραρχική Ταξινόμηση

Η συνάρτηση HCPC ταξινομεί τα δεδομένα με βάση τα αποτελέσματα είτε της ΑΚΣ είτε της ΠΠΑ και εξ ορισμού χρησιμοποιεί την Ευκλείδεια μετρική και τη μέθοδο συνένωσης Ward. Είναι δυνατό, όμως, να παραμετροποιηθεί για άλλες μετρικές. Για παράδειγμα, εάν θέλουμε με βάση τα αποτελέσματα της ΠΠΑ να ταξινομήσουμε τα δεδομένα (π.χ. στον Πίνακα 1β) εκτελούμε την εντολή:

mydata.hcpc <- HCPC(mydata.mca)

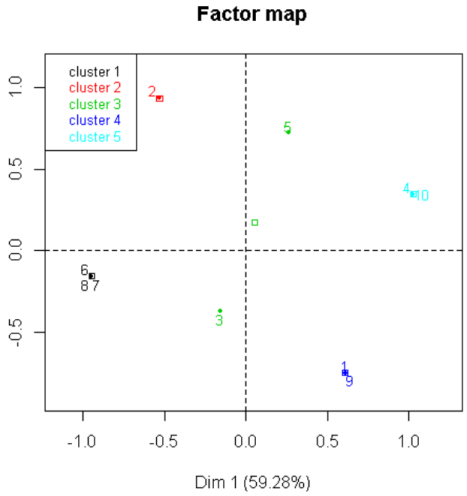
με την εκτέλεση της οποίας εμφανίζεται στο χώρο των γραφημάτων (Plots) δενδρόγραμμα με την ακόλουθη μορφή:



Διάγραμμα 7: Δενδρόγραμμα Αυτόματης Ταξινόμησης

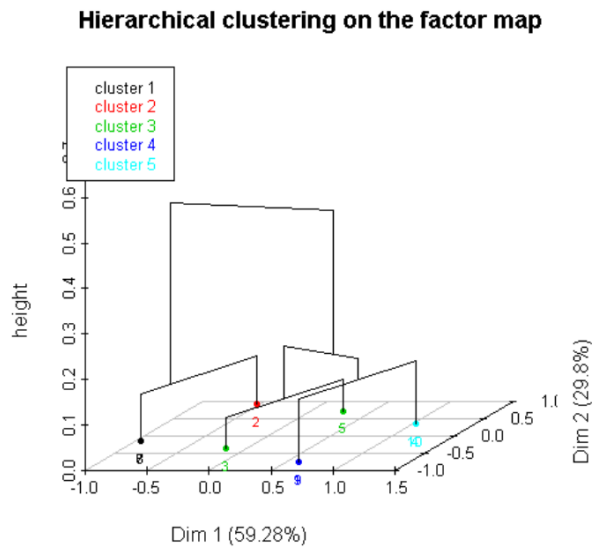
Μετά το σχηματισμό του δενδρογράμματος, η εκτέλεση της ρουτίνας διακόπτεται και ο χρήστης καλείται να επιλέξει το σημείο τομής του. Προτείνεται με έντονη γραμμή ένα σημείο με βάση τα δεδομένα στην είσοδο. Η τομή του δενδρογράμματος που προτείνεται από το λογισμικό, βασίζεται στο κριτήριο της αύξησης της εσωταξικής αδράνειας (βλ. π.χ. Παπαδημητρίου, 2007). Στο παράδειγμα η εφαρμογή προτείνει την τομή στο ύψος των 5 ομάδων (Διάγραμμα 7).

Εάν επιλέξουμε με τον δείκτη του ποντικιού την προτεινόμενη ομαδοποίηση, το πρώτο γράφημα που δημιουργείται είναι αυτό του 1ου Παρ. επιπέδου με τις ομαδοποιήσεις των σημείων.



