Στους γονείς μου.

Περιεχόμενα

ΠΕΡΙΛΗΨΗ 6
ΚΕΦΑΛΑΙΟ 1. ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ 8
1.1 ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ 8
1.2 ΕΚΤΙΜΩΝΤΑΣ ΤΙΣ ΠΑΡΑΜΕΤΡΟΥΣ ΤΟΥ ΜΟΝΤΕΛΟΥ 10
1.3 ΟΙ ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΚΑΙ ΤΑ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ 13
1.4 ΕΡΜΗΝΕΙΑ ΤΗΣ ΣΥΝΟΛΙΚΗΣ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ ΤΟΥ ΜΟΝΤΕΛΟΥ 14
1.5 ΕΠΙΛΟΓΗ ΤΟΥ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ 17
1.6 Η ΕΦΑΡΜΟΓΗ ΤΟΥ CENTERING (ΚΕΝΤΡΟΠΟΙΗΣΗ) ΚΑΙ ΤΟΥ SCALING (ΤΥΠΟΠΟΙΗΣΗ) 18
ΚΕΦΑΛΑΙΟ 2. ΤΟ ΦΑΙΝΟΜΕΝΟ ΤΗΣ ΠΟΛΥΣΥΓΡΑΜΜΙΚΟΤΗΤΑΣ 22
2.1 ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑ ΚΑΙ ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ 24
ΤΕΤΡΑΓΩΝΩΝ 24
2.2 ΠΕΡΙΠΤΩΣΕΙΣ ΕΜΦΑΝΙΣΗΣ ΤΗΣ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ 27
2.3 ΔΙΑΓΝΩΣΗ ΠΟΛΥΣΥΓΡΓΑΜΜΙΚΟΤΗΤΑΣ 31
2.4 ΜΕΘΟΔΟΙ ΑΝΤΙΜΕΤΩΠΙΣΗΣ ΤΗΣ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ 34
2.5 Η ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΟΡΥΦΟΓΡΑΜΜΗΣ 36
(RIDGE REGRESSION) 36
2.5.1 Θεώρημα Gauss – Markov 37
2.5.2 Παρουσίαση του < μηχανισμού > της Ridge Regression 38
ΚΕΦΑΛΑΙΟ 3. PCA (ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ) 41
3.1 ΔΙΑΔΙΚΑΣΙΑ ΠΟΥ ΑΚΟΛΟΥΘΕΙΤΑΙ ΣΤΗΝ PCA 42
3.2 ΓΕΩΜΕΤΡΙΚΗ ΕΡΜΗΝΕΙΑ ΤΗΣ ΜΕΘΟΔΟΥ PCA 44
3.3 ΕΠΙΛΟΓΗ ΤΟΥ ΑΡΙΘΜΟΥ ΤΩΝ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ ΠΟΥ ΘΑ ΕΞΕΤΑΣΟΥΜΕ 53
3.4 Η PCA ΩΣ ΜΕΘΟΔΟΣ ΜΕΙΩΣΗΣ ΔΙΑΣΤΑΣΗΣ 55
3.5 Η ΜΕΘΟΔΟΣ ΙΔΙΟΑΝΑΛΥΣΗΣ ΓΙΑ ΤΗΝ PCA 60
SVD(Singular Value Decomposition) 60
3.6 ΣΥΜΠΕΡΑΣΜΑΤΑ ΓΙΑ ΤΗ ΜΕΘΟΔΟ PCA 62
3.7 ΕΡΜΗΝΕΙΑ ΤΗΣ ΜΕΘΟΔΟΥ PCA 62

ΚΕΦΑΛΑΙΟ 4. Η ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ (PLS) 65
4.1 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΥ PLS ΚΑΙ ΤΟΥ PLSR ΜΟΝΤΕΛΟΥ 65
4.1.1 Το PLSR μοντέλο 66
4.2 ΕΡΜΗΝΕΙΑ ΤΟΥ ΜΟΝΤΕΛΟΥ PLSR 69
4.3 ΓΕΩΜΕΤΡΙΚΗ ΕΡΜΗΝΕΙΑ ΤΟΥ ΜΟΝΤΕΛΟΥ PLSR 70
4.4 Ο ΑΡΙΘΜΟΣ ΤΩΝ Y-ΜΕΤΑΒΛΗΤΩΝ ΚΑΘΕ ΦΟΡΑ 71
4.5 Ο ΑΡΙΘΜΟΣ ΤΩΝ PLS ΣΥΝΙΣΤΩΣΩΝ, A 72
4.5.1 Περιγραφή του στατιστικού PRESS 73
4.5.2 Περιγραφή του Cross Validation 76
4.6 ΟΙ PLSR ΑΛΓΟΡΙΘΜΟΙ 79
4.6.1 Ο αλγόριθμος NIPALS 79
4.7 ΤΥΠΙΚΑ ΣΦΑΛΜΑΤΑ ΚΑΙ ΔΙΑΣΤΗΜΑΤΑ 81
ΕΜΠΙΣΤΟΣΥΝΗΣ 81
4.7.1 Η τεχνική Jack-knifing 82
4.8 ΠΑΡΑΔΟΧΕΣ ΣΤΙΣ ΟΠΟΙΕΣ ΒΑΣΙΖΕΤΑΙ Η PLSR 83
4.9 ΤΡΟΠΟΙ ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΡΜΗΝΕΙΑΣ ΕΝΟΣ PLSR ΜΟΝΤΕΛΟΥ 90
4.10 Η PLSC (PARTIAL LEAST SQUARES CORRELATION) 91
4.10.1 Περιγραφή της λειτουργίας της PLSC 92
ΚΕΦΑΛΑΙΟ 5. ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ (PRINCIPAL COMPONENT REGRESSION) ΚΑΙ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ (PARTIAL LEAST SQUARES REGRESSION) 95
5.1 ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ 96
5.2 ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ 104
ΠΑΡΑΡΤΗΜΑ 1ο (ΠΙΝΑΚΑΣ ΔΕΔΟΜΕΝΩΝ ΘΕΩΡΙΑΣ) 120
ΠΑΡΑΡΤΗΜΑ 2ο (ΠΙΝΑΚΕΣ ΔΕΔΟΜΕΝΩΝ ΠΡΑΚΤΙΚΩΝ ΕΦΑΡΜΟΓΩΝ) 122
ΕΥΡΕΤΗΡΙΟ ΟΡΩΝ ΚΑΙ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ 125
ΒΙΒΛΙΟΓΡΑΦΙΑ 126

Ευρετήριο πινάκων και διαγραμμάτων.

Πινακας 1. Πινακας σχεδιασμου. 10

Πινακας 2. Αναλυση διασπορας 16

Διαγραμμα 1. Πολυσυγγραμικοτητα λογω συσχετισμενων μεταβλητων 28

Διαγραμμα 2. Γραφικη παραστση αναλογιας οικογενειακου εισοδηματος/μεγεθος σπιτιου 29

Πινακας 3 . Πινακας συσχετισης 31

Πινακας 4. Δεικτες καταστασης ιδιοτιμων 34

Πινακας 5. Αρχικες και μετασχηματισμενες μεταβλητες 45

Διαγραμμα 3. Μετασχηματισμενες μεταβλητες και προβολη των νεων σημειων στον νεο αξονα X_1* 46

Πινακας 6. Μετασχηματισμενες μεταβλητες και η νεα μεταβλητη x_1* για κλιση του νεου αξονα ιση με 10°. 47

Πινακας 7. Διασπορα νεων μεταβλητων / νεων αξονων. 48

Διαγραμμα 4. Γωνια θ του αξονα X_1* στον X_1 49

Πινακας 8. Μετασχηματισμενες αρχικες μεταβλητες και νεες μεταβλητες x_1* και x_2* για τους νεους αξονες με κλιση 43.261° 50

Διαγραμμα 5. Μετασχηματισμενα δεδομενα / νεοι αξονες. 51

Διαγραμμα 6. Χαρακτηριστικες ριζες η ιδιοτιμες για καθε κυρια συνιστωσα (scree plot). 54

Διαγραμμα 7. Τα δεδομενα της PLSR συλλεγονται σε 2 πινακες X και Y. 66

Διαγραμμα 8. Η γεωμετρικη αναπαρασταση της PLSR. Ο X πινακας αναπαριστατai με N σημεια στον K-διαστατο χωρο οπου καθε στηλη x_k του X οριζει εναν αξονα συντεταγμενων. 71

Πινακας 9. Το δευτερο μισο του παραπανω πινακα, ο πινακας συσχετισης, περιεχει τους ανα ζευγη συντελεστες συσχετισης μεταξυ των δεδομενων μας. Οι PIE και PIF αποτελουν συμφωνα με τους El Tayar, Fauchere και Pliska αντιστοιχα, τις σταθερες λιποφιλικοτητας της AA πλευρικης αλυσιδας, ενω η DGR αποτελει την εκλυομενη ενεργεια μιας πλευρικης αλυσιδας AA συμφωνα με τους Radzicka και Woldenden. Η μεταβλητη SAC αποτελει την προσβασιμη απο το νερο επιφανεια της AA, η MR τη μοριακη διαθλαστικοτητα, η Lam ειναι μια παραμετρος πολωσης ενω τελος η μεταβλητη Vol ειναι ο υπολογισμενος μοριακος ογκος της AA. 84

Διαγραμμα 8. Τα PLS scores u_1 και t_1 του AA παραδειγματος, 1η αναλυση..... 85

Διαγραμμα 9. Τα PLS scores u_1 και t_1 του AA παραδειγματος , 2η αναλυση 86

Διαγραμμα 10. Τα PLS scores t_1 και t_2 του AA παραδειγματος, 2η αναλυση 87

Διαγραμμα 11. Τα PLS βαρη w^* και c για τις πρωτες δυο διαστασεις του AA παραδειγματος, 2η αναλυση. 88

Διαγραμμα 12. Οι PLS συντελεστες για $A=2$ κυριες συνιστωσες,2η αναλυση. Οι μπαρες του σχηματος παρουσιαζουν 95% διαστηματα εμπιστοσυνης συμφωνα με την τεχνικη jack-knifing. 89

ΠΕΡΙΛΗΨΗ

Πολλές φορές δυο ή περισσότερες ποσοτικές μεταβλητές εξετάζονται ταυτόχρονα προκειμένου να προσδιοριστεί η οποιαδήποτε σχέση υπάρχει μεταξύ τους ή αλλιώς για την πρόβλεψη μιας από τις υπόλοιπες μεταβλητές. Έστω λοιπόν ότι θέλουμε να μελετήσουμε ένα πρόβλημα οικονομικής φύσεως, όπου η εξαρτημένη μεταβλητή Y εκφράζει το μισθό ενός εργαζομένου, ενώ οι ανεξάρτητες μεταβλητές $X_1, X_2, \ldots X_k$ εκφράζουν τα χρόνια προϋπηρεσίας του, το φύλο του, την μόρφωση του κλπ. Η μέθοδος που εφαρμόζεται σε αυτές τις περιπτώσεις καλείται ανάλυση Παλινδρόμησης.

Η συνηθέστερη μέθοδος της ανάλυσης Παλινδρόμησης που ακολουθείται για την κατασκευή ενός τέτοιου στατιστικού μοντέλου είναι η μέθοδος ελαχίστων τετραγώνων. Παρ' όλα αυτά, μέσω της συγκεκριμένης μεθόδου ερχόμαστε συχνά αντιμέτωποι με σημαντικά προβλήματα. Τέτοια παρουσιάζονται σε περιπτώσεις όπου μια ή περισσότερες απ' τις ανεξάρτητες μεταβλητές μας $X_1, X_2, \ldots X_k$, συνδέονται μεταξύ τους με κάποια γραμμική σχέση. Τότε η μέθοδος ελαχίστων τετραγώνων αδυνατεί να προσαρμόσει τα δεδομένα μας με τέτοιον τρόπο ώστε να πάρουμε αξιόπιστα στατιστικά μοντέλα. Το συγκεκριμένο φαινόμενο καλείται Πολυσυγγραμμικότητα και εμφανίζεται σε διάφορους κλάδους των σύγχρονων επιστημών. Προκειμένου να αντιμετωπιστεί, οι στατιστικοί οδηγήθηκαν στην κατασκευή νέων μεθόδων.

Τέτοιες είναι η Παλινδρόμηση Κυρίων Συνιστωσών (Principal Component Regression) καθώς και μια αρκετά πρόσφατη τεχνική Παλινδρόμησης, η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων (Partial Least Squares Regression). Στην παρούσα εργασία, σκοπός είναι η παρουσίαση των δυο παραπάνω μεθόδων.

Για το λόγο αυτό η εργασία έχει χωριστεί σε πέντε ενότητες. Στην πρώτη ενότητα κάνουμε μια ανασκόπηση της Πολλαπλής Γραμμικής Παλινδρόμησης και αναφερόμαστε στις τεχνικές scaling (τυποποίηση) και centering (κεντροποίηση) των μεταβλητών μας, οι οποίες θα εφαρμοστούν αργότερα. Ακολούθως στη δεύτερη ενότητα παραθέτουμε το πρόβλημα της Πολυσυγγραμμικότητας και τη συσχέτιση του με τη μέθοδο ελαχίστων τετραγώνων.

Στην Τρίτη ενότητα, προσπαθούμε να κάνουμε μια αναλυτική παρουσίαση της Ανάλυσης Κυρίων Συνιστωσών μέσω της οποίας θα οδηγηθούμε στην Παλινδρόμηση Κυρίων Συνιστωσών, στο μηχανισμό λειτουργίας της και στις μεθόδους που καλείται να χρησιμοποιήσει προκειμένου να αντιμετωπίσει το φαινόμενο της Πολυσυγγραμμικότητας. Κατόπιν της μεθόδου αυτής,

παρουσιάζεται ενδελεχώς η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων, που αποτελεί και τη βασική μέθοδο που θα ερευνήσουμε στην παρούσα εργασία. Αναφερόμαστε εκτενώς στη λειτουργία της, στους αλγόριθμους που χρησιμοποιεί και την καθιστούν ένα χρήσιμο εργαλείο στα χέρια κάθε στατιστικού.

Τέλος η πέμπτη και τελευταία ενότητα, συνιστά την πρακτική εφαρμογή των δυο προαναφερθέντων μεθόδων. Έτσι, τρία προβλήματα στα οποία το πρόβλημα της Πολυσυγγραμμικότητας κάνει την εμφάνιση του μελετώνται μέσω της Παλινδρόμησης Κυρίων Συνιστωσών και της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων.

ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ : Παλινδρόμηση, Γραμμικό μοντέλο, Πολυσυγγραμμικότητα, PCA, Ανάλυση Κυρίων Συνιστωσών, PCR, Παλινδρόμηση Κυρίων Συνιστωσών, PLSR, Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων, Minitab, R

ΚΕΦΑΛΑΙΟ 1. ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η Παλινδρόμηση αποτελεί μία από τις σημαντικότερες στατιστικές μεθόδους. Πολλά φαινόμενα των σύγχρονων επιστημών μοντελοποιούνται μέσω αυτής. Σε προβλήματα οικονομικής, βιολογικής φύσεως καθώς και σε άλλες επιστήμες απαραίτητος θεωρείται ο καθορισμός της σχέσης μεταξύ μεταβλητών, δηλαδή ο σχεδιασμός ενός μοντέλου, μέσω του οποίου ερμηνεύεται ικανοποιητικά το εκάστοτε δοθέν πρόβλημα. Το μοντέλο αυτό προκύπτει θεωρώντας μια εξαρτημένη μεταβλητή Y και ένα πλήθος ανεξάρτητων μεταβλητών X_i όπου τα Y, X_i συνδέονται με γραμμικό τρόπο. Έτσι δημιουργείται ένα μοντέλο που σχετίζεται με τα δεδομένα μας (παρατηρήσεις) και μέσω αυτού κάνουμε προβλέψεις για μελλοντικές τιμές του Y. Η συγκεκριμένη διαδικασία ονομάζεται Γραμμική Παλινδρόμηση. Σκοπός της είναι ο καθορισμός του βέλτιστου για τα δεδομένα μας μοντέλου.

1.1 ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Αρχικά, θα αναφερθούμε στο απλό γραμμικό μοντέλο και ακολούθως θα επεκταθούμε στη γενική περίπτωση του πολλαπλού γραμμικού μοντέλου.

Απλό γραμμικό μοντέλο ονομάζεται το μοντέλο μέσω του οποίου διερευνάται η σχέση μεταξύ δυο μεταβλητών, Y εξαρτημένης και X ανεξάρτητης. Αυτό περιγράφεται με τη συλλογή ενός δείγματος μεγέθους n από έναν πληθυσμό και καταγράφοντας για κάθε άτομο του δείγματός μας τις τιμές των δυο μεταβλητών X και Y. Με βάση τα ζεύγη τιμών που δημιουργούνται :

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

μπορούμε να διερευνήσουμε τη σχέση μεταξύ των μεταβλητών X, Y. Έτσι θεωρώντας το απλούστερο μοντέλο που ερμηνεύει μια τέτοια σχέση (απλό γραμμικό μοντέλο) τα $\mathrm{X}_i, \mathrm{Y}_i$ συνδέονται με την εξής σχέση :

$$\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{X_i} + \varepsilon_\mathrm{i} \qquad \mathrm{i} = 1, 2, \ldots, \mathrm{n} \tag{1.1}$$

όπου β_0, β_1 δυο άγνωστες σταθερές και $\varepsilon_1, \varepsilon_2, \ldots\ldots, \varepsilon_n$ ανεξάρτητες τυχαίες μεταβλητές. Ονομάζονται σφάλματα και θεωρούμε ότι ακολουθούν την Κανονική κατανομή $\mathrm{N}\left(0, \sigma^2\right)$ με σ^2 άγνωστο.

Συχνά όμως σε διάφορα προβλήματα, η μεταβλητή απόκρισης Y μπορεί να συνδέεται με περισσότερες από μια επεξηγηματικές μεταβλητές $\mathrm{X}_1, \mathrm{X}_2, \ldots., \mathrm{X}_k$ Ομοίως με το απλό γραμμικό μοντέλο, μπορούμε να χρησιμοποιήσουμε ένα νέο μοντέλο που καλείται πολλαπλό γραμμικό μοντέλο, το οποίο διερευνά την εξάρτηση της Y από τις $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_k$ μεταβλητές. Θα έχει την ακόλουθη μορφή :

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ \ldots. \ + \beta_\mathrm{k} \mathrm{X}_\mathrm{k} + \ \varepsilon \tag{1.2}$$

όπου $\beta_0, \beta_1, \ldots \beta_\mathrm{k}$ άγνωστες σταθερές.

Για τη διερεύνηση της εξάρτησης αυτής, λαμβάνουμε απ' τον προς μελέτη πληθυσμό μας, δείγμα μεγέθους n και για κάθε άτομο του δείγματος καταγράφουμε τις τιμές των μεταβλητών αυτών.

Δηλαδή, για το i-άτομο καταγράφουμε τις τιμές $(\mathrm{Y_i}, \mathrm{X_{i1}}, \mathrm{X_{i2}}, \ldots, \mathrm{X_{in}})$ οδηγώντας μας στο μοντέλο (Montgomery et al., 2006):

$$\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{X_{i1}} + \beta_2 \mathrm{X_{i2}} + \ldots.. + \beta_\mathrm{k} \mathrm{X_{ik}} + \ \varepsilon_\mathrm{i} \tag{1.3}$$

όπου $i = \ 1, 2, \ldots, \mathrm{n}.$

Τα ε_i καλούνται ομοίως σφάλματα. Πρόκειται για ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την Κανονική κατανομή $N(0, \sigma^2)$. Αντιθέτως, οι μεταβλητές $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_k$, οι οποίες καλούνται επεξηγηματικές δεν είναι τυχαίες.

Έτσι με βάση τα παραπάνω, το μοντέλο μας γράφεται με τη μορφή πινάκων ως:

$$\mathrm{Y} = \mathrm{X}\beta + \ \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2k} \\ 1 & . & . & . & . & . & . \\ 1 & . & . & . & . & . & . \\ 1 & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Πίνακας 1.Πίνακας σχεδιασμού.

1.2 ΕΚΤΙΜΩΝΤΑΣ ΤΙΣ ΠΑΡΑΜΕΤΡΟΥΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Η εκτίμηση των παραμέτρων του μοντέλου, δηλαδή των $\beta_0, \beta_1, \ldots \beta_k$ μπορεί να γίνει με δυο μεθόδους :

- Τη μέθοδο των ελαχίστων τετραγώνων (least squares method). Η βασική αρχή της μεθόδου των ελαχίστων τετραγώνων είναι ότι επιλέγονται οι παράμετροι $\beta_0, \beta_1, \ldots \beta_k$ που ελαχιστοποιούν το άθροισμα των τετραγώνων των παρατηρημένων υπολοίπων ε_i, όπου $i = 1, 2, \ldots, n$.
- Τη μέθοδο μέγιστης πιθανοφάνειας (maximum likelihood method). Η μέθοδος μέγιστης πιθανοφάνειας βασίζεται στην υπόθεση της κανονικότητας, δηλαδή στο γεγονός ότι τα υπόλοιπα ε_i ακολουθούν την Κανονική κατανομή με μέση τιμή ίση με 0 και διασπορά ίση με σ^2 δηλαδή : $\varepsilon_i \sim N(0, \sigma^2)$.

➢ Η ΜΕΘΟΔΟΣ ΤΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Όπως παραπάνω αναφέρθηκε σκοπός της μεθόδου των ελαχίστων τετραγώνων είναι η ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων. Ας υποθέσουμε λοιπόν ότι $S(\beta)$ η συνάρτηση ελαχίστων τετραγώνων για την οποία ισχύει (Οικονόμου και Καρώνη , 2010):

$$S(\beta) = \sum_{i=1}^{n} \varepsilon^2_i \qquad (1.4)$$

όπου β το διάνυσμα των παραμέτρων του μοντέλου.

Στόχος της μεθόδου είναι η εύρεση εκείνου του β για το οποίο η ανωτέρω συνάρτηση λαμβάνει την ελάχιστη τιμή της. Ειδικότερα :

$$S(\beta) = \sum_{i=1}^{n} \varepsilon^2_i = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta)$$
$$= Y^T Y - \beta^T X^T X \beta - 2Y^T X \beta . \qquad (1.5)$$

όπου Y το $(n \times 1)$ διάνυσμα στήλη της μεταβλητής απόκρισης.

Απαραίτητη προϋπόθεση για την ελαχιστοποίηση της $S(\beta)$ είναι :

$$\frac{dS(\beta)}{d\beta} = 0 \tag{1.6}$$

Δηλαδή :

$$2(X^T X)\beta - 2X^T Y = 0$$
$$(X^T X)\beta = X^T Y \tag{1.7}$$

Εάν ο πίνακας $X^T X$ είναι αντιστρέψιμος, δηλαδή εάν ισχύει ότι $\left|X^T X\right| \neq 0$, τότε η εκτιμήτρια του διανυσμάτων των παραμέτρων ισούται με :

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.8}$$

Επομένως η χρήση της μεθόδου των ελαχίστων τετραγώνων προϋποθέτει την ύπαρξη του $(X^T X)^{-1}$. Αυτό παρατηρείται όταν οι μεταβλητές X_j είναι μεταξύ τους γραμμικά ανεξάρτητες δηλαδή αν καμία στήλη του πίνακα X δεν μπορεί να γραφτεί ως γραμμικός συνδυασμός των υπολοίπων. Αν το παραπάνω δεν ισχύει, τότε εμφανίζεται το φαινόμενο της Πολυσυγγραμμικότητας το οποίο θα εκτεθεί αναλυτικά στο επόμενο κεφάλαιο της παρούσης.

Έχοντας παραθέσει τη μέθοδο ελαχίστων τετραγώνων, αξίζει να αναφερθεί στο σημείο αυτό ότι τα ε_i θεωρούνται ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την $N(0,\sigma^2)$. Άρα το διάνυσμα ε, θα έχει από κοινού σ.π.π $\mathrm{N}(0,\sigma^2\mathrm{I}_n)$ ακολουθώντας την πολυδιάστατη Κανονική κατανομή, όπου I_n ο μοναδιαίος πίνακας με διάσταση ίση με n. Επομένως όμοια και το τυχαίο διάνυσμα:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}$$

θα ακολουθεί την πολυδιάστατη Κανονική κατανομή $\mathrm{N}(\mathrm{X}\beta,\sigma^2\mathrm{I}_n)$. Επομένως θα έχει συνάρτηση πιθανοφάνειας :

$$L=(\beta,\sigma^2)= \quad f(\mathrm{y}_1,\mathrm{y}_2,...,\mathrm{y}_\mathrm{n};\beta,\sigma^2)= \qquad (1.9)$$
$$\frac{1}{(2\pi)^{\frac{n}{2}}}\times\frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}}\times e^{-\frac{1}{2\sigma^2}\left(y-X\beta\right)^T\left(y-X\beta\right)}$$

Ακολούθως, για να βρω την ε.μ.π του β, αρκεί ομοίως με την μέθοδο των ελαχίστων τετραγώνων να ελαχιστοποιήσουμε τη σχέση (1.6). Επομένως αν ισχύει η (1.7), τότε οι εκτιμήτριες μέγιστης πιθανοφάνειας για το $\beta=[\beta_0,\beta_1,\ldots,\beta_\mathrm{k}]^\mathrm{T}$ δίνονται από την (1.8).

Η ε.μ.π του σ^2 δίνεται και ως εξής :

$$\hat{\sigma}^2=\frac{1}{n}(\mathrm{Y}\text{-}\mathrm{X}\hat{\beta})^\mathrm{T}(Y-X\hat{\beta})=\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Y}_\mathrm{i}-\hat{\beta}_0-\hat{\beta}_1\mathrm{X}_{\mathrm{i}1}-\ldots-\hat{\beta}_\mathrm{k}\mathrm{X}_{\mathrm{ik}})^2 \qquad (1.10)$$

Οι προβλέψεις των Y_i που παίρνουμε (ονομάζονται αλλιώς Y predicted) λαμβάνονται ως οι εκτιμήσεις των $E(\mathrm{Y}_i)=\beta_0+\beta_1\mathrm{X}_{\mathrm{i}1}+\ldots+\beta_\mathrm{k}\mathrm{X}_{\mathrm{ik}}$ και ισούνται με :

$$\hat{\mathrm{Y}}_\mathrm{i}=\hat{\beta}_0+\hat{\beta}_1\mathrm{X}_{\mathrm{i}1}+\ldots+\hat{\beta}_\mathrm{k}\mathrm{X}_{\mathrm{ik}} \quad \mathrm{i}=1,2,\ldots,\mathrm{n} \qquad (1.11)$$

ή αλλιώς με τη μορφή πινάκων ως :

$$\hat{Y}=\mathrm{X}\hat{\beta}=\mathrm{X}(\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}\mathrm{Y}=HY \qquad (1.12)$$

όπου $H=\mathrm{X}\left(\mathrm{X}^\mathrm{T}\mathrm{X}\right)^{-1}\mathrm{X}^\mathrm{T}$.

Ο πίνακας H καλείται πίνακας προβολής (hat matrix). Λαμβάνουμε επίσης τα εκτιμημένα σφάλματα (residuals) ως :

$$\hat{\varepsilon}_\mathrm{i}=Y-\hat{Y}_\mathrm{i} \quad \mathrm{i}=1,2,\ldots,n \qquad (1.13)$$

ή ομοίως σε μορφή πινάκων :

$$\hat{\varepsilon}=Y-\hat{Y}=Y-X\hat{\beta}=Y-X\left(X^TX\right)^{-1}X^TY=(I_n-\mathrm{H})Y \qquad (1.14)$$

1.3 ΟΙ ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΚΑΙ ΤΑ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Έχουμε ότι $\varepsilon \sim N(0, \sigma^2 I_n)$. Τότε προκύπτει ότι (Draper & Smith , 1998):

$$\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \sim N(\beta, \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}) \qquad (1.15)$$

δηλαδή το $\hat{\beta}$ ακολουθεί την πολυδιάστατη Κανονική κατανομή με μέσο :

$$\beta = [\beta_0, \beta_1, \ldots, \beta_{\mathrm{k}}]^{\mathrm{T}}$$ και πίνακα διασποράς $\sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$.

Επιπροσθέτως έχουμε ότι :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mathrm{Y_i} - \hat{Y}_{\mathrm{i}})^2 = \frac{1}{\sigma^2}(\mathrm{Y} - \mathrm{X}\hat{\beta})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\hat{\beta}) = \frac{1}{\sigma^2}\mathrm{Y}^{\mathrm{T}}(\mathrm{I}_n - H)\mathrm{Y} \sim \chi^2_{\mathrm{n}-k-1} \qquad (1.16)$$

Αντικαθιστούμε τώρα ως εκτιμήτρια του σ^2, την αμερόληπτη :

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(\mathrm{Y_i} - \hat{Y}_{\mathrm{i}})^2 = S^2 \qquad (1.17)$$

Έστω τώρα $\mathrm{c_{ii}}$, $\mathrm{i} = 0,1,2,..,\mathrm{k}$ να είναι τα διαγώνια στοιχεία του πίνακα $\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}$ έχουμε δηλαδή ότι :

$$\hat{\beta}_{\mathrm{i}} \sim N(\beta_{\mathrm{i}}, \sigma^2\mathrm{c_{ii}}) \qquad \mathrm{i} = 0,1,2\ldots,k \qquad (1.18)$$

Επειδή όμως

$$\frac{n-k-1}{\sigma^2}S^2 \sim \chi^2_{\mathrm{n}-k-1} \qquad (1.19)$$

προκύπτει ότι :

$$\frac{\hat{\beta}_i - \beta_i}{S\sqrt{c_{ii}}} \sim t_{n-k-1} \quad i = 0,1,2,\ldots,k \tag{1.20}$$

Σύμφωνα με τα παραπάνω, τα διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου μας με ε.σ $1-\alpha$, δίνονται από :

$$(\hat{\beta}_i - S\sqrt{c_{ii}}\, t_{n-k-1}(\alpha/2), \hat{\beta}_i + S\sqrt{c_{ii}}\, t_{n-k-1}(\alpha/2)) \quad i = 0,1,2,\ldots,k \tag{1.21}$$

Για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \beta_i = 0$, οι περιοχές απόρριψης σε ε.σ α θα είναι :

$$K: \ |T_i| \ > t_{n-k-1}(\alpha/2) \tag{1.22}$$

όπου

$$T_i = \frac{\hat{\beta}_i}{S\sqrt{c_{ii}}} \quad i = 0,1,2,\ldots k \tag{1.23}$$

Εάν απορρίψουμε τη μηδενική υπόθεση $H_0 : \beta_i = 0$ για κάποιο i, αυτό ερμηνεύεται με το ότι η μεταβλητή απόκρισης Y εξαρτάται από την X_i μεταβλητή.

1.4 ΕΡΜΗΝΕΙΑ ΤΗΣ ΣΥΝΟΛΙΚΗΣ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Η δειγματική διασπορά των παρατηρήσεων ορίζεται ως εξής :

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{1.24}$$

όπου το πρώτο άθροισμα συμβολίζεται με SST, το δεύτερο με SSE και το τρίτο με SSR, εκ των οποίων το SST ερμηνεύει την ολική μεταβλητότητα των $\mathrm{Y_i}$, το SSR τη μεταβλητότητα των προβλέψεων (Y predicted).

Τέλος το SSE ερμηνεύει τη μεταβλητότητα των $\mathrm{Y_i}$ σε σχέση με τις αντίστοιχες τιμές που έχουμε προβλέψει. Το ποσοστό της μεταβλητότητας των $\mathrm{Y_i}$, το οποίο ερμηνεύεται από το μοντέλο μας υπολογίζεται απ τον συντελεστή προσδιορισμού R^2.

$$\mathrm{R}^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} \tag{1.25}$$

Ισχύει ότι :

$$\frac{SSE}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mathrm{Y_i} - \hat{Y}_\mathrm{i})^2 \sim \chi^2_{\mathrm{n-k-1}} \tag{1.26}$$

Επιπροσθέτως είναι γνωστό, ότι εάν :

$$\beta_1 = \beta_2 = \ldots = \beta_\mathrm{k} = 0 \tag{1.27}$$

ισχύουν τα ακόλουθα :

$$\frac{SSR}{\sigma^2} \sim \chi^2_{\mathrm{k}} \tag{1.28}$$

$$\frac{SST}{\sigma^2} \sim \chi^2_{\mathrm{n-1}} \tag{1.29}$$

δηλαδή, στην περίπτωση όπου $\beta_1 = \beta_2 = \ldots = \beta_\mathrm{k} = 0$, τότε :

$$F = \frac{\frac{SSR}{\sigma^2} \Big/ k}{\frac{SSE}{\sigma^2} \Big/ n-k-1} = \frac{SSR \Big/ k}{SSE \Big/ n-k-1} \sim \mathrm{F}_{k,\mathrm{n}-k-1} \tag{1.30}$$

αφού $\mathrm{SSR}, \mathrm{SSE}$ ανεξάρτητες.

Με βάση τα ανωτέρω, κάνουμε έναν έλεγχο για τη μηδενική υπόθεση :

$$\mathrm{H}_0 : \beta_0 = \beta_1 \ldots = \beta_\mathrm{k} = 0 \qquad (1.31)$$

που σημαίνει ότι η Y μεταβλητή είναι ανεξάρτητη από τις $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_k$ μεταβλητές. Η υπόθεση H_0 θα απορρίπτεται όταν για το παραπάνω στατιστικό F ισχύει :

$$\mathrm{F} = \frac{SSR/k}{SSE/n-k-1} > \mathrm{F}_{\mathrm{k,n-k-1}}(\alpha) \qquad (1.32)$$

όπου α επίπεδο σημαντικότητας και $\mathrm{F}_{\mathrm{k,n-k-1}}(\alpha)$ το α -σημείο της F κατανομής με βαθμό ελευθερίας k και $n-k-1$ αντίστοιχα. Ο πίνακας ανάλυσης διασποράς του παραπάνω μοντέλου (ANOVA) είναι ο ακόλουθος :

ΜΟΝΤΕΛΟ	ΑΘΡΟΙΣΜΑ	ΒΑΘΜΟΙ ΕΛΕΥΘΕΡΙΑΣ	ΜΕΣΟ ΑΘΡΟΙΣΜΑ ΤΕΤΡΑΓΩΝΩΝ	ΕΛΕΓΧΟΣ-F
Λόγω παλινδρόμησης	$SSR = \sum_{i=1}^{n} (\hat{Y}_\mathrm{i} - \bar{Y})^2$	k	$\mathrm{MSR} = \mathrm{SSR}/k$	$\frac{MSR}{MSE}$
Λόγω υπολοίπων	$SSE = \sum_{i=1}^{n} (\mathrm{Y}_\mathrm{i} - \hat{Y}_\mathrm{i})^2$	n-k-1	$\mathrm{MSE} = \mathrm{SSE}/n-k-1$	
Ολικό	$SST = \sum_{i=1}^{n} (\mathrm{Y}_\mathrm{i} - \bar{Y})^2$	n-1		

Πίνακας 2. Ανάλυση διασποράς

1.5 ΕΠΙΛΟΓΗ ΤΟΥ ΒΕΛΤΙΣΤΟΥ ΜΟΝΤΕΛΟΥ

Συχνά, στην Πολλαπλή Γραμμική Παλινδρόμηση παρουσιάζεται το φαινόμενο της ύπαρξης μεγάλου αριθμού ανεξαρτήτων μεταβλητών. Σκοπός του παρόντος κεφαλαίου, είναι να διαπιστώσουμε ποιες απ' αυτές τις μεταβλητές επηρεάζουν τη μεταβλητή απόκρισης Y. Αυτό μπορεί να πραγματοποιηθεί μελετώντας όλα τα δυνατά μοντέλα και ακολούθως επιλέγοντας το πλέον κατάλληλο. Για παράδειγμα έστω ότι το προς εξέταση πολλαπλό γραμμικό μοντέλο μας, περιέχει 3 ανεξάρτητες μεταβλητές X_1, X_2, X_3. Άρα τα υποψήφια γραμμικά μοντέλα μας είναι τα ακόλουθα 7:

1. $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
2. $Y = \beta_0 + \beta_1 X_2 + \varepsilon$
3. $Y = \beta_0 + \beta_1 X_3 + \varepsilon$
4. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
5. $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$
6. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$
7. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Σύμφωνα με τα γνωστά, πλέον κατάλληλο θεωρείται το μοντέλο με το μεγαλύτερο συντελεστή προσδιορισμού $R = 1 - SSE/SST$, παρατηρείται όμως, ότι αυτό δεν ισχύει. Οφείλεται στο γεγονός ότι με την πρόσθεση ανεξάρτητων μεταβλητών στο μοντέλο μας, το R^2 αυξάνεται ή παραμένει σταθερό.

Αυτό συμβαίνει διότι με την πρόσθεση ανεξάρτητων μεταβλητών, το SSE (άθροισμα τετραγώνων λόγω σφαλμάτων) μειώνεται ή παραμένει σταθερό, ενώ το SST (ολικό άθροισμα τετραγώνων) παραμένει σταθερό. Για αυτό το λόγο, συχνά χρησιμοποιείται ο προσαρμοσμένος συντελεστής προσδιορισμού R^2 adjusted,

όπου
$$R^2 \text{ adjusted} = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} \qquad (1.33)$$

επομένως καλύτερο μοντέλο κρίνεται αυτό με το μεγαλύτερο συντελεστή. Αξίζει να τονιστεί ότι το SST παραμένει σταθερό σε όλα τα υποψήφια μοντέλα, επομένως καλύτερο μοντέλο είναι αυτό με το μικρότερο $SSE/(n-k-1)$.