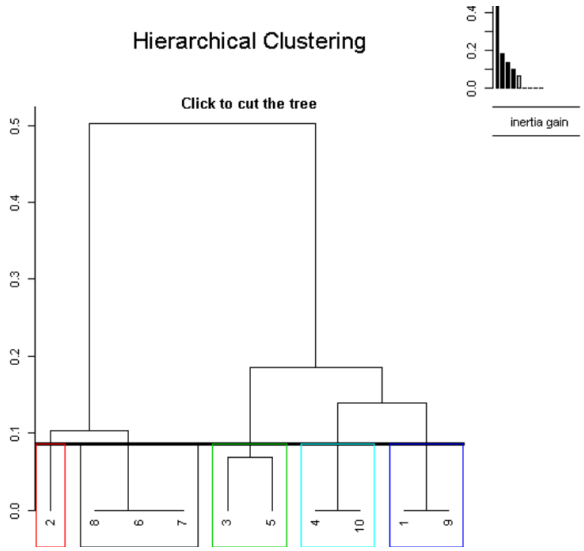
Διάγραμμα 8: Οι ομαδοποιήσεις της AIT στο 1ο παραγοντικό επίπεδο

Στο επόμενο γράφημα (Διάγραμμα 9) απεικονίζεται το δενδρόγραμμα της ταξινόμησης επί του 1ου παρ. επιπέδου.



Διάγραμμα 9: Το δενδρόγραμμα της AIT επί του 1ου παραγοντικού επιπέδου

Τέλος, κατασκευάζεται και το δενδρόγραμμα με τις ομαδοποιήσεις των σημείων περιγεγραμμένες με έγχρωμα παραλληλόγραμμα (Διάγραμμα 10).



Διάγραμμα 10: Το δενδρόγραμμα της AIT με περιγεγραμμένες τις επιλεγμένες ομάδες

Ο κατάλογος με τα εξαγόμενα της AIT εμφανίζονται πληκτρολογώντας mydata.hcpc στην κονσόλα (Εξαγόμενα 11).

Επίλογος

Το περιβάλλον εργασίας της R και του R-Studio για την ανάλυση και οπτικοποίηση ποσοτικών και ποιοτικών δεδομένων προσφέρεται τόσο για ερευνητική όσο και για διδακτική χρήση στο χώρο της Π.Α.Δ. Πρόσθετα σχετικά πακέτα και βιβλιοθήκες αναπτύσσονται με γρήγορους ρυθμούς και πλέον, τόσο ο αναλυτής, όσο

και ο ειδικός του πεδίου έχουν στη διάθεσή τους ευρύτατο φάσμα επιλογών επεξεργασίας, ανάλυσης και απεικόνισης δεδομένων. Είναι στις προτεραιότητες της Ελληνικής κοινότητας των επιστημόνων της Ανάλυσης Δεδομένων τόσο ο εξελληνισμός συγκεκριμένων πακέτων, όσο και η ανάπτυξη νέων εφαρμογών, προσαρμοσμένων στις αυξανόμενες ανάγκες εμπειρικής έρευνας, τόσο στις Κοινωνικές όσο και στις Ανθρωπιστικές Επιστήμες. Για το σκοπό αυτό κατασκευάστηκε από την Ελληνική Εταιρία Ανάλυσης Δεδομένων σχετικός ιστοχώρος με διαρκή ενημέρωση αναφορικά κυρίως με τα ζητήματα λογισμικού (https://sites.google.com/view/gsda-soft).

Abstract

Modern tools for R, an open source "statistical" language, include various routines for Multidimensional Data Analysis (M.D.A.). In this work the installation and basic functions of R and its R-Studio graphical GUI are shown, along with the set-up and usage of R packages for M.D.A. that comply with the French School algorithms and results' structure (interpretation tables and graphs).

Βιβλιογραφία

Καραπιστόλης, Δ. (2003). Το λογισμικό MAD, *Τετράδια Ανάλυσης Δεδομένων* (*Data Analysis Bulletin*), τ.2, σσ.133-144.

Κουτσουπιάς, Ν. (2002). S-PRO: Εφαρμογή υλοποίησης μεθόδων της ανάλυσης δεδομένων, *Τετράδια Ανάλυσης Δεδομένων* (*Data Analysis Bulletin*), τ.1, σσ.118-128.

Μενεξές, Γ. (2013). *Ανάλυση Δεδομένων (ΑΔ) και Στατιστική: Διαλεκτική και Συμπληρωματικότητα,* https://users.auth.gr/gmenexes/Presentations/DataAnalysis.pdf

Παπαδημητρίου, Γ. (2007). *Η Ανάλυση Δεδομένων*, Τυπωθήτω.