1.6 Η ΕΦΑΡΜΟΓΗ ΤΟΥ CENTERING (ΚΕΝΤΡΟΠΟΙΗΣΗ) ΚΑΙ ΤΟΥ SCALING (ΤΥΠΟΠΟΙΗΣΗ)

Σε πολλά στατιστικά προβλήματα, προκειμένου τα μοντέλα μας να ερμηνευθούν στον καλύτερο δυνατό βαθμό, χρήσιμη είναι η διαδικασία του centering (κεντροποίηση) καθώς και του scaling (τυποποίηση). Χρησιμοποιείται σε προβλήματα χημειομετρίας, οικονομετρίας και σε ποικίλους επιστημονικούς κλάδους και βρίσκει εφαρμογή στην Παλινδρόμηση Κυρίων Συνιστωσών και στην Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων που θα μελετήσουμε στη συνέχεια.

Σκοπός του centering είναι στο τελικό μας μοντέλο να αφαιρείται ο σταθερός όρος. Αυτό συμβαίνει αφαιρώντας από κάθε μεταβλητή μας $\mathrm{X_j}$ τον αντίστοιχο μέσο $\bar{x}$. Έστω τώρα το μοντέλο :

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots + \beta_k \mathrm{X}_k + \varepsilon$$

είναι δυνατόν οι μεταβλητές που υπεισέρχονται στο μοντέλο μας να εκφράζονται σε διαφορετικές μονάδες π.χ X_1 σε κιλά, X_2 σε γραμμάρια και X_k σε τόνους.

Επομένως κρίνεται απαραίτητη η διαδικασία του scaling, καθώς οι ανεξάρτητες αλλά και οι εξαρτημένες μεταβλητές οφείλουν να είναι τυποποιημένες προκειμένου οι συντελεστές τους να είναι αδιάστατα μεγέθη. Παρακάτω περιγράφεται ο « μηχανισμός » του scaling. Θα αναφερθούμε σε δυο τεχνικές scaling, τις ακόλουθες (Montgomery et al., 2006, p. 105) :

SCALING TO UNIT NORMAL

Έστω $\mathrm{x_{ij}}$ η κάθε παρατήρηση για τις επεξηγηματικές μεταβλητές του δείγματος με $\mathrm{i} = 1, 2, \ldots, n$ το πλήθος των παρατηρήσεων και $\mathrm{j} = 1, 2, \ldots, \mathrm{k}$ το πλήθος των μεταβλητών. Θεωρούμε μια νέα μεταβλητή έστω X^* η οποία προκύπτει ως εξής :

$$\mathrm{X^*_{ij}} = \frac{\mathrm{x_{ij}} - \bar{x}_\mathrm{j}}{\mathrm{s_j}} \qquad (1.34)$$

όπου $\mathrm{s_j}$ η τυπική απόκλιση όταν

$$s^2{}_j = \frac{\sum_{i=1}^{n} (\mathrm{x_{ij}} - \bar{x}_\mathrm{j})^2}{n-1} \qquad (1.35)$$

η διασπορά των παρατηρήσεων. Ομοίως, αν y_i οι παρατηρήσεις για την y-μεταβλητή του δείγματος με $i = 1, 2, \ldots, n$ το πλήθος των παρατηρήσεων, θεωρούμε μια νέα μεταβλητή, έστω Y^* η οποία προκύπτει ως εξής :

$$Y^*_i = \frac{y_i - \bar{y}}{s} \tag{1.36}$$

όπου s η τυπική απόκλιση της y-μεταβλητής όταν :

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \quad . \tag{1.37}$$

Το μοντέλο που προκύπτει σύμφωνα με τα παραπάνω είναι το εξής :

$$y^*_i = b^*_1 x^*_{i1} + b^*_2 x^*_{i2} + \ldots. + b^*_k x^*_{ik} + \varepsilon^*_i \tag{1.38}$$

ως αποτέλεσμα της παραπάνω διαδικασίας, η νέα μορφή της εκτιμήτριας των ελάχιστων τετραγώνων για το μοντέλο μας, είναι η ακόλουθη :

$$\hat{b}^* = (X^{*T} X^*)^{-1} X^{*T} Y^* \tag{1.39}$$

SCALING TO UNIT LENGTH

Έστω x_{ij} η κάθε παρατήρηση για τις επεξηγηματικές μεταβλητές του δείγματος με $i = 1, 2, \ldots, n$ το πλήθος των παρατηρήσεων και $j = 1, 2, \ldots, k$ το πλήθος των μεταβλητών. Για τη συγκεκριμένη τεχνική, όπως και στην περίπτωση του scaling to unit normal θα θεωρήσουμε μια νέα μεταβλητή, έστω Z η οποία προκύπτει ως εξής (Montgomery et al., 2006, p. 106):

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{jj}^{1/2}} \tag{1.40}$$

όπου $$s_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \quad (1.41)$$

Ομοίως, αν y_i οι παρατηρήσεις για την y-μεταβλητή του δείγματος με $i = 1, 2, \ldots, n$ το πλήθος των παρατηρήσεων, θεωρούμε μια νέα μεταβλητή έστω W η οποία προκύπτει ως εξής :

$$W_i = \frac{y_i - \bar{y}}{s_y^{1/2}} \quad (1.42)$$

όπου $$s_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (1.43)$$

κάνοντας χρήση της παραπάνω διαδικασίας έχουμε ότι οι νέες μεταβλητές Z_{ij} που θεωρούμε έχουν μέση τιμή ίση με το μηδέν δηλαδή $\bar{Z}_{ij}$=0 και επιπροσθέτως έχουν μήκος ίσο με τη μονάδα καθώς ισχύει :

$$\sqrt{\sum_{i=1}^{n} (Z_{ij} - \bar{Z}_{ij})^2} = 1 \quad (1.44)$$

Τελικά, το μοντέλο που προκύπτει σύμφωνα με τα παραπάνω είναι το εξής :

$$w_i = b_1 z_{i1} + b_2 z_{i2} + \ldots + b_k z_{ik} + \varepsilon_i \quad (1.45)$$

Αποτέλεσμα της παραπάνω διαδικασίας, η νέα μορφή της εκτιμήτριας των ελαχίστων τετραγώνων για το μοντέλο μας είναι η ακόλουθη :

$$\hat{b} = (Z^T Z)^{-1} Z^T W \quad (1.46)$$

Εξαιτίας εφαρμογής της συγκεκριμένης μεθόδου scaling που μόλις περιγράψαμε, προκύπτει επίσης ότι ο πίνακας $Z^T Z$ διαθέτει δυο χαρακτηριστικά γνωρίσματα. Πρόκειται για έναν πίνακα συμμετρικό καθώς και για έναν πίνακα που

περιέχει τους συντελεστές συσχέτισης μεταξύ των επεξηγηματικών μας μεταβλητών. Πράγματι ο $Z^{\mathrm{T}}Z$ είναι της μορφής :

$$Z^{\mathrm{T}}Z=\begin{pmatrix}1 & \dots & r_{1k}\\ \vdots & \ddots & \vdots\\ r_{1k} & \cdots & 1\end{pmatrix} \tag{1.47}$$

με τα διαγώνια στοιχεία του ίσα με τη μονάδα και τα μη διαγώνια στοιχεία του ίσα με $r_{ij}\quad i=1,2,...,n\quad j=1,2,.p$, καλούνται συντελεστές συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών $\mathrm{X_i}$ και $\mathrm{X_j}$. Ισχύει :

$$r_{ij}=\frac{\sum_{l=1}^{n}(\mathrm{x_{lj}}-\bar{x}_{\mathrm{j}})(x_{lj}-\bar{x}_{\mathrm{j}})}{(\mathrm{s_{jj}s_{ii}})^{1/2}}=\frac{\mathrm{s_{ij}}}{(\mathrm{s_{jj}s_{ii}})^{1/2}} \tag{1.48}$$

το διάνυσμα-στήλη $Z^{\mathrm{T}}W$ της μορφής :

$$Z^{\mathrm{T}}W=[r_{1y}r_{2y}.....r_{ky}]^{\mathrm{T}} \tag{1.49}$$

περιέχει τους συντελεστές συσχέτισης μεταξύ της ανεξάρτητης μεταβλητής $\mathrm{X_j}$ και της εξαρτημένης μεταβλητής Y. Ισχύει :

$$r_{jy}=\frac{\sum_{l=1}^{n}(\mathrm{x_{lj}}-\bar{x}_{\mathrm{j}})(y_l-\bar{y})}{(\mathrm{s_{jj}s_y})^{1/2}}=\frac{\mathrm{s_{jy}}}{(\mathrm{s_{jj}s_y})^{1/2}} \tag{1.50}$$

Οι μέθοδοι scaling που εκθέσαμε παραπάνω σχετίζονται μεταξύ τους με την ακόλουθη σχέση :

$$\mathrm{X}^{*\mathrm{T}}\,\mathrm{X}^{*}=(n-1)Z^{\mathrm{T}}Z \tag{1.51}$$

Παρατηρώντας τις σχέσεις (1.38) και (1.45) συμπεραίνουμε ότι οι δυο μέθοδοι τυποποίησης μας οδηγούν στην ίδια εκτιμήτρια των παραμέτρων, ας την καλούμε από εδώ και πέρα $\hat{b}$.

ΚΕΦΑΛΑΙΟ 2. ΤΟ ΦΑΙΝΟΜΕΝΟ ΤΗΣ ΠΟΛΥΣΥΓΡΑΜΜΙΚΟΤΗΤΑΣ

Συχνά, στην Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression) είναι πιθανόν, μια ή περισσότερες ανεξάρτητες μεταβλητές $\mathrm{X_j}$, να είναι γραμμικά εξαρτημένες. Πρόκειται για ένα συχνό φαινόμενο που παρουσιάζεται σε βιολογικές, οικονομικές, χημικές έρευνες και οδηγεί στη λανθασμένη εξαγωγή εκτιμήσεων. Μπορεί να αντιμετωπιστεί με την αφαίρεση κάποιων ανεξάρτητων μεταβλητών απ' το μοντέλο μας.

Έστω λοιπόν ότι κάνουμε ένα πείραμα, συλλέγουμε τις παρατηρήσεις μας και οδηγούμαστε στο μοντέλο της Πολλαπλής Γραμμικής Παλινδρόμησης. Χρησιμοποιούμε το εξής πολλαπλό γραμμικό μοντέλο :

$$Y = X\beta + \varepsilon$$

όπου Y είναι ένα $n\times 1$ διάνυσμα των αποκρίσεων, X ένας πίνακας $n\times p$ $(p = k+1)$ των μεταβλητών Παλινδρόμησης, β ένα $p\times 1$ διάνυσμα των συντελεστών της Παλινδρόμησης και ε ένα $n\times 1$ διάνυσμα των σφαλμάτων μας όπου $\varepsilon_{\mathrm{i}} \sim N(0,\sigma^2)$.

Παρουσιάζονται τώρα δυο πιθανές περιπτώσεις (Montgomery et al., 2006 , pp. 323-324):

1) Οι ανεξάρτητες μεταβλητές του μοντέλου μας, $\mathrm{X_j}$, δε σχετίζονται μεταξύ τους.

Στην περίπτωση αυτή λέγονται ορθογώνιες. Τότε το εσωτερικό γινόμενο μεταξύ των στηλών του πίνακα X ισούται με 0 ενώ οι στήλες είναι κάθετες μεταξύ τους. Επομένως ο πίνακας $\mathrm{X^TX}$ είναι διαγώνιος. Οπότε τα β_{j}, τα οποία εξάγονται μέσω της Πολλαπλής Γραμμικής Παλινδρόμησης εξαρτώνται μόνο απ' την μεταβλητή τους , $\mathrm{X_j}$.

Με αυτό τον τρόπο λαμβάνουμε ένα "επιθυμητό" μοντέλο διαπιστώνοντας ότι η επίδραση της κάθε μεταβλητής είναι εμφανής, έτσι οδηγούμαστε σε ασφαλείς προβλέψεις για τις μελλοντικές τιμές του Y.

Πρακτικά όμως, η παραπάνω περίπτωση σπάνια εμφανίζεται στα προβλήματα Παλινδρόμησης που καλούμαστε να λύσουμε, τα οποία προκύπτουν

από την παρατήρηση και τη συλλογή δεδομένων. Το ενδεχόμενο, οι μεταβλητές μας να χαρακτηρίζονται από ορθογωνιότητα μεταξύ τους δεν είναι και το πλέον πιθανό. Αντιθέτως κάτι τέτοιο μπορεί να παρατηρηθεί σε προσχεδιασμένα πειράματα, στα οποία όντως έχει επιδιωχθεί η ορθογωνιότητα.

Αξίζει να τονιστεί στο σημείο αυτό, ότι πολλές φορές όταν οι μεταβλητές μας είναι ασθενώς συσχετισμένες μεταξύ τους, δηλαδή όταν η έλλειψη ορθογωνιότητας δεν είναι ιδιαιτέρως σοβαρή, το μοντέλο μας και η ποιότητα των μελλοντικών του προβλέψεων δεν απειλείται σημαντικά.

2) Οι ανεξάρτητες μεταβλητές του μοντέλου μας, X_j, είναι γραμμικά εξαρτημένες μεταξύ τους.

Σε αυτή την περίπτωση, προκύπτουν μοντέλα Παλινδρόμησης τα οποία να μας δίνουν λανθασμένες και μη ασφαλείς μελλοντικές προβλέψεις και άρα οδηγούμαστε σε εσφαλμένα συμπεράσματα. Υποθέτουμε ότι X ο πίνακας σχεδιασμού και $X_1, X_2, \ldots, X_p$ οι στήλες του πίνακα αυτού, τότε υπάρχουν σταθερές, $\alpha_0, \alpha_1, \ldots, \alpha_p$ όχι όλες 0 και άρα ισχύει :

$$\alpha_0 + \sum_{i=1}^{p} \alpha_k X_k = 0 \qquad (2.1)$$

Εάν η παραπάνω σχέση ισχύει απολύτως, τότε λέμε ότι το μοντέλο μας παρουσιάζει τέλεια Πολυσυγγραμμικότητα. Τότε, ο πίνακας σχεδιασμού δεν έχει βαθμό p αλλά μικρότερο του p και το ίδιο συμβαίνει για τον $X^T X$. Ο πίνακας $X^T X$ καλείται τώρα ιδιάζων (singular), ισχύει επομένως $\| X^T X \| = 0$ και άρα δεν μπορεί να υπολογιστεί ο αντίστροφος του, $(X^T X)^{-1}$. Λόγω της παραπάνω αδυναμίας υπολογισμού του αντιστρόφου, δεν καθίσταται εφικτός ο υπολογισμός των συντελεστών Παλινδρόμησης του μοντέλου.

Στα πλαίσια όμως που εμείς θα ασχοληθούμε, η σχέση (2.1) δεν ισχύει απολύτως, αλλά οι μεταβλητές μας χαρακτηρίζονται από σχεδόν γραμμική εξάρτηση μεταξύ τους (near – linear – dependency). Ισχύει δηλαδή :

$$\alpha_0 + \sum_{i=1}^{p} \alpha_k X_k \cong 0 \qquad (2.2)$$

και ο αντίστροφος πίνακας $(X^T X)^{-1}$ μπορεί να υπολογιστεί.

Παρ' όλα αυτά όμως στην περίπτωση ύπαρξης της Πολυσυγγραμμικότητας, η μέθοδος ελαχίστων τετραγώνων αποτυγχάνει και επομένως παρουσιάζεται εύλογα το πρόβλημα του τρόπου υπολογισμού των συντελεστών Παλινδρόμησης για το μοντέλο. Τούτο όμως απαιτεί την εφαρμογή διαφορετικών μεθόδων.

2.1 ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑ ΚΑΙ ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Όπως παραπάνω εκτέθηκε, η Πολυσυγγραμμικότητα παρουσιάζει πολλές σοβαρές συνέπειες στη μέθοδο των ελαχίστων τετραγώνων για τον υπολογισμό των συντελεστών Παλινδρόμησης. Έστω ότι στο μοντέλο που θα μελετήσουμε έχουμε δυο επεξηγηματικές μεταβλητές και επομένως παίρνουμε το ακόλουθο μοντέλο :

$$\mathrm{Y} = \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \varepsilon \qquad (2.3)$$

όπου τα Y, X_1, X_2 έχουν υποστεί την παραπάνω δεύτερη διαδικασία τυποποίησης, scaling to unit length. Οι κανονικές εξισώσεις ελαχίστων τετραγώνων που λαμβάνουμε είναι (Montgomery et al., 2006 , pp. 326-328) :

$$(Z^{\mathrm{T}} Z)\hat{b} = Z^{\mathrm{T}} W \qquad (2.4)$$

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} [\hat{b}_1 \, \hat{b}_2]^{\mathrm{T}} = [r_{1y} r_{2y}]^{\mathrm{T}} \qquad (2.5)$$

όπου r_{12} η συσχέτιση μεταξύ των X_1 και X_2 και r_{jy} η συσχέτιση μεταξύ του X_j και του Y με $j = 1, 2$. Τώρα ο αντίστροφος του $(Z^{\mathrm{T}} Z)$ είναι ο :

$$C = (Z^{\mathrm{T}} Z)^{-1} = \begin{pmatrix} \dfrac{1}{(1 - r_{12}{}^2)} & \dfrac{-r_{12}}{(1 - r_{12}{}^2)} \\ \dfrac{-r_{12}}{(1 - r_{12}{}^2)} & \dfrac{1}{(1 - r_{12}{}^2)} \end{pmatrix} \qquad (2.6)$$

οι εκτιμήτριες των συντελεστών Παλινδρόμησης είναι :

$$\hat{b}_1 = \frac{r_{1y} - r_{12}r_{2y}}{(1-r_{12}^{\ 2})} \quad \text{και} \quad \hat{b}_2 = \frac{r_{2y} - r_{12}r_{1y}}{(1-r_{12}^{\ 2})} \ . \tag{2.7}$$

Εάν υπάρχει ισχυρή Πολυσυγγραμμικότητα μεταξύ των X_1 και X_2, τότε ο συντελεστής συσχέτισης r_{12} είναι υψηλός. Απ' την σχέση (2.6), παρατηρούμε ότι καθώς :

$|r_{12}| \to 1$ τότε $\text{var}(\hat{b}_j) = C_{jj}\sigma^2 \to \infty$

καθώς και ότι :

$\text{cov}(\hat{b}_1, \hat{b}_2) = C_{12}\sigma^2 \to \pm\infty$, ανάλογα με το αν :
$|r_{12}| \to +1$ ή αν $|r_{12}| \to -1$.

όπου $\text{var}(\hat{b}_j) = C_{jj}\sigma^2$ ο τύπος που μας δίνει τη διασπορά των παραμέτρων του μοντέλου μας και $\text{cov}(\hat{b}_i, \hat{b}_j) = C_{ij}\sigma^2$ ο τύπος που μας δίνει την συνδιασπορά τους.

Επομένως από τα παραπάνω συμπεραίνεται ότι η ισχυρή Πολυσυγγραμμικότητα μεταξύ των X_1 και X_2 οδηγεί στην ύπαρξη υψηλών διασπορών και συνδιασπορών για την εκτιμήτρια των ελαχίστων τετραγώνων των συντελεστών παλινδρόμησης.

Το τελευταίο σημαίνει ότι λαμβάνοντας διαφορετικά δείγματα, παίρνουμε πολύ διαφορετικές εκτιμήσεις για τις παραμέτρους του μοντέλου μας. Όταν υπάρχουν δυο ή περισσότερες ανεξάρτητες μεταβλητές, η Πολυσυγγραμμικότητα παρουσιάζει τις ίδιες επιδράσεις. Τα διαγώνια στοιχεία του πίνακα $C = (Z^{\mathrm{T}}Z)^{-1}$ είναι :

$$C_{jj} = \frac{1}{1-R_j^{\ 2}} \qquad j = 1, 2, \ldots, p \tag{2.8}$$

όπου $R_j^{\ 2}$ ο συντελεστής προσδιορισμού της Παλινδρόμησης του X_j με τις υπόλοιπες $p-1$ το πλήθος επεξηγηματικές μεταβλητές. Εάν υπάρχει ισχυρή Πολυσυγγραμμικότητα μεταξύ της X_j και ενός υποσυνόλου των υπολοίπων μεταβλητών, τότε η τιμή του συντελεστή προσδιορισμού προσεγγίζει τη μονάδα.

Έτσι καθώς η διασπορά των $\hat{b}_j$ δίνεται από τον τύπο :

$$\mathrm{var}(\hat{b}_j) = C_{jj}\sigma^2 = (1 - R_j^{\,2})^{-1}\sigma^2 \qquad (2.9)$$

Συμπερασματικά από τα παραπάνω αναφερθέντα, προκύπτει ότι οι υψηλές διασπορές που παρουσιάζουν οι συντελεστές Παλινδρόμησης οδηγούν στο φαινόμενο της Πολυσυγγραμμικότητας. Απ' το σημείο αυτό και έπειτα θα χρησιμοποιούμε όχι την τυποποιημένη εκτιμήτρια που μέχρι τώρα χρησιμοποιούμε, αλλά την εκτιμήτρια ελαχίστων τετραγώνων :

$$\hat{\beta}_{\mathrm{j}}$$

Επιπροσθέτως, η Πολυσυγγραμμικότητα τείνει να παράγει εκτιμήτριες ελαχίστων τετραγώνων με πολύ υψηλή απόλυτη τιμή. Για να διαπιστώσουμε το παραπάνω αρκεί να θεωρήσουμε το τετράγωνο της απόστασης του $\hat{\beta}$ από το πραγματικό διάνυσμα β. Δηλαδή :

$$L_1^{\,2} = (\hat{\beta} - \beta)^T(\hat{\beta} - \beta) \qquad (2.10)$$

με το αναμενόμενο τετράγωνο της απόστασης να είναι :

$$E(L_1^{\,2}) = E(\hat{\beta} - \beta)^T(\hat{\beta} - \beta) = \sum_{j=1}^{p} E(\hat{\beta}_j - \beta_j)^2 = \sum_{j=1}^{p} \mathrm{var}(\hat{\beta}_j) = \sigma^2 \mathrm{Tr}(X^T X)^{-1} \qquad (2.11)$$

όπου το ίχνος του πίνακα (συντομογραφία Tr) είναι το άθροισμα των διαγώνιων στοιχείων. Σε περίπτωση ύπαρξης Πολυσυγγραμμικότητας, μερικές από τις ιδιοτιμές του πίνακα $X^T X$ είναι πιθανόν να είναι μικρές. Καθώς το ίχνος ενός πίνακα ισούται με το άθροισμα των ιδιοτιμών η παραπάνω σχέση μετασχηματίζεται στην :

$$E(L_1^{\,2}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \qquad (2.12)$$

όπου $\lambda_j > 0 \quad j = 1, 2, \ldots, p$ οι ιδιοτιμές του πίνακα $X^T X$.

Συνεπώς εάν υπάρχει Πολυσυγγραμμικότητα, τουλάχιστον μια από τις λ_j θα είναι μικρή και η σχέση (2.12), καταδεικνύει ότι η απόσταση του $\hat{\beta}$ από το πραγματικό διάνυσμα παραμέτρων β θα είναι μεγάλη. Ισοδύναμα αποδεικνύεται ότι :

$$E(L_1^2) = E(\hat{\beta} - \beta^T)(\hat{\beta} - \beta) = E(\hat{\beta}^T\hat{\beta} - 2\hat{\beta}^T\beta + \beta^T\beta) \qquad (2.13)$$

ή

$$\mathrm{E}(\hat{\beta}^T\hat{\beta}) = \beta^T\beta + \sigma^2 Tr(X^T X)^{-1} \qquad (2.14)$$

Επομένως το διάνυσμα $\hat{\beta}$ είναι γενικά μεγαλύτερο από το διάνυσμα β, το οποίο ενισχύει την άποψη ότι λόγω της ύπαρξης Πολυσυγγραμμικότητας η μέθοδος ελαχίστων τετραγώνων μας δίνει εκτιμήσεις πολύ μεγάλες κατά απόλυτη τιμή. Από τα παραπάνω, γίνεται εύκολα κατανοητό ότι όσο πιο ισχυρή είναι η Πολυσυγγραμμικότητα, τόσο πιο επισφαλείς και αναξιόπιστες είναι οι εκτιμήσεις που λαμβάνουμε μέσω του μοντέλου.

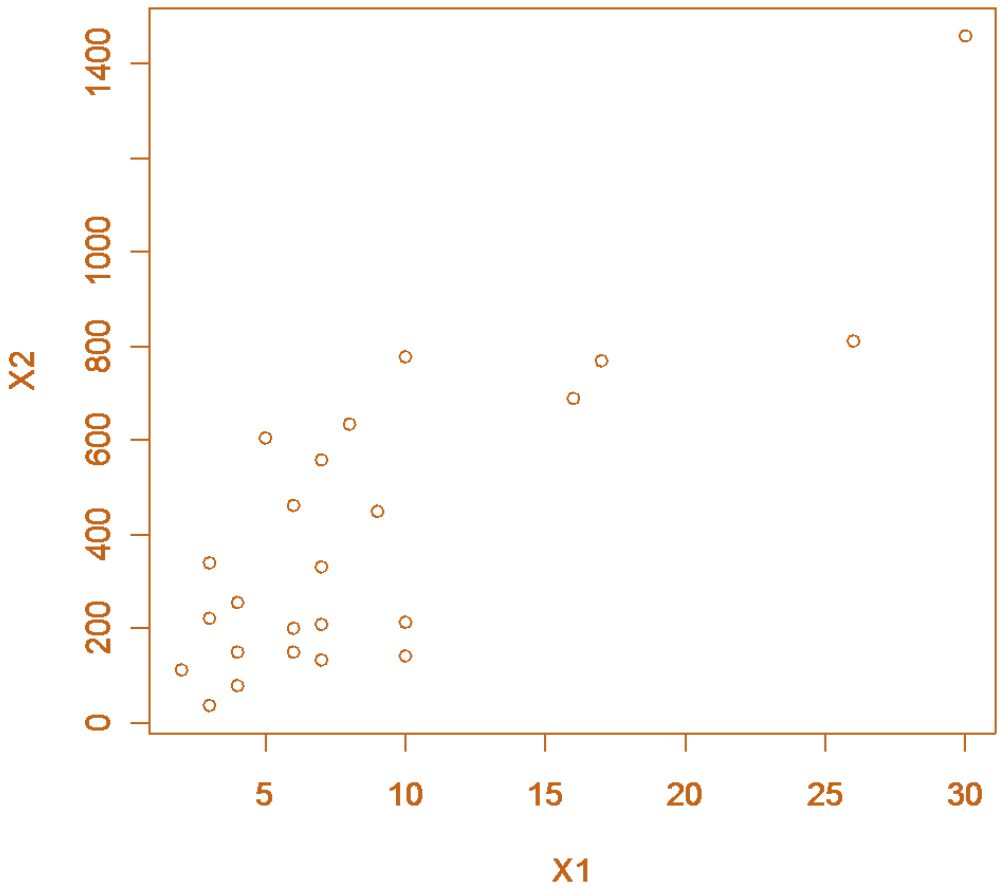
2.2 ΠΕΡΙΠΤΩΣΕΙΣ ΕΜΦΑΝΙΣΗΣ ΤΗΣ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ

Στο σημείο αυτό, θα παραθέσουμε τις περιπτώσεις στις οποίες εμφανίζεται το φαινόμενο της Πολυσυγγραμμικότητας. Σύμφωνα με τους Montgomery, Peck και Vining (2006) μπορούμε να οδηγηθούμε στο φαινόμενο της Πολυσυγγραμμικότητας ως απόρροια των παρακάτω αιτιών :

1) **Η μέθοδος συλλογής των δεδομένων.**

Ειδικότερα, όταν λαμβάνουμε υπόψη μας μερικές και όχι όλες τις επεξηγηματικές μεταβλητές ενός προβλήματος. Στο σημείο αυτό για την καλύτερη κατανόηση του παραπάνω θα εξετάσουμε τα δεδομένα του προβλήματος "χρόνος παράδοσης"[1]. Χρησιμοποιώντας το στατιστικό πακέτο της R, θα δημιουργήσουμε γραφική παράσταση μεταξύ της X_1 μεταβλητής (αριθμός κουτιών) και της X_2 μεταβλητής (απόσταση).

[1] Βλ. Παράρτημα 1, Πίνακας 1, Δεδομένα από Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 314, New Jersey : John Wiley & Sons, Inc.



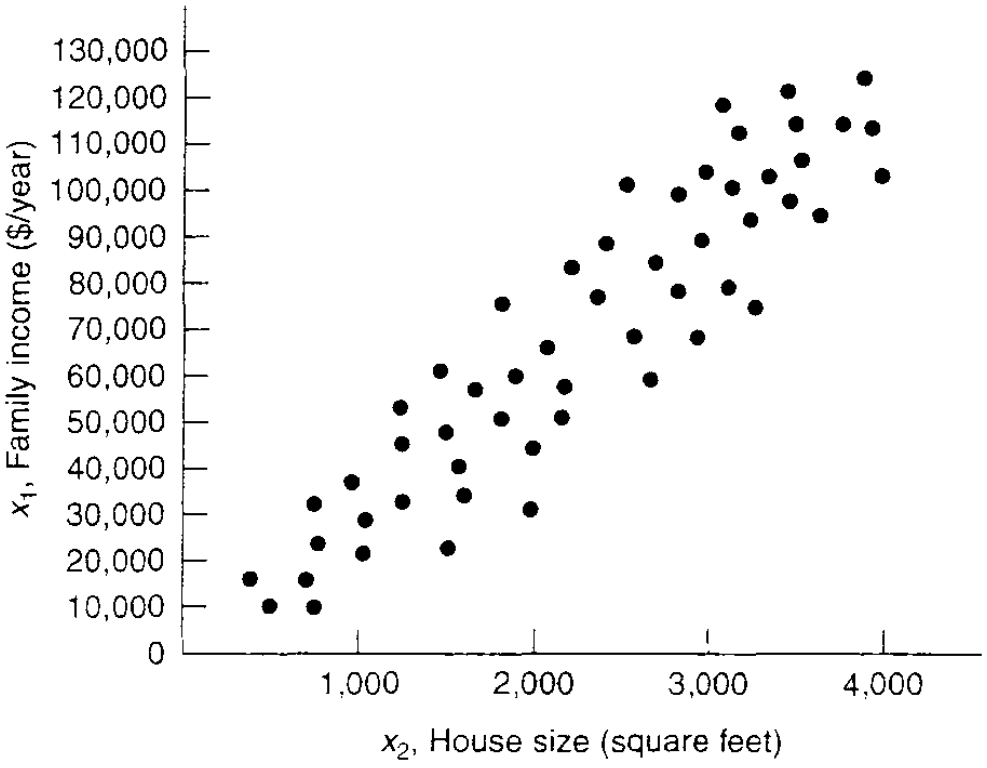
Διάγραμμα 1. Πολυσυγγραμικότητα λόγω συσχετισμένων μεταβλητών

Στην ανωτέρω γραφική παράσταση γίνεται εύκολα κατανοητό ότι τα ζεύγη των παρατηρήσεων βρίσκονται πάνω σε μια προσεγγιστικά ευθεία γραμμή. Μάλιστα ο συντελεστής συσχέτισης των δυο μεταβλητών ισούται με $r_{X1,X2} = 0.824215$ που επιβεβαιώνει τη γραμμική συσχέτιση μεταξύ τους. Απ' τα δεδομένα που χρησιμοποιήσαμε φαίνεται ότι σε παρατηρήσεις με μικρό αριθμό κουτιών αντιστοιχούν μικρές αποστάσεις, ενώ σε παρατηρήσεις με μεγάλο αριθμό κουτιών αντιστοιχούν μεγάλες αποστάσεις. Προκύπτει δηλαδή ότι οι μεταβλητές μας, X_1 και X_2 είναι θετικά συσχετισμένες. Αν η συσχέτιση μεταξύ τους είναι ισχυρή εμφανίζεται το φαινόμενο της Πολυσυγγραμμικότητας.

Αξίζει να τονιστεί, ότι στο πρόβλημα "χρόνου παράδοσης" θα μπορούσαμε να είχαμε συλλέξει δεδομένα που σε μικρό αριθμό κουτιών θα αντιστοιχούσαν μεγάλες αποστάσεις. Δεν υπάρχει κάτι στη φυσική δομή του προβλήματος που να αποτρέπει κάτι τέτοιο. Αυτό σημαίνει ότι το συγκεκριμένο δείγμα που έχουμε συλλέξει είναι η αιτία της εμφάνισης της Πολυσυγγραμμικότητας. Προφανώς λοιπόν, η μέθοδος συλλογής των δεδομένων διαδραματίζει σημαντικό ρόλο.

2) **Οι περιορισμοί των δεδομένων.**

Πολλές φορές τα δεδομένα που αντλούμε απ' τον προς μελέτη πληθυσμό μας, είναι ικανά να οδηγήσουν στην εμφάνιση γραμμικής σχέσης μεταξύ των μεταβλητών μας. Θεωρούμε το παράδειγμα (Montgomery και Peck), όπου υπολογίζεται το κατά πόσον η κατανάλωση ρεύματος σε ένα σπίτι εξαρτάται από το εισόδημα της οικογένειας (μεταβλητή x_1) καθώς και από το μέγεθος του σπιτιού (μεταβλητή x_2). Το διάγραμμα των δυο μεταβλητών είναι το ακόλουθο :



Διάγραμμα 2. Γραφική παράσταση αναλογίας οικογενειακού εισοδήματος/μέγεθος σπιτιού[2]

[2] Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 325, New Jersey : John Wiley & Sons, Inc.

Από το παραπάνω διάγραμμα παρατηρούμε ότι τα ζεύγη των δεδομένων ακολουθούν προσεγγιστικά μια ευθεία γραμμή, γεγονός που καταδεικνύει την ύπαρξη Πολυσυγγραμμικότητας. Είναι άλλωστε αναμενόμενο καθώς οικογένειες με υψηλότερα εισοδήματα θα έχουν και μεγαλύτερου μεγέθους σπίτια, συνεπώς και μεγαλύτερη κατανάλωση σε ρεύμα. Το παράδειγμα αυτό ενισχύει την άποψη ότι πολλές φορές οι **περιορισμοί** των δεδομένων μας στον πληθυσμό που εξετάζουμε οδηγούν αναπόφευκτα στο φαινόμενο της Πολυσυγγραμμικότητας. Στην περίπτωση αυτή η μέθοδος συλλογής των δεδομένων μας δεν παίζει κανένα απολύτως ρόλο και η Πολυσυγγραμμικότητα δεν δύναται να εξαλειφθεί. Συγκεκριμένου τύπου περιορισμοί εμφανίζονται συνήθως σε προβλήματα χημικής φύσεως και παραγωγής.

3) **η ύπαρξη ενός υπερορισμένου μοντέλου (overfitted)**.

Πρόκειται για το μοντέλο εκείνο, στο οποίο ο αριθμός των ανεξαρτήτων μεταβλητών είναι κατά πολύ μεγαλύτερος απ' τον αριθμό των παρατηρήσεων που διαθέτουμε. Εμφανίζεται κυρίως σε επιστημονικούς τομείς της βιολογίας και της χημείας και για την αντιμετώπιση του συνίσταται η αφαίρεση ενός αριθμού μεταβλητών.

4) **η επιλογή του μοντέλου που θα χρησιμοποιήσουμε**

Μπορεί να μας οδηγήσει στο φαινόμενο της Πολυσυγγραμμικότητας. Η προσθήκη πολυωνυμικών όρων στο μοντέλο μας ενισχύει το παραπάνω. Έτσι εάν το εύρος μιας μεταβλητής, έστω x, είναι μικρό, τότε η προσθήκη ενός όρου, πχ x^2 επηρεάζει σημαντικά το τρέχων μοντέλο μας και οδηγεί στην Πολυσυγγραμμικότητα. Για να αντιμετωπιστεί το παραπάνω, συνήθως συνίσταται η επιλογή ενός κατάλληλου υποσυνόλου των επεξηγηματικών μεταβλητών (regressors)

2.3 ΔΙΑΓΝΩΣΗ ΠΟΛΥΣΥΓΡΓΑΜΜΙΚΟΤΗΤΑΣ

Στο υποκεφάλαιο αυτό χρήσιμο κρίνεται να αναφερθούμε στις βασικές μεθόδους διάγνωσης της Πολυσυγγραμμικότητας.

- **Ένα πολύ απλό μέσο διάγνωσης του φαινομένου της Πολυσυγγραμμικότητας, όπως έχει ήδη αναφερθεί είναι μέσω του πίνακα συσχέτισης** $X^T X$.