Chambers, J.M. (1998). *Programming with Data: A Guide to the S Language*. Springer.

Iodice D'Enza, A., Markos, A., & Buttarazzi, D. (2017. Package 'fpc'. Flexible procedures for clustering. http://cran.r-project.org/web/packages/idm/index.html

Dray, S. & Dufour, A.B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software.* 22(4): 1-20.

Everitt, B. & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R.* Springer Science & Business Media.

Gandrud, C. (2013). *Reproducible research with R and R studio*. CRC Press.

Hatzilias, S. & Koutsoupias, N. (2017). Exploring Cases of Attention Deficit - Hyperactivity Disorder (ADHD). *Book of Abstracts, 9th PanHellenic Conference on Data Analysis (GSDA17)*.

Hewson, P. (2015). CRAN task view: Multivariate statistics.

Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example using R*. CRC press.

Ihaka, R. & Gentleman, R. (1996). "R: A Language for Data Analysis and Graphics". *Journal of Computational and Graphical Statistics.* **5** (3): 299–314.

Jones, J. (2009, July 25). Why Use R?, Retrieved from http://monkeysuncle.stanford.edu/?p=367

Karakos, A. (2012). PRAXITELE: The New Generation of Data Analysis Tools. *Software Engineering*, 2(5), 186-194.

Kelley, K., Lai, K., & Wu, P. J. (2008). 34 Using R for Data Analysis: A Best Practice for Research. Ανακτήθηκε από: http://www3.nd.edu/~kkelley/publications/chapters/Kelley_Lai_Wu_Using_R_2008.pdf

Koutsoupias, N. & Boutsiouki, S. (2017). Exploring European Cultural Goods Trade: A Multidimensional Perspective, *Proceedings: ICIB17* (to appear).

Le Roux, B. & Rouanet, H. (2010). *Multiple correspondence analysis* (Vol. 163). Sage.

Le, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software,* 25(1), 1-18.

Leisch, F. & Gruen, B. (2016). CRAN task view: Cluster analysis & finite mixture models.

Level, (2017, May 31). *How Big Companies Are Using R for Data Analysis*, Retrieved from http://www.northeastern.edu/levelblog/2017/05/31/big-companies-using-r-data-analysis/

Markos, A., Iodice D'Enza, A. & van de Velden, M. (2017). Package 'clustrd'.

Markos, A., Menexes, G. & Papadimitriou, I. (2010). The CHIC Analysis Software v1.0. In H. Loracek-Junge & C. Weihs (eds.), Classification as a Tool for Research, *Proceedings of the 11th IFCS Conference*. Berlin: Springer, 409-416.

Murtagh, F. (1985). Multidimensional clustering algorithms. *Compstat Lectures,* Vienna: Physika Verlag.

Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R*. CRC Press.

Nenadic, O. & Greenacre, M. (2007). Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software* 20(3),1-13.

Pagès, J. (2015). *Multiple Factor Analysis by Example Using R.*. Chapman & Hall/CRC.

Pagès, J. (2016). *Multiple factor analysis by example using R*. CRC Press.

Peng, R. (2012). *Exploratory data analysis with R.* Lulu.com.

Struyf, A., Hubert, M. & Rousseeuw, P. J. (1997). Integrating robust clustering techniques in S-PLUS. *Computational Statistics & Data Analysis*, 26(1), 17-37.

Studio, R. (2013). *RStudio: Integrated development environment for R Version 0.98. 501.* R Studio, Boston, MA. http://www.rstudio.org.

Team, R. C. (2000). *R language definition.* Vienna, Austria: R foundation for statistical computing.

Alberti, G. (2015). CAinterprTools: An R package to help interpreting Correspondence Analysis' results. *SoftwareX*, 1, 26-31.

De Leeuw J, Mair P (2009). Gifi Methods for Optimal Scaling in R: The Package homals. *Journal of Statistical Software*, 31(4), 1–20.

Beh, E.J., Lombardo R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. John Wiley & Sons.

Pardo CE, Del Campo PC (2007). Combination of Factorial Methods and Cluster Analysis in R: The Package FactoClass. *Revista Colombiana de Estadistica*, *30*(2), 231–245.

ΠΑΡΑΡΤΗΜΑ

Εξαγόμενα Ε1

```
$row.w
 [1] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
$col.w
[1] 1 1 1 1 1 1
$scale.unit
[1] TRUE
$ncp
[1] 5
$centre
[1]  0.5080  0.9290  0.8654 11.1900 37.6000  7.8900
$ecart.type
[1]  0.19528441  0.66597973  0.02828144  6.31323214 23.60169485  2.30453900
$X
     IM   EX    HD   ED CO   UN
CY 0.59 0.91 0.830  2.7 37  9.6
FR 0.64 0.63 0.853 16.2 32  6.1
DE 0.49 0.71 0.873  6.4 18  8.5
GR 0.43 0.44 0.813  5.9 52  6.5
IT 0.37 1.99 0.895 20.5 13  5.7
MT 0.45 0.66 0.860  3.3 90 12.9
PT 0.41 0.40 0.877 16.0 56  8.3
SI 0.33 0.78 0.890 17.8 23 10.0
ES 0.35 0.35 0.853  7.4 48  5.9
UK 1.02 2.42 0.910 15.7  7  5.4
$row.w.init
 [1] 1 1 1 1 1 1 1 1 1 1
$call
PCA(X = mdata)
```

Εξαγόμενα Ε2

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1  3.3130401              55.217335                           55.21733
comp 2  1.0598388              17.663981                           72.88132
comp 3  0.9308753              15.514588                           88.39590
comp 4  0.3412295               5.687158                           94.08306
comp 5  0.2225335               3.708892                           97.79195
comp 6  0.1324828               2.208046                          100.00000
```

Εξαγόμενα Ε3

```
$coord
         Dim.1       Dim.2       Dim.3      Dim.4       Dim.5
CY -1.20937519 -1.36628449  0.22559850 -0.7355756 -0.01493893
FR  0.51799775 -0.25994788 -0.89297736  0.6549115 -0.67103586
DE -0.08906468 -0.15376184  0.08856413 -1.0159646 -0.43383271
GR -1.67455183 -0.80320774 -1.38712513  0.1149981  0.33036941
IT  2.39245402  1.18634055 -0.35167573 -0.2704281  0.95887411
MT -2.65579136 -0.04007951  2.01792839  0.4641441  0.40256042
PT -0.46466003  1.12818756  0.05533160  0.8169865 -0.31246463
SI  0.38649521  1.71522576  0.51901466 -0.5295449 -0.48548453
ES -1.00330225  0.12814749 -1.14296920  0.1175278  0.24920632
UK  3.79979836 -1.53461989  0.86831014  0.3829451 -0.02325361
$cos2
         Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
CY 0.356416191 0.4549017323 0.012402457 0.131852963 5.438437e-05
FR 0.125380652 0.0315752900 0.372610997 0.200419503 2.104099e-01
DE 0.005204326 0.0155113865 0.005145993 0.677190143 1.234804e-01
GR 0.503458649 0.1158303544 0.345460101 0.002374366 1.959595e-02
IT 0.689866119 0.1696275418 0.014906044 0.008814165 1.108156e-01
MT 0.613054854 0.0001396228 0.353934524 0.018724796 1.408554e-02
PT 0.095075286 0.5604793498 0.001348166 0.293918424 4.299310e-02
SI 0.037772817 0.7439328366 0.068116196 0.070908255 5.959941e-02
ES 0.346761009 0.0056570092 0.450024071 0.004758260 2.139365e-02
UK 0.814267755 0.1328152009 0.042520257 0.008270262 3.049489e-05
$contrib
         Dim.1       Dim.2       Dim.3      Dim.4       Dim.5
CY  4.41464130 17.61336948  0.54674012 15.8565259  0.01002867
FR  0.80989564  0.63757714  8.56622353 12.5695192 20.23466303
DE  0.02394332  0.22307829  0.08426053 30.2489706  8.45763837
GR  8.46389949  6.08717705 20.66996713  0.3875560  4.90460650
IT 17.27668881 13.27941430  1.32859711  2.1431720 41.31689829
MT 21.28929196  0.01515671 43.74415269  6.3133393  7.28226815
PT  0.65169434 12.00944060  0.03288933 19.5606492  4.38738998
SI  0.45088059 27.75893173  2.89379492  8.2178670 10.59144727
ES  3.03834356  0.15494600 14.03387378  0.4047948  2.79076091
UK 43.58072101 22.22090869  8.09950086  4.2976059  0.02429883
$dist
      CY       FR       DE       GR       IT       MT       PT       SI
ES       UK
2.025734 1.462893 1.234591 2.360025 2.880456 3.391910 1.506959 1.988632
1.703792 4.210919
```

Εξαγόμενα Ε4