Ο πίνακας αυτός έχει προκύψει μέσω της διαδικασίας του scaling. Έτσι τα μη-διαγώνια στοιχεία του εκφράζουν τους συντελεστές συσχέτισης των αντίστοιχων μεταβλητών. Εάν οι μεταβλητές μας, έστω X_i και X_j είναι γραμμικά εξαρτημένες τότε ο συντελεστής συσχέτισης τους $|r_{ij}| \to +1$. Για την καλύτερη περιγραφή του ανωτέρου, θα χρησιμοποιήσουμε τον πίνακα συσχέτισης $X^T X$ των δεδομένων του προβλήματος της Ακετυλήνης (Acetylene data).[3]

$$\begin{bmatrix} 1.000 & 0.224 & -0.958 & -0.132 & 0.443 & 0.205 & -0.271 & 0.031 & -0.577 \\ & 1.000 & -0.240 & 0.039 & 0.192 & -0.023 & -0.148 & 0.498 & -0.224 \\ & & 1.000 & 0.194 & 0.661 & -0.274 & 0.501 & -0.018 & 0.765 \\ & & & 1.000 & -0.265 & -0.975 & 0.246 & 0.398 & 0.274 \\ & & & & 1.000 & 0.323 & -0.972 & 0.126 & -0.972 \\ & & & & & 1.000 & -0.279 & -0.374 & 0.358 \\ & & & & & & 1.000 & -0.124 & 0.874 \\ & & & & & & & 1.000 & -0.158 \\ & & & & & & & & 1.000 \end{bmatrix}$$

Πίνακας 3 .Πίνακας συσχέτισης [4] $X^T X$

Παρατηρώντας τον παραπάνω πίνακα, διακρίνουμε την υψηλή συσχέτιση μεταξύ της X_1 και της X_3 μεταβλητής, καθώς $r_{13} = -0.958$. Δηλαδή οι μεταβλητές αυτές είναι αρνητικά ισχυρά συσχετισμένες. Η παραπάνω διαδικασία όμως είναι χρήσιμη μόνο για την εξέταση γραμμικών εξαρτήσεων μεταξύ ζευγών μεταβλητών. Δυστυχώς, όταν περισσότερες από δυο μεταβλητές είναι γραμμικά εξαρτημένες, δεν μας παρέχεται καμία διαβεβαίωση ότι οι συσχετίσεις ανά ζεύγη r_{ij} θα είναι

[3] Βλ. Παράρτημα 2, Πίνακας 1, Δεδομένα από Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 333, New Jersey : John Wiley & Sons, Inc.

[4] Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 333, New Jersey : John Wiley & Sons, Inc.

μεγάλες. Δηλαδή είναι πιθανόν παρ' όλο της ύπαρξης Πολυσυγγραμμικότητας κάτι τέτοιο να μην καθίσταται ευδιάκριτο.

- **Μια άλλη μέθοδος εξέτασης της ύπαρξης Πολυσυγγραμμικότητας είναι ο παράγοντας VIF (Variance Inflation Factor).**

Όπως έχει ήδη ειπωθεί, τα διαγώνια στοιχεία του πίνακα $C=(Z^{\mathrm{T}}Z)^{-1}$ είναι χρήσιμα για τη διάγνωση Πολυσυγγραμμικότητας. Πράγματι το j-στο διαγώνιο στοιχείο του παραπάνω πίνακα μπορεί να γραφτεί ως :

$$C_{jj}=(1-R_j^{\,2})^{-1} \qquad (2.15)$$

όπου $R_j^{\,2}$ είναι ο συντελεστής προσδιορισμού της Παλινδρόμησης μεταξύ της X_j και των υπολοίπων $p-1$ μεταβλητών. Εάν η X_j μεταβλητή είναι σχεδόν ορθογώνια με τις υπόλοιπες $p-1$ μεταβλητές, τότε ο $R_j^{\,2}$ είναι μικρός και το C_{jj} πλησιάζει τη μονάδα, ενώ όταν η X_j είναι εξαρτημένη με ένα υποσύνολο των υπολοίπων μεταβλητών, τότε το $R_j^{\,2}$ πλησιάζει τη μονάδα και αντίστοιχα το $VIF=C_{jj}=(1-R_j^{\,2})^{-1}$ είναι μεγάλο. Καθώς η διασπορά του j-στου συντελεστή Παλινδρόμησης ισούται με $C_{jj}\sigma^2$ μπορούμε να δούμε το C_{jj} ως τον παράγοντα μέσω του οποίου, λόγω της γραμμικής εξάρτησης μεταξύ των μεταβλητών, αυξάνεται η διασπορά του $\hat{b}_j$.

Καλούμε επομένως ως VIF την ποσότητα :

$$VIF=C_{jj}=(1-R_j^{\,2})^{-1} \qquad (2.16)$$

με την ορολογία να αποδίδεται στον Marquadt (1970). Ο VIF δηλαδή μετρά για κάθε όρο του μοντέλου τη συνδυασμένη επίδραση των εξαρτήσεων ανάμεσα στις μεταβλητές στη διασπορά του συγκεκριμένου όρου. Μία η περισσότερες ενδείξεις υψηλού VIF καταδεικνύουν την ύπαρξη Πολυσυγγραμμικότητας. Η πρακτική εμπειρία δείχνει ότι εάν κάποια απ' τις τιμές του VIF ξεπερνά το 5 ή το 10 αποτελεί ένδειξη αδυναμίας εκτίμησης των συντελεστών της Παλινδρόμησης εξαιτίας της παρουσίας Πολυσυγγραμμικότητας.

- **Η ανάλυση των ιδιοτιμών του πίνακα X^TX μπορεί να χρησιμοποιηθεί προκειμένου να εκτιμηθεί το μέγεθος της Πολυσυγγραμμικότητας στα δεδομένα μας.**

Εάν υπάρχει μια η περισσότερες γραμμικές εξαρτήσεις στα δεδομένα μας, τότε μια ή περισσότερες ιδιοτιμές θα είναι μικρές θα τείνουν δηλαδή στο μηδέν. Πολλοί αναλυτές προτιμούν να εξετάζουν τον αριθμό κατάστασης (condition number) του πίνακα $X^T X$ που ορίζεται ως :

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \qquad (2.17)$$

εάν ο παραπάνω αριθμός κατάστασης είναι μικρότερος του 100, τότε δεν παρίσταται πρόβλημα Πολυσυγγραμμικότητας. Εάν παίρνει τιμές μεταξύ 100 και 1000 υπάρχει σοβαρή Πολυσυγγραμμικότητα, ενώ εάν ξεπεράσει το 1000 τότε υπάρχει πολύ έντονο πρόβλημα Πολυσυγγραμμικότητας. Οι δείκτες κατάστασης (condition indices) για τον X^TX ορίζονται ως :

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j} \qquad j = 1, 2, \ldots, p \qquad (2.18)$$

Λαμβάνοντας πάλι υπόψην μας τα δεδομένα του προβλήματος της Ακετυλήνης, υπολογίζουμε τις ιδιοτιμές του πίνακα συσχέτισης $X^T X$. Αυτές είναι: $\lambda_1 = 4,2048$, $\lambda_2 = 2,1626$, $\lambda_3 = 1,1384$, $\lambda_4 = 1,0413$, $\lambda_5 = 0,3845$, $\lambda_6 = 0,0495$, $\lambda_7 = 0,0136$, $\lambda_8 = 0,0051$, $\lambda_9 = 0,0001$. Υπάρχουν τέσσερις ιδιοτιμές με μικρές τιμές, οπότε ελλοχεύει ο κίνδυνος ύπαρξης Πολυσυγγραμμικότητας. Ο αριθμός κατάστασης είναι :

$$\kappa = \frac{4,2048}{0,0001} = 42.028 >> 1000 \qquad (2.19)$$

που υποδεικνύει έντονο πρόβλημα Πολυσυγγραμμικότητας.

Οι δείκτες κατάστασης για όλες τις ιδιοτιμές είναι σύμφωνα με τον προηγούμενο τύπο :

$\lambda_1 = 4,2048$ $\lambda_2 = 2,1626$ $\lambda_3 = 1,1384$	$\lambda_4 = 1,0413$ $\lambda_5 = 0,3845$ $\lambda_6 = 0,0495$	$\lambda_7 = 0,0136$ $\lambda_8 = 0,0051$ $\lambda_9 = 0,0001$
$\kappa_1 = \frac{4,2048}{4,2048} = 1$	$\kappa_4 = \frac{4,2048}{1,0413} = 4,04$	$\kappa_7 = \frac{4,2048}{0,0136} = 309,18$
$\kappa_2 = \frac{4,2048}{2,1626} = 1,94$	$\kappa_5 = \frac{4,2048}{0,3845} = 10,94$	$\kappa_8 = \frac{4,2048}{0,0051} = 824,47$
$\kappa_3 = \frac{4,2048}{1,1384} = 3,69$	$\kappa_6 = \frac{4,2048}{0,0495} = 84,96$	$\kappa_9 = \frac{4,2048}{0,0001} = 42.048$

Πίνακας 4. Δείκτες κατάστασης ιδιοτιμών[5]

Βλέπουμε ότι ένας δείκτης συγκεκριμένα ο κ_9 είναι πολύ μεγαλύτερος του 1000, ενώ δυο άλλοι δείκτες, οι κ_7 και κ_8 ξεπερνούν αρκετά το 100. Συνεπώς καταλήγουμε στο συμπέρασμα ότι υπάρχει τουλάχιστον μια έντονη γραμμική εξάρτηση στα δεδομένα του προβλήματος μας.

2.4 ΜΕΘΟΔΟΙ ΑΝΤΙΜΕΤΩΠΙΣΗΣ ΤΗΣ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ

Πολλές τεχνικές έχουν προταθεί για την αντιμετώπιση των προβλημάτων λόγω Πολυσυγγραμμικότητας. Οι γενικότερες προσεγγίσεις συνιστούν την συλλογή περαιτέρω δεδομένων, τον επαναπροσδιορισμό του μοντέλου καθώς και την εκτίμηση διαφορετικών εκτιμητικών μεθόδων πέρα των ελαχίστων τετραγώνων. Τέτοιες είναι η Παλινδρόμηση Κορυφογραμμής (ridge regression), η Παλινδρόμηση Κυρίων Συνιστωσών (principal component regression ή αλλιώς PCR) καθώς και η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων (partial least square regression ή αλλιώς PLSR). Οι δυο τελευταίες θα παρουσιαστούν εκτενώς στα επόμενα κεφάλαια, ενώ παρακάτω θα κάνουμε μια μικρή αναφορά στην πρώτη.

[5] Δείκτες κατάστασης ιδιοτιμών από Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 336, New Jersey : John Wiley & Sons, Inc.

- **Συλλογή επιπλέον δεδομένων**

Η συλλογή επιπλέον δεδομένων έχει προταθεί ως η καλύτερη μέθοδος αντιμετώπισης της Πολυσυγγραμμικότητας σύμφωνα με τους Farrar και Glauber (1967). Τα επιπλέον δεδομένα πρέπει να συλλεχθούν με τέτοιο τρόπο ώστε να αποκλειστεί η Πολυσυγγραμμικότητα απ' τα δεδομένα μας. Για παράδειγμα, ας συλλογιστούμε το πρόβλημα "χρόνος παράδοσης" που είχαμε προεπεξεργαστεί. Είχαμε καταλήξει στο συμπέρασμα ότι οι μεταβλητές X_1 (αριθμός κουτιών) και X_2 (απόσταση) σχετίζονται μεταξύ τους, κάτι που φάνηκε και στη γραφική παράσταση.

Αυτό θα μπορούσε να είχε αποφευχθεί εάν είχαμε συλλέξει επιπλέον δεδομένα όπου σε μικρό αριθμό κουτιών θα αντιστοιχούσε μεγάλη απόσταση καθώς και σε μεγάλο αριθμό κουτιών θα αντιστοιχούσε μικρή απόσταση. Προσθέτοντας τα επιπλέον δεδομένα, οι δυο μεταβλητές θα σταματούσαν να συνδέονται γραμμικά μεταξύ τους και η Πολυσυγγραμμικότητα θα εξαφανιζόταν.

Δυστυχώς όμως η συλλογή επιπλέον δεδομένων δεν είναι πάντα εφικτή εξαιτίας οικονομικών περιορισμών αλλά και λόγω του ότι η διαδικασία που μελετούσαμε δεν είναι πλέον διαθέσιμη. Συχνά παρατηρείται σε χημικές διαδικασίες όπου μετά από σύντομο χρονικό διάστημα η διαδικασία λαμβάνει τέλος. Ακόμη και αν επιπλέον δεδομένα είναι διαθέσιμα, μπορεί να είναι ακατάλληλη η εισαγωγή τους καθώς μπορεί μια ή περισσότερες νέες παρατηρήσεις που θα εισάγουμε να ξεπερνούν το εύρος των συντελεστών Παλινδρόμησης, να πρόκειται δηλαδή για σημεία επίδρασης. Η παρουσία των παραπάνω μπορεί να αποβεί ακατάλληλη για το προσαρμοσμένο μοντέλο μας.

Τελικά η συλλογή επιπρόσθετων δεδομένων μπορεί να αποβεί άχρηστη όσον αφορά την προσπάθεια που κάνουμε για να εξαλείψουμε την Πολυσυγγραμμικότητα, σε περιπτώσεις όπου τα δεδομένα μας υπόκεινται σε κάποιο φυσικό περιορισμό. Χαρακτηριστικό τέτοιο παράδειγμα είναι η σχέση μεταξύ των μεταβλητών οικογενειακό εισόδημα και μέγεθος σπιτιού που αναφέραμε προηγουμένως, καθώς είναι αναμενόμενο άτομα με μεγαλύτερο οικογενειακό εισόδημα να έχουν και μεγαλύτερο σπίτι. Η συλλογή επιπλέον δεδομένων στην περίπτωση αυτή θα μας ήταν ουσιαστικά ανούσια αφού θα παρέμενε αυτούσια η γραμμική εξάρτηση μεταξύ τους (Montgomery et al., 2006, pp. 341-344).

- **Επαναπροσδιορισμός του μοντέλου**

Η Πολυσυγγραμμικότητα συχνά προκαλείται με την επιλογή του μοντέλου, όταν δυο υψηλά συσχετισμένες μεταβλητές υπεισέρχονται στην εξίσωση Παλινδρόμησης. Σε αυτή την περίπτωση ένας επαναπροσδιορισμός της εξίσωσης Παλινδρόμησης θα ελάττωνε την επίδραση της Πολυσυγγραμμικότητας. Μια προσέγγιση για το παραπάνω είναι να ξαναορίσουμε τις μεταβλητές. Έστω για παράδειγμα, ότι στο μοντέλο μας έχουμε τρεις μεταβλητές X_1, X_2, X_3 που είναι γραμμικά εξαρτημένες. Τότε μπορούμε να βρούμε μια συνάρτηση όπως :

$$X = X_1 X_2 X_3 \text{ ή } X = \frac{(X_1 + X_2)}{X_3} \qquad (2.20)$$

η οποία να συντηρεί τις πληροφορίες των αρχικών μεταβλητών αλλά και ταυτόχρονα να μειώνει την Πολυσυγγραμμικότητα. Μια άλλη προσέγγιση επαναπροσδιορισμού του μοντέλου μας είναι η διαγραφή κάποιων μεταβλητών.

Έτσι αν X_1, X_2, X_3 είναι γραμμικά εξαρτημένες διαγράφοντας μια απ' αυτές έστω την τελευταία, μπορεί να συμβάλλει στην αντιμετώπιση της Πολυσυγγραμμικότητας. Παρ' όλα αυτά, μπορεί να μη μας δώσει μια ικανοποιητική λύση καθώς είναι πιθανό η μεταβλητή που διαγράψαμε να είχε σημαντική ερμηνευτική ισχύ για το μοντέλο μας.

Άρα η εξάλειψη μεταβλητών από το μοντέλο μπορεί να καταστρέψει την ικανότητα πρόβλεψης του μοντέλου μας, συνεπώς να καταλήξουμε σε αβάσιμα και λανθασμένα συμπεράσματα.

2.5 Η ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΟΡΥΦΟΓΡΑΜΜΗΣ (RIDGE REGRESSION)

Σε περιπτώσεις που παρουσιάζεται Πολυσυγγραμμικότητα, λύση καλείται να δώσει η Παλινδρόμηση Κορυφογραμμής. Καλείται έτσι, καθώς τα μαθηματικά που χρησιμοποιούνται σχετίζονται με τη μέθοδο της Ridge–ανάλυσης η οποία είχε προηγουμένως χρησιμοποιηθεί απ' τον Hoerl για την περιγραφή της συμπεριφοράς των δευτεροβάθμιων επιφανειών (Montgomery et al., 2006, p. 345).

Όταν τα δεδομένα μας χαρακτηρίζονται από Πολυσυγγραμμικότητα, η κλασική μέθοδος των ελαχίστων τετραγώνων αποτυγχάνει να εκτιμήσει σε ικανοποιητικό βαθμό τους συντελεστές Παλινδρόμησης των μεταβλητών μας, με αποτέλεσμα οι εκτιμήσεις να μην είναι αξιόπιστες. Οφείλεται στο γεγονός ότι στη μέθοδο ελαχίστων τετραγώνων, πρέπει η $\hat{\beta}$ να είναι αμερόληπτη εκτιμήτρια του β εκτιμήτριας των συντελεστών Παλινδρόμησης, η οποία όμως δεν θα είναι αμερόληπτη αλλά μεροληπτική. Θα ισχύει δηλαδή :

$$\mathrm{E}(\hat{\beta}) \neq \beta \qquad (2.21)$$

Ακολούθως παρατίθεται μια συνοπτική περιγραφή του θεωρήματος Gauss – Markov το οποίο ισχυρίζεται ότι η εκτιμήτρια ελαχίστων τετραγώνων για τις παραμέτρους β του γενικού γραμμικού μοντέλου :

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \tag{2.22}$$

έχει τη μικρότερη διασπορά απ' όλες τις δυνατές γραμμικές αμερόληπτες εκτιμήτριες. Παρ' όλα αυτά, αυτό δεν μας διασφαλίζει ότι η διασπορά θα είναι και μικρή.

2.5.1 Θεώρημα Gauss – Markov

Έστω ότι έχουμε την εκτίμηση του γραμμικού συνδυασμού των παραμέτρων $\mu = \mathrm{a}^{\mathrm{T}}\mathrm{b}$. Αρχικά, θα δείξουμε ότι η εκτιμήτρια ελαχίστων τετραγώνων του παραπάνω γραμμικού συνδυασμού είναι αμερόληπτη (Montgomery et al., 2006, pp. 558-560). Πράγματι :

$$\hat{\mu} = \mathrm{a}^{\mathrm{T}}\hat{b} = \mathrm{a}^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \tag{2.23}$$

έχουμε τώρα ότι :

$$\mathrm{E}(\hat{\mu}) = E(\mathrm{a}^{\mathrm{T}} b) = E[\mathrm{a}^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}] = \mathrm{a}^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}b = \mathrm{a}^{\mathrm{T}}b \tag{2.24}$$

αφού

$$(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X} = I \tag{2.25}$$

Επομένως, η εκτιμήτρια $\hat{\mu} = \mathrm{a}^{\mathrm{T}} b$ είναι αμερόληπτη.

Σύμφωνα με το θεώρημα Gauss–Markov, για κάθε άλλη γραμμική εκτιμήτρια, έστω $\hat{\mu}^* = \lambda^{\mathrm{T}}\mathrm{c}$, που είναι αμερόληπτη για την $\mathrm{a}^{\mathrm{T}} b$, δηλαδή ισχύει :

$$\mathrm{E}(\lambda^{\mathrm{T}}\mathrm{c}) = \mathrm{a}^{\mathrm{T}} b \tag{2.26}$$

θα ισχύει το εξής :

$$Var(\mathrm{a}^{\mathrm{T}}\, b) \leq Var(\lambda^{\mathrm{T}}\mathrm{c}) \qquad (2.27)$$

που επαληθεύει ότι η διασπορά της εκτιμήτριας των ελαχίστων τετραγώνων είναι η μικρότερη δυνατή.

2.5.2 Παρουσίαση του < μηχανισμού > της Ridge Regression

Έστω, ότι έχουμε το μοντέλο Παλινδρόμησης της μορφής :

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ \ldots\ + \beta_{\mathrm{p}} \mathrm{X}_{\mathrm{p}} + \ \varepsilon \qquad (2.28)$$

όπου p το πλήθος των επεξηγηματικών μεταβλητών. Η εκτιμήτρια Κορυφογραμμής (Ridge estimator) η οποία παρουσιάστηκε απ' τους Hoerl και Kennard (1970), είναι η εξής :

$$\hat{\beta}_{\mathrm{R}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \rho \mathrm{I})^{-1} \mathrm{X}^{\mathrm{T}}\mathrm{Y} \qquad (2.29)$$

με $\hat{\beta}_{\mathrm{R}} = (\hat{\beta}_{\mathrm{R1,}}\, \hat{\beta}_{\mathrm{R2}}, \ldots., \hat{\beta}_{\mathrm{Rp}})$ να είναι το διάνυσμα των εκτιμητριών.

Ο αριθμός ρ, όπου $\rho \geq 0$ είναι μια σταθερά η οποία επιλέγεται κάθε φορά και καλείται παράμετρος μεροληψίας (biasing parameter). Όταν $\rho = 0$, τότε η ridge–εκτιμήτρια ταυτίζεται με τη γνωστή εκτιμήτρια των ελαχίστων τετραγώνων. Η εκτιμήτρια Κορυφογραμμής αποδεικνύεται ότι είναι ένας γραμμικός συνδυασμός της εκτιμήτριας ελαχίστων τετραγώνων αφού:

$$\hat{\beta}_{\mathrm{R}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \rho \mathrm{I})^{-1} \mathrm{X}^{\mathrm{T}}\mathrm{Y} = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \rho \mathrm{I})^{-1} (\mathrm{X}^{T} X)\beta = \mathrm{Z}_{\mathrm{K}}\, \beta \qquad (2.30)$$

Οφείλεται στο γεγονός ότι :

$$\mathrm{E}(\hat{\beta}_{\mathrm{R}}) = \mathrm{E}(\mathrm{Z}_{\mathrm{K}}\,\beta) = \mathrm{Z}_{\mathrm{K}}\,\beta \qquad (2.31)$$

με $\hat{\beta}_{\mathrm{R}}$ να είναι μεροληπτική εκτιμήτρια του $\hat{\beta}$. Ο πίνακας συνδιασποράς της $\hat{\beta}_{\mathrm{R}}$ είναι ο ακόλουθος :

$$Var(\hat{\beta}_{\mathrm{R}}) = \sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X}+\rho\mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}X(\mathrm{X}^{\mathrm{T}}X+\rho I)^{-1} \qquad (2.32)$$

με το μέσο τετραγωνικό σφάλμα της ridge–εκτιμήτριας να είναι :

$$MSE(\hat{\beta}_{\mathrm{R}}) = Var(\hat{\beta}_{\mathrm{R}}) + [bias(\hat{\beta}_{\mathrm{R}})]^2 = \sigma^2 Tr[(\mathrm{X}^{\mathrm{T}}\mathrm{X}+\rho\mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}X(\mathrm{X}^{\mathrm{T}}X+\rho I)^{-1} + \rho^2\beta^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}X+\rho I)^{-1}\beta$$
$$= \sigma^2\sum_{j=1}^{p}\frac{\lambda_{\mathrm{j}}}{(\lambda_{\mathrm{j}}+k)^2} + \rho^2\beta^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}X+\rho I)^{-2}\beta \qquad (2.33)$$

με $\lambda_1, \lambda_2, \ldots, \lambda_{\mathrm{p}}$ να είναι οι ιδιοτιμές του πίνακα $X^T X$.

Σύμφωνα με την παραπάνω ισότητα, αύξηση του k προκαλεί αύξηση του δεύτερου όρου της μεροληψίας, ενώ η ταυτόχρονη αύξηση του ρ, προκαλεί μείωση του πρώτου όρου, αυτού της διασποράς. Σκοπός της Ridge regression είναι η επιλογή μιας τέτοιας τιμής του ρ ώστε η μείωση στον όρο της διασποράς να είναι μεγαλύτερη απ' την αύξηση στον όρο που εκφράζει τη μεροληψία.

Αν επιτύχουμε κάτι τέτοιο, το μέσο τετραγωνικό σφάλμα της ridge εκτιμήτριας θα είναι μικρότερο απ' τη διασπορά της εκτιμήτριας των ελαχίστων τετραγώνων. Πράγματι αυτό απεδείχθη από τους Hoerl και Kennard (1970). Συγκεκριμένα, απέδειξαν την ύπαρξη μιας μη μηδενικής τιμής για το ρ για την οποία ισχύει :

$$\mathrm{MSE}(\hat{\beta}_{\mathrm{R}}) < \mathrm{Var}(\hat{\beta}) \qquad (2.34)$$

με την προϋπόθεση ότι το $\beta^{\mathrm{T}}\beta$ είναι φραγμένο. Σχετικά με το άθροισμα τετραγώνων των υπολοίπων έχουμε :

$$\mathrm{SSE} = (\mathrm{Y} - \mathrm{X}\hat{\beta}_{\mathrm{R}})^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\hat{\beta}_{\mathrm{R}})$$

$$= (\mathrm{Y} - \mathrm{X}\hat{\beta}\ )^{\mathrm{T}}(\mathrm{Y} - \mathrm{X}\hat{\beta}\ ) + (\hat{\beta}_{\mathrm{R}} - \hat{\beta})^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\ (\hat{\beta}_{\mathrm{R}} - \hat{\beta}) \qquad (2.35)$$

από την παραπάνω ισότητα βλέπουμε ότι αύξηση του ρ προκαλεί και αύξηση του SSE.

Όμως γνωρίζουμε ότι το συνολικό άθροισμα τετραγώνων SST διατηρείται σταθερό. Άρα προκύπτει ότι το άθροισμα τετραγώνων λόγω Παλινδρόμησης θα μειώνεται. Συνεπώς, όσο το ρ αυξάνεται τόσο θα μειώνεται και ο R^2 (συντελεστής προσδιορισμού), κάτι που μεταφράζεται στο ότι απ' την εκτιμήτρια Κορυφογραμμής μπορεί να μη πάρουμε την καλύτερη προσαρμογή για τα δεδομένα μας, αλλά θα πάρουμε σίγουρα ένα σταθερό σύνολο εκτιμήσεων για τις παραμέτρους μας.

ΚΕΦΑΛΑΙΟ 3. PCA (ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ)

Πολλές φορές, σε περιπτώσεις όπου έχουμε δεδομένα με πολλές μεταβλητές, η κλασική μέθοδος Παλινδρόμησης αποτυγχάνει να τα ερμηνεύσει ορθά. Αυτό συμβαίνει όταν το πλήθος των μεταβλητών p είναι μεγάλο. Ένα άλλο συχνό φαινόμενο που μπορεί να παρουσιαστεί είναι οι ανεξάρτητες μεταβλητές να συσχετίζονται μεταξύ τους (Abdi & Williams, 2010).

Λύση στο συγκεκριμένο πρόβλημα καλείται να δώσει η Ανάλυση Κυρίων Συνιστωσών. Πρόκειται για μια πολυμεταβλητή στατιστική τεχνική, η οποία χρησιμοποιείται σε διάφορα επιστημονικά προβλήματα, όπως στη χημειομετρία, στις εφαρμογές αναγνώρισης προτύπων καθώς και στη φασματοσκοπία και σε ποικίλες βιομηχανικές δραστηριότητες.

Σκοπός της είναι να εξάγει τις σημαντικές πληροφορίες από ένα πλαίσιο δεδομένων και να τις εκφράσει μέσω ενός πλήθους νέων ορθογώνιων μεταβλητών που ονομάζονται κύριες συνιστώσες. Στηρίζεται σε αναδιάταξη των αρχικών μας δεδομένων πραγματοποιώντας ένα μαθηματικό μετασχηματισμό στην αρχική μας μήτρα.

Χρησιμοποιείται τόσο σε περιπτώσεις όπου το σύνολο δεδομένων μας αποτελείται από λίγες το πλήθος μεταβλητές, όσο και για την εξέταση και ερμηνεία συνόλου δεδομένων πολλών μεταβλητών, όπως συμβαίνει στα φασματοσκοπικά δεδομένα.

Η PCA είναι ενδεχομένως η δημοφιλέστερη πολυμεταβλητή στατιστική τεχνική. Η προέλευση της ανάγεται στον Pearson ή στον Cauchy και επίσης στους Cayley, Silvester, Hamilton. Η τωρινή της μορφή όμως οφείλεται στον Hotelling ο οποίος και επινόησε τον όρο principal components (Abdi & Williams, 2010).

Έστω ότι το πρόβλημα μας περιγράφεται από p μεταβλητές. Επομένως για να ερμηνευθεί η ολική μεταβλητότητα του μοντέλου απαιτούνται όλες οι μεταβλητές. Μέσω της PCA, απώτερος στόχος αποτελεί η ερμηνεία σχεδόν όλης αυτής της μεταβλητότητας από έναν σχετικά μικρό αριθμό k ορθογωνίων μεταβλητών που ονομάζονται κύριες συνιστώσες (principal components). Έτσι η πληροφορία που περιέχεται στις αρχικές p μεταβλητές, περιέχεται σχεδόν όλη στις k νέες συνιστώσες. Τότε οι k συνιστώσες μπορούν να αντικαταστήσουν τις αρχικές p μεταβλητές.

Με άλλα λόγια, το αρχικό σύνολο δεδομένων, αποτελούμενο από n το πλήθος παρατηρήσεις-μετρήσεις των p μεταβλητών, συρρικνώνεται σε ένα νέο σύνολο δεδομένων από n το πλήθος μετρήσεων για κάθε απ' τις k συνιστώσες. Καταλήγουμε επομένως σε ένα σύνολο k ασυσχέτιστων συνιστωσών.

Τα πλεονεκτήματα της μεθόδου αυτής συνίστανται στο ότι τα δεδομένα μας μελετώνται πλέον στον IR^k και όχι στον IR^p ($p > k$), καθώς και αν η διάσταση είναι μικρή πχ. $\mathrm{k} = 2$ ή $\mathrm{k} = 3$ τότε μπορούμε να παραστήσουμε τα δεδομένα μας γραφικά εξάγοντας χρήσιμα για τη μελέτη μας συμπεράσματα.

Ακολούθως εφόσον έχει εφαρμοστεί η Ανάλυση Κυρίων Συνιστωσών, το επόμενο βήμα είναι χρησιμοποιώντας το πλήθος των επεξηγηματικών μεταβλητών του δεδομένου προβλήματός να εφαρμόσουμε την Παλινδρόμηση Κυρίων

Συνιστωσών (PCR), πρακτική εφαρμογή της οποίας παρατίθεται στο 5^{ο} Κεφάλαιο της παρούσης.

- **ΣΤΟΧΟΙ ΤΗΣ PCA** (Abdi & Williams, 2010)

1) Η **εξαγωγή** των σημαντικότερων πληροφοριών από τον πίνακα δεδομένων.
2) Η "**συρρίκνωση** " του μεγέθους του συνόλου δεδομένων.
3) Η **απλοποίηση** της περιγραφής του συνόλου δεδομένων.
4) Η **ανάλυση** της δομής των παρατηρήσεων και των μεταβλητών.

Όπως προαναφέρθηκε, μέσω της PCA, οι p το πλήθος μεταβλητές αντικαθίστανται από k το πλήθος νέες μεταβλητές που καλούνται κύριες συνιστώσες. Οι τελευταίες αποτελούν γραμμικούς συνδυασμούς των p τυχαίων μεταβλητών έστω $X_1, X_2, \ldots, X_p$.

Γραφικά, μπορούμε να περιστρέψουμε το αρχικό σύστημα με $X_1, X_2, \ldots, X_p$ άξονες συντεταγμένων, σε ένα νέο σύστημα συντεταγμένων που προκύπτει απ' το προηγούμενο και παριστάνεται απ' τις k νέες συνιστώσες που υπολογίστηκαν. Η φυσική ερμηνεία αυτής της διαδικασίας είναι ότι οι καινούριοι άξονες εκφράζουν τις διευθύνσεις με την υψηλότερη μεταβλητότητα των δεδομένων μας.

Σημειώνεται, ότι οι νέες k κύριες συνιστώσες (principal components) εξαρτώνται μόνο απ' τον πίνακα συνδιασπορών ή τον πίνακα συσχετίσεων P των $X_1, X_2, \ldots, X_p$

3.1 ΔΙΑΔΙΚΑΣΙΑ ΠΟΥ ΑΚΟΛΟΥΘΕΙΤΑΙ ΣΤΗΝ PCA

Κατά την εφαρμογή της μεθόδου ακολουθούμε συγκεκριμένα βήματα. Αυτά περιγράφονται παρακάτω :

1) Υπολογισμός για το σύνολο δεδομένων της μήτρας συνδιασποράς (covariance matrix).

2) Εξαγωγή των ιδιοδιανυσμάτων (eigenvectors) και των αντίστοιχων τους ιδιοτιμών (eigenvalues) από τη μήτρα συνδιασποράς, κάνοντας χρήση μεθόδων ιδιοανάλυσης που θα περιγραφούν παρακάτω.

3) Εκτίμηση της σημαντικότητας των ιδιοδιανυσμάτων με βάση της αντίστοιχης ιδιοτιμής τους.

4) Δημιουργία της μήτρας φορτίων (loading matrix) η οποία περιέχει τα σημαντικότερα από τα παραπάνω διανύσματα.

5) Δημιουργία της μήτρας αποτελεσμάτων (scores matrix) η οποία περιέχει τις προβολές του αρχικού συνόλου δεδομένων μας πάνω στο νέο χώρο διανυσμάτων που έχουμε δημιουργήσει.

Ακολούθως θα προβούμε σε μια μαθηματική περιγραφή-ανάλυση των παραπάνω βημάτων. Έστω ότι το σύνολο δεδομένων μας περιγράφεται από p το πλήθος μεταβλητές με n μετρήσεις. Συνεπώς, ο πίνακας X που περιέχει τις μετρήσεις μας είναι διάστασης $n \times p$:

$$\mathrm{X}_{\mathrm{n}\times\mathrm{p}} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{3.1}$$

βασική λειτουργία που επιτελεί η μέθοδος που εξετάζουμε είναι να ποσοτικοποιεί την αλληλεπίδραση κάθε ζεύγους μεταβλητών. Αυτό επιτυγχάνεται με βάση τη μήτρα συνδιασποράς που αναφέρεται στο σύνολο δεδομένων μας. Πράγματι έστω, ότι έχουμε δυο μεταβλητές έστω X και Y με μέση τιμή αντίστοιχα για κάθε μεταβλητή :

$$E(X) = a \quad \text{και} \quad E(Y) = b \tag{3.2}$$

επομένως, η ζητούμενη μήτρα συνδιασποράς (covariance matrix) δίνεται από την παρακάτω σχέση :

$$\mathrm{cov}(X,Y) = E((X-a)(Y-b)) = \mathrm{E}(\mathrm{X}\cdot\mathrm{Y}) - a\cdot b \tag{3.3}$$

Επόμενο βήμα είναι η εξαγωγή των ιδιοδιανυσμάτων (eigenvectors) και των αντίστοιχων τους ιδιοτιμών (eigenvalues) από τη μήτρα συνδιασποράς που υπολογίσαμε παραπάνω. Κάνοντας χρήση μεθόδων ιδιοανάλυσης όπως π.χ είναι η SVD (Singular Value Decomposition), υπολογίζουμε τα p το πλήθος ιδιοδιανύσματα και τις p το πλήθος αντίστοιχες ιδιοτιμές τους.

Ακολούθως κάνοντας χρήση συγκεκριμένων κριτηρίων αποφασίζουμε ποια από αυτά τα ιδιοδιανύσματα, k το πλήθος, θα χρησιμοποιηθούν στον καταρτισμό της μήτρας φορτίων (loading matrix) που καλούμαστε να δημιουργήσουμε.

Η μήτρα φορτίων που δημιουργούμε έχει σαν στήλες τα παραπάνω ιδιοδιανύσματα που επιλέξαμε να χρησιμοποιήσουμε. Αυτά καλούνται πλέον κύριες συνιστώσες (principal components) ο ρόλος των οποίων έχει περιγραφεί. Άρα η μήτρα φορτίων έστω V , θα είναι ένας πίνακας διαστάσεων $p \times k$:

$$\mathrm{V}_{\mathrm{p\times k}} = \begin{pmatrix} v_{11} & \cdots & v_{1k} \\ \vdots & \ddots & \vdots \\ v_{p1} & \cdots & v_{pk} \end{pmatrix} \qquad (3.4)$$

Τελευταίο βήμα είναι η κατασκευή της μήτρας των αποτελεσμάτων (factor scores matrix), η οποία περιέχει τις προβολές του αρχικού συνόλου δεδομένων πάνω στο νέο χώρο διανυσμάτων που έχουμε δημιουργήσει. Πράγματι η μήτρα αποτελεσμάτων προκύπτει ως το γινόμενο του αρχικού πίνακα δεδομένων X με τη μήτρα φορτίων μας V , όπως φαίνεται παρακάτω :

$$\mathrm{U}_{n\times k} = \mathrm{X}_{\mathrm{n\times p}} \cdot \mathrm{V}_{\mathrm{p\times k}} \qquad (3.5)$$

όπου

$$\mathrm{U}_{n\times k} = \begin{pmatrix} u_{11} & \cdots & u_{1k} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nk} \end{pmatrix} \qquad (3.6)$$

3.2 ΓΕΩΜΕΤΡΙΚΗ ΕΡΜΗΝΕΙΑ ΤΗΣ ΜΕΘΟΔΟΥ PCA

Στο σημείο αυτό θα προσπαθήσουμε να δώσουμε μια ερμηνεία της Ανάλυσης Κυρίων Συνιστωσών με τη βοήθεια της γεωμετρίας, βασιζόμενοι στον Subhash Sharma (1996), καθορίζοντας εναλλακτικούς άξονες και σχηματίζοντας νέες μεταβλητές με τη βοήθεια πινάκων. Στον παρακάτω πίνακα παρουσιάζουμε ένα σύνολο δεδομένων αποτελούμενο από 12 παρατηρήσεις και 2 μεταβλητές τις οποίες και μετασχηματίζουμε προκειμένου να αποκτήσουν μηδενικό μέσο όρο.

Αριθμός παρατήρησης	x_1		x_2	
	Αρχική μεταβλητή	Μετασχηματισμένη μεταβλητή	Αρχική μεταβλητή	Μετασχηματισμένη μεταβλητή
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Μέσος όρος	8	0	3	0
Διασπορά	23.091	23.091	21.091	21.091

Πίνακας 5. Αρχικές και μετασχηματισμένες μεταβλητές.[6]

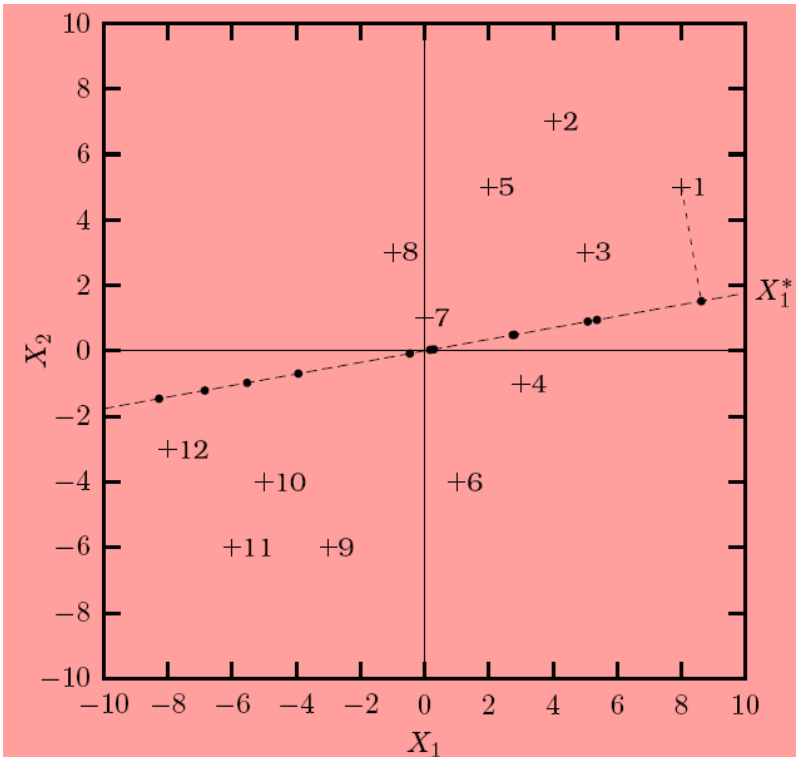
Από τον πίνακα βλέπουμε ότι η διασπορά της μεταβλητής x_1 είναι ίση με $23,091$ ενώ η διασπορά της x_2 είναι ίση με $21,091$ καθώς και η συνολική διασπορά ισούται με $(23,091+21,091)=44,182$. Το ποσοστό της συνολικής διασποράς που ερμηνεύει η x_1 ισούται με $\frac{23,091}{44,182}\cong 0,52$ ή περίπου το 52%. Αντίστοιχα η μεταβλητή x_2 ερμηνεύει το υπόλοιπο 48%. Οι πίνακες συνδιασποράς και συσχέτισης των δυο μεταβλητών είναι αντίστοιχα οι ακόλουθοι :

$$C=\begin{pmatrix}23,091 & 16,455\\ 16,455 & 21,091\end{pmatrix} \quad \text{και} \quad R=\begin{pmatrix}1,000 & 0,746\\ 0,746 & 1,000\end{pmatrix} \qquad (3.7)$$

με τις μεταβλητές x_1 και x_2 να συσχετίζονται με συντελεστή συσχέτισης $r=0,746$.

[6] Βλ δεδομένα στο Table 4.1 από Sharma S. (1996). Principal Component Analysis, *Applied Multivariate Techniques.* p. 59, New York : John Wiley & Sons, Inc. Μετατροπή πίνακα: Σταυρινίδης Σταύρος Κων/νος.

Το ακόλουθο σχήμα αποτελεί μια γραφική παράσταση των μετασχηματισμένων μας μεταβλητών.



Διάγραμμα 3. Μετασχηματισμένες μεταβλητές και προβολή των νέων σημείων στον νέο άξονα $X_1{}^*$.[7]

Έστω τώρα ένας νέος άξονας $X_1{}^*$ που σχηματίζει γωνία θ μοιρών με τον άξονα X_1. Η συντεταγμένη των σημείων ως προς το νέο άξονα $X_1{}^*$ λαμβάνεται ύστερα από προβολή των σημείων (παρατηρήσεων) στον άξονα $X_1{}^*$. Η νέα αυτή συντεταγμένη που προκύπτει είναι ένας γραμμικός συνδυασμός των συντεταγμένων κάθε σημείου σε σχέση με το ζεύγος των αρχικών αξόνων και εκφράζεται με την ακόλουθη εξίσωση :

$$x_1{}^* = \cos\theta \times x_1 + \sin\theta \times x_2 \qquad (3.8)$$

[7] Βλ Figure 4.1 στο Sharma S. (1996). Principal Component Analysis, *Applied Multivariate Techniques.* p. 60, New York : John Wiley & Sons, Inc.

με x_1^* να είναι η συντεταγμένη της κάθε παρατήρησης σε σχέση με τον άξονα X_1^* ενώ x_1 και x_2 οι συντεταγμένες κάθε παρατήρησης ως προς τους άξονες X_1 και X_2. Εύκολα συμπεραίνουμε ότι η μεταβλητή x_1^*, η οποία είναι ένας γραμμικός συνδυασμός των αρχικών μας μεταβλητών, μπορεί να θεωρηθεί ως μια νέα μεταβλητή. Για μια δεδομένη δοθείσα γωνία , έστω $\theta = 10$, δηλαδή εάν ο άξονας X παρουσιάζει κλίση 10° με τη βοήθεια της παραπάνω εξίσωσης υπολογίζουμε τη νέα μας μεταβλητή x_1^*:

$$x_1^* = 0{,}985 \times x_1 + 0{,}174 \times x_2 \qquad (3.9)$$

Οι νέες τιμές της μεταβλητής x_1^* παρουσιάζονται στον ακόλουθο πίνακα καθώς και στο παραπάνω διάγραμμα :

	Μετασχηματισμένες μεταβλητές		
Παρατήρηση	x_1	x_2	x_1^*
1	8	5	8.747
2	4	7	5.155
3	5	3	5.445
4	3	-1	2.781
5	2	5	2.838
6	1	-4	0.290
7	0	1	0.174
8	-1	3	-0.464
9	-3	-6	-3.996
10	-5	-4	-5.619
11	-6	-6	-6.951
12	-8	-3	-8.399
Μέσος όρος	0.000	0.000	0.000
Διασπορά	23.091	21.091	28.659

Πίνακας 6. Μετασχηματισμένες μεταβλητές και η νέα μεταβλητή x_1^* για κλίση του νέου άξονα ίση με 10°.[8]

[8]Βλ δεδομένα Table 4.2 από Sharma S. (1996). Principal Component Analysis, *Applied Multivariate Techniques.* p. 61, New York : John Wiley & Sons, Inc.Μετατροπή πίνακα: Σταυρινίδης Σταύρος-Κων/νος.

σημειώνεται ότι οι συντεταγμένες της πρώτης παρατήρησης είναι 8 και 5 ενώ η πρώτη τιμή της νέας μεταβλητής είναι $8,747$. Έχουμε δηλαδή :

$$0,985\times 8 + 0,174\times 5 = 8,747 \qquad (3.10)$$

επιπροσθέτως στο παραπάνω πίνακα βλέπουμε ότι η νέα μας μεταβλητή έχει μηδενικό μέσο όρο, καθώς και ότι η διασπορά της ισούται με $28,659$ δηλαδή ερμηνεύει το $\frac{28,659}{44,182} = 0,6487$ ή περίπου το 65% της συνολικής διασποράς.

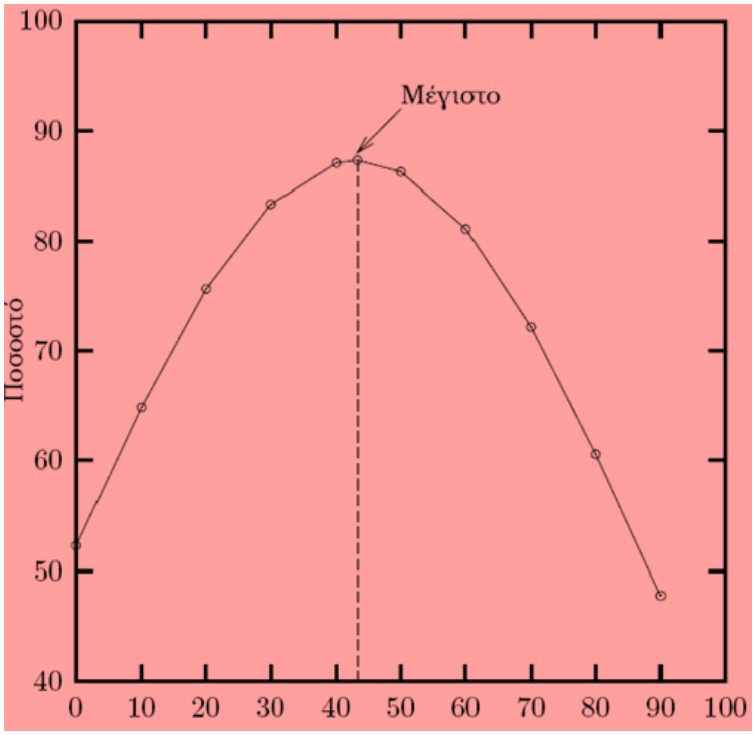
Παρατηρούμε ότι η διασπορά της ξεπερνά αυτές των αρχικών μας μεταβλητών. Έστω τώρα ότι ο άξονας X_1^* σχηματίζει γωνία ίση με 20° με τον άξονα X_1, προφανώς και η νέα μεταβλητή x_1^* θα έχει τώρα νέες τιμές. Στον παρακάτω πίνακα παρουσιάζεται το ποσοστό της συνολικής διασποράς που εκφράζουν οι νέες μεταβλητές καθώς αυξάνουμε σταδιακά την κλίση του άξονα X_1^* και κατόπιν υπολογίζουμε τις αντίστοιχες τιμές της x_1^*.

Γωνία θ με τον X_1	Ολική διασπορά	Διασπορά που εκφράζει η x_1^*	Ποσοστό (%)
0	44.182	23.091	52.263
10	44.182	28.659	64.866
20	44.182	33.434	75.676
30	44.182	36.841	83.387
40	44.182	38.469	87.072
43.261	44.182	38.576	87.312
50	44.182	38.122	86.282
60	44.182	35.841	81.117
70	44.182	31.902	72.195
80	44.182	26.779	60.597
90	44.182	21.091	47.772

Πίνακας 7. Διασπορά νέων μεταβλητών / νέων αξόνων.[9]

[9] Βλ δεδομένα Table 4.3 από Sharma S. (1996). Principal Component Analysis, *Applied Multivariate Techniques.* p. 61, New York : John Wiley & Sons, Inc.Μετατροπή πίνακα: Σταυρινίδης Σταύρος-Κων/νος.

Ακολούθως το παρακάτω διάγραμμα υποδεικνύει ότι αρχικά αύξηση της κλίσης του X_1^* σηματοδοτεί και αύξηση της συνολικής διασποράς που ερμηνεύει η x_1^* ενώ μετά από μια συγκεκριμένη τιμή, η διασπορά που ερμηνεύει η x_1^* μειώνεται καθώς συνεχίζουμε να αυξάνουμε τη γωνία μεταξύ X_1^* και X_1.



Διάγραμμα 4. Γωνία θ του άξονα X_1^* ***στον*** X_1 [10]

όπως συνάγεται από το διάγραμμα, υπάρχει ένας και μόνο ένας νέος άξονας του οποίου η αντίστοιχη νέα μεταβλητή ερμηνεύει τη μέγιστη διασπορά των δεδομένων μας. Ο συγκεκριμένος νέος άξονας έχει κλίση $43,261^\circ$ με τον άξονα X_1. Η αντίστοιχη x_1^* μεταβλητή υπολογίζεται από την ακόλουθη εξίσωση:

$$x_1^* = \cos 43{,}261 \times x_1 + \sin 43{,}261 \times x_2 = 0{,}728\, x_1 + 0{,}685 x_2 \qquad (3.11)$$

[10] Βλ Figure 4.2 στο Sharma S. (1996). Principal Component Analysis, *Applied Multivariate Techniques.* p. 62, New York : John Wiley & Sons, Inc. Μετατροπή διαγράμματος: Σταυρινίδης Σταύρος-Κων/νος.

Στον παρακάτω πίνακα παρουσιάζουμε τις νέες μεταβλητές x_1^* και x_2^* οι οποίες έχουν προκύψει καθώς οι άξονες σχηματίζουν γωνία $43,261^\circ$.

	Μετασχηματισμένες μεταβλητές		Νέες μεταβλητές	
Παρατήρηση	x_1	x_2	x_1^*	x_2^*
1	8	5	9.253	-1.841
2	4	7	7.710	2.356
3	5	3	5.697	-1.242
4	3	-1	1.499	-2.784
5	2	5	4.883	2.271
6	1	-4	-2.013	-3.598
7	0	1	0.685	0.728
8	-1	3	1.328	2.870
9	-3	-6	-6.297	-2.313
10	-5	-4	-6.382	0.514
11	-6	-6	-8.481	-0.257
12	-8	-3	-7.882	3.298
Μέσος όρος	0.000	0.000	0.000	0.000
Διασπορά	23.091	21.091	38.576	5.606

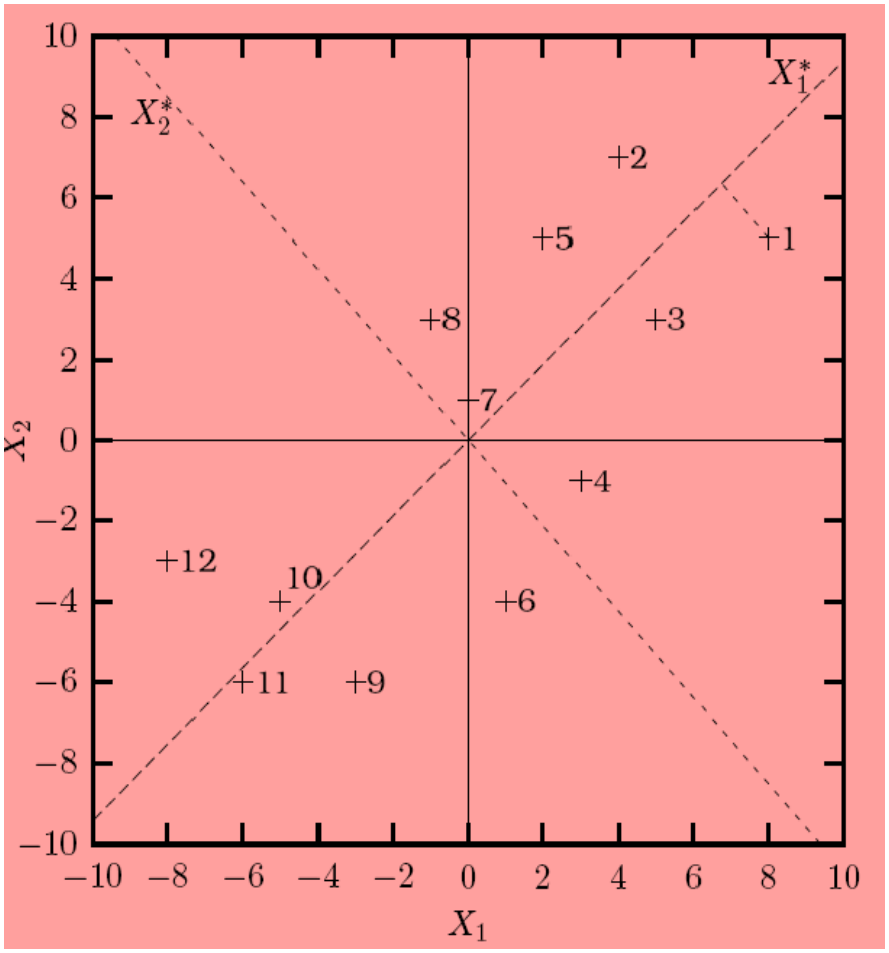
Πίνακας 8. Μετασχηματισμένες αρχικές μεταβλητές και νέες μεταβλητές x_1^* ***και*** x_2^* ***για τους νέους άξονες με κλίση*** 43.261° [11]

Οι πίνακες διασποράς και συσχέτισης των δυο μεταβλητών είναι αντίστοιχα :

$$C = \begin{pmatrix} 38,576 & 0,000 \\ 0,000 & 5,606 \end{pmatrix} \qquad R = \begin{pmatrix} 1,000 & 0,000 \\ 0,000 & 1,000 \end{pmatrix} \tag{3.12}$$

[11] Βλ δεδομένα Table 4.4 από Sharma S. (1996). Principal Component Analysis, *Applied Multivariate Techniques.* p. 62, New York : John Wiley & Sons, Inc.Μετατροπή πίνακα: Σταυρινίδης Σταύρος-Κων/νος.

Οι νέοι άξονες που ορίστηκαν φαίνονται στο ακόλουθο διάγραμμα :



Διάγραμμα 5. Μετασχηματισμένα δεδομένα / νέοι άξονες.[12]

Εν κατακλείδι, από τον πίνακα 8 και το διάγραμμα 5 συμπεραίνουμε τα εξής:

- Η διευθέτηση των σημείων (παρατηρήσεων) στο δισδιάστατο χώρο δεν αλλάζει, δηλαδή τα σημεία παρουσιάζονται είτε ως προς τους παλιούς είτε ως προς τους νέους άξονες.

[12] Βλ Figure 4.3 στο Sharma S. (1996). Principal Component Analysis, *Applied Multivariate Techniques.* p. 63, New York : John Wiley & Sons, Inc.

- Οι προβολές των σημείων (παρατηρήσεων) στους αρχικούς άξονες δίνουν τις τιμές των αρχικών μεταβλητών, ενώ οι προβολές των σημείων πάνω στους νέους άξονες δίνουν τις τιμές των νέων μεταβλητών. Οι νέοι άξονες ή μεταβλητές καλούνται principal components (κύριες συνιστώσες) και οι τιμές των νέων μεταβλητών καλούνται principal component scores (τιμές των κυρίων συνιστωσών).

- Κάθε μια απ' τις νέες μεταβλητές (x_1* και x_2*) είναι γραμμικός συνδυασμός των αρχικών μεταβλητών και παραμένει μετασχηματισμένη με μηδενικό μέσο όρο (mean corrected).

- Οι διασπορές των x_1* και x_2* είναι αντίστοιχα $38,576$ και $5,606$. Η συνολική διασπορά των δυο αυτών μεταβλητών ισούται με $(38,576+5,606)=44,182$ που είναι ίση με τη συνολική διασπορά των αρχικών μας μεταβλητών x_1 και x_2, άρα η συνολική διασπορά δεν αλλάζει. Οφείλεται στο ότι δεν αλλάζει η διευθέτηση των σημείων στο χώρο.

- Τα ποσοστά της συνολικής διασποράς που ερμηνεύουν οι x_1* και x_2* είναι $87,31\%$ και $12,69\%$ αντίστοιχα. Η πρώτη νέα μεταβλητή x_1* ερμηνεύει το μεγαλύτερο ποσοστό της συνολικής διασποράς απ' ότι οι δυο αρχικές μεταβλητές. Η άλλη νέα μας μεταβλητή x_2* ερμηνεύει το ποσοστό της διασποράς που δεν ερμηνεύεται από την x_1*. Συνολικά όμως οι δυο νέες μεταβλητές εκφράζουν την ολική διασπορά των δεδομένων μας.

- Οι νέες μεταβλητές έχουν μηδενικό συντελεστή συσχέτισης, είναι επομένως ασυσχέτιστες μεταξύ τους.

Η γεωμετρική αναπαράσταση της Ανάλυσης Κυρίων Συνιστωσών μπορεί εύκολα να επεκταθεί και σε περισσότερες από δυο μεταβλητές. Ειδικότερα, ένα σύνολο δεδομένων αποτελούμενο από p το πλήθος μεταβλητές μπορεί να παρουσιαστεί γραφικά σε έναν p-διάστατο χώρο σε σχέση με τους αρχικούς p άξονες ή τους p νέους άξονες.

Ο πρώτος νέος άξονας, X_1*, αντιστοιχεί στην πρώτη νέα μεταβλητή x_1* η οποία ερμηνεύει το μέγιστο ποσοστό της συνολικής διασποράς. Ακολούθως σχηματίζουμε το δεύτερο νέο άξονα, που είναι κάθετος στον πρώτο, του οποίου η αντίστοιχη νέα μεταβλητή x_2* ερμηνεύει το μέγιστο της διασποράς που δεν εξηγεί η πρώτη μεταβλητή και δεν σχετίζεται μαζί της.

Η διαδικασία αυτή συνεχίζεται μέχρι να οριστούν όλοι οι p το πλήθος άξονες και οι p νέες μεταβλητές να ερμηνεύουν το μέγιστο της κάθε φορά υπολείπουσας διασποράς υπό την προϋπόθεση όμως οι νέες μεταβλητές να παραμένουν

ασυσχέτιστες μεταξύ τους. Τέλος ο αριθμός των νέων μεταβλητών που σχηματίζονται οφείλει πάντα να ισούται με τον αριθμό των αρχικών μεταβλητών.

3.3 ΕΠΙΛΟΓΗ ΤΟΥ ΑΡΙΘΜΟΥ ΤΩΝ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ ΠΟΥ ΘΑ ΕΞΕΤΑΣΟΥΜΕ

Μετά την εφαρμογή της Ανάλυσης Κυριών Συνιστωσών, επόμενο βήμα είναι να επιλέξουμε τον αριθμό των κυρίων συνιστωσών που θα διατηρήσουμε στη μελέτη μας και ακολούθως θα εξετάσουμε. Η απόφαση αυτή εξαρτάται απ' το ποσοστό της πληροφορίας που αποφασίζουμε να χάσουμε (πληροφοριακό κόστος), με την εξάλειψη ενός αριθμού κυρίων συνιστωσών. Διάφορα κριτήρια έχουν αναπτυχθεί για το σκοπό αυτό και είναι τα ακόλουθα :

Ιδιοτιμή μεγαλύτερη της μονάδας

Βρίσκει εφαρμογή στην περίπτωση που τα δεδομένα μας έχουν υποστεί τυποποίηση. Τότε επιλέγουμε να μελετήσουμε τις κύριες συνιστώσες ,των οποίων οι ιδιοτιμές είναι μεγαλύτερες της μονάδας (eigenvalue–greater-than–one-rule). Στηρίζεται στο ότι μια κύρια συνιστώσα με αντίστοιχη ιδιοτιμή ίση με τη μονάδα ερμηνεύει τη μέση διασπορά στα δεδομένα μας. Άρα οι κύριες συνιστώσες με ιδιοτιμές μεγαλύτερες της μονάδας ερμηνεύουν μεγαλύτερο ποσοστό της συνολικής διασποράς.

Παρ' όλα αυτά, σύμφωνα με τον Cliff (1988) η εφαρμογή του κριτηρίου αυτού μπορεί να οδηγήσει σε επιλογή περισσότερων ή λιγότερων κυρίων συνιστωσών από ότι απαιτούνται. Γι' αυτό δεν πρέπει να εφαρμόζεται τυφλά, αλλά σε συνδυασμό με άλλα κριτήρια.

Ποσοστό της διασποράς που ερμηνεύεται

Το συγκεκριμένο κριτήριο καθίσταται χρήσιμο όταν εκ των προτέρων γνωρίζουμε το ποσοστό της συνολικής διασποράς που οφείλει να ερμηνεύεται μέσω των κυρίων συνιστωσών που έχουμε επιλέξει. Συχνά το ποσοστό αυτό είναι υποκειμενικού χαρακτήρα και εξαρτάται από τον αναλυτή των δεδομένων μας αλλά και από τις ανάγκες αλλά και τη φύση της μελέτης που εφαρμόζουμε. Συχνά όμως συνηθίζεται να επιλέγουμε κύριες συνιστώσες που ερμηνεύουν ένα ποσοστό της τάξεως του 80%.

Γνωρίζοντας όμως ότι η συνολική διασπορά ισούται με το άθροισμα των ιδιοτιμών του πίνακα συσχέτισης, ο υπολογισμός του επιθυμητού ποσοστού της συνολικής διασποράς πραγματοποιείται με ευκολία. Έτσι το ποσοστό που ερμηνεύουν οι k πρώτες κύριες συνιστώσες ισούται με :

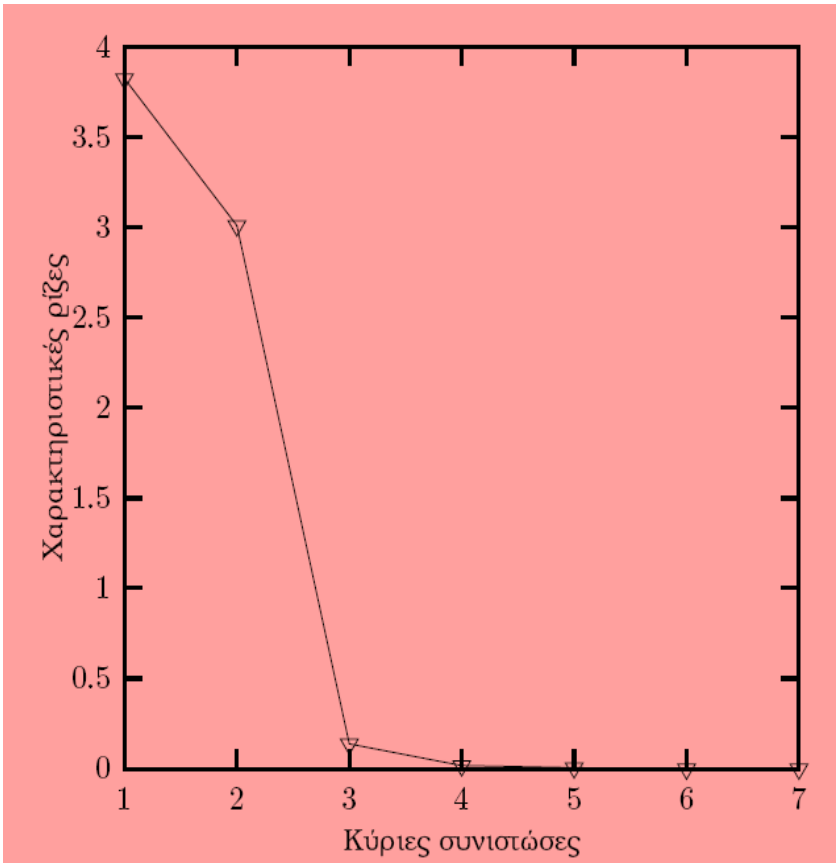
$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \qquad (3.13)$$

όπου p ο αριθμός όλων των ιδιοτιμών του πίνακα συσχέτισης.

Γραφική παράσταση των ιδιοτιμών ή χαρακτηριστικών ριζών (scree plot).

Το συγκεκριμένο κριτήριο, το οποίο προτάθηκε από τον Cattell (1966) είναι ένα απ' τα πλέον δημοφιλή. Στον άξονα X τοποθετούνται οι κύριες συνιστώσες και στον άξονα Y οι ιδιοτιμές (χαρακτηριστικές ρίζες) του πίνακα συσχέτισης. Το γράφημα που προκύπτει είναι μια καμπύλη της οποίας η κλίση διαρκώς μειώνεται και τείνει να πάρει τη μορφή ευθείας όσο ο αριθμός των κυρίων συνιστωσών αυξάνεται.

Οι κύριες συνιστώσες που τελικά επιλέγονται στο μοντέλο μας είναι αυτές που βρίσκονται πριν το σημείο της ξαφνικής απότομης μείωσης της κλίσης της καμπύλης, όπως φαίνεται στο ακόλουθο διάγραμμα :



Διάγραμμα 6. Χαρακτηριστικές ρίζες ή ιδιοτιμές για κάθε κύρια συνιστώσα (scree plot).

επομένως επιλέγουμε τις **δυο** κύριες συνιστώσες οι οποίες βρίσκονται πριν απ' το σημείο στο οποίο η καμπύλη αρχίζει να σχηματίζει ευρεία γωνία (elbow).

3.4 Η PCA ΩΣ ΜΕΘΟΔΟΣ ΜΕΙΩΣΗΣ ΔΙΑΣΤΑΣΗΣ

Παρατηρούμε , ότι έχοντας ξεκινήσει με έναν αρχικό πίνακα X, διαστάσεων $n\times p$, εφαρμόζοντας την παραπάνω διαδικασία, καταλήγουμε σε έναν πίνακα U που καλείται μήτρα αποτελεσμάτων με διαστάσεις $n\times k$ όπου :

$$\mathrm{U}_{n\times k}=\begin{pmatrix} u_{11} & \cdots & u_{1k} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nk} \end{pmatrix} \qquad (3.14)$$

καταλήγουμε επομένως σε έναν πίνακα μικρότερων διαστάσεων με λιγότερες στήλες–μεταβλητές. Η ανωτέρω αποτελεί βασική ιδιότητα της μεθόδου PCA και στην πλειονότητα των περιπτώσεων πραγματοποιείται με ελάχιστο πληροφοριακό κόστος. Γίνεται εύκολα κατανοητό ότι η διαδικασία μείωσης της διάστασης οδηγεί σε απώλεια πληροφορίας, καθώς κάθε στήλη που «χάνεται» λόγω της μείωσης της διάστασης ταυτίζεται με την ταυτόχρονη απώλεια n το πλήθος στοιχείων στην μήτρα αποτελεσμάτων $\mathrm{U}_{n\times k}$ που τελικά λαμβάνουμε.

Αξίζει να αναφερθεί ότι το ποσοστό της πληροφορίας που εκφράζεται από κάθε ιδιοδιάνυσμα της μήτρας φορτίου (loading matrix), εκφράζεται από το μέγεθος της ιδιοτιμής του. Δηλαδή ιδιοδιανύσματα στα οποία αντιστοιχούν μεγαλύτερες ιδιοτιμές εκφράζουν υψηλότερο ποσοστό της πληροφορίας.

Προκειμένου να αποφασίσουμε ποια ιδιοδιανύσματα θεωρούνται σημαντικά ή μη δημιουργούμε την ακόλουθη ποσότητα η οποία καλείται σχετικό μέτρο της ιδιοτιμής. Έτσι, έστω ότι έχουμε το m-στο ιδιοδιάνυσμα. Το σχετικό μέτρο της ιδιοτιμής του συγκεκριμένου ιδιοδιανύσματος δίνεται απ' τον λόγο :

$$\frac{\lambda_m}{\sum_{i=1}^{p}\lambda_i} \qquad (3.15)$$

Για την καλύτερη κατανόηση της PCA, ο γράφων θα παραθέσει και στη συνέχεια αποδείξει τρεις θεωρητικές προτάσεις :

Πρόταση 1

Αν Σ ο πίνακας συνδιασποράς του τυχαίου διανύσματος $X^T=(X_1,X_2,\ldots,X_p)$ και αν ο πίνακας Σ έχει ιδιοτιμές – ιδιοδιανύσματα τα ζεύγη $(\lambda_1,e_1),(\lambda_2,e_2)\ldots,(\lambda_p,e_p)$ όπου $\lambda_1\geq\lambda_2\geq\ldots\lambda_p\geq 0$ τότε η i-κύρια συνιστώσα δίνεται από $Z_i=e_i{}^TX=e_{1i}X_1+e_{2i}X_2+\ldots.+e_{pi}X_p \quad i=1,2,\ldots,p$ όπου $\mathrm{var}(Z_i)=e_i{}^T\Sigma e_i=\lambda_i$ $i=1,2,..,p$ και $\mathrm{cov}(Z_i,Z_j)=e_i{}^T\Sigma e_j=0 \quad i\neq j$

Θα δείξουμε δηλαδή ότι οι κύριες συνιστώσες είναι μεταξύ τους ασυσχέτιστες και έχουν διασπορές ίσες με τις ιδιοτιμές του πίνακα Σ.

Απόδειξη

Έχουμε ότι για $\lambda\neq 0$,

$$\max\frac{\lambda^T\Sigma\lambda}{\lambda^T\lambda}=\lambda_1$$

όταν $\lambda=e_1$

όμως έχουμε ότι $e_1{}^Te_1=1$ **αφού τα ιδιοδιανύσματα είναι κανονικοποιημένα.**

άρα για $\lambda\neq 0$**,**

$$\max\frac{\lambda^T\Sigma\lambda}{\lambda^T\lambda}=\lambda_1=e_1{}^T\Sigma e_1=\mathrm{var}(Z_1)$$

με τον ίδιο τρόπο έχουμε για $\lambda\perp e_1,e_2,\ldots,e_j$**,**

$$\max\frac{\lambda^T\Sigma\lambda}{\lambda^T\lambda}=\lambda_{j+1} \qquad j=1,2,\ldots,p-1$$

όταν $\lambda=e_{j+1}$**, με** $e_{j+1}{}^Te_i=0$**, για** $i=1,2,\ldots,j$ **και** $i=1,2,\ldots,j$

έχουμε

$$\frac{e_{j+1}^T \Sigma e_{j+1}}{e_{j+1}^T e_{j+1}} = e_{j+1}^T \Sigma e_{j+1} = \mathrm{var}(Z_{j+1})$$

Όμως $$e_{j+1}^T \Sigma e_{j+1} = \lambda_{j+1} e_{j+1}^T e_{j+1} = \lambda_{j+1} \Rightarrow \mathrm{var}(Z_{j+1}) = \lambda_{j+1}$$

θα δείξουμε τώρα, ότι $\mathrm{cov}(Z_i, Z_j) = 0$ **με την προϋπόθεση ότι** $e_i^T e_j = 0 \quad i \neq j$

Πράγματι, αυτό συμβαίνει αν οι ιδιοτιμές του πίνακα συνδιασπορών Σ είναι διακεκριμένες. Αν αυτό συμβαίνει, τότε όλα τα e_i , $i = 1, 2, .., p$ είναι ορθογώνια. Αν αυτό δεν συμβαίνει, δηλαδή οι ιδιοτιμές δεν είναι όλες διακεκριμένες, μπορούμε καταλλήλως να επιλέξουμε τα e_i ώστε να είναι ορθογώνια. Άρα για δυο οποιαδήποτε διανύσματα e_i και e_j έχουμε $e_i^T e_j = 0 \quad i \neq j$. Ισχύει όμως ότι :

$$\Sigma e_j = \lambda e_j$$

άρα

$$\mathrm{cov}\left(Z_i, Z_j\right) = e_i^T \Sigma e_j = e_i^T \lambda_j e_j = \lambda_j e_i^T e_j = 0$$

για κάθε $i \neq j$

Πρόταση 2

Έστω ότι $X^T = [X_1, X_2, \ldots, X_p]$, το διάνυσμα των $X_1, X_2, \ldots, X_p$ μεταβλητών και ο πίνακας συνδιασπορών του Σ, ο οποίος έχει ιδιοτιμές $\lambda_1, \lambda_2, \ldots, \lambda_p$ με $\lambda_1 \geq \lambda_2 \geq \ldots . \lambda_p \geq 0$ και αντίστοιχα ιδιοδιανύσματα $e_1, e_2, \ldots e_p$. Οι κύριες συνιστώσες μας είναι : $Z_1 = e_1^T X, Z_2 = e_2^T X, \ldots .., Z_p = e_p^T X$. Ισχύει τότε :

$$\sigma_{11} + \sigma_{22} + \ldots . + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{var}(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} \mathrm{var}(Z_i)$$

Απόδειξη

Ο πίνακας Σ μπορεί να διαγωνιοποιηθεί ως : $\Sigma = P\Lambda P^T$, με Λ να είναι ο διαγώνιος πίνακας των ιδιοτιμών και $P = [e_1, e_2, \ldots, e_p]$ με $P^T P = PP^T = 1$. Τότε ισχύει :

$$tr(\Sigma) = tr(P\Lambda P^T) = tr(\Lambda P^T P) = tr(\Lambda) = \lambda_1 + \lambda_2 + \ldots + \lambda_p.$$

άρα προκύπτει ότι :

$$\sum_{i=1}^{p} var(X_i) = tr(\Sigma) = tr(\Lambda) = \sum_{i=1}^{p} var(Z_i)$$

η συνολική διασπορά ισούται με :

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

άρα τώρα , μπορούμε να υπολογίσουμε το ποσοστό της συνολικής διασποράς που αντιστοιχεί στην j-κύρια συνιστώσα και ισούται με :

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad j = 1, 2, \ldots, p$$

Συνήθως στην PCA, λίγες κύριες συνιστώσες (το πολύ 3 συνήθως) εξηγούν μεγάλο ποσοστό της συνολικής διασποράς. Δηλαδή, οι αρχικές p μεταβλητές, αντικαθίστανται με αυτές τις συνιστώσες και δεν παρατηρείται παρά ελάχιστη απώλεια πληροφορίας. Επιπροσθέτως, τα μεγέθη e_{ji} μετρούν την σημαντικότητα της j-μεταβλητής στην i-κύρια συνιστώσα. Το e_{ji} είναι ανάλογο του συντελεστή συσχέτισης μεταξύ των X_j και Z_i.