```
$coord
         Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
IM  0.5643692 -0.72828755  0.2473679  0.20823141 -0.21458117
EX  0.8574399 -0.20639550  0.2887711 -0.07138773  0.35081118
HD  0.7666422  0.38842062  0.4450844  0.03469882 -0.03931191
ED  0.7539255  0.55392043 -0.1104274  0.25260025 -0.09788261
CO -0.8460617  0.02653487  0.2306794  0.44668853  0.17222519
UN -0.6223826  0.16862525  0.7230410 -0.16802206 -0.11239230
$cor
         Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
IM  0.5643692 -0.72828755  0.2473679  0.20823141 -0.21458117
EX  0.8574399 -0.20639550  0.2887711 -0.07138773  0.35081118
HD  0.7666422  0.38842062  0.4450844  0.03469882 -0.03931191
ED  0.7539255  0.55392043 -0.1104274  0.25260025 -0.09788261
CO -0.8460617  0.02653487  0.2306794  0.44668853  0.17222519
UN -0.6223826  0.16862525  0.7230410 -0.16802206 -0.11239230
$cos2
         Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
IM 0.3185126 0.5304027484 0.06119087 0.043360321 0.046045078
EX 0.7352032 0.0425991043 0.08338873 0.005096209 0.123068483
HD 0.5877403 0.1508705799 0.19810016 0.001204008 0.001545426
ED 0.5684036 0.3068278379 0.01219421 0.063806886 0.009581006
CO 0.7158204 0.0007040994 0.05321301 0.199530642 0.029661517
UN 0.3873601 0.0284344745 0.52278828 0.028231411 0.012632030
```

```
$contrib
       Dim.1       Dim.2     Dim.3      Dim.4     Dim.5
IM  9.613908 50.04560374  6.573477 12.7070854 20.691298
EX 22.191194  4.01939451  8.958100  1.4934843 55.303341
HD 17.740210 14.23523781 21.281064  0.3528441  0.694469
ED 17.156557 28.95042388  1.309972 18.6991133  4.305421
CO 21.606149  0.06643457  5.716449 58.4740344 13.329010
UN 11.691982  2.68290549 56.160938  8.2734385  5.676461
```

Εξαγόμενα Ε5

```
$vs
[1] 1.8201758 1.0294847 0.9648188 0.5841485 0.4717346 0.3639818
$U
aaaaa             [,1]        [,2]        [,3]       [,4]        [,5]
 [1,] -0.66442767 -1.32715370  0.23382475 -1.2592270 -0.03166808
 [2,]  0.28458665 -0.25250290 -0.92553895  1.1211387 -1.42248596
 [3,] -0.04893191 -0.14935806  0.09179353 -1.7392231 -0.91965419
 [4,] -0.91999454 -0.78020363 -1.43770536  0.1968644  0.70032896
 [5,]  1.31440819  1.15236341 -0.36449926 -0.4629441  2.03265586
 [6,] -1.45908505 -0.03893163  2.09151028  0.7945652  0.85336207
 [7,] -0.25528305  1.09587593  0.05734922  1.3985939 -0.66237376
 [8,]  0.21233949  1.66610119  0.53794005 -0.9065245 -1.02914757
 [9,] -0.55121172  0.12447731 -1.18464652  0.2011951  0.52827653
[10,]  2.08759960 -1.49066793  0.89997227  0.6555613 -0.04929385
$V
           [,1]       [,2]       [,3]        [,4]        [,5]
[1,]  0.3100630 -0.7074292  0.2563879  0.35646999 -0.45487688
[2,]  0.4710753 -0.2004843  0.2993008 -0.12220819  0.74366216
[3,]  0.4211913  0.3772961  0.4613140  0.05940068 -0.08333481
[4,]  0.4142047  0.5380560 -0.1144540  0.43242471 -0.20749508
[5,] -0.4648241  0.0257749  0.2390910  0.76468317  0.36508916
[6,] -0.3419354  0.1637958  0.7494060 -0.28763586 -0.23825325
```

Εξαγόμενα Ε6