Πρόταση 3

Έστω ότι $Z_1 = e_1^T X, Z_2 = e_2^T X, \ldots, Z_p = e_p^T X$ οι κύριες συνιστώσες που προκύπτουν από τον πίνακα συνδιασπορών Σ. Τότε οι συντελεστές συσχέτισης μεταξύ των μεταβλητών X_j και των κυρίων συνιστωσών Z_i δίνονται απ' τον τύπο :

$$\rho_{Xi,Zj} = \frac{e_{ji}\sqrt{\lambda_j}}{\sqrt{\sigma_{ii}}}$$

όπου $\lambda_1 + \lambda_2 + \ldots + \lambda_p$ οι ιδιοτιμές του πίνακα Σ με αντίστοιχα ιδιοδιανύσματα τα $e_1, e_2, \ldots, e_p$.

Απόδειξη

Έστω $\lambda_i = [0, \ldots, 0, 1, 0, \ldots 0]$

ώστε $X_i = \lambda_i^T X$

και $\mathrm{cov}(X_i, Z_j) = \mathrm{cov}(\lambda_i^T X, e_j^T X) = \lambda_i^T \Sigma e_j$

επειδή ως γνωστόν $\Sigma e_j = \lambda_j e_j$

ισχύει $\mathrm{Cov}(X_i, Z_j) = \lambda_i^T \lambda_j e_j = \lambda_j e_{ji}$

όμως $\mathrm{var}(Z_j) = \lambda_j$

και $\mathrm{var}(X_i) = \sigma_{ii.}$

έχουμε $$\rho_{\mathrm{Xi,Zj}} = \frac{\mathrm{cov}(\mathrm{X}_i, Z_j)}{\sqrt{\mathrm{var}(\mathrm{X}_i)}\sqrt{\mathrm{var}(Z_j)}} = \frac{\lambda_j \mathrm{e}_{\mathrm{ji}}}{\sqrt{\lambda_j}\sqrt{\sigma_{\mathrm{ii}}}} = \frac{\mathrm{e}_{\mathrm{ji}}\sqrt{\lambda_j}}{\sqrt{\sigma_{\mathrm{ii}}}} \quad \mathrm{i,j} = 1,2,\ldots,\mathrm{p}.$$

3.5 Η ΜΕΘΟΔΟΣ ΙΔΙΟΑΝΑΛΥΣΗΣ ΓΙΑ ΤΗΝ PCA

SVD(Singular Value Decomposition)

Πρόκειται για ένα χρήσιμο μαθηματικό εργαλείο, το οποίο μπορεί να μεταφραστεί ως Παραγοντοποίηση ιδιαζουσών τιμών. Αποτελεί μια γενίκευση της eigen-decomposition και συνιστά τη συνηθέστερη μέθοδο ιδιοανάλυσης για την PCA. Στο σημείο αυτό, θα δοθεί μια αναλυτική περιγραφή της διαδικασίας αυτής. Η εν λόγω διαδικασία παραγοντοποιεί έναν οποιοδήποτε πίνακα X, διαστάσεων $n \times p$, σε ένα γινόμενο τριών πινάκων, δυο ορθογώνιων και ενός διαγώνιου, στην παρακάτω μορφή (Abdi & Williams, 2010) :

$$\mathrm{X} = P\mathrm{DV}^{\mathrm{T}} \tag{3.16}$$

με τον πίνακα P να αποτελείται από τα (κανονικοποιημένα) ιδιοδιανύσματα του πίνακα XX^{T}, $(P^{\mathrm{T}}P = I)$. Οι στήλες του P καλούνται τα αριστερά ιδιάζοντα διανύσματα (left singular vectors) του πίνακα X.

Αντίστοιχα ο πίνακας V αποτελείται από τα (κανονικοποιημένα) ιδιοδιανύσματα του πίνακα $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ $(\mathrm{V}^{\mathrm{T}}\mathrm{V} = I)$. Οι στήλες του V καλούνται τα δεξιά ιδιάζοντα διανύσματα (right singular vectors) του πίνακα X. Τέλος ο διαγώνιος πίνακας D αποτελείται από τις ιδιάζουσες τιμές (singular values), όπου $\mathrm{D} = \Lambda^{1/2}$, με Λ να είναι ο διαγώνιος πίνακας των ιδιοτιμών του XX^{T} ή του $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ μιας και είναι ταυτόσημοι. Η παραπάνω μορφή μπορεί να γραφτεί ως :

$$X = \sum_{i=1}^{L} \mathrm{d}_{\mathrm{l}} p_{\mathrm{l}} \mathrm{v}_{\mathrm{l}}^{\mathrm{T}} \tag{3.17}$$

όπου L είναι ο βαθμός του πίνακα X και d_{l}, p_{l} και v_{l} είναι αντίστοιχα η l-στη singular τιμή, το αριστερό και το δεξί singular διάνυσμα του X. Βάσει του

παραπάνω, φαίνεται ότι ο πίνακας X μπορεί να ανακατασκευαστεί ως ένα άθροισμα l πινάκων βαθμού 1. Δηλαδή, ο πρώτος απ' αυτούς τους πίνακες δίνει τη βέλτιστη ανακατασκευή του X από έναν πίνακα βαθμού 1. Ομοίως το άθροισμα των πρώτων δυο πινάκων τη βέλτιστη ανακατασκευή του X από έναν πίνακα βαθμού 2 κλπ.

Θα παρουσιάσουμε τώρα τον τρόπο εύρεσης των κυρίων συνιστωσών χρησιμοποιώντας την SVD ενός πίνακα X των p μεταβλητών με n παρατηρήσεις-μετρήσεις για κάθε μια.

Όπως έχει αναφερθεί πάνω, η PCA υπολογίζει νέες μεταβλητές, τις κύριες συνιστώσες, που είναι γραμμικοί συνδυασμοί των αρχικών μας μεταβλητών, με τον περιορισμό η πρώτη κύρια συνιστώσα να ερμηνεύει την μέγιστη διασπορά. Ακολούθως, η δεύτερη κύρια συνιστώσα να είναι ορθογώνια στην πρώτη και να ερμηνεύει ομοίως τη μέγιστη δυνατή διασπορά. Οι τιμές αυτών των νέων μεταβλητών για τις παρατηρήσεις μας καλούνται factor scores και μπορούν να ερμηνευθούν γεωμετρικά ως οι προβολές των παρατηρήσεων μας στις κύριες συνιστώσες.

Εφαρμόζοντας την SVD στον πίνακα X παίρνουμε $\mathrm{X} = P\mathrm{D}\mathrm{V}^{\mathrm{T}}$. Ο πίνακας U (factor scores matrix) διαστάσεων $n \times l$ προκύπτει βάσει της SVD μεθόδου ως εξής:

$$\mathrm{U} = \mathrm{PD} \tag{3.18}$$

με τον πίνακα V, να περιέχει τους συντελεστές των γραμμικών συνδυασμών που χρειάζονται για τον υπολογισμό των factor scores.

Ο πίνακας U μπορεί επίσης να ερμηνευθεί ως πίνακας προβολής αφού πολλαπλασιάζοντας τους X, V παίρνουμε τις τιμές των προβολών των παρατηρήσεων στις κύριες συνιστώσες. Πράγματι φαίνεται απ' το παρακάτω :

$$\mathrm{U} = \mathrm{PD} = \mathrm{PDV}^{\mathrm{T}}\mathrm{V} = \mathrm{XV} \tag{3.19}$$

όπου ο πίνακας V, καλείται πίνακας φορτίων (loading matrix) και επειδή είναι ορθοκανονικός μπορεί να ερμηνευθεί ως πίνακας των συνημίτονων κατεύθυνσης (direction cosines). Επομένως, σύμφωνα με την (3.19) :

ο πίνακας $$\mathrm{X} = \mathrm{UV}^{\mathrm{T}} \tag{3.20}$$

ισούται δηλαδή με το γινόμενο του πίνακα U με τον πίνακα V.

3.6 ΣΥΜΠΕΡΑΣΜΑΤΑ ΓΙΑ ΤΗ ΜΕΘΟΔΟ PCA

Όπως έχουμε αναφέρει το πρώτο στάδιο της Ανάλυσης Κυρίων Συνιστωσών, είναι ο υπολογισμός της μήτρας συνδιασπορών (covariance matrix). Ακολούθως, είδαμε ότι με τη βοήθεια μεθόδων ιδιοανάλυσης (Singular Value Decomposition), η μήτρα αυτή παραγοντοποιείται σε δυο νέες, τη μήτρα αποτελεσμάτων (factor scores matrix) καθώς και στη μήτρα φορτίων (loading matrix).

Η μήτρα των αποτελεσμάτων περιέχει την πληροφορία σχετικά με τις σχέσεις που παρουσιάζονται ανάμεσα στα δείγματα που μελετάμε. Αντιθέτως, η μήτρα των φορτίων περιέχει τις πληροφορίες για τις ποσοτικές αλληλεπιδράσεις που υπάρχουν μεταξύ των μεταβλητών . Μέσω των δυο τελευταίων οδηγούμαστε στον καθορισμό των ανεξάρτητων αξόνων ή διανυσμάτων που ερμηνεύουν με τον πλέον κατάλληλο τρόπο τα δεδομένα μας. Οι άξονες αυτοί που έχουμε καθορίσει καλούνται κύρια διανύσματα και προκύπτουν ως γραμμικοί συνδυασμοί των αρχικών μας μεταβλητών.

Το πλεονέκτημα της παραπάνω διαδικασίας των κυρίων διανυσμάτων, συνίσταται στο ότι χρησιμοποιώντας λίγες το πλήθος απ' τις μεταβλητές που εξάγουμε (κύριες συνιστώσες), ερμηνεύουμε το μεγαλύτερο μέρος της πληροφορίας των δεδομένων μας. Έτσι η απώλεια πληροφορίας είναι ελάχιστη. Με άλλα λόγια δεν χρειάζεται να απεικονίσουμε τα δεδομένα μας στα πολυάριθμα διαγράμματα δυο μεταβλητών τα οποία οφείλονται στις αρχικές μας παρατηρήσεις για τα δείγματα που εξετάζουμε.

Αντί αυτού υπολογίζουμε τις προβολές των αρχικών μας παρατηρήσεων, στο χώρο που ορίζουν τα κύρια διανύσματα που έχουμε καθορίσει. Συνεπώς, κάνοντας τις γραφικές παραστάσεις σε δυο διαστάσεις των κυρίων διανυσμάτων, ερμηνεύουμε το μεγαλύτερο ποσοστό πληροφορίας που περιέχουν τα δεδομένα μας.

3.7 ΕΡΜΗΝΕΙΑ ΤΗΣ ΜΕΘΟΔΟΥ PCA

Στην παρούσα υποενότητα, θα ερμηνεύσουμε τη διαδικασία που σχετίζεται με τις κύριες συνιστώσες. Οι τελευταίες αποτελούν γραμμικούς συνδυασμούς των αρχικών μας μεταβλητών και η νέα αυτή κατηγορία μεταβλητών που υπολογίζουμε μας βοηθά στη γενίκευση συμπεριφορών σχετικά με χαρακτηριστικά γνωρίσματα που περιέχει το δείγμα που μελετάμε. Αυτό μεταφράζεται στο ότι σε κάθε συνιστώσα αποδίδεται ένα διαφορετικό λανθάνον χαρακτηριστικό.

Γι' αυτό το λόγο, πολλές φορές στη βιβλιογραφία, οι κύριες συνιστώσες συναντώνται ως λανθάνουσες μεταβλητές (latent variables). Σημαντικός για την ερμηνεία των κυρίων συνιστωσών είναι ο ρόλος που διαδραματίζει η μήτρα φορτίων που έχουμε αναφέρει. Τα φορτία που αυτή περιέχει, ερμηνεύονται ως οι

συνεισφορές (contributions) που έχουν οι αρχικές μας μεταβλητές στο γραμμικό συνδυασμό. Το παραπάνω μας βοηθάει στον εντοπισμό των μηχανισμών που συγκροτούν τη φύση των δεδομένων μας.

Πράγματι, τα υψηλά φορτία των διαφορετικών μεταβλητών ως προς ένα συγκεκριμένο λανθάνων χαρακτηριστικό ή γνώρισμα, υποδεικνύουν ότι οι συγκεκριμένες μεταβλητές είναι συσχετισμένες μεταξύ τους. Αντιθέτως η συσχέτιση μεταξύ των μεταβλητών δε δύναται να αποδειχθεί εάν τα φορτία τους ως προς ένα χαρακτηριστικό είναι μικρά.

Όσον αφορά τα pc–scores που αντλούμε απ' την ανάλυση μας, κάνουμε τις ακόλουθες παρατηρήσεις. Καταρχάς, μελετάμε τον πίνακα των αποτελεσμάτων, καθώς γνωρίζουμε ότι όλες οι κύριες συνιστώσες είναι ορθογώνιες μεταξύ τους και επομένως τα αποτελέσματα παρουσιάζουν μηδενικό συντελεστή συσχέτισης για τις κύριες συνιστώσες.

Επομένως, βάσει των γραφημάτων που κάνουμε, αποκτούμε πληροφορίες για τις σχέσεις ανάμεσα στα δείγματα. Σε περίπτωση που διακρίνουμε την ύπαρξη σημείων που βρίσκονται απομακρυσμένα, τότε μπορούμε να υποθέσουμε την ύπαρξη σφαλμάτων κατά τη μέτρηση των δεδομένων μας. Αν συμβεί κάτι τέτοιο, μπορούμε να τα εξαλείψουμε απ' το σύνολο των δεδομένων μας και ακολούθως να εφαρμόσουμε ξανά τη μέθοδο της Ανάλυσης Κυρίων Συνιστωσών, λαμβάνοντας αυτή τη φορά καλύτερα και πιο αξιόπιστα αποτελέσματα.

Η διαδικασία αυτή μπορεί να επαναληφθεί έως ότου όλες οι πιθανές ανωμαλίες που παρατηρούμε να ερμηνευθούν και να οδηγηθούμε σε ένα λογικό και εύλογο τελικό γράφημα. Έχοντας αποκλείσει πλέον απ' το δείγμα μας κάθε ανώμαλη μέτρηση, μπορούμε να ερμηνεύσουμε τη σχέση των αρχικών μεταβλητών με τις κύριες συνιστώσες. Κάθε κύρια συνιστώσα ερμηνεύεται με ένα πλήθος συσχετισμένων μεταβλητών που συνεισφέρουν σε αυτήν. Όμως όλες οι κύριες συνιστώσες διέπονται απ' την ορθογωνιότητα και άρα δε σχετίζονται μεταξύ τους. Επιπλέον η κάθε κύρια συνιστώσα ερμηνεύει ένα συγκεκριμένο ποσοστό επί της συνολικής διασποράς του δείγματος μας και κάθε μεταβλητή μας μέσω της μήτρας φορτίου, μας παρουσιάζει το ποσοστό που και αυτή συνεισφέρει στην ολική διασπορά.

Επομένως, προκειμένου να αποφασίσουμε και να κρίνουμε τις συνεισφορές αυτές καλούμαστε να ορίσουμε με βάση το μέγεθος του φορτίου, τα διάφορα «επίπεδα» σημαντικότητας των συνεισφορών. Γνωρίζοντας ότι τα φορτία της παραπάνω μήτρας αντιπροσωπεύουν τους συντελεστές διασποράς ανάμεσα σε κάθε μεταβλητή και κάθε κύρια συνιστώσα, οδηγούμαστε στην ποσοτικοποίηση της συσχέτισης μεταξύ κάθε μεταβλητής με κάθε συνιστώσα. Αυτό γίνεται με βάση υποκειμενικά κριτήρια που εξαρτώνται κάθε φορά απ' τον αναλυτή .

Μπορούμε αυθαίρετα να υποθέσουμε ότι φορτία μεταξύ 0–0.5 τιμών ορίζουν χαμηλή συσχέτιση μεταξύ μεταβλητής και κύριας συνιστώσας, φορτία μεταξύ 0.5-0.7 τιμών ορίζουν μέτριου επιπέδου συσχέτιση και τέλος φορτία 0.7–1 υποδεικνύουν υψηλού επιπέδου συσχέτιση. Καθοριστικό ρόλο, στη παρούσα ερμηνεία παίζει και το πρόσημο του φορτίου, καθώς θετικό πρόσημο (+), υποδηλώνει την ευθεία συσχέτιση μεταβλητής – κύριας συνιστώσας, ενώ αρνητικό πρόσημο (-) υποδηλώνει την αντίστροφη συσχέτιση.

Έστω τώρα ότι μια μεταβλητή μας παρουσιάζει υψηλό φορτίο μόνο για μια συγκεκριμένη συνιστώσα. Αυτό ερμηνεύεται με το ότι η μεταβλητή αυτή δεν έχει

να κάνει καθόλου με τις υπόλοιπες μεταβλητές και η πηγή της συνδιασποράς της είναι μοναδική.

Στο σημείο αυτό θα προσπαθήσουμε να απαριθμήσουμε περιληπτικά τα πλεονεκτήματα της μεθόδου PCA, τα οποία την κατατάσσουν ανάμεσα στις σημαντικότερες και πλέον δημοφιλείς πολυμεταβλητές στατιστικές τεχνικές. Ειδικότερα συνοψίζονται ως εξής :

- Ταχεία και αποτελεσματική μαθηματική τεχνική.
- Οδηγεί σε νέες μεταβλητές (κύριες συνιστώσες) ορθογώνιες μεταξύ τους, συνεπώς ασυσχέτιστες που διευκολύνει την ερμηνεία τους.
- Προκαλεί μείωση των προς εξέταση μεταβλητών μας, σε περιπτώσεις μεγάλου πλήθους δεδομένων, με ελάχιστη έως μηδαμινή απώλεια πληροφορίας.
- Μοντελοποιεί τα προβλήματα σε σύγχρονους επιστημονικούς και κλάδους της βιομηχανίας.
- Μας δίνει τη δυνατότητα εύρεσης διαφορών και ομοιοτήτων κατά την επεξεργασία των δεδομένων μας.

ΚΕΦΑΛΑΙΟ 4. Η ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ (PLS)

Η έρευνα στην επιστήμη και την τεχνολογία περιλαμβάνει τη χρήση εύκολων στη μέτρηση μεταβλητών προκειμένου να εξηγηθεί ή να προβλεφθεί η συμπεριφορά άλλων μεταβλητών (αποκρίσεων). Όταν οι μεταβλητές είναι λίγες τον αριθμό ή μη συγγραμμικές, τότε χρησιμοποιείται η Πολλαπλή Γραμμική Παλινδρόμηση. Όταν τουλάχιστον μια απ' τις παραπάνω προϋποθέσεις δεν ισχύει, χρησιμοποιείται η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων, στα αγγλικά καλείται PLS (Partial Least Square Regression).

Πρόκειται για μια πρόσφατη τεχνική η οποία συνδυάζει χαρακτηριστικά τόσο της Παλινδρόμησης Κυρίων Συνιστωσών όσο και της Πολλαπλής Γραμμικής Παλινδρόμησης. Πρωτοχρησιμοποιήθηκε στις πολιτικές επιστήμες από τον Herman Wold (1966), συγκεκριμένα στον τομέα της οικονομίας. Αργότερα εφαρμόστηκε στην υπολογιστική χημεία, στην οργανοληπτική αξιολόγηση και στις μέρες μας βρίσκει εφαρμογή στη νευροαπεικόνιση (brain imaging).

Ο όρος PLS αναφέρεται σε δυο συσχετιζόμενες μεθόδους, τη συμμετρική PLS ή PLSC (Partial Least Squares Correlation) και στην ασύμμετρη PLS ή PLSR (Partial Least Squares Regression). Στην παρούσα μελέτη θα ασχοληθούμε εκτενώς με τη δεύτερη καθώς η PLSC βρίσκει αποκλειστικά εφαρμογή στη νευροαπεικόνιση (neuroimaging).

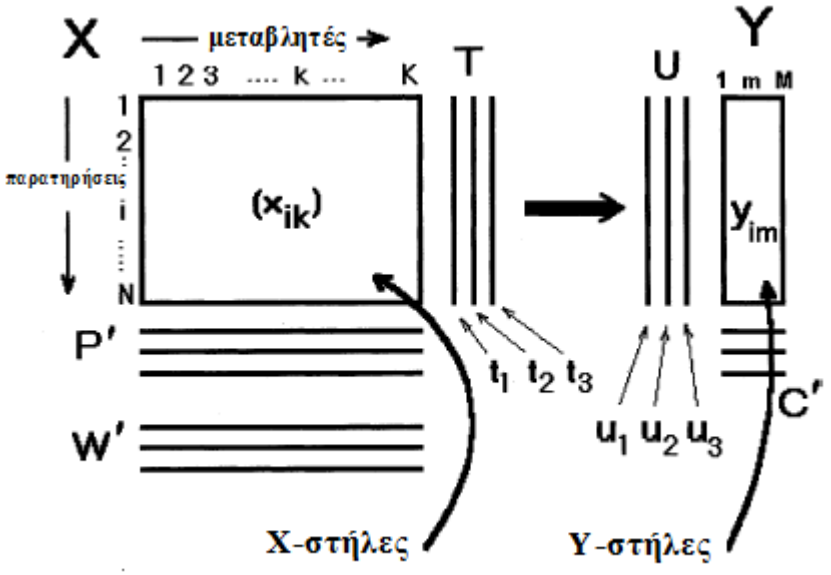
4.1 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΥ PLS ΚΑΙ ΤΟΥ PLSR ΜΟΝΤΕΛΟΥ

Ως γνωστόν η Πολλαπλή Γραμμική Παλινδρόμηση χρησιμοποιείται όταν τα δεδομένα μας περιγράφονται από πολλές μεταβλητές. Στην περίπτωση που οι μεταβλητές μας είναι αρκετά περισσότερες απ' τις παρατηρήσεις μας, λαμβάνουμε ένα μοντέλο που ναι μεν προσαρμόζεται στα πειραματικά δεδομένα, αλλά αποτυγχάνει να προβλέψει τα νέα δεδομένα. Το φαινόμενο αυτό καλείται over-fitting.

Σε αντίθεση με την Παλινδρόμηση Κυρίων Συνιστωσών, η οποία βρίσκει συνιστώσες του πίνακα X οι οποίες ερμηνεύουν καταλλήλως τον ίδιο πίνακα, η Παλινδρόμηση Μερικών Ελαχίστων τετραγώνων βρίσκει συνιστώσες του X, οι οποίες προβλέπουν με τον πλέον κατάλληλο τρόπο τον Y. Βασική προϋπόθεση είναι οι συγκεκριμένες συνιστώσες να ερμηνεύουν τη μέγιστη δυνατή συνδιασπορά μεταξύ των X και Y. Επομένως, στόχος της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων είναι να εξάγει τις λανθάνουσες μεταβλητές (latent variables), οι οποίες ερμηνεύουν τη μέγιστη της διασποράς στην απόκριση ενώ παράλληλα οδηγεί στην καλή μοντελοποίηση των αποκρίσεων. Γι' αυτό το λόγο το ακρώνυμο PLS καλείται επίσης projection to latent structures (Wold et al., 2001).

4.1.1 Το PLSR μοντέλο

Έστω ότι τα δεδομένα μας περιγράφονται από K ανεξάρτητες μεταβλητές ή αλλιώς προβλέπουσες (predictors) των N παρατηρήσεων που συλλέγονται σε έναν πίνακα X και από M εξαρτημένες μεταβλητές (αποκρίσεις) των N παρατηρήσεων συγκεντρωμένες σε έναν πίνακα Y. Πριν την ανάλυση των δεδομένων μας, αυτά συχνά υπόκεινται σε διαδικασία μετασχηματισμού (scaling), όπως έχουμε ήδη περιγράψει. Στην παρακάτω περιγραφή που θα παραθέσουμε θα είναι μετασχηματισμένα (Wold et al., 2001).



Διάγραμμα 7. Τα δεδομένα της PLSR συλλέγονται σε 2 πίνακες X ***και*** Y ***.***[13]

Στόχος του μοντέλου Μερικών Ελαχίστων Τετραγώνων είναι η εύρεση νέων μεταβλητών που καλούνται $X-scores$, οι οποίες είναι εκτιμήτριες των λανθανουσών μεταβλητών. Οι νέες μεταβλητές $X-scores$ συμβολίζονται με t_a $a=1,2,..,A$, όπου A ο αριθμός των συνιστωσών, είναι προβλέπουσες του πίνακα Y και ταυτόχρονα μοντελοποιούν τον πίνακα X .

Τα $X-scores$ είναι A τον αριθμό και ορθογώνια. Εκτιμώνται ως γραμμικοί συνδυασμοί των αρχικών X_k μεταβλητών με τους συντελεστές ή "βάρη" όπως

[13] Βλ Figure 1 στο Wold S. and Sjostrom M. and Eriksson L. (2001). PLS-Regression: a basic tool of chemometrics*, Chemometrics and Intelligent Laboratory Systems*, vol.58 p. 113. Μετατροπή σχήματος: Σταυρινίδης Σταύρος-Κων/νος.

αλλιώς ονομάζονται w^*_{ka}, $a=1,2,..,A$. Συχνά, οι συντελεστές αυτοί μπορεί να αναγράφονται και ως r_{ka}.

Ακολουθούν οι παραπάνω τύποι τόσο σε μορφή στοιχείων όσο και σε μορφή πινάκων (μέσα σε παρένθεση).

$$t_{ia} = \sum_k X_{ik} W^*_{ka} \qquad (T = XW^*) \tag{4.1}$$

Τα $X-scores$ έχουν τις ακόλουθες ιδιότητες :

Α) Πολλαπλασιάζονται με τα φορτία (loadings) p_{ak}, καλές "περιλήψεις" του X, ώστε τα X-υπόλοιπα στην παρακάτω εξίσωση να είναι μικρά :

$$X_{ik} = \sum_\alpha t_{ia} p_{ak} + e_{ik} \qquad (X = TP^{\mathrm{T}} + \mathrm{E}) \tag{4.2}$$

Όταν τα δεδομένα μας περιέχουν πάνω από μια εξαρτημένη μεταβλητή ($\mathrm{M}>1$), τότε τα αντίστοιχα Y-scores, τα οποία συμβολίζονται με u_a, $a=1,2,..,A$, πολλαπλασιάζονται με τα "βάρη" c_{am}, τα οποία είναι καλές "περιλήψεις" του Y, ώστε τα υπόλοιπα g_{ik} στην παρακάτω εξίσωση να είναι μικρά:

$$y_{im} = \sum_a u_{ia} c_{am} + g_{im} \qquad (Y = UC^{\mathrm{T}} + \mathrm{G}) \tag{4.3}$$

Β) Τα $X-scores$ είναι καλές προβλέπουσες του Y όπως φαίνεται στην ακόλουθη εξίσωση :

$$y_{im} = \sum_a t_{ia} c_{am} + f_{im} \qquad (Y = TC^{\mathrm{T}} + \mathrm{F}) \tag{4.4}$$

Τα Y-υπόλοιπα f_{im} αποτελούν τα στοιχεία του πίνακα υπολοίπων F, και ταυτόχρονα εκφράζουν τις αποκλίσεις μεταξύ των παρατηρηθέντων και των προβλεπόμενων τιμών. Η παραπάνω εξίσωση, λόγω της (4.1) μπορεί να γραφτεί ως ένα πολλαπλό γραμμικό μοντέλο Παλινδρόμησης. Πράγματι γράφεται με την ακόλουθη μορφή :

$$y_{im} = \sum_k x_{ik} w^*_{ka} \sum_a c_{am} + f_{im} = \sum_k x_{ik} b_{km} + f_{im}$$
$$(Y = XW^*C^{\mathrm{T}} + F = XB + F) \tag{4.5}$$

όπου b_{km} οι συντελεστές της PLS Παλινδρόμησης. Μπορούν να γραφτούν ως :

$$b_{km} = \sum_a w^*_{ka} c_{am} \qquad (B = W^* C^{\mathrm{T}}) \tag{4.6}$$

Οι παραπάνω b συντελεστές δεν είναι ανεξάρτητοι εκτός κι αν ο αριθμός των PLSR συνιστωσών, A, ισούται με τον αριθμό των X μεταβλητών k. Στην ειδική περίπτωση όπου έχουμε μόνο μια Y μεταβλητή δηλαδή $\mathrm{M}=1$ και ο πίνακας $X^{\mathrm{T}}\mathrm{X}$ είναι διαγώνιος, τότε απουσιάζει η δομή συσχέτισης στον πίνακα X και το μοντέλο μας εκφυλίζεται σε ένα μοντέλο Πολλαπλής Γραμμικής Παλινδρόμησης. Τότε οι συντελεστές Παλινδρόμησης τόσο για την Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων όσο και για την Πολλαπλή Γραμμική Παλινδρόμηση ισούνται με $w_1 c_1^{\mathrm{T}}$ καθώς η λύση προκύπτει για μια μόνο συνιστώσα.

Επιστρέφουμε τώρα στην περιγραφή του μοντέλου της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων. Μετά από κάθε συνιστώσα, a, ο πίνακας X υφίσταται ελάττωση (deflation), καθώς αφαιρείται η ποσότητα $t^*_{ia} p_{ak}$ από το x_{ik}. Σε μορφή πινάκων αφαιρούμε το $t_a p_a^{\mathrm{T}}$ από τον X. Έτσι το μοντέλο Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων εκφράζεται σε βάρη w_a, τα οποία αναφέρονται στα υπόλοιπα προηγούμενης διάστασης, E_{a-1}, αντί να σχετίζονται με τις X μεταβλητές. Τότε αντί της εξίσωσης (4.1) μπορούμε να χρησιμοποιήσουμε (Wold et al., 2001):

$$t_{ia} = \sum_k w_{ka} e_{ik,a-1} \qquad (t_a = E_{a-1} W_a) \tag{4.7a}$$

$$e_{ik,a-1} = e_{ik,a-2} - t_{i,a-1} p_{a-1,k} \qquad (E_{a-1} = E_{a-2} - t_{a-1} p_{a-1}^{\mathrm{T}}) \tag{4.7b}$$

$$e_{ik,0} = X_{ik} \qquad (E_0 = X) \tag{4.7c}$$

Τα βάρη w, μπορούν να μετασχηματιστούν στα w^* τα οποία σχετίζονται άμεσα με τον X, σύμφωνα με τη σχέση (4.1). Τα w και w^* σχετίζονται μεταξύ τους σύμφωνα με την :

$$W^* = W(P^{\mathrm{T}} W)^{-1} \tag{4.8}$$

Όσον αφορά τον πίνακα Y, αυτός μπορεί να ελαττωθεί αφαιρώντας την ποσότητα $t_a c_a^{\mathrm{T}}$, κάτι που όμως δεν είναι απαραίτητο καθώς τα αποτελέσματα που λαμβάνουμε παραμένουν ίδια. Παρατηρώντας τον παραπάνω αλγόριθμο που μόλις περιγράψαμε, προκύπτουν τα εξής συμπεράσματα :

το πρώτο διάνυσμα βάρους, w_1, είναι το πρώτο ιδιοδιάνυσμα του συνδυασμένου πίνακα συνδιασποράς $X^{\mathrm{T}}YY^{\mathrm{T}}X$ και τα επόμενα διανύσματα βάρους, με a συνιστώσες, είναι τα ιδιοδιανύσματα που αντιστοιχούν στις ελαττωμένες μορφές του ίδιου πίνακα που έχει τώρα τη μορφή :

$$Z_a^{\mathrm{T}}YY^{\mathrm{T}}Z_a^{\mathrm{T}} \qquad (4.9)$$

όπου $Z_a = Z_{a-1} - T_{a-1}P_{a-1}^{\mathrm{T}}$.

Ομοίως, το πρώτο διάνυσμα των $X-scores$, t_1 ,είναι ένα ιδιοδιάνυσμα του $XX^{\mathrm{T}}YY^{\mathrm{T}}$ και τα επόμενα διανύσματα των $X-scores$, δηλαδή τα t_a, είναι ιδιοδιανύσματα του πίνακα $Z_a Z_a^{\mathrm{T}}YY^{\mathrm{T}}$.

Οι σχέσεις αυτές των ιδιοδιανυσμάτων υποδεικνύουν ότι τα διανύσματα των βαρών w_a σχηματίζουν ένα ορθοκανονικό σύνολο και ότι τα διανύσματα των $X-scores$, t_a, είναι ορθογώνια μεταξύ τους. Αντιθέτως τα διανύσματα των φορτίων, p_a, όπως και τα $Y-scores$, u_a , δεν είναι ορθογώνια μεταξύ τους. Τα u_a όμως και τα p_a είναι ορθογώνια με τα t_a και τα w_a αντίστοιχα.

4.2 ΕΡΜΗΝΕΙΑ ΤΟΥ ΜΟΝΤΕΛΟΥ PLSR

Όπως έχουμε ήδη περιγράψει, το Μοντέλο Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων σχηματίζει νέες X μεταβλητές, τις εκτιμήσεις t_a , οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μας X μεταβλητών και ακολούθως χρησιμοποιεί τις t_a ως προβλέπουσες του πίνακα Y.

Έτσι βάσει του αλγορίθμου που περιγράψαμε στην προηγούμενη ενότητα καθορίζονται οι παράμετροι t, u, w ή w^*, p, c. Όσον αφορά τα $X-scores$ και τα $Y-scores$, t_a και u_a αντίστοιχα, περιέχουν τις πληροφορίες σχετικά με τις συνιστώσες καθώς και τις ομοιότητες και τις διαφορές μεταξύ τους σε αντιστοιχία πάντα με το δοθέν πρόβλημα. Σχετικά τώρα με τα βάρη w_a ή w^*_a και c_a , η συνεισφορά τους στη μελέτη συνίσταται στα ακόλουθα (Wold et al., 2001):

Παρέχουν την πληροφορία σχετικά με τον τρόπο που οι μεταβλητές συνδυάζονται προκειμένου να σχηματίσουν την ποσοτική σχέση ανάμεσα στους X και Y πίνακες. Έτσι κατορθώνουν να ερμηνεύσουν τα t_a και u_a. Επιπλέον ανάλογα με την τιμή τους, τα βάρη είναι απαραίτητα για την κατανόηση του ποιες X μεταβλητές είναι σημαντικές για την ερμηνεία του μοντέλου μας. Τέτοιες μεταβλητές έχουν αντίστοιχες υψηλές τιμές των w_a. Παρόμοιες τιμές των w_a, υποδεικνύουν ότι οι αντίστοιχες μεταβλητές τους παρέχουν παρόμοιες πληροφορίες.

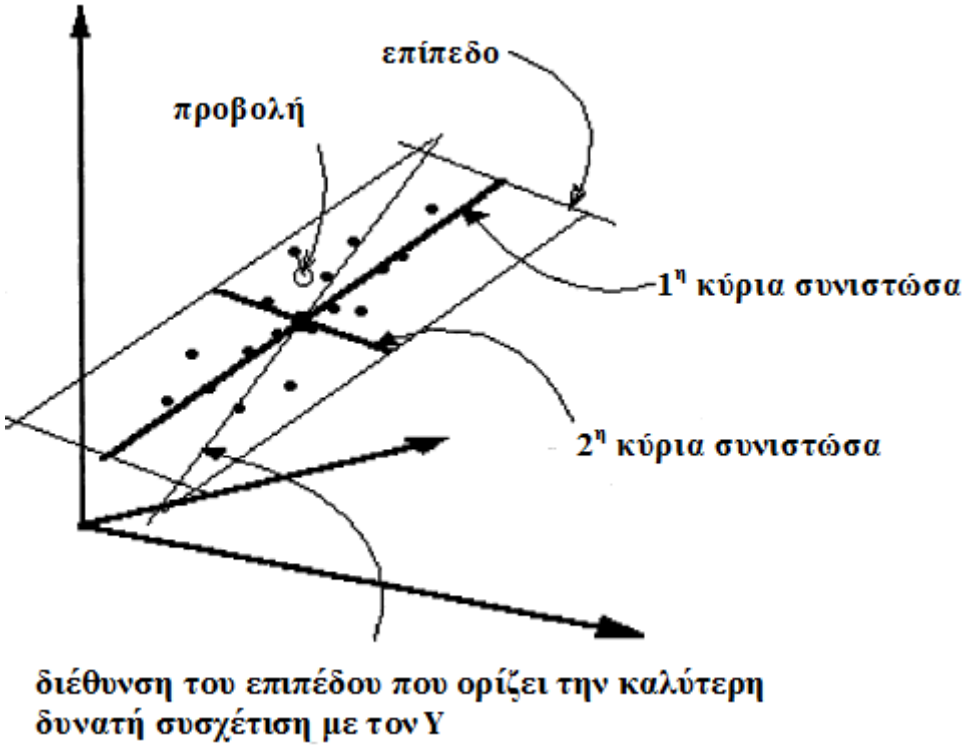
Τέλος τα βάρη w_a, εκφράζουν τόσο τις "θετικές" συσχετίσεις ανάμεσα στους X και Y, όσο και τις "συσχετίσεις αποζημίωσης" (compensation correlations) που είναι απαραίτητες για την πρόβλεψη του Y από τον X με βάση τη δευτερεύουσα διασπορά στον X. Η τελευταία είναι οποιαδήποτε πληροφορία που περιέχεται στον X και δε σχετίζεται με τον πίνακα Y.

Όσον αφορά τώρα τα υπόλοιπα (residuals), πρόκειται για το μέρος των δεδομένων μας το οποίο αδυνατεί να εξηγήσει το μοντέλο μας. Παρ' όλα αυτά, τα υπόλοιπα παρέχουν διαγνωστικό ενδιαφέρον. Συγκεκριμένα, υψηλές τιμές των Y υπολοίπων δείχνουν ότι το μοντέλο είναι ασθενές, δεν προσαρμόζεται δηλαδή κατάλληλα στα δεδομένα. Στην περίπτωση αυτή, συνίσταται η εφαρμογή ενός διαγράμματος κανονικής πιθανότητας (normal probability plot) των υπολοίπων μιας Y μεταβλητής, προκειμένου να διαπιστωθεί η ύπαρξη ακραίων τιμών (outliers) στη σχέση μεταξύ των T και Y.

Η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων σε αντιστοιχία με τα Y υπόλοιπα, περιέχει και υπόλοιπα όσον αφορά τον πίνακα X. Πρόκειται για το μέρος των δεδομένων που δε χρησιμοποιούνται στη μοντελοποίηση του πίνακα Y. Είναι χρήσιμα για τη διαπίστωση ύπαρξης ακραίων τιμών στον X χώρο.

4.3 ΓΕΩΜΕΤΡΙΚΗ ΕΡΜΗΝΕΙΑ ΤΟΥ ΜΟΝΤΕΛΟΥ PLSR

Η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων είναι μια μέθοδος προβολής και συνεπώς έχει μια απλή γεωμετρική ερμηνεία ως η προβολή ενός πίνακα X, πλήθος από N σημεία σε έναν χώρο διάστασης K, σε ένα υπερεπίπεδο διάστασης A, με τέτοιο τρόπο ώστε οι συντεταγμένες της προβολής, t_a, όπου $a = 1, 2, ..., A$, να είναι καλές προβλέπουσες του Y, όπως φαίνεται στο ακόλουθο διάγραμμα :



Διάγραμμα 8. Η γεωμετρική αναπαράσταση της PLSR. Ο X πίνακας αναπαρίσταται με N σημεία στον K -διάστατο χώρο όπου κάθε στήλη x_k του X ορίζει έναν άξονα συντεταγμένων.[14]

Η κατεύθυνση του επιπέδου εκφράζεται με τη μορφή κλίσεων, p_{ak} , της κάθε κατεύθυνσης κάθε συνιστώσας του επιπέδου σε αντιστοιχία με κάθε άξονα συντεταγμένων x_k . Η κλίση αυτή, που μόλις αναφέραμε, είναι το συνημίτονο της γωνίας μεταξύ της κάθε κατεύθυνσης και κάθε άξονα συντεταγμένων.

Με τον τρόπο αυτό, η Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων, αναπτύσσει ένα υπερεπίπεδο διάστασης A στον X χώρο, έτσι ώστε το επίπεδο αυτό να προσεγγίζει τον X και ταυτόχρονα οι θέσεις των προβαλλόμενων σημείων των δεδομένων του επιπέδου, που περιγράφονται από τα $X-scores$, t_{ia} , να σχετίζονται με τις τιμές των αποκρίσεων Y_{im} .

4.4 Ο ΑΡΙΘΜΟΣ ΤΩΝ Y-ΜΕΤΑΒΛΗΤΩΝ ΚΑΘΕ ΦΟΡΑ

Βασική ιδιότητα της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων είναι η ικανότητα της να μοντελοποιεί και να αναλύει πολλές Y μεταβλητές ταυτόχρονα, κάτι που μας δίνει μια καλύτερη εικόνα απ' ότι όταν έχουμε ένα ξεχωριστό μοντέλο

[14] Ό.π . Figure 2, p. 115.

για κάθε μια Y μεταβλητή. Στην περίπτωση όμως που έχουμε συσχετισμένες Y μεταβλητές, αυτές οφείλουν να αναλυθούν ταυτόχρονα στο ίδιο μοντέλο. Εάν όμως οι μετρήσεις των Y μεταβλητών, αφορούν διαφορετικά αντικείμενα, δηλαδή είναι σχεδόν ανεξάρτητες, τότε το PLSR μοντέλο περιέχει πολλές συνιστώσες και επομένως είναι αρκετά δύσκολο να ερμηνευθεί. Τότε συνίσταται η ξεχωριστή μοντελοποίηση των Y μεταβλητών, λαμβάνοντας έτσι ένα σύνολο απλούστερων μοντέλων λιγότερων διαστάσεων, που είναι ευκολότερα στην ερμηνεία (Wold et al., 2001).

4.5 Ο ΑΡΙΘΜΟΣ ΤΩΝ PLS ΣΥΝΙΣΤΩΣΩΝ, A

Όπως σε κάθε μοντέλο Παλινδρόμησης, έτσι και στην Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων, απαραίτητος θεωρείται ο καθορισμός της σωστής πολυπλοκότητας (complexity) του μοντέλου μας. Λόγω της ύπαρξης πολλών και συσχετισμένων μεταξύ τους X μεταβλητών, ελλοχεύει ο κίνδυνος να πάρουμε ένα υπερορισμένο μοντέλο, δηλαδή ένα μοντέλο ναι μεν καλά προσαρμοσμένο στα δεδομένα μας, αλλά με λίγη έως μηδαμινή ικανότητα πρόβλεψης. Επομένως απαραίτητη είναι η κατασκευή ενός ελέγχου όπου θα διαπιστώνεται η σημαντικότητα στην πρόβλεψη κάθε PLS συνιστώσας. Μάλιστα, όταν οι εξαγόμενες συνιστώσες παύουν να είναι σημαντικές θα σταματάμε. Αυτός ο έλεγχος δύναται να πραγματοποιηθεί μέσω της τεχνικής του CV (Cross Validation).

Πρόκειται για μια ευρέως διαδεδομένη τεχνική, σημαντική για την ανάλυση της PLSR, μέσω της οποίας τα δεδομένα χωρίζονται σε έναν αριθμό ομάδων, έστω G και ακολούθως αναπτύσσεται ένα πλήθος παράλληλων μοντέλων με μειωμένα δεδομένα, καθώς μια ομάδα διαγράφεται κάθε φορά. Στην περίπτωση που έχουμε $G = N$, δηλαδή ο αριθμός των ομάδων συμπίπτει με τον αριθμό των παρατηρήσεων, τότε συνίσταται η αποφυγή εφαρμογής της συγκεκριμένης τεχνικής (Wold et al., 2001).

Έτσι κάθε φορά που αναπτύσσεται ένα μοντέλο μέσω της CV τεχνικής, υπολογίζονται οι διαφορές μεταξύ των αρχικών και των προβλεπόμενων Y τιμών για τα δεδομένα που έχουν διαγραφεί. Ακολούθως υπολογίζεται το άθροισμα των τετραγώνων αυτών των διαφορών, απ' όλα τα παράλληλα μοντέλα, το οποίο δίνει το PRESS (Predicted-Residual-Sum-Of-Squares). Το PRESS εκτιμά την ικανότητα πρόβλεψης του μοντέλου μας και το οποίο θα περιγράψουμε στη συνέχεια. Η χρήση της τεχνικής της CV μπορεί να πραγματοποιηθεί επαναληπτικά (sequential mode). Τότε το CV εφαρμόζεται σε κάθε συνιστώσα μετά την άλλη, αλλά η διαδικασία της αποφλοίωσης (peeling off) που περιγράφτηκε στη σχέση (4.7b) πραγματοποιείται μόνο μια φορά, στους αρχικούς πίνακες που περιέχουν όλα τα δεδομένα (full data matrices).

Στη συνέχεια οι εναπομείναντες πίνακες υπολοίπων E και F διαχωρίζονται σε ομάδες, προκειμένου να εφαρμοστεί η CV για την επόμενη συνιστώσα. Μετά από κάθε συνιστώσα, υπολογίζεται ο λόγος $PRESS_a / SS_{a-1}$, όπου μια συνιστώσα

χαρακτηρίζεται σημαντική εάν ο παραπάνω λόγος είναι μικρότερος από περίπου 0.9. Εναλλακτικά χρησιμοποιούνται οι υποθέσεις των αποτελεσμάτων των Waking και Morris. Η παραπάνω ποσότητα SS_{a-1} επεξηγεί το προσαρμοσμένο άθροισμα τετραγώνων των υπολοίπων πριν την τρέχουσα συνιστώσα, εδώ την a. Οι υπολογισμοί των παραπάνω λόγων συνεχίζονται μέχρι μια συνιστώσα να κριθεί ασήμαντη (non-significant).

Μια άλλη δυνατή επιλογή είναι η χρήση του "total CV". Στην περίπτωση αυτή, αρχικά χωρίζουμε τα δεδομένα μας σε ομάδες και ακολούθως υπολογίζουμε το $PRESS$ για κάθε συνιστώσα, πραγματοποιώντας ξεχωριστή διαδικασία αποφλοίωσης των πινάκων κάθε ομάδας CV. Τότε χρησιμοποιείται εκείνο το μοντέλο με τον αντίστοιχο αριθμό συνιστωσών που δίνει τη μικρότερη τιμή του λόγου $PRESS/(N-A-1)$.

Συνοψίζοντας, το $PRESS$ υπολογίζεται και με τους δυο παραπάνω τρόπους προκειμένου να αποφασίσουμε τον αριθμό των συνιστωσών που θα διατηρήσουμε και θα μελετήσουμε. Αυτό μπορεί να εκφραστεί ομοίως με το Q^2 (cross validated R^2), όπου $Q^2=1-PRESS/SS$, με SS να είναι το άθροισμα τετραγώνων του Y διορθωμένου για το μέσο (corrected for the mean). Η παραπάνω ποσότητα μπορεί να συγκριθεί με το $R^2=(1-RSS/SS)$, όπου το RSS είναι το προσαρμοσμένο άθροισμα τετραγώνων των υπολοίπων. Προφανώς, σε περίπτωση που έχουμε έναν αριθμό Y μεταβλητών, υπολογίζονται οι ποσότητες R_m^2 και Q_m^2 για κάθε μια απ' τις Y μεταβλητές, y_m. Στο σημείο αυτό θα επιχειρήσουμε να δώσουμε μια αναλυτική περιγραφή του $PRESS$ καθώς και του Cross-Validation, τα οποία αναφέρθηκαν παραπάνω.

4.5.1 Περιγραφή του στατιστικού PRESS

Η χρήση του στατιστικού $PRESS$, όπως έχει τονιστεί, συνίσταται στη μέτρηση της εγκυρότητας ενός μοντέλου Παλινδρόμησης (regression model validity) αλλά και στη δυνατότητα απόδοσης της πρόβλεψης (potential performance in prediction) (Montgomery et al., 2006). Το $PRESS$ υπόλοιπο δίνεται απ' τη σχέση :

$$e_{(i)} = y_i - \hat{y}_{(i)} \qquad (4.10)$$

όπου $\hat{y}_{(i)}$ είναι η προβλεπόμενη τιμή ενός προσαρμοσμένου μοντέλου, όταν έχει παρακρατηθεί η i-στη παρατήρηση. Τότε :

$$PRESS = \sum_{i=1}^{n} \mathrm{e}_{(i)}{}^{2} = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2 \qquad (4.11)$$

Παρ' όλο που αρχικά φαίνεται ότι για τον υπολογισμό του $PRESS$, απαιτείται η προσαρμογή n αριθμού Παλινδρομήσεων, είναι εφικτό να υπολογίσουμε την τιμή του λαμβάνοντας τα αποτελέσματα μιας προσαρμογής ελαχίστων τετραγώνων σε όλες τις n παρατηρήσεις. Για την καλύτερη περιγραφή, έστω $\hat{\beta}_{(i)}$ το διάνυσμα των συντελεστών Παλινδρόμησης έχοντας κατακρατήσει την i-στη παρατήρηση. Τότε έχουμε :

$$\hat{\beta}_{(i)} = [\mathrm{X}_{(i)}{}^{\mathrm{T}} \mathrm{X}_{(i)}]^{-1} \mathrm{X}_{(i)}{}^{\mathrm{T}} y_{(i)} \qquad (4.12)$$

όπου $\mathrm{X}_{(i)}$ και $y_{(i)}$ τα X και $w^{*}c$ διανύσματα με την i-στη παρατήρηση να έχει παρακρατηθεί. Έτσι το i-στο $PRESS$ υπόλοιπο μπορεί να γραφτεί ως :

$$\mathrm{e}_{(i)} = y_i - \hat{y}_{(i)} = y_i - x_i^{T} \hat{\beta}_{(i)} = y_i - x_i^{T} (\mathrm{X}_{(i)}{}^{\mathrm{T}} \mathrm{X}_{(i)})^{-1} \mathrm{X}_{(i)}{}^{\mathrm{T}} y_{(i)} \qquad (4.13)$$

Υπάρχει μια στενή σχέση μεταξύ των $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$ και $[\mathrm{X}_{(i)}{}^{\mathrm{T}} \mathrm{X}_{(i)}]^{-1}$ πινάκων. Συγκεκριμένα ισχύει :

$$[\mathrm{X}_{(i)}{}^{\mathrm{T}} \mathrm{X}_{(i)}]^{-1} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} + \frac{(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} x_i x_i{}^{\mathrm{T}} (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}}{1 - h_{ii}} \qquad (4.14)$$

με $h_{ii} = x_i^{\mathrm{T}} (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} x_i$ να είναι τα διαγώνια στοιχεία του πίνακα Η. Χρησιμοποιώντας την (4.14) έχουμε :

$$\begin{aligned} e_{(i)} &= y_i - x_i^{\mathrm{T}} [(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} + \frac{(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} x_i x_i{}^{\mathrm{T}} (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}}{1 - h_{ii}}] \mathrm{X}_{(i)}{}^{\mathrm{T}} y_{(i)} \\ &= y_i - x_i^{\mathrm{T}} (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \mathrm{X}_{(i)}{}^{\mathrm{T}} y_{(i)} - \frac{x_i^{\mathrm{T}} (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} x_i x_i{}^{\mathrm{T}} (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} \mathrm{X}_{(i)}{}^{\mathrm{T}} y_{(i)}}{1 - h_{ii}} \end{aligned}$$

$$= \frac{(1-h_{ii})y_i - (1-h_{ii})x_i^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}_{(i)}{}^{\mathrm{T}}y_{(i)} - h_{ii}x_i^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}_{(i)}{}^{\mathrm{T}}y_{(i)}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - x_i^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}_{(i)}{}^{\mathrm{T}}y_{(i)}}{1-h_{ii}}.$$

Επειδή όμως $\mathrm{X}^{\mathrm{T}}y = \mathrm{X}_{(i)}{}^{\mathrm{T}}y_{(i)} + x_i y_i$ η παραπάνω ισότητα γράφεται :

$$e_{(i)} = \frac{(1-h_{ii})y_i - x_i^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}(\mathrm{X}^{\mathrm{T}}y - x_i y_i)}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - x_i^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}y + x_i^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}x_i y_i}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - x_i^{\mathrm{T}}\hat{\beta} + h_{ii}y_i}{1-h_{ii}}$$

$$= \frac{y_i - x_i^{\mathrm{T}}\hat{\beta}}{1-h_{ii}} \qquad (4.15)$$

Έτσι ο αριθμητής της παραπάνω σχέσης αποτελεί το σύνηθες (ordinary) υπόλοιπο e_i το οποίο προκύπτει από μια προσαρμογή ελαχίστων τετραγώνων σε όλες τις n παρατηρήσεις. Συνεπώς το i-στο $PRESS$ υπόλοιπο ισούται με :

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \qquad (4.16)$$

Επομένως, καθώς το $PRESS$ είναι το άθροισμα τετραγώνων των υπολοίπων των $PRESS$ υπολοίπων, ένας απλός υπολογιστικός τύπος είναι :

$$PRESS = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2 \qquad (4.17)$$

4.5.2 Περιγραφή του Cross Validation

Πρόκειται για μια συνήθη στατιστική τεχνική, η οποία περιλαμβάνει επαναχρησιμοποίηση (reusing) καθώς και εκ νέου δειγματοληψία (resampling) των δεδομένων μας. Είναι μια μέθοδος εκτίμησης του σφάλματος πρόβλεψης όταν προσαρμόζουμε μια συνάρτηση που σχετίζεται με δυο ή περισσότερες μεταβλητές. Για να δούμε την τιμή του Cross Validation, πρέπει πρώτα να εκτιμήσουμε τη διαφορά μεταξύ του προσαρμοσμένου σφάλματος (fitted error) και του σφάλματος πρόβλεψης (predictive error) (Alun, 2009).

Για παράδειγμα, καθώς εφαρμόζουμε την Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων του Y στον X, ένα δείγμα παρατηρήσεων αποτελούμενο από δυο μεταβλητές, $(X_1,Y_1),....,(X_n,Y_n)$, ελαχιστοποιούμε το προσαρμοσμένο σφάλμα :

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-(\hat{a}+\hat{b}X_i))^2 \qquad (4.18)$$

Παρ' όλα αυτά, εάν βασικό μας μέλημα αποτελεί η πρόβλεψη της επόμενης τιμής του Y, δηλαδή της Y_{n+1}, καθώς μας δίνεται μόνο η επόμενη τιμή του X, δηλαδή η X_{n+1}, η συνάρτηση που καλούμαστε να ελαχιστοποιήσουμε είναι η :

$$Y_{n+1}-(\hat{a}+\hat{b}X_{n+1}))^2 \qquad (4.19)$$

όπου τα $\hat{a}$ και $\hat{b}$ έχουν επιλεγεί δίχως τη χρησιμοποίηση των τιμών των X_{n+1} και Y_{n+1}. Μια συχνά χρησιμοποιούμενη τεχνική εκτίμησης του σφάλματος πρόβλεψης είναι ο διαχωρισμός του δείγματος στη μέση. Το πρώτο μισό θα χρησιμοποιηθεί για τον υπολογισμό του προσαρμοσμένου σφάλματος ενώ το δεύτερο μισό θα χρησιμοποιηθεί για την εκτίμηση του σφάλματος πρόβλεψης. Έτσι θα έχουμε ένα δείγμα για προσαρμογή (fitting) και ένα για επικύρωση (validation), γνωστά και ως training set και testing set αντίστοιχα (Alun, 2009).

Ένα τέτοιο εγχείρημα όμως μπορεί να είναι πολυδάπανο. Πρέπει όμως να προσαρμόσουμε ένα καλύτερο μοντέλο συμπεριλαμβάνοντας όλα μας τα δεδομένα. Συνεπώς εμφανίζεται το ερώτημα του αν μπορούμε ακόμα να πάρουμε μια εκτίμηση για το σφάλμα πρόβλεψης. Γι' αυτόν ακριβώς το σκοπό έχει επινοηθεί η τεχνική του Cross Validation. Η τεχνική αυτή ακολουθεί τα εξής στάδια :

- Δημιουργία ενός δείγματος δυο μεταβλητών $(X_1,Y_1),....,(X_n,Y_n)$
- Για $i=1$ έως n

-αφήνει το σημείο (X_i,Y_i) εκτός του δείγματος,

-προσαρμόζει το μοντέλο $Y = f_{-i}(X)$ χρησιμοποιώντας τα εναπομείναντα σημεία,

-υπολογίζει την προβλεπόμενη τιμή του αποκλεισμένου σημείου :

$$\hat{Y}_{-i} = f_{-i}(X_i) \tag{4.20}$$

-εκτιμά το σφάλμα πρόβλεψης χρησιμοποιώντας :

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_{-i})^2 \tag{4.21}$$

Παρατηρούμε ότι το προσαρμοσμένο και το σφάλμα πρόβλεψης μοιάζουν αρκετά. Η τεχνική του Cross Validation δε χρησιμοποιείται κυρίως για την εκτίμηση των παραμέτρων μιας προσαρμογής, αλλά για την επιλογή μεταξύ διαφορετικών μοντέλων. Για παράδειγμα ας υποθέσουμε ότι έχουμε να εξετάσουμε δυο μοντέλα, μια γραμμική και μια τετραγωνική προσαρμογή όπως φαίνεται ακολούθως :

$$\begin{aligned} Y &= a + bX \\ Y &= a + bX + cX^2 \end{aligned} \tag{4.22}$$

Είναι γνωστό ότι το προσαρμοσμένο σφάλμα για το τετραγωνικό μοντέλο θα είναι πάντα μικρότερο απ' το αντίστοιχο σφάλμα του γραμμικού μοντέλου. Απεναντίας το σφάλμα πρόβλεψης ενδέχεται να μην είναι μικρότερο και αυτό μπορεί να χρησιμοποιηθεί ως κριτήριο επιλογής του μοντέλου.

Ειδικότερα, στο 5^{ο} Κεφάλαιο της παρούσης, κατά την πρακτική εφαρμογή των μεθόδων Παλινδρόμησης PCR και PLSR, καλούμαστε να κρίνουμε με βάση την κανονική εκτιμήτρια CV και τη διορθωμένη εκτιμήτρια adjCV των Mevik και Cederkvist (2004) την εγκυρότητα του κάθε μοντέλου που έχουμε προσαρμόσει. Συγκεκριμένα, τα validation results που μας επιστρέφει η R είναι με την RMSEP μορφή (root mean squared error of prediction).

Στο σημείο αυτό θα παραθέσουμε τις εκτιμήτριες MSEP (mean squared error of prediction) των Mevik και Cederkvist (2004) με τη διαφορά ότι δε λαμβάνεται η τετραγωνική ρίζα τους, δεν είναι δηλαδή root mean squared. Έστω ότι έχουμε ένα σύνολο δεδομένων $L = \{(x_i, y_i)\}$ αποτελούμενο από n_L παρατηρήσεις και μια προβλέπουσα (predictor) f_L πάνω στο σύνολο L, με το σύνολο L να αποτελεί ένα τυχαίο δείγμα μιας συγκεκριμένης κατανομής.

Για την κανονική εκτιμήτρια CV κάνουμε τα εξής :

Χωρίζουμε το σύνολο δεδομένων L, τυχαία σε K το πλήθος τμήματα L_k, όπου $k = 1, 2, ..., K$ ιδίου μεγέθους με f_k να είναι η προβλέπουσα πάνω στο $L \setminus L_k$,

δηλαδή σε όλες τις παρατηρήσεις που δεν ανήκουν στο L_k. Η K-fold cross validation εκτιμήτρια είναι η :

$$MSEP_{cv.K} = \frac{1}{n_L}\sum_{k=1}^{K}\sum_{i\in L_k}(f_k(x_i)-y_i)^2 \qquad (4.23)$$

όπου το εσωτερικό άθροισμα λαμβάνεται στις παρατηρήσεις του k-στου τμήματος. Η εκτιμήτρια αυτή συχνά καλείται mean squared error of cross-validation (MSECV). Πολλοί συγγραφείς ορίζουν την K-fold cross-validation εκτιμήτρια ως :

$$(1/K)\sum_{k=1}^{K}(1/\#L_k)\sum_{i\in L_k}(f_k(x_i)-y_i)^2 \qquad (4.24)$$

όπου $\#L_k$ είναι το μέγεθος του k-στου τμήματος. Παρ' όλα αυτά, η διαφορά είναι μηδαμινή εάν τα τμήματα είναι παραπλήσιου μεγέθους. Η μεροληψία της $MSEP_{CV.K}$ είναι της τάξης του $(K-1)^{-1}n_L^{-1}$.

Η τεχνική leave-one-out cross-validation που βρίσκει εφαρμογή στο 5^{Ο} Κεφάλαιο, είναι η K-fold cross-validation τεχνική με $K = n_L$. Έχει αποδειχθεί ότι τεχνική leave-one-out cross-validation είναι ασυμπτωτικά άριστη για επιλογή του βέλτιστου μοντέλου στην Κλασική Παλινδρόμηση (OLS) και αυτό διότι ο MSEP του επιλεγμένου μοντέλου προσεγγίζει το ελάχιστο (minimal) MSEP. Απεναντίας δεν είναι ασυμπτωτικά ευσταθής τεχνική για την επιλογή μεταβλητών καθώς τείνει να συμπεριλαμβάνει υπερβολικά πολλές εξ αυτών. Οφείλεται στο γεγονός ότι ο MSEP συνήθως αυξάνει ελαφρώς όταν συμπεριλαμβάνονται μερικές μη-σημαντικές μεταβλητές, ενώ αυξάνει πολύ καθώς αφαιρούνται σημαντικές μεταβλητές.

Για τη διορθωμένη εκτιμήτρια adjCV έχουμε :

Ως γνωστόν, στην K-fold cross-validation τεχνική οι προβλέπουσες (predictors) εφαρμόζονται πάνω στα υποσύνολα του L και επομένως αναμένεται να λειτουργούν χειρότερα απ' ότι αν εφαρμόζονταν σε όλο το L, ειδικά όταν $K << n_L$. Αυτό μπορεί να οδηγήσει σε ένα υπερτιμημένο MSEP. Για το λόγο αυτό χρησιμοποιείται η τεχνική Adjusted K-fold cross-validation, της οποίας η προσαρμογή είναι η εξής :

$$MSEP_{adj} = MSEP_{app} - \frac{1}{n_L}\sum_{k=1}^{K}\frac{n_k}{n_L}\sum_{i\notin L_k}(f_k(x_i)-y_i)^2 \qquad (4.25)$$

όπου $MSEP_{app}$ (apparent $MSEP$), η mean square error εκτίμητρια η οποία χρησιμοποιεί το σύνολο L και δίνεται από τη σχέση :

$$MSEP_{app} = \frac{1}{n_L}\sum_{i=1}^{n_L}(f_L(x_i)-y_i)^2 \qquad (4.26)$$

Με το άθροισμα να λαμβάνεται πάνω στο L. Επιστρέφουμε τώρα στην περιγραφή της τεχνικής Adjusted K-fold cross-validation. Στη σχέση (4.25), n_k είναι το μέγεθος του k-στου τμήματος και το εσωτερικό άθροισμα λαμβάνεται στο $L \setminus L_k$. Αυτή είναι η διαφορά στο apparent MSEP μεταξύ της προβλέπουσας σε όλο το L και του σταθμισμένου μέσου όρου των προβλεπουσών πάνω στο $L \setminus L_k$.

Η adjusted cross-validation εκτίμητρια είναι :

$$MSEP_{cv.K} = \frac{1}{n_L} \sum_{k=1}^{K} \sum_{i \in L_k} (f_k(x_i) - y_i)^2 \qquad (4.27)$$

4.6 ΟΙ PLSR ΑΛΓΟΡΙΘΜΟΙ

Στην παράγραφο αυτή, θα επιχειρήσουμε να περιγράψουμε τους αλγόριθμους που χρησιμοποιούνται για τον υπολογισμό του PLSR μοντέλου. Διάφορες παραλλαγές έχουν αναπτυχθεί για διαφορετικά είδη δεδομένων. Η πλειοψηφία αυτών όμως ανέχεται μέτριο ποσοστό ελλειπόντων δεδομένων. Έτσι ο αλγόριθμος NIPALS, τον οποίο θα παραθέσουμε παρακάτω χρησιμοποιεί τους αρχικούς πίνακες X και Y, οι οποίοι όμως έχουν ήδη υποστεί διαδικασία τυποποίησης (scaling και centering), ενώ ο αλγόριθμος kernel τους πίνακες διασποράς-συνδιασποράς, $X^{\mathrm{T}}X$, $Y^{\mathrm{T}}Y$ και $X^{\mathrm{T}}Y$. Εναλλακτικά ο kernel–αλγόριθμος χρησιμοποιεί τους πίνακες XX^{T} και YY^{T}, κάτι το οποίο όμως είναι επικίνδυνο όταν ο αριθμός των παρατηρήσεων N διαφέρει σημαντικά απ' τον αριθμό των μεταβλητών K και M. Ακολούθως παραθέτουμε τον αλγόριθμο NIPALS, ο οποίος προτάθηκε απ' τον Wold, τον Sjostrom και τον Eriksson (2001).

4.6.1 Ο αλγόριθμος NIPALS

Ο αλγόριθμος NIPALS ξεκινά με προαιρετικά μετασχηματισμένα δεδομένα X και Y τα οποία είναι ενδεχομένως τυποποιημένα και κεντροποιημένα. Σε περίπτωση μιας μόνο Y-μεταβλητής ο αλγόριθμος παύει να είναι επαναληπτικός. Τα βήματα που ακολουθούνται είναι τα εξής :

1° βήμα

Πάρε ένα αρχικό διάνυσμα y, που συνήθως είναι μια στήλη του Y και ονόμασε το u.

2ο βήμα
Υπολόγισε τα X -βάρη , w , όπου :

$$w = \frac{X^{\mathrm{T}} u}{u^{\mathrm{T}} u} \tag{4.28}$$

(εδώ το w μπορεί να τροποποιηθεί στη νόρμα w , με $\|w\| = 1.0$).

3ο βήμα
Υπολόγισε τα $X - scores$, t , όπου :

$$t = Xw \tag{4.29}$$

4ο βήμα
Υπολόγισε τα Y -βάρη, c , όπου :

$$c = \frac{Y^{\mathrm{T}} t}{t^{\mathrm{T}} t} \tag{4.30}$$

5ο βήμα
Υπολόγισε ένα νέο πλήθος $Y - scores$, u , όπου :

$$u = \frac{Yc}{c^{\mathrm{T}} c} \tag{4.31}$$

6ο βήμα
Έλεγξε τη σύγκλιση του t .Υπολόγισε δηλαδή την ποσότητα :

$$\|t_{old} - t_{new}\| \Big/ \|t_{new}\| \tag{4.32}$$

και εάν η ποσότητα αυτή είναι μικρότερη από μια προκαθορισμένη μικρή ποσότητα, ε , όπου ε της τάξεως του 10^{-6} ή του 10^{-8}, τότε η σύγκλιση επιτυγχάνεται. Εάν δεν συγκλίνει επέστρεψε στο 2ο βήμα, ενώ αν συγκλίνει συνέχισε στο 7ο βήμα και μετά στο 1ο βήμα. Στην περίπτωση ύπαρξης μιας μόνο

Y μεταβλητής, δηλαδή $\mathrm{M}=1$, η διαδικασία συγκλίνει σε μια μόνο επανάληψη και ακολουθείται το 7^{ο} βήμα.

7^{ο} βήμα

Αφαίρεσε (deflate, peel off) την τρέχουσα συνιστώσα από τα X και Y και κατόπιν χρησιμοποίησε τους deflated πίνακες αυτούς ως τους X και Y πίνακες στην επόμενη συνιστώσα. Πρέπει να σημειωθεί εδώ, ότι η διαδικασία αυτή όσον αφορά τον Y πίνακα είναι προαιρετική καθώς τα αποτελέσματα παραμένουν σε κάθε περίπτωση τα ίδια.

$$p = \frac{X^{\mathrm{T}}t}{t^{\mathrm{T}}t} \qquad (4.33)$$

$$X = X - tp^{\mathrm{T}} \qquad (4.34)$$

$$Y = Y - tc^{\mathrm{T}} \qquad (4.35)$$

8^{ο} βήμα

Συνέχισε με την επόμενη συνιστώσα (πίσω στο 1^{ο} βήμα) μέχρι η τεχνική Cross-Validation να υποδείξει ότι δεν υπάρχει περαιτέρω σημαντική πληροφορία στον X σχετικά με τον Y.

4.7 ΤΥΠΙΚΑ ΣΦΑΛΜΑΤΑ ΚΑΙ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Πολλές προσπάθειες έχουν γίνει για την θεωρητική άντληση διαστημάτων εμπιστοσύνης των PLSR παραμέτρων, οι περισσότερες των οποίων βασιζόμενες σε παραδοχές Παλινδρόμησης θεωρούν το PLSR μοντέλο ως ένα μεροληπτικό μοντέλο παλινδρόμησης. Μόλις πρόσφατα, χάρη στον Burnham το ζήτημα αυτό μελετήθηκε ενδελεχώς, με την Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων ως ένα μοντέλο παλινδρόμησης λανθανουσών μεταβλητών.

Ένας τρόπος εκτίμησης των τυπικών σφαλμάτων και των διαστημάτων εμπιστοσύνης είναι μέσω του jack-knifing. Συστάθηκε απ' τον Wold και πρόσφατα επεξεργάστηκε απ' τους H.Martens και M.Martens (2000). Κεντρική ιδέα είναι ότι η διασπορά που ερμηνεύεται απ' τις παραμέτρους των υπομοντέλων, χάρη στην

εφαρμογή της τεχνικής Cross Validation, χρησιμοποιείται για την άντληση των τυπικών αποκλίσεων ή τυπικών σφαλμάτων όπως συνηθέστερα καλούνται. Ακολούθως με τη βοήθεια της t-κατανομής κατασκευάζονται τα διαστήματα εμπιστοσύνης (confidence intervals). Επειδή όλες οι PLSR παράμετροι (scores, loadings) είναι γραμμικοί συνδυασμοί των αρχικών πιθανώς αποφλοιωμένων δεδομένων (deflated data), οι εκτιμήσεις των παραμέτρων αυτών ακολουθούν κατά προσέγγιση την Κανονική κατανομή και επομένως η τεχνική jack-knifing αποδίδει σωστά (Wold et al., 2001). Στο σημείο αυτό, θα κάνουμε μια σύντομη περιγραφή της τεχνικής του jack-knifing.

4.7.1 Η τεχνική Jack-knifing

Πρόκειται για μια συνήθη στατιστική τεχνική, η οποία περιλαμβάνει επαναχρησιμοποίηση καθώς και εκ νέου δειγματοληψία των δεδομένων μας. Ειδικότερα αποτελεί μια μέθοδο ελάττωσης της μεροληψίας καθώς και αξιολόγησης της διασποράς μιας εκτιμήτριας (Alun, 2009).

Έστω τώρα ότι τα δεδομένα μας είναι οι μεταβλητές $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$. Η τεχνική αυτή αντί να δημιουργήσει ένα σύνολο τυχαίων δειγμάτων απ' τα δεδομένα μας, παράγει n δείγματα μεγέθους $n-1$ αποκλείοντας μια παρατήρηση κάθε φορά. Ακολουθεί τα εξής στάδια :

- Κατασκευάζει ένα δείγμα $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_n$.
- Υπολογίζει μια συνάρτηση των δεδομένων, την $\hat{\theta}(\mathrm{X})$ η οποία εκτιμά μια παράμετρο θ του μοντέλου.
- Για $i = 1$ έως n

-κατασκευάζει ένα jackknife δείγμα $X^{-i} = \{\mathrm{X}_1, ...\mathrm{X}_{i-1}, \mathrm{X}_{i+1}, ..., \mathrm{X}_n\}$ αποκλείοντας την i-στη παρατήρηση,

-υπολογίζει την $\hat{\theta}_{-i}$ εφαρμόζοντας την διαδικασία εκτίμησης στο jackknife δείγμα.

- Υπολογίζει την jackknife εκτιμήτρια $\hat{\theta}_* = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}$ και την jackknife εκτιμήτρια της διασποράς $\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{-i} - \hat{\theta}_*)^2$.

Η τεχνική jackknife μπορεί να χρησιμοποιηθεί προκειμένου να εκτιμήσουμε τη διασπορά μιας εκτιμήτριας αλλά και για να εκτιμήσουμε τη μεροληψία της. Η εκτιμήτρια της μεροληψίας ισούται με :

$$(n-1)(\hat{\theta}_* - \hat{\theta}) \tag{4.36}$$

Έτσι λαμβάνουμε τη διορθωμένη μεροληπτική jackknife εκτιμήτρια :

$$\tilde{\theta} = \hat{\theta} - (n-1)(\hat{\theta}_* - \hat{\theta}) = n\hat{\theta} - (n-1)\hat{\theta}_* \tag{4.37}$$

4.8 ΠΑΡΑΔΟΧΕΣ ΣΤΙΣ ΟΠΟΙΕΣ ΒΑΣΙΖΕΤΑΙ Η PLSR

Όπως κάθε ανάλυση δεδομένων, έτσι και η PLSR βασίζεται στην παραδοχή της ομοιογένειας (homogeneity). Πρόκειται για την κατάσταση εκείνη, όπου το πρόβλημα ή η διαδικασία που μελετάμε, οφείλει να βρίσκεται σε μια σταθερή κατάσταση καθ' όλη τη διάρκεια και ο μηχανισμός επίδρασης του X στον Y να μην διαφοροποιείται. Επομένως, οφείλουν να παρατηρούνται ορισμένα όρια στη μεταβλητότητα και την ποικιλομορφία τους (Wold et al., 2001).

Σκοπός της PLSR ανάλυσης συνεπώς, είναι η πραγματοποίηση διαγνωστικών ελέγχων προκειμένου να διαπιστωθεί αν και σε ποιο βαθμό τα παραπάνω εκπληρούνται. Η πρόσφατη έρευνα στην εφαρμοσμένη στατιστική συμπεριλαμβάνει τη δημιουργία τέτοιων διαγνωστικών μεθόδων και ορισμένες εξ' αυτών βρίσκουν εφαρμογή στην PLSR. Επιπρόσθετοι μέθοδοι του ιδίου περιεχομένου απαιτούν την κατασκευή διαγραμμάτων όπου χρησιμοποιούνται τα $X-scores$ καθώς και τα $loadings$ (φορτία) στα οποία έχει ήδη γίνει νύξη.