```
$X
   ΦΥΛΟ ΔΙΑΓΝΩΣΗ ΗΛΙΚΙΑ
1     Α      ΔΥΣ   ΠΑΙΔ
2     Κ     ΔΕΠΥ     ΕΦ
3     Α     ΔΕΠΥ   ΠΑΙΔ
4     Κ      ΔΥΣ   ΠΑΙΔ
5     Κ     ΔΕΠΥ   ΠΑΙΔ
6     Α     ΔΕΠΥ     ΕΦ
7     Α     ΔΕΠΥ     ΕΦ
8     Α     ΔΕΠΥ     ΕΦ
9     Α      ΔΥΣ   ΠΑΙΔ
10    Κ      ΔΥΣ   ΠΑΙΔ
$marge.col
        Α         Κ      ΔΕΠΥ       ΔΥΣ        ΕΦ      ΠΑΙΔ
0.2000000 0.1333333 0.2000000 0.1333333 0.1333333 0.2000000
$marge.row
  1   2   3   4   5   6   7   8   9  10
0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
$ncp
[1] 3
$row.w
 [1] 1 1 1 1 1 1 1 1 1 1
$excl
NULL
$call
MCA(X = mydata)
$Xtot
   Α Κ ΔΕΠΥ ΔΥΣ ΕΦ ΠΑΙΔ
1  1 0    0   1  0    1
2  0 1    1   0  1    0
3  1 0    1   0  0    1
4  0 1    0   1  0    1
5  0 1    1   0  0    1
6  1 0    1   0  1    0
7  1 0    1   0  1    0
8  1 0    1   0  1    0
9  1 0    0   1  0    1
10 0 1    0   1  0    1
$N
[1] 30
$quali
[1] 1 2 3
```

Εξαγόμενα Ε7

```
      eigenvalue percentage of variance cumulative percentage of variance
dim 1  0.5928378              59.28378                           59.28378
dim 2  0.2979561              29.79561                           89.07939
dim 3  0.1092061              10.92061                          100.00000
```

Εξαγόμενα Ε8

```
$coord
        Dim 1      Dim 2      Dim 3
1   0.6093035 -0.7501033  0.1026578
2  -0.5304828  0.9367021  0.2517923
3  -0.1576414 -0.3731975 -0.7089002
4   1.0275228  0.3458183  0.2163997
5   0.2605778  0.7227242 -0.5951583
6  -0.9487021 -0.1592196  0.1380504
7  -0.9487021 -0.1592196  0.1380504
8  -0.9487021 -0.1592196  0.1380504
9   0.6093035 -0.7501033  0.1026578
10  1.0275228  0.3458183  0.2163997
$contrib
        Dim 1      Dim 2     Dim 3
1   6.2622658 18.8838244  0.965021
2   4.7468639 29.4476557  5.805480
3   0.4191839  4.6743925 46.017535
4  17.8093064  4.0136898  4.288115
5   1.1453522 17.5304450 32.435314
6  15.1818185  0.8508261  1.745133
7  15.1818185  0.8508261  1.745133
8  15.1818185  0.8508261  1.745133
9   6.2622658 18.8838244  0.965021
10 17.8093064  4.0136898  4.288115
```

```
$cos2
        Dim 1      Dim 2      Dim 3
1  0.39308909 0.59575237 0.01115854
2  0.23024622 0.71788156 0.05187222
3  0.03727621 0.20891455 0.75380924
4  0.86383888 0.09784663 0.03831449
5  0.07189498 0.55305557 0.37504945
6  0.95297892 0.02684211 0.02017897
7  0.95297892 0.02684211 0.02017897
8  0.95297892 0.02684211 0.02017897
9  0.39308909 0.59575237 0.01115854
10 0.86383888 0.09784663 0.03831449
```

Εξαγόμενα Ε9