Στο σημείο αυτό, θα παραθέσουμε ένα αριθμητικό παράδειγμα χημικού ενδιαφέροντος, όπου θα γίνει εφαρμογή της μεθόδου Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων. Θα εξεταστεί η ποσοτική σχέση δομής-δραστηριότητας ενός συνόλου πεπτιδίων προκειμένου να οδηγηθούμε στην ανάπτυξη, ερμηνεία και τελειοποίηση ενός PLSR μοντέλου. Για το λόγο αυτό, το πρόβλημα μας καλείται QSAR (Quantitative Structure-Activity Relationship). Σε κάθε QSAR πρόβλημα, η μεταβλητή απόκρισης Y περιγράφει τις χημικές ιδιότητες, σε αντίθεση με τις τον πίνακα X, ο οποίος περιέχει την ποσοτική περιγραφή της διασποράς στη χημική δομή μεταξύ των ενώσεων που θα μελετήσουμε.

Το ακόλουθο παράδειγμα περιγράφεται μέσω μιας μόνο Y μεταβλητής και επτά X μεταβλητών. Σκοπός μας είναι η κατανόηση της διασποράς της Y-μεταβλητής DDGTS, η οποία περιγράφει την παραγόμενη ενέργεια μιας πρωτεΐνης (σύνθεση μιας μονάδας του βακτηριοφάγου T4 από την τρυπτοφάνη), όταν η θέση 49 υπόκειται σε μετατροπή προκειμένου να περιέχει ένα απ' τα 19 κωδικοποιημένα αμινοξέα AA πλην της αργινίνης. Τα 19 αμινοξέα περιγράφονται

από επτά υψηλά συσχετισμένες X μεταβλητές όπως φαίνεται στον πίνακα 10 που ακολουθεί.

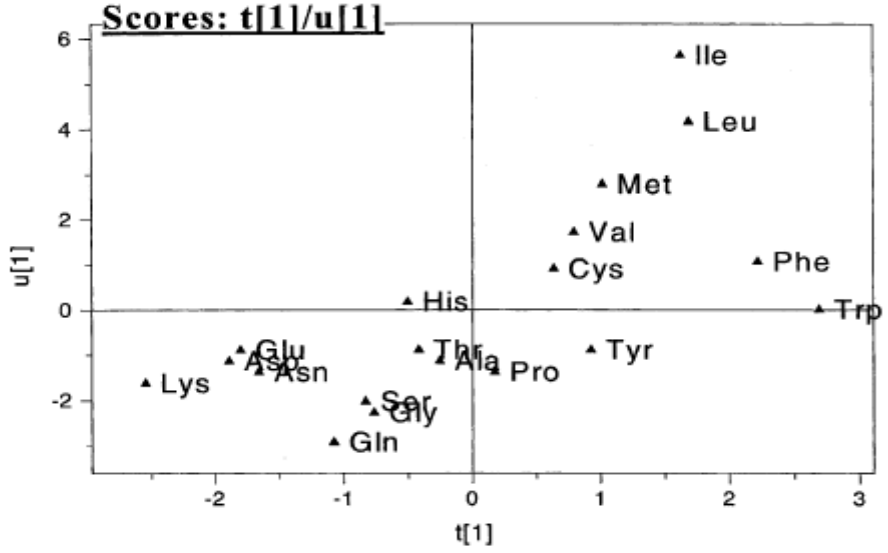
	PIE	PIF	DGR	SAC	MR	Lam	Vol	DDGTS
(1) Ala	0.23	0.31	−0.55	254.2	2.126	−0.02	82.2	8.5
(2) Asn	−0.48	−0.60	0.51	303.6	2.994	−1.24	112.3	8.2
(3) Asp	−0.61	−0.77	1.20	287.9	2.994	−1.08	103.7	8.5
(4) Cys	0.45	1.54	−1.40	282.9	2.933	−0.11	99.1	11.0
(5) Gln	−0.11	−0.22	0.29	335.0	3.458	−1.19	127.5	6.3
(6) Glu	−0.51	−0.64	0.76	311.6	3.243	−1.43	120.5	8.8
(7) Gly	0.00	0.00	0.00	224.9	1.662	0.03	65.0	7.1
(8) His	0.15	0.13	−0.25	337.2	3.856	−1.06	140.6	10.1
(9) Ile	1.20	1.80	−2.10	322.6	3.350	0.04	131.7	16.8
(10) Leu	1.28	1.70	−2.00	324.0	3.518	0.12	131.5	15.0
(11) Lys	−0.77	−0.99	0.78	336.6	2.933	−2.26	144.3	7.9
(12) Met	0.90	1.23	−1.60	336.3	3.860	−0.33	132.3	13.3
(13) Phe	1.56	1.79	−2.60	366.1	4.638	−0.05	155.8	11.2
(14) Pro	0.38	0.49	−1.50	288.5	2.876	−0.31	106.7	8.2
(15) Ser	0.00	−0.04	0.09	266.7	2.279	−0.40	88.5	7.4
(16) Thr	0.17	0.26	−0.58	283.9	2.743	−0.53	105.3	8.8
(17) Trp	1.85	2.25	−2.70	401.8	5.755	−0.31	185.9	9.9
(18) Tyr	0.89	0.96	−1.70	377.8	4.791	−0.84	162.7	8.8
(19) Val	0.71	1.22	−1.60	295.1	3.054	−0.13	115.6	12.0
Correlation matrix								
PIE	1.000	0.967	−0.970	0.518	0.650	0.704	0.533	0.645
PIF	0.967	1.000	−0.968	0.416	0.555	0.750	0.433	0.711
DGR	−0.970	−0.968	1.000	−0.463	−0.582	−0.704	−0.484	−0.648
SAC	0.518	0.416	−0.463	1.000	0.955	−0.230	0.991	0.268
MR	0.650	0.555	−0.582	0.955	1.000	−0.027	0.945	0.290
Lam	0.704	0.750	−0.704	−0.230	−0.027	1.000	−0.221	0.499
Vol	0.533	0.433	−0.484	0.991	0.945	−0.221	1.000	0.300
DDGTS	0.645	0.711	−0.648	0.268	0.290	0.499	0.300	1.000

Πίνακας 9. Το δεύτερο μισό του παραπάνω πίνακα, ο πίνακας συσχέτισης, περιέχει τους ανά ζεύγη συντελεστές συσχέτισης μεταξύ των δεδομένων μας. Οι PIE και PIF αποτελούν σύμφωνα με τους El Tayar, Fauchere και Pliska αντίστοιχα, τις σταθερές λιποφιλικότητας της ΑΑ πλευρικής αλυσίδας, ενώ η DGR αποτελεί την εκλυόμενη ενέργεια μιας πλευρικής αλυσίδας ΑΑ σύμφωνα με τους Radzicka και Woldenden. Η μεταβλητή SAC αποτελεί την προσβάσιμη από το νερό επιφάνεια της ΑΑ, η MR τη μοριακή διαθλαστικότητα, η Lam είναι μια παράμετρος πόλωσης ενώ τέλος η μεταβλητή Vol είναι ο υπολογισμένος μοριακός όγκος της ΑΑ.[15]

➢ 1ο στάδιο της PLSR ανάλυσης

[15] Βλ Table 1 στο Wold S. and Sjostrom M. and Eriksson L. (2001).PLS-Regression: a basic tool of chemometrics*, Chemometrics and Intelligent Laboratory Systems*, vol.58 p. 112.

Εφαρμόζοντας την PLSR ανάλυση των ΑΑ δεδομένων μας, λαμβάνουμε μια μόνο σημαντική συνιστώσα που ερμηνεύει το 43% της Y διασποράς με συντελεστές $R^2 = 0.435$ και $Q^2 = 0.299$. Αντίθετα, μέσω της Πολλαπλής Γραμμικής Παλινδρόμησης λαμβάνουμε $R^2 = 0.788$, το οποίο όμως ισοδυναμεί με την PLSR ανάλυση σε 7 όμως κύριες συνιστώσες. Η Πολλαπλή Γραμμική Παλινδρόμησης παρ' όλα αυτά μας δίνει έναν συντελεστή $Q^2 = -0.215$, το οποίο μας οδηγεί στο εύλογο συμπέρασμα ότι το μοντέλο μας είναι ασθενές. Λαμβάνοντας επομένως μια σημαντική PLS συνιστώσα, το μοναδικό σημαντικό score-διάγραμμα (score plot) που παίρνουμε είναι το ακόλουθο μεταξύ $u_1 - t_1$.



Διάγραμμα 8. Τα PLS scores u_1 και t_1 του ΑΑ παραδείγματος, 1η ανάλυση[16]

Παρατηρούμε ότι τα αρωματικά αμινοξέα, Trp, Phe και ενδεχομένως Tyr παρουσιάζουν μια αρκετά χειρότερη προσαρμογή από τα υπόλοιπα κάτι το οποίο αποτελεί ένδειξη ανομοιογένειας στα δεδομένα μας. Προκειμένου να το ελέγξουμε οδηγούμαστε σε μια 2^{η} ανάλυση των δεδομένων μας, αυτή τη φορά με ένα μειωμένο σύνολο όμως δεδομένων με $N = 16$ χωρίς τα αρωματικά αμινοξέα.

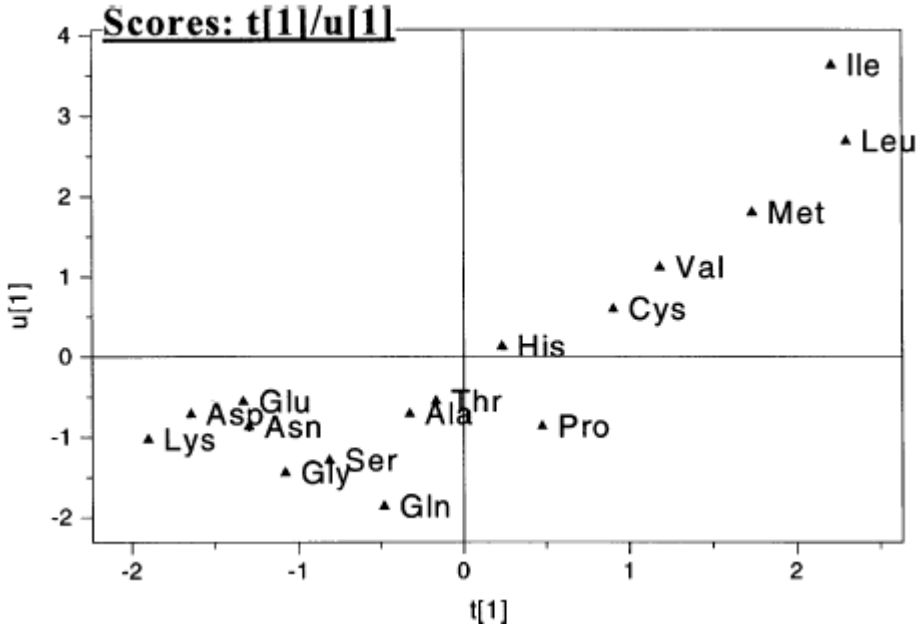
➢ 2^{ο} στάδιο της PLSR ανάλυσης

Χρησιμοποιώντας τώρα το νέο σύνολο δεδομένων με $N = 16$ και κάνοντας πάλι την PLSR ανάλυση, λαμβάνουμε ένα ουσιαστικά καλύτερο αποτέλεσμα με

[16] Ο.π. Figure 3, p. 120.

$A = 2$ κύριες συνιστώσες αυτή τη φορά, $R^2 = 0.783$ και $Q^2 = 0.706$. Η Πολλαπλή Γραμμική Παλινδρόμηση μας δίνει τώρα $R^2 = 0.872$ και $Q^2 = 0.608$. Τα παραπάνω αποτελέσματα που παίρνουμε υποδεικνύουν ότι το νέο σύνολο δεδομένων μας είναι περισσότερο ομοιογενές και επομένως μπορεί να μοντελοποιηθεί σε καλύτερο βαθμό απ' ότι προηγουμένως.

Το ακόλουθο διάγραμμα 9, μεταξύ $u_1 - t_1$, παρουσιάζει όμως μια ισχυρή καμπυλότητα, υποδεικνύοντας την ύπαρξη τετραγωνικών όρων στη λιποφιλικότητα και πιθανώς στην πόλωση.



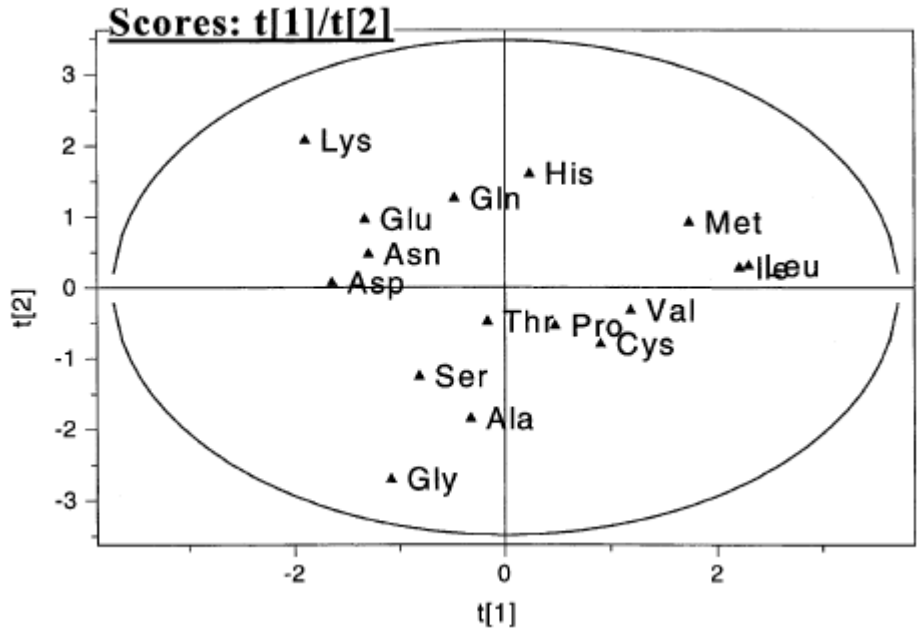
Διάγραμμα 9. Τα PLS scores u_1 και t_1 του ΑΑ παραδείγματος , 2η ανάλυση[17]

Στην τελική ανάλυση, τα τετράγωνα των τεσσάρων αυτών μεταβλητών συμπεριλήφθηκαν στο μοντέλο, παίρνοντας ακόμα καλύτερα αποτελέσματα. Δυο κύριες συνιστώσες και μια ακόμα συνοριακής σημασίας υιοθετήθηκαν. Οι R^2 και Q^2 τιμές των συντελεστών για $A = 2$ είναι 0.90 και 0.80, ενώ για $\mathrm{A} = 3$ είναι 0.925 και 0.82 αντίστοιχα. Στο σημείο αυτό θα προβούμε σε μια ερμηνεία των PLSR παραμέτρων t, w και c.

Όπως έχει ήδη προαναφερθεί τα $X - scores, t_a$, μας παρουσιάζουν τις ομοιότητες και διαφορές που υπάρχουν μεταξύ των μεταβλητών.

Στο ακόλουθο διάγραμμα 10 μεταξύ των $X - scores$ $t_1 - t_2$, παρατηρείται ότι τα 16 αμινοξέα είναι ομαδοποιημένα σύμφωνα με την πόλωση από πάνω αριστερά έως κάτω δεξιά και μέσα σε κάθε ομάδα σύμφωνα με το μέγεθος και τη λιποφιλικότητα τους.

[17] Ο.π. Figure 4, p. 120.

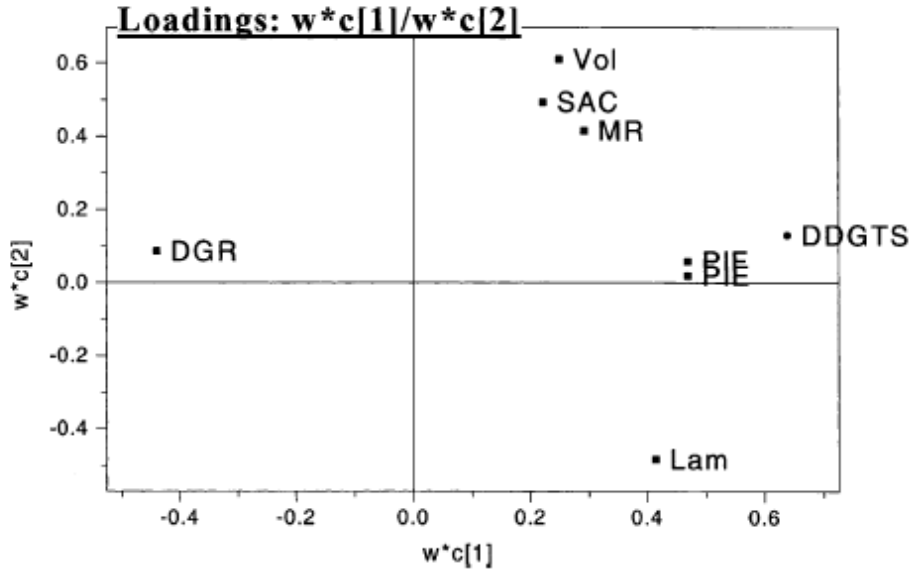


Διάγραμμα 10. Τα PLS scores t_1 ***και*** t_2 ***του ΑΑ παραδείγματος, 2η ανάλυση***[18]

Προκειμένου να ερμηνεύσουμε το PLSR μοντέλο, η συνηθέστερη τεχνική είναι να κάνουμε το διάγραμμα των βαρών w^* και c των διαστάσεων ενός μοντέλου μεταξύ τους. Ένας εναλλακτικός τρόπος είναι να κάνουμε το διάγραμμα των w και c με τα αποτελέσματα και την ερμηνεία τους να μη διαφέρει καθόλου. Τα διαγράμματα μας φανερώνουν τον τρόπο με τον οποίο οι X μεταβλητές συνδυάζονται μεταξύ τους προκειμένου να σχηματίσουν τα $X-scores, t_a$. Οι σημαντικές X μεταβλητές για την a-κύρια συνιστώσα απέχουν αρκετά απ' τη αρχή κατά μήκος του a-στου άξονα στο διάγραμμα wc.

Με την ίδια λογική οι σημαντικές Y μεταβλητές για την a-κύρια συνιστώσα παρουσιάζουν υψηλούς συντελεστές c_{am} και απέχουν αρκετά απ' τη αρχή κατά μήκος του a-στου άξονα στο ίδιο διάγραμμα. Το διάγραμμα των βαρών (weight plot) ή αλλιώς διάγραμμα φορτίων (loading plot) για το παράδειγμα που εξετάζουμε φαίνεται ακολούθως :

[18]Ο.π. Figure 5, p. 121.

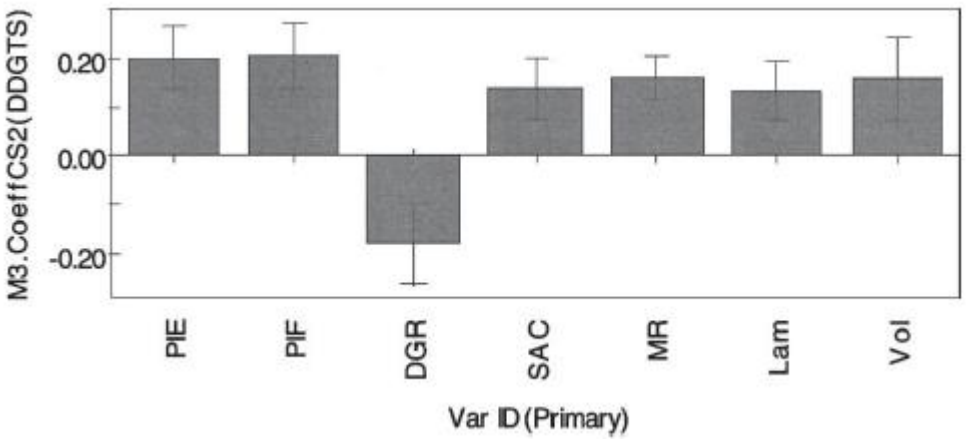


Διάγραμμα 11. Τα PLS βάρη w^* ***και*** c ***για τις πρώτες δυο διαστάσεις του ΑΑ παραδείγματος, 2η ανάλυση.***[19]

Ερμηνεύοντας το παραπάνω διάγραμμα, παρατηρούμε ότι η πρώτη κύρια συνιστώσα αντιπροσωπεύεται από τη λιποφιλικότητα και την πόλωση με τις μεταβλητές PIF, PIE, Lam να βρίσκονται στη θετική πλευρά και τη μεταβλητή DGR να βρίσκεται στην αρνητική. Η δεύτερη κύρια συνιστώσα αντιπροσωπεύεται από ένα συνονθύλευμα του μεγέθους και της πόλωσης με τις μεταβλητές MR, Vol και SAC να είναι στη θετική πάνω πάνω πλευρά την ώρα που η μεταβλητή Lam κείτεται στην αρνητική και κάτω κάτω πλευρά.

Οι c τιμές της απόκρισης y, είναι ανάλογες της διασποράς του Y που ερμηνεύεται μέσω της αντίστοιχης διάστασης, ορίζοντας ένα σημείο για κάθε απόκριση. Στην περίπτωση μας δηλαδή, όπου έχουμε μια το πλήθος απόκριση, το αντίστοιχο σημείο (DDGTS) κείτεται μακριά δεξιά στο πρώτο τεταρτημόριο του παραπάνω διαγράμματος. Η σημασία μιας δοθείσας X μεταβλητής για το Y είναι ανάλογη της απόστασης της απ' την αρχή στο χώρο των φορτίων (loading space). Τα μήκη αυτά, αντιστοιχούν στους PLS συντελεστές Παλινδρόμησης με δυο διαστάσεις, όπως φαίνεται στο Διάγραμμα 12.

[19] Ο.π. Figure 6, p. 121.



Διάγραμμα 12. Οι PLS συντελεστές για $A=2$ ***κύριες συνιστώσες,2η ανάλυση. Οι μπάρες του σχήματος παρουσιάζουν*** 95% ***διαστήματα εμπιστοσύνης σύμφωνα με την τεχνική jack-knifing.***[20]

- **Συμπεράσματα του ανωτέρω παραδείγματος**

Η 1η PLSR ανάλυση και τα συνακόλουθα score plot διαγράμματα, μας υπέδειξαν την ύπαρξη εσωτερικών ομοιογενειών στο σύνολο των δεδομένων μας. Έτσι λάβαμε ένα καλύτερα προσαρμοσμένο μοντέλο για τα μη αρωματικά αμινοξέα. Η ύπαρξη καμπυλότητας για στο score-διάγραμμα των $u_1 - t_1$, μας οδήγησε στο συμπέρασμα τετραγωνικών όρων, κάτι το οποίο μας έδωσε ένα επιθυμητό τελικό μοντέλο, όπου μόνο τα τετράγωνα στις μεταβλητές λιποφιλικότητας είναι σημαντικά σε αυτό.

Εάν δεν είχαμε αφαιρέσει από το τελικό μας μοντέλο τα αρωματικά αμινοξέα, θα είχαμε οδηγηθεί στην ανάπτυξη ενός ξεχωριστού μοντέλου για αυτά. Το παραπάνω μας καταστεί σαφές, τη μεγάλη διαφορά που υπάρχει μεταξύ των αρωματικών και μη αμινοξέων.

[20] Ο.π. Figure 7, p. 122.

4.9 ΤΡΟΠΟΙ ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΡΜΗΝΕΙΑΣ ΕΝΟΣ PLSR ΜΟΝΤΕΛΟΥ

Στην παράγραφο αυτή, θα προσπαθήσουμε να παραθέσουμε ένα προς ένα τα βήματα που απαιτείται να ακολουθήσουμε για την ορθή ανάπτυξη και ερμηνεία ενός μοντέλου Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων (Wold et al., 2001).

- Αρχικά πρέπει να κατανοήσουμε τη μορφή του προβλήματος που καλούμαστε να μοντελοποιήσουμε. Δηλαδή, τις μεταβλητές απόκρισης Y που μας ενδιαφέρει να μελετήσουμε καθώς και τις X μεταβλητές.
- Επόμενο βήμα είναι να πάρουμε καλά δεδομένα τόσο για τις Y όσο και για τις X μεταβλητές. Σε περίπτωση ύπαρξης πολλών μεταβλητών απόκρισης, οι πληροφορίες που αντλούμε είναι περισσότερες καθώς μπορούν αρχικά να μελετηθούν ξεχωριστά εφαρμόζοντας τη γνωστή τεχνική της PCA, δίνοντας μας μια άποψη για την ποσότητα της συστηματικής διασποράς των Y και για το ποιες απ' αυτές τις μεταβλητές απόκρισης πρέπει να αναλυθούν ταυτόχρονα.
- Σχετικά με το μοντέλο τώρα, αρχικά πρέπει να καθοριστεί η σωστή πολυπλοκότητα του, δηλαδή ο αριθμός των κυρίων συνιστωσών που θα διατηρήσουμε στο μοντέλο.
- Η ποιότητα προσαρμογής του μοντέλου στα δεδομένα μας, ελέγχεται μέσω των συντελεστών R^2 και Q^2. Όταν έχουμε περισσότερες από μια μεταβλητές απόκρισης, τότε συνίσταται η χρήση των $R_m{}^2$ και $Q_m{}^2$ για κάθε μια απ' τις μεταβλητές απόκρισης ξεχωριστά. Ο συντελεστής R^2 μας δίνει ένα άνω φράγμα για το πόσο καλά τα δεδομένα ερμηνεύονται απ' το μοντέλο μας αλλά και για την ποιότητα πρόβλεψης των νέων παρατηρήσεων, ενώ το Q^2 μας δίνει ένα κάτω φράγμα για τα παραπάνω.
- Τα $u-t$ score-διαγράμματα για τις πρώτες δυο ή τρεις διαστάσεις του μοντέλου μας υποδεικνύουν την ενδεχόμενη ύπαρξη ακραίων τιμών, καμπυλότητας αλλά και ομαδοποίησης στα δεδομένα. Τα $t-t$ score-διαγράμματα μας υποδεικνύουν την τυχόν ύπαρξη εσωτερικής ομοιογένειας στα δεδομένα μας αλλά και ενδεχόμενη ομαδοποίηση σε αυτά. Τέλος τα w^*c weight-διαγράμματα μας δίνουν μια ερμηνεία των παραπάνω.

- Εάν διαπιστωθούν προβλήματα όπως χαμηλές τιμές των R^2 και Q^2 συντελεστών, ακραίες τιμές, ομαδοποίηση, ή καμπυλότητα στα score-διαγράμματα, συνίσταται η κατασκευή διαγραμμάτων των

υπολοίπων μέσω των οποίων αντλούμε περαιτέρω πληροφορίες για της πηγές του προβλήματος. Ακραίες τιμές οφείλουν να ελεγχθούν για την ορθότητα των δεδομένων μας και να εξαιρεθούν απ' την ανάλυση εάν είναι απαραίτητο. Η καμπυλότητα των $u-t$ διαγραμμάτων μπορεί να βελτιωθεί με τον μετασχηματισμό των δεδομένων μας, ενδεχομένως λογαριθμίζοντας ή εισάγοντας τετραγωνικούς ή κυβικούς όρους στο μοντέλο. Ύστερα απ' τον μετασχηματισμό των δεδομένων, την τροποποίηση του μοντέλου, την ομαδοποίηση του αλλά και τη διαγραφή των ακραίων τιμών επιστρέφουμε στο 1^{ο} βήμα.

- Εάν δεν διαπιστωθεί κανένα πρόβλημα απ' τα παραπάνω, μπορούμε τότε να οδηγηθούμε σε μια ήπια μείωση του μοντέλου διαγράφοντας τους ασήμαντους όρους, δηλαδή αυτούς με μικρούς συντελεστές Παλινδρόμησης. Εφ' όσον γίνουν τα παραπάνω, το τελικό μας μοντέλο έχει αναπτυχθεί.

4.10 Η PLSC (PARTIAL LEAST SQUARES CORRELATION)

Όπως έχει ήδη αναφερθεί η PLSC αποτελεί κατηγορία της PLS. Πρόκειται για την τεχνική, σκοπός της οποίας είναι η μελέτη και η ανάλυση της σχέσης μεταξύ των X και Y πινάκων, σε αντίθεση με την PLSR η οποία στοχεύει με βάση τον πίνακα X να εκτιμήσει τον Y πίνακα. Βρίσκει εφαρμογή κατά γενική ομολογία στον τομέα της νευροαπεικόνισης (neuroimaging) (Krishnan et al., 2010). Ο πίνακας X διαστάσεων $(N \times \mathrm{K})$, όπου N το πλήθος των παρατηρήσεων και K το πλήθος των ανεξάρτητων μεταβλητών, αντιστοιχεί στην εγκεφαλική δραστηριότητα (brain activity) ενώ ο πίνακας Y διαστάσεων $(N \times \mathrm{M})$ περιέχει τις μεταβλητές συμπεριφοράς ή σχεδιασμού (behavioral or design variables), όπου N ομοίως το πλήθος των παρατηρήσεων και M οι μεταβλητές απόκρισης. Το εσωτερικό γινόμενο της j-στήλης του X και της i-στήλης του Y υπολογίζει τη σχέση μεταξύ τους. Όταν οι δυο στήλες που απαρτίζουν το εσωτερικό γινόμενο έχουν υποστεί τη διαδικασία του centering, τότε αυτό εκφράζει τη συνδιασπορά τους. Αντίθετα, όταν οι στήλες είναι κανονικοποιημένες (normalized) είτε όταν το άθροισμα τετραγώνων των τιμών της κάθε στήλης ισούται με τη μονάδα, τότε το εσωτερικό γινόμενο ερμηνεύει τη συσχέτιση μεταξύ τους.

Επομένως η PLSC πρόκειται για μια συμμετρική τεχνική. Οφείλεται στο γεγονός ότι οι ρόλοι των X και Y είναι συμμετρικοί καθώς τόσο η συνδιασπορά όσο και η συσχέτιση δεν επηρεάζονται από τη σειρά των μεταβλητών. Ανάλογα με το περιεχόμενο του Y πίνακα, η τεχνική της PLSC διαιρείται σε τέσσερις κατηγορίες. Πρόκειται για την behavior PLSC, την task PLSC, την seed PLSC και την multi-table PLSC. Συγκεκριμένα στην πρώτη κατηγορία, ο πίνακας Y περιέχει τις μεταβλητές συμπεριφοράς (behavioral variables), στη δεύτερη κατηγορία περιέχει

τις μεταβλητές σχεδιασμού ή αντιθέσεων (contrast or design variables), στην τρίτη τις μεταβλητές της voxel-δραστηριότητας από τις περιοχές που μας ενδιαφέρουν (ROIs) και τέλος στην multi-table PLSC ο πίνακας Y περιέχει τρεις κατηγορίες μεταβλητών, τις behavioral, design και ROIs (Krishnan et al., 2010). Περαιτέρω ανάλυση δε θα γίνει καθώς ξεπερνά τις περιοχές ενδιαφέροντος της παρούσης.

4.10.1 Περιγραφή της λειτουργίας της PLSC

Στο σημείο αυτό θα γίνει μια περιγραφή της λειτουργίας της PLSC μεθόδου βασιζόμενοι στη γνώστη μέθοδο ιδιοανάλυσης SVD για την οποία έχει γίνει λόγος στο Κεφάλαιο 3. Χαρακτηριστικό γνώρισμα της PLSC μεθόδου είναι ότι στηρίζεται στην κατασκευή ενός πίνακα R ο οποίος συνιστά το cross-product γινόμενο των X και Y πινάκων ισούται δηλαδή με :

$$R = Y^T X \tag{4.38}$$

Τις περισσότερες των περιπτώσεων πρόκειται για έναν πίνακα συσχέτισης μιας και οι X και Y είναι κεντροποιημένοι και κανονικοποιημένοι. Μέσω του μετασχηματισμού της SVD, ο πίνακας R γράφεται ως το γινόμενο τριών πινάκων ως εξής :

$$R = P \Delta V^T \tag{4.39}$$

με τα αριστερά ιδιάζοντα διανύσματα που αποτελούν τον P πίνακα να αναπαριστούν τα προφίλ συμπεριφοράς ή σχεδιασμού και τα δεξιά ιδιάζοντα διανύσματα του V πίνακα να εκφράζουν τα voxels ή τις εικόνες του εγκεφάλου (brain images), με τη σημαντική παρατήρηση ότι οι P και V ερμηνεύουν με τον πλέον κατάλληλο τρόπο τον πίνακα R.

Σύμφωνα με τον Bookstein τα ανωτέρω ιδιάζοντα διανύσματα καλούνται στα πλαίσια της PLSC και saliences με τις ονομασίες αυτές να είναι συνώνυμες. Μέσω της διαδικασίας αυτής, παράγονται οι λανθάνουσες μεταβλητές ως γραμμικοί συνδυασμοί των αρχικών μεταβλητών ως εξής :

$$L_X = XV \quad \text{και} \quad L_Y = YP \tag{4.40}$$

όπου ο πίνακας των λανθάνουσων μεταβλητών του X, ο $(N \times L)$πίνακας L_X δηλαδή, καλείται brain scores και ο αντίστοιχος L_Y πίνακας, διαστάσεων $(N \times L)$ των λανθάνουσων μεταβλητών του Y καλείται behavior ή design scores. Μάλιστα

ένα ζευγάρι διανυσμάτων $l_{X,i}$ (η i-στη στήλη του L_X) και $l_{Y,i}$ (η j-στη στήλη του L_Y) αντικατοπτρίζουν την εκάστοτε σχέση εγκεφαλικής δραστηριότητας και συμπεριφοράς.

Ειδικότερα, στόχος της PLSC είναι η εύρεση ζευγών λανθάνουσων διανυσμάτων $l_{X,i}$ και $l_{Y,i}$ που όχι μόνο να έχουν τη μέγιστη δυνατή συνδιασπορά αλλά επιπροσθέτως να ικανοποιούν και δυο περαιτέρω περιορισμούς. Πρώτον τα ζεύγη των λανθάνουσων διανυσμάτων διαφορετικών δεικτών να είναι ασυσχέτιστα μεταξύ τους και δεύτερον οι συντελεστές οι οποίοι είναι υπεύθυνοι για τον υπολογισμό των λανθάνουσων διανυσμάτων να είναι κανονικοποιημένοι. Σε μαθηματική μορφή τα προαναφερθέντα εκφράζονται ως εξής :

εύρεση $l_{X,i} = Xv_i$ και $l_{Y,i} = Yu_i$ ώστε :

$$\mathrm{cov}(l_{X,i}, l_{Y,i}) = \max \qquad (4.41)$$

με την προϋπόθεση ότι :

$$l^T{}_{X,i} l_{Y,i'} = 0 \quad \text{όταν } i \neq i' \qquad (4.42)$$

και

$$v_i^T v_i = u_i^T u_i = 1\,. \qquad (4.43)$$

Μάλιστα, βάσει της SVD, η συνδιασπορά των λανθανουσών διανυσμάτων $l_{X,i}$ και $l_{Y,i}$ ισούται με την αντίστοιχη ιδιάζουσα τιμή που υπολογίζεται ως :

$$l_{X,i}{}^T l_{Y,i} = \delta_i\,. \qquad (4.44)$$

Δηλαδή, με $i = 1$, έχουμε τη μέγιστη δυνατή συνδιασπορά μεταξύ του ζεύγους των λανθανουσών μεταβλητών. Όταν $i = 2$, έχουμε ομοίως τη μέγιστη δυνατή συνδιασπορά με τον περιορισμό όμως ότι το δεύτερο ζεύγος των λανθανουσών μεταβλητών είναι ασυσχέτιστο με το πρώτο. Ομοίως για μεγαλύτερες τιμές του i.

ΚΕΦΑΛΑΙΟ 5. ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ (PRINCIPAL COMPONENT REGRESSION) ΚΑΙ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ (PARTIAL LEAST SQUARES REGRESSION)

Στην ενότητα αυτή θα παραθέσουμε τρεις πρακτικές εφαρμογές για την καλύτερη κατανόηση της PCR και της PLSR που έχουμε ήδη περιγράψει. Παρατίθεται μια εφαρμογή που αφορά την Παλινδρόμηση Κυρίων Συνιστωσών και δυο εφαρμογές της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων. Πριν την εφαρμογή των παραπάνω μεθόδων θα εφαρμοστεί η κλασική μέθοδος Πολλαπλής Γραμμικής Παλινδρόμησης (OLS) όπου θα αποδειχθεί η Πολυσυγγραμμικότητα των δεδομένων και επομένως η ακαταλληλότητα χρήσης της συγκεκριμένης μεθόδου.

Κατά τη διάρκεια των παραπάνω εφαρμογών, χρησιμοποιούμε το στατιστικό πακέτο της R για την PCR και την PLSR, ενώ για την OLS το Minitab. Μέσω αυτών και των επιλογών τους, θα προσπαθήσουμε να προσαρμόσουμε ένα βέλτιστο κάθε φορά για τα δεδομένα μας μοντέλο και να εξάγουμε χρήσιμα κάθε φορά συμπεράσματα. Στην περίπτωση της εφαρμογής της Παλινδρόμησης Κυρίων Συνιστωσών, θα χρησιμοποιήσουμε τον αλγόριθμο της SVD (Singular Value Decomposition) για την ανάλυση των δεδομένων μας, ενώ στις δυο εφαρμογές της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων, αυθαίρετα θα προτιμήσουμε να κάνουμε χρήση του κλασικού orthogonal scores αλγόριθμου NIPALS των Martens και Naes (1989), τη λειτουργία του οποίου έχουμε προηγουμένως αναλύσει με λεπτομέρεια.