```
$coord
           Dim 1      Dim 2       Dim 3
A     -0.3864143 -0.7178553 -0.04510506
K      0.5796215  1.0767829  0.06765759
ΔΕΠΥ -0.7086200  0.2468825 -0.32182829
ΔΥΣ   1.0629300 -0.3703237  0.48274243
ΕΦ    -1.0963526  0.2102411  0.50379519
ΠΑΙΔ  0.7309018 -0.1401607 -0.33586346
$contrib
          Dim 1     Dim 2      Dim 3
A      5.037331 34.590077  0.3725921
K      7.555997 51.885115  0.5588881
ΔΕΠΥ 16.940291  4.091271 18.9684382
ΔΥΣ  25.410436  6.136906 28.4526573
ΕΦ   27.033567  1.977978 30.9884545
ΠΑΙΔ 18.022378  1.318652 20.6589697
$cos2
          Dim 1      Dim 2       Dim 3
A     0.2239740 0.77297426 0.003051699
K     0.2239740 0.77297426 0.003051699
ΔΕΠΥ 0.7532134 0.09142643 0.155360169
ΔΥΣ  0.7532134 0.09142643 0.155360169
ΕΦ   0.8013261 0.02946753 0.169206394
ΠΑΙΔ 0.8013261 0.02946753 0.169206394
$v.test
          Dim 1      Dim 2      Dim 3
A     -1.419777 -2.6375686 -0.1657266
K      1.419777  2.6375686  0.1657266
ΔΕΠΥ -2.603636  0.9071041 -1.1824726
ΔΥΣ   2.603636 -0.9071041  1.1824726
ΕΦ    -2.685505  0.5149833  1.2340411
ΠΑΙΔ  2.685505 -0.5149833 -1.2340411
$eta2
              Dim 1      Dim 2       Dim 3
ΦΥΛΟ      0.2239740 0.77297426 0.003051699
ΔΙΑΓΝΩΣΗ 0.7532134 0.09142643 0.155360169
ΗΛΙΚΙΑ    0.8013261 0.02946753 0.169206394
```

Εξαγόμενα Ε10

```
$vs
[1] 7.699596e-01 5.458535e-01 3.304634e-01 1.666749e-16 6.018988e-17 2.443016e-
17
$U
             [,1]       [,2]       [,3]
 [1,]  0.7913448 -1.3741843  0.3106479
 [2,] -0.6889749  1.7160319  0.7619370
 [3,] -0.2047398 -0.6836953 -2.1451698
 [4,]  1.3345151  0.6335369  0.6548370
 [5,]  0.3384305  1.3240259 -1.8009807
 [6,] -1.2321452 -0.2916892  0.4177479
 [7,] -1.2321452 -0.2916892  0.4177479
 [8,] -1.2321452 -0.2916892  0.4177479
 [9,]  0.7913448 -1.3741843  0.3106479
[10,]  1.3345151  0.6335369  0.6548370
$V
           [,1]       [,2]       [,3]
[1,] -0.5018631 -1.3151060 -0.1364903
[2,]  0.7527946  1.9726590  0.2047355
[3,] -0.9203339  0.4522870 -0.9738696
[4,]  1.3805009 -0.6784305  1.4608043
[5,] -1.4239092  0.3851602  1.5245111
[6,]  0.9492728 -0.2567735 -1.0163407
```

Εξαγόμενα Ε11

```
**Results for the Hierarchical Clustering on Principal Components**
   name                       description
1  "$data.clust"              "dataset with the cluster of the individuals"
2  "$desc.var"                "description of the clusters by the variables"
3  "$desc.var$test.chi2"      "description of the cluster var. by the categorical
var."
4  "$desc.axes$category"      "description of the clusters by the categories."
5  "$desc.axes"               "description of the clusters by the dimensions"
6  "$desc.axes$quanti.var"    "description of the cluster var. by the axes"
7  "$desc.axes$quanti"        "description of the clusters by the axes"
8  "$desc.ind"                "description of the clusters by the individuals"
9  "$desc.ind$para"           "parangons of each clusters"
10 "$desc.ind$dist"           "specific individuals"
11 "$call"                    "summary statistics"
12 "$call$t"                  "description of the tree"
```