Εναλλακτικά, το στατιστικό πακέτο της R, μας παρέχει τη δυνατότητα επιλογής είτε του Kernel αλγόριθμου των Dayal και Mac-Gregor (1997) είτε του SIMPLS αλγόριθμου του de Jong (1993). Μέσω των υπολογιστικών εφαρμογών μας, θα επιβεβαιώσουμε την άποψη ότι δικαιολογημένα τόσο η PCR όσο και η PLSR καλούνται μέθοδοι συρρίκνωσης σύμφωνα με τους Hastie, Tibshirani και Friedman (2001).

5.1 ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

Για την εφαρμογή αυτή, θα κάνουμε χρήση των δεδομένων του προβλήματος "περιεκτικότητα σε πρωτεΐνη"[21]. Αρχικά θα εισάγουμε τα δεδομένα μας, δηλαδή τη μεταβλητή απόκρισης Y και τις επεξηγηματικές μεταβλητές $X_1 - X_6$ στο Minitab προκειμένου να προσαρμόσουμε το πολλαπλό γραμμικό μοντέλο και να δείξουμε το πρόβλημα της Πολυσυγγραμμικότητας. Αυτό γίνεται επιλέγοντας από τη γραμμή εντολών τις εντολές Stat -> Regression -> Regression και ακολούθως στο πλαίσιο Response εισάγουμε την εξαρτημένη μεταβλητή Y και στο πλαίσιο Predictors τις ανεξάρτητες μεταβλητές $X_1, X_2, X_3, X_4, X_5, X_6$. Τα αποτελέσματα που λαμβάνουμε είναι τα ακόλουθα :

Regression Analysis: Y versus X1; X2; X3; X4; X5; X6

```
The regression equation is
Y = 23,6 + 0,0212 X1 + 0,0060 X2 + 0,235 X3 - 0,235 X4 + 0,0116 X5 - 0,0393
X6
```

Predictor	Coef	SE Coef	T	P	VIF
Constant	23,642	9,651	2,45	0,025	
X1	0,02115	0,07536	0,28	0,782	**2955,9**
X2	0,00595	0,08447	0,07	0,945	**2709,1**
X3	0,23514	0,07808	3,01	0,008	**2650,6**
X4	-0,23465	0,05634	-4,16	0,001	**1400,6**
X5	0,011593	0,006110	1,90	0,075	**34,8**
X6	-0,03933	0,04137	-0,95	0,355	**233,8**

```
S = 0,220587   R-Sq = 98,2%   R-Sq(adj) = 97,6%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	45,4068	7,5678	155,53	0,000
Residual Error	17	0,8272	0,0487		
Total	23	46,2340			

Source	DF	Seq SS
X1	1	10,1737
X2	1	28,8768
X3	1	0,7482
X4	1	5,3579
X5	1	0,2061

[21] Βλ. Παράρτημα 2, Πίνακας 2, Δεδομένα από Fearn T., (1983). A Misure of Ridge Regression in the Calibration of a Near Infrared Reflectance Instrument, *Journal of the Royal Statistical Society*, vol. 32, (1), p. 74.

```
X6       1    0,0440


Unusual Observations

Obs    X1         Y       Fit   SE Fit   Residual   St Resid
  4   450   11,6700   11,2738   0,1562     0,3962      2,54R

R denotes an observation with a large standardized residual.
```

Παρατηρώντας τα παραπάνω αποτελέσματα εύκολα καταλήγουμε στο συμπέρασμα ότι η παρουσία του φαινομένου της Πολυσυγγραμμικότητας η οποία εντοπίζεται στους πολύ υψηλούς παράγοντες VIF's συντελεί στην αδυναμία ορθής εκτίμησης των συντελεστών της Παλινδρόμησης.

Οφείλεται στο γεγονός ότι όπως έχει τονιστεί στο Κεφάλαιο 2 της παρούσης, παράγοντες VIF μεγαλύτεροι του 5 ή του 10 καταδεικνύουν την ύπαρξη Πολυσυγγραμμικότητας. Επομένως βάσει των ανωτέρω οδηγούμαστε στην εφαρμογή της μεθόδου PCR. Για το σκοπό αυτό εισάγουμε τα δεδομένα μας στην R.

Προκειμένου να είμαστε σε θέση να προσαρμόσουμε το μοντέλο μας, θα δημιουργήσουμε έναν πίνακα δεδομένων data.frame στον οποίο θα εισάγουμε τις επεξηγηματικές μας μεταβλητές, όπου κάθε στήλη θα αντιπροσωπεύεται από μια μεταβλητή. Ακολούθως ο πίνακας αυτός θα οριστεί ως pcrdata. Πράγματι, έχουμε :

```
X<-cbind(X1,X2,X3,X4,X5,X6)
pcrdata<-as.data.frame(X)
```

Με την εντολή cor(X) θα υπολογίσουμε τον πίνακα συσχέτισης των επεξηγηματικών μεταβλητών. Λαμβάνουμε :

```
       X1        X2        X3        X4        X5        X6
X1 1.0000000 0.9938643 0.9958238 0.9945777 0.9370514 0.9887293
X2 0.9938643 1.0000000 0.9993365 0.9800670 0.9245668 0.9875778
X3 0.9958238 0.9993365 1.0000000 0.9842661 0.9337063 0.9885750
X4 0.9945777 0.9800670 0.9842661 1.0000000 0.9544739 0.9889840
X5 0.9370514 0.9245668 0.9337063 0.9544739 1.0000000 0.9493275
X6 0.9887293 0.9875778 0.9885750 0.9889840 0.9493275 1.0000000
```

όπου η ένδειξη υψηλής συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών είναι εμφανής. Στο σημείο αυτό με την εντολή library(pls) στην R, φορτώνουμε το πακέτο που περιλαμβάνει την pcr και την plsr, ώστε να είμαστε σε θέση να το τρέξουμε. Πλέον είμαστε σε θέση να προσαρμόσουμε ένα PCR μοντέλο. Αυτό γίνεται με την ακόλουθη εντολή :

```
test<-pcr(Y~X,data=pcrdata)
```

εναλλακτικά με την εντολή :

```
test<-pcr(Y~X,data=pcrdata,validation="LOO")
```

προσαρμόζουμε ένα μοντέλο PCR, όπως προηγουμένως, αλλά αυτή τη φορά συμπεριλαμβάνονται οι leave-one-out (LOO) cross-validated προβλέψεις των Lachenbruch και Mickey (1968). Μέσω του ορίσματος "LOO", η τεχνική του Cross-Validation αφήνει ένα ζεύγος παρατηρήσεων κάθε φορά εκτός του δείγματος (βλ. σελ. 76 της παρούσης). Πληκτρολογώντας στην R τώρα την εντολή :

```
summary(test)
```

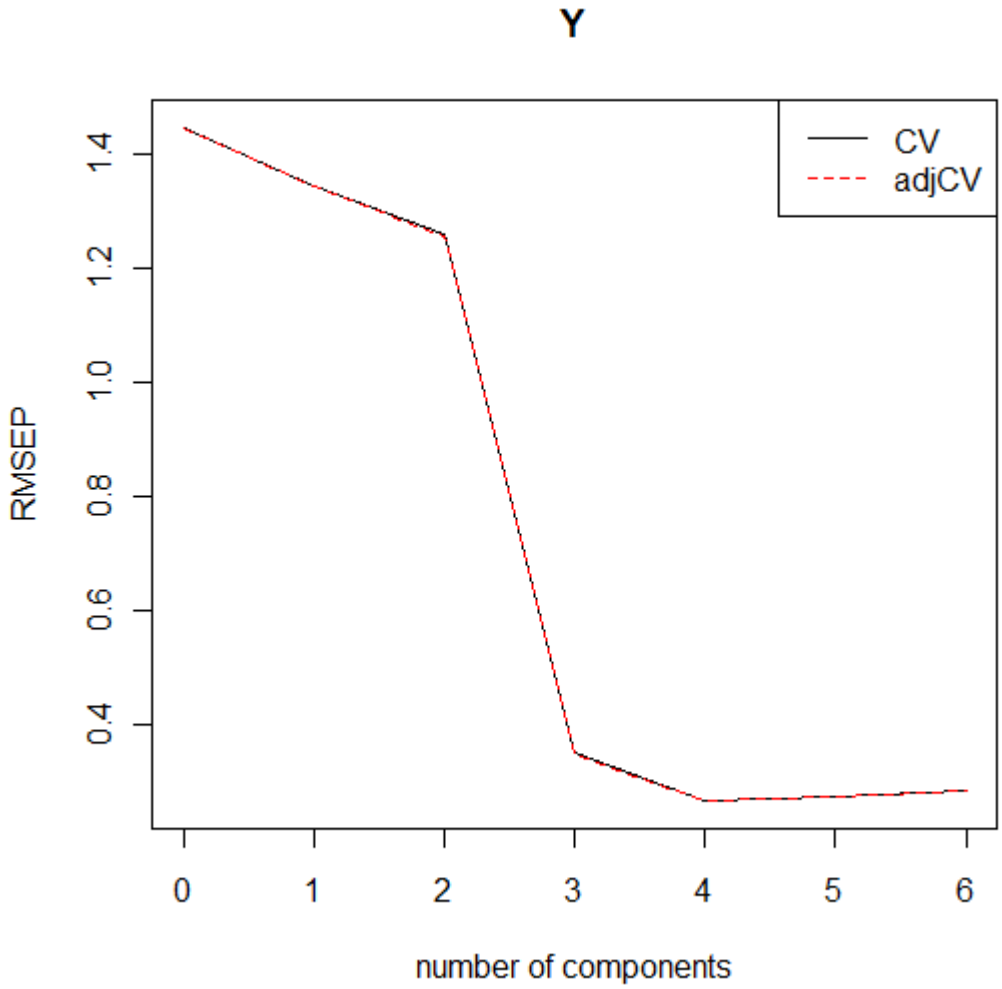
μπορούμε να δούμε τα αποτελέσματα της προσαρμογής μας. Λαμβάνουμε :

```
Data:   X dimension: 24 6
        Y dimension: 24 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 24 leave-one-out segments.
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          1.448    1.346    1.259   0.3501   0.2672   0.2745   0.2850
adjCV       1.448    1.344    1.256   0.3486   0.2657   0.2728   0.2828

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X   96.99    99.63    99.92    99.99   100.00   100.00
Y   19.96    34.40    96.08    97.97    98.03    98.21
```

Η μέθοδος SVD, παρατηρούμε ότι έχει εξάγει 6 κύριες συνιστώσες καθώς έχουμε εισάγει 6 επεξηγηματικές μεταβλητές. Τα validation results είναι με την RMSEP μορφή (root mean squared error of prediction). Παίρνουμε δυο εκτιμήτριες cross-validation, την κανονική CV εκτιμήτρια και τη διορθωμένη εκτιμήτρια adjCV των Mevik και Cederkvist (2004) (βλ. σελ. 77-79 της παρούσης). Μπορούμε να κρίνουμε τα RMSEP κάνοντας το διάγραμμα τους, με την εντολή :

```
plot(RMSEP(test),legendpos="topright")
```

όπου παίρνουμε το ακόλουθο διάγραμμα των RMSEP (1) ως συναρτήσεις του αριθμού των κυρίων συνιστωσών. Με το όρισμα “legendpos“ , προσθέτουμε μια λεζάντα στη θέση που αυτό υποδεικνύει.



1. Διάγραμμα RMSEP-αριθμού Κυρίων Συνιστωσών

Απ’ το παραπάνω διάγραμμα 1 συμπεραίνουμε ότι αρκεί να κρατήσουμε **τρεις κύριες συνιστώσες**, καθώς παίρνουμε $RMSEP \cong 0,3$. Αξίζει στο σημείο αυτό, να τονίσουμε ότι πολλές φορές η μέθοδος PCR απαιτεί περισσότερες κύριες συνιστώσες απ’ τη μέθοδο PLSR, προκειμένου να επιτύχει το ίδιο σφάλμα πρόβλεψης. Από τη στιγμή που έχουμε επιλέξει τον αριθμό των συνιστωσών που θα χρησιμοποιήσουμε περαιτέρω στην ανάλυση μας, μπορούμε μέσω των διαθέσιμων διαγραμμάτων της μεθόδου να αντλήσουμε επιπλέον δεδομένα χρήσιμα για τη μελέτη μας. Αυτό γίνεται κάνοντας τα διαγράμματα προβλέψεων, scores, loadings.

Ένα προσαρμοσμένο μοντέλο χρησιμοποιείται συχνά για την πρόβλεψη των τιμών νέων παρατηρήσεων. Με την ακόλουθη εντολή, θα προβλέψουμε τις τιμές

για τις 24 παρατηρήσεις της Y μεταβλητής, χρησιμοποιώντας τρεις κύριες συνιστώσες :

```
predict(test,ncomp=3,newdata=pcrdata)
```

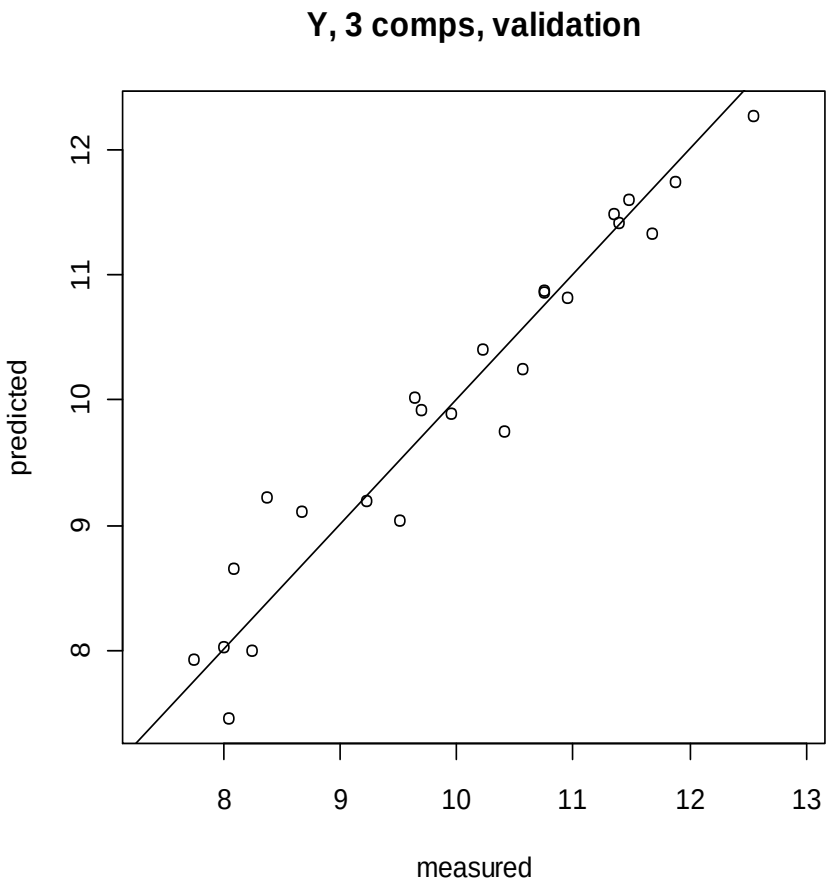
Έτσι παίρνουμε τις ακόλουθες τιμές :

```
, , 3 comps

         Y
1   9.191808
2   8.022263
3  10.843187
4  11.501543
5   9.825749
6   9.096501
7   9.069872
8   7.868458
9   7.563292
10 11.410756
11  9.899167
12  8.031025
13 10.278436
14 10.383108
15 11.770547
16  8.421981
17 12.406522
18  8.971017
19  9.970179
20 11.469366
21  9.898929
22 10.844616
23 10.865221
24 11.586460
```

Για το διάγραμμα προβλέψεων η εντολή είναι :

```
plot(test,ncomp=3,asp=1,line=TRUE)
```

και παίρνουμε το ακόλουθο διάγραμμα (2) :

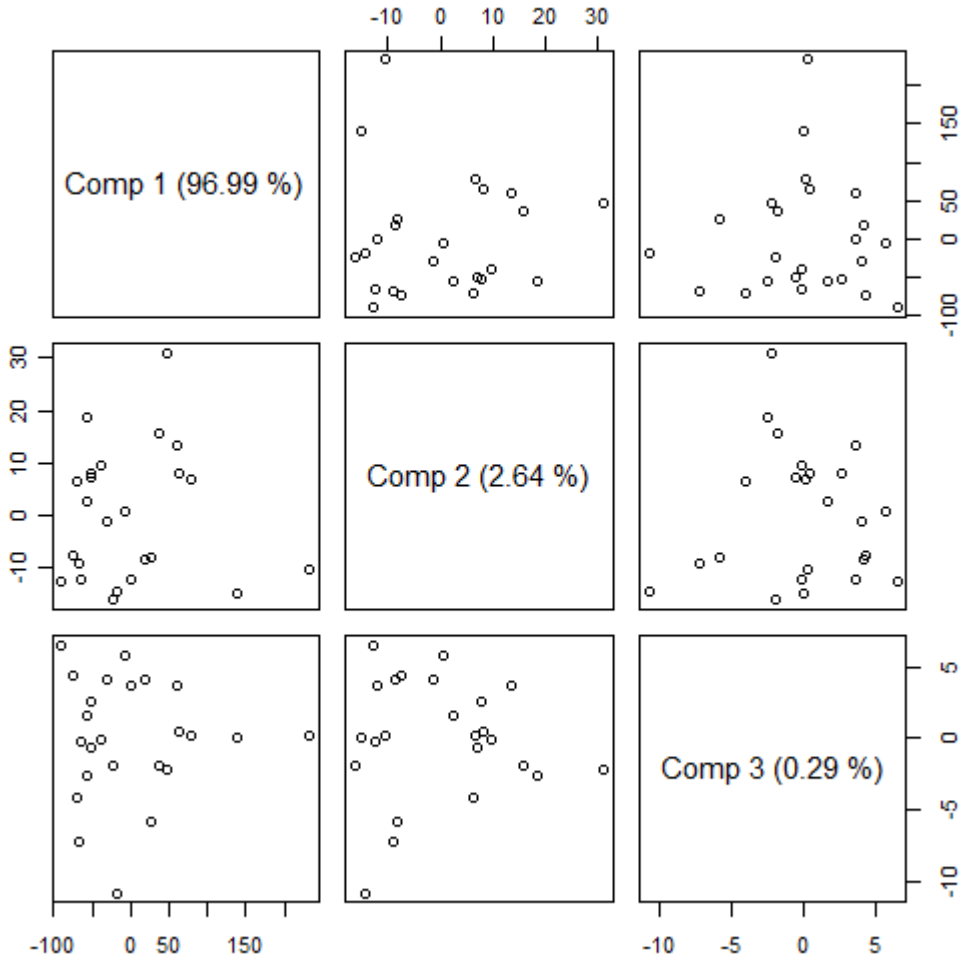


2. Διάγραμμα Προβλέψεων

το οποίο μας δίνει τις cross-validated προβλέψεις με τις τρεις συνιστώσες που επιλέξαμε ενάντια στις τιμές που μετρήθηκαν (measured values). Παρατηρούμε, ότι τα σημεία ακολουθούν ικανοποιητικά την προσαρμοσμένη ευθεία, κάτι το οποίο αποτρέπει την ύπαρξη καμπυλότητας ή άλλων ανωμαλιών. Με την εντολή:

```
plot(test,plottype="scores",comps=1:3)
```

λαμβάνουμε τα ανά ζεύγη διαγράμματα των score τιμών για τις τρεις πρώτες κύριες συνιστώσες (3).



3. *Διάγραμμα των score-τιμών*

Χρησιμοποιούνται για τη διαπίστωση τυχόν ύπαρξης ομαδοποίησης ή ακραίων τιμών στα δεδομένα μας. Παρατηρώντας το παραπάνω διαπιστώνουμε ότι δεν υπάρχει καθαρή ένδειξη είτε ομαδοποίησης είτε ακραίων τιμών. Οι αριθμοί στην παρένθεση δίπλα από κάθε συνιστώσα επεξηγούν το ποσοστό της X διασποράς που ερμηνεύει η κάθε συνιστώσα. Οι ερμηνευμένες διασπορές μπορούν να αντληθούν απευθείας μέσω της εντολής :

```
explvar(test)
```

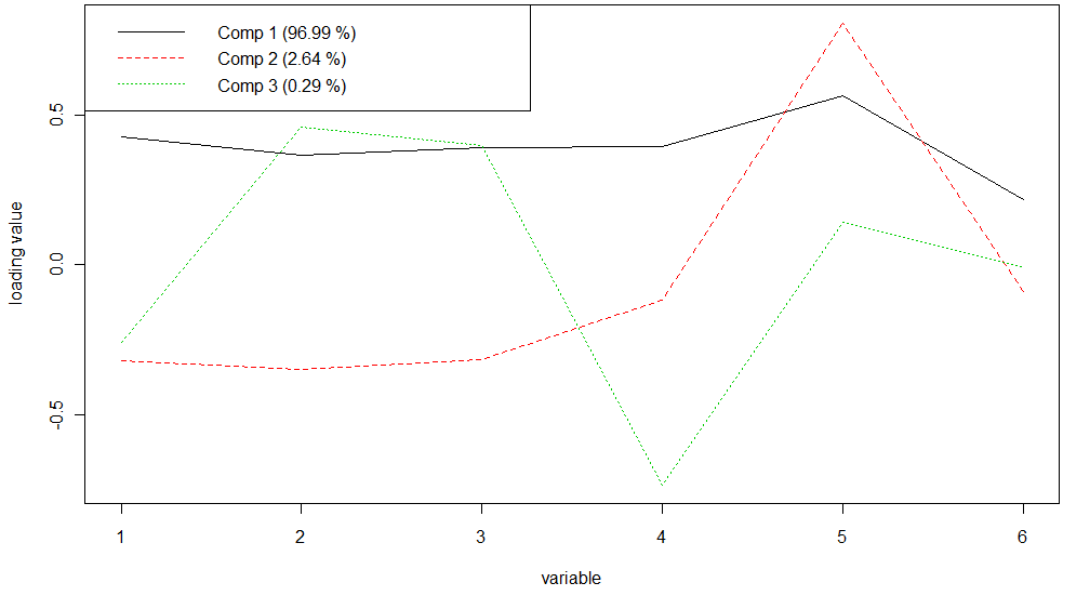
Λαμβάνουμε έτσι :

```
     Comp 1      Comp 2      Comp 3      Comp 4      Comp 5      Comp 6
96.993140366  2.636384561  0.291313581  0.073397857  0.003571867
0.002191767
```

Τέλος το διάγραμμα φορτίων (loadings), (4) χρησιμοποιείται ως επί το πλείστον, για ερμηνευτικούς σκοπούς, όπως για τον εντοπισμό φασματικών κορυφών (spectral peaks) που μπορούν να υποδείξουν λανθασμένη εξαγωγή των παραμέτρων του μοντέλου.

Γίνεται με την ακόλουθη εντολή :

```
plot(test, "loadings", comps=1:3, legendpos="topleft")
```



4. Διάγραμμα φορτίων-αριθμού Κυρίων Συνιστωσών

5.2 ΕΦΑΡΜΟΓΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

1η εφαρμογή

Θα χρησιμοποιήσουμε τα δεδομένα του προβλήματος " χρωματογράφου "[22] Αρχικά, εισάγουμε τα δεδομένα μας, δηλαδή τις δυο μεταβλητές απόκρισης Y_1 και Y_2 και τις επεξηγηματικές μεταβλητές $X_1 - X_6$ στο Minitab. Προσαρμόζοντας το πολλαπλό γραμμικό μοντέλο Παλινδρόμησης με μια μεταβλητή απόκρισης κάθε φορά λαμβάνουμε τα ακόλουθα :

Regression Analysis: Y1 versus X1; X2; X3; X4; X5; X6

```
The regression equation is
Y1 = 0,02 - 0,21 X1 - 1,34 X2 + 4,08 X3 + 0,080 X4 + 1,39 X5 + 5,03 X6
```

Predictor	Coef	SE Coef	T	P	**VIF**
Constant	0,018	3,153	0,01	0,996	
X1	-0,207	1,737	-0,12	0,909	**64,7**
X2	-1,342	1,613	-0,83	0,437	**31,5**
X3	4,077	3,299	1,24	0,263	**44,9**
X4	0,0804	0,8870	0,09	0,931	**10,5**
X5	1,394	2,647	0,53	0,617	**16,7**
X6	5,030	2,740	1,84	0,116	**9,0**

```
S = 0,495027   R-Sq = 87,2%   R-Sq(adj) = 74,3%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	9,9809	1,6635	6,79	0,017
Residual Error	6	1,4703	0,2451		
Total	12	11,4512			

Source	DF	Seq SS
X1	1	3,5957

[22] Βλ. Παράρτημα 2, Πίνακας 3, Δεδομένα από Pietrogrande et al., (1989). Principal component analysis in structure-retention and retention-activity studies of benzodiazepines. *Chemometrics and Intelligent Laboratory Systems*, vol.5, p. 258.

```
X2        1  0,4301
X3        1  1,7189
X4        1  0,4371
X5        1  2,9733
X6        1  0,8258
```

Regression Analysis: Y2 versus X1; X2; X3; X4; X5; X6

```
The regression equation is
Y2 = - 0,66 - 1,12 X1 + 2,13 X2 + 1,38 X3 + 1,10 X4 - 1,07 X5 + 5,53 X6


Predictor     Coef  SE Coef      T      P   VIF
Constant   -0,665    3,471  -0,19  0,854
X1         -1,124    1,912  -0,59  0,578  64,7
X2          2,129    1,775   1,20  0,276  31,5
X3          1,379    3,630   0,38  0,717  44,9
X4         1,0978   0,9762   1,12  0,304  10,5
X5         -1,074    2,914  -0,37  0,725  16,7
X6          5,529    3,016   1,83  0,116   9,0


S = 0,544838   R-Sq = 68,0%   R-Sq(adj) = 36,0%


Analysis of Variance

Source           DF      SS      MS     F      P
Regression        6  3,7861  0,6310  2,13  0,190
Residual Error    6  1,7811  0,2968
Total            12  5,5672


Source  DF  Seq SS
X1       1  0,7861
X2       1  0,0128
X3       1  0,0015
X4       1  0,8734
X5       1  1,1145
X6       1  0,9977
```

Οι υψηλοί παράγοντες VIF (>5 ή >10) μας υποδεικνύουν την ύπαρξη Πολυσυγγραμμικότητας στα δεδομένα μας. Επομένως οδηγούμαστε στην εφαρμογή της μεθόδου PLSR. Εισάγουμε τα δεδομένα μας στην R.

Στο σημείο αυτό προκειμένου να προσαρμόσουμε το μοντέλο μας, όπως και στο παράδειγμα της PCR, θα δημιουργήσουμε τον πίνακα δεδομένων X και θα οριστεί ως pls1data.

```
X<-cbind(X1,X2,X3,X4,X5,X6)
Y<-cbind(Y1,Y2)
pls1data<-as.data.frame(X)
```

Προκειμένου να πάρουμε τον πίνακα συσχέτισης των επεξηγηματικών μεταβλητών $X_1, X_2, X_3, X_4, X_5, X_6$ χρησιμοποιούμε την εντολή cor(X). Λαμβάνουμε :

	X1	X2	X3	X4	X5	X6
X1	1.0000000	0.9702836	0.9648497	-0.7415419	-0.7987749	-0.7982624
X2	0.9702836	1.0000000	0.9637072	-0.8124216	-0.7946506	-0.8119116
X3	0.9648497	0.9637072	1.0000000	-0.8031938	-0.7321493	-0.7745361
X4	-0.7415419	-0.8124216	-0.8031938	1.0000000	0.7777294	0.6771946
X5	-0.7987749	-0.7946506	-0.7321493	0.7777294	1.0000000	0.8848018
X6	-0.7982624	-0.8119116	-0.7745361	0.6771946	0.8848018	1.0000000

όπου παρατηρείται η υψηλή συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών. Ομοίως με το προηγούμενο παράδειγμα φορτώνουμε το πακέτο pls από την R.

```
library(pls)
```

Με την εντολή :

```
pls.options(plsralg="oscorespls")
```

επιλέγουμε να χρησιμοποιήσουμε τον αλγόριθμο NIPALS για την προσαρμογή του PLSR μοντέλου. Σε αντίθετη περίπτωση, η R θα έκανε αυτόματα χρήση του kernel αλγόριθμου.

Τώρα, είμαστε έτοιμοι να προσαρμόσουμε το PLSR μοντέλο. Αυτό γίνεται με την εντολή :

```
test1<-plsr(Y~X,data=pls1data,validation="LOO")
```

και με την εντολή :

```
summary(test1)
```

παίρνουμε τα αποτελέσματα της προσαρμογής μας.

```
Data:   X dimension: 13 6
        Y dimension: 13 2
Fit method: oscorespls
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 13 leave-one-out segments.

Response: Y1
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          1.017   0.9008   0.8418   0.6237   0.6861   0.8764    1.191
adjCV       1.017   0.8947   0.8224   0.6137   0.6716   0.8544    1.153
```

Response: Y2

	(Intercept)	1 comps	2 comps	**3 comps**	4 comps	5 comps	6 comps
CV	0.709	0.6865	0.6437	**0.6544**	1.006	1.061	1.257
adjCV	0.709	0.6829	0.6308	**0.6460**	0.984	1.033	1.219

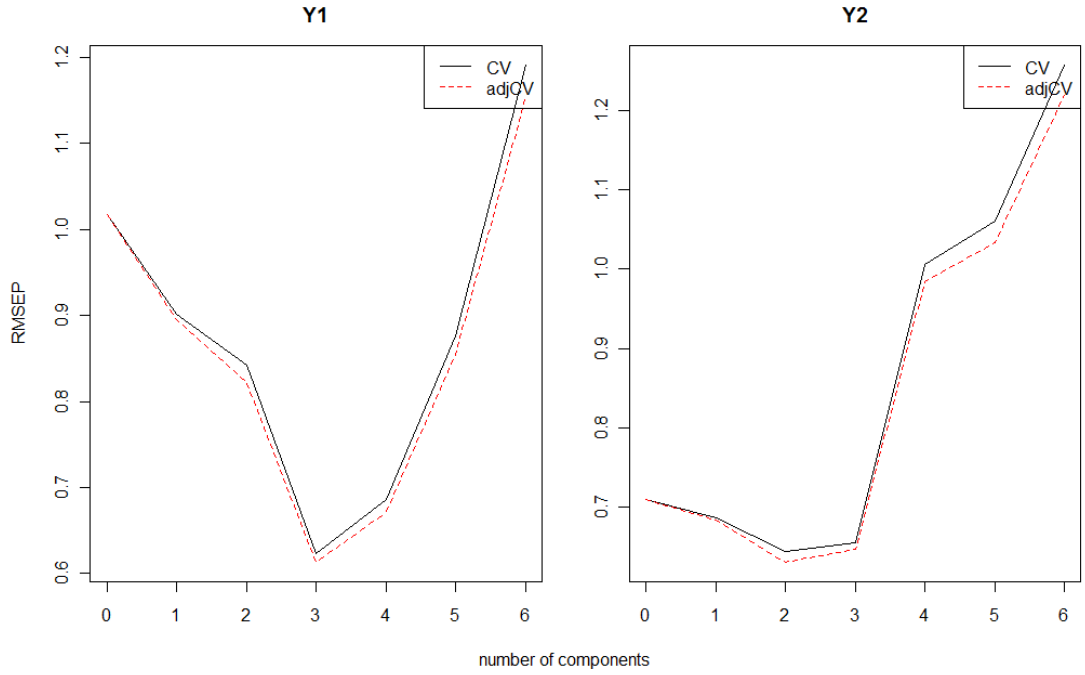
TRAINING: % variance explained

	1 comps	2 comps	**3 comps**	4 comps	5 comps	6 comps
X	88.92	93.45	**98.83**	99.42	99.91	100.00
Y1	37.93	65.82	**82.80**	85.55	86.07	87.16
Y2	21.31	52.50	**54.22**	58.93	65.41	68.01

Η μέθοδος oscorespls παρατηρούμε ότι έχει εξάγει 6 κύριες συνιστώσες, καθώς έχουμε συμπεριλάβει 6 επεξηγηματικές μεταβλητές. Τα validation results είναι με την RMSEP μορφή (root mean squared error of prediction). Παίρνουμε δυο εκτιμήτριες cross-validation, την κανονική CV εκτιμήτρια και τη διορθωμένη εκτιμήτρια adjCV. Μπορούμε να κρίνουμε τα RMSEP κάνοντας το διάγραμμα τους (5), με την εντολή :

```
plot(RMSEP(test1),legendpos="topright")
```

και παίρνουμε τα ακόλουθα διαγράμματα :



5. Διάγραμμα RMSEP-αριθμού Κυρίων Συνιστωσών

Παρατηρούμε ότι για τρεις κύριες συνιστώσες το RMSEP για τη μεταβλητή απόκρισης Y_1 είναι το μικρότερο δυνατό. Όσο για την Y_2, είναι μικρότερο για δυο κύριες συνιστώσες. Παρ' όλα αυτά παρατηρώντας το ανωτέρω διάγραμμα,

φαίνεται ότι για την Y_2 το RMSEP για χρήση τριών κυρίων συνιστωσών προσεγγίζει την αντίστοιχη τιμή για τις δυο κύριες συνιστώσες. Συγκεκριμένα για την Y_1 παίρνουμε $CV = 0.6237$ ενώ για την Y_2 παίρνουμε $CV = 0.6544$. Επομένως αποφασίζουμε να διατηρήσουμε **τρεις κύριες συνιστώσες** στο τελικό μας μοντέλο προκειμένου να επιτυγχάνεται ταύτιση του αριθμού των κυρίων συνιστωσών τόσο για την Y_1 όσο και για την Y_2.

Για να προβλέψουμε τις τιμές των νέων παρατηρήσεων των Y_1 και Y_2 μεταβλητών απόκρισης του μοντέλου δίνουμε την εντολή :

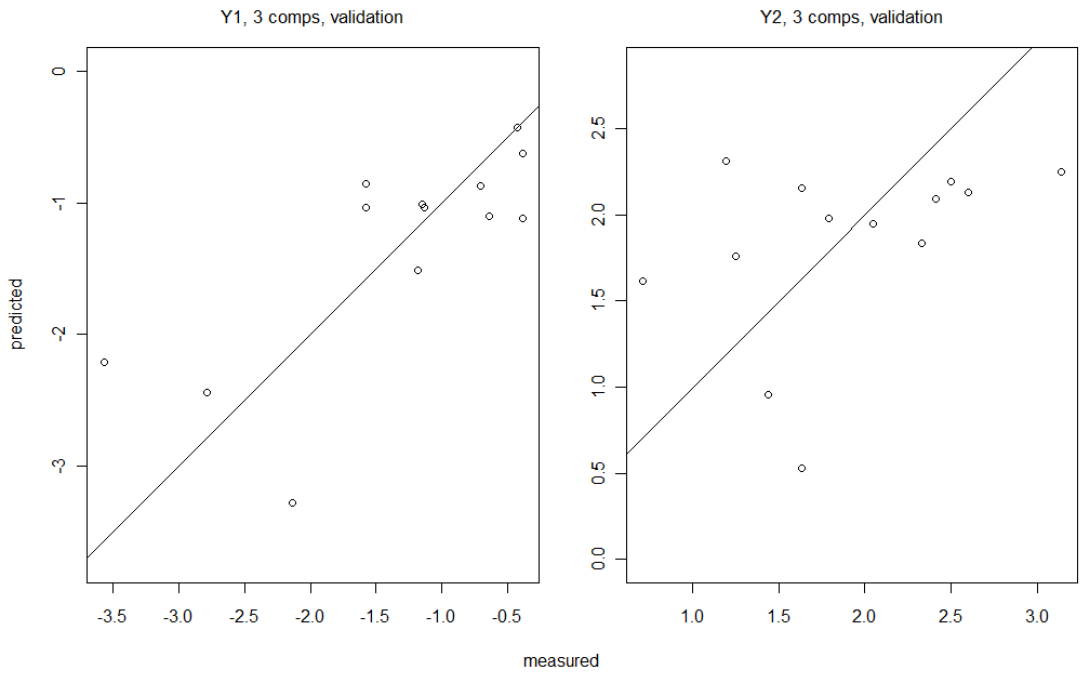
```
predict(test1, ncomp=3, newdata=pls1data)
```

και παίρνουμε τις ακόλουθες προβλέψεις :

```
, , 3 comps

        Y1       Y2
1  -0.8340380 2.3190662
2  -1.1276245 2.0732235
3  -1.0067520 1.9329637
4  -1.4413254 1.6220362
5  -0.7990170 1.9811136
6  -0.9806329 2.1513505
7  -0.4140545 2.7259029
8  -2.8473129 0.9968231
9  -1.0257917 1.9498743
10 -0.5237122 2.2837381
11 -0.9645514 2.1961867
12 -2.8409495 1.2472389
13 -2.8742381 1.1904822
```

Για να πάρουμε το διάγραμμα προβλέψεων (6), η αντίστοιχη εντολή είναι :

```
plot(test1,ncomp=3,asp=1,line=TRUE)
```



6. Διάγραμμα Προβλέψεων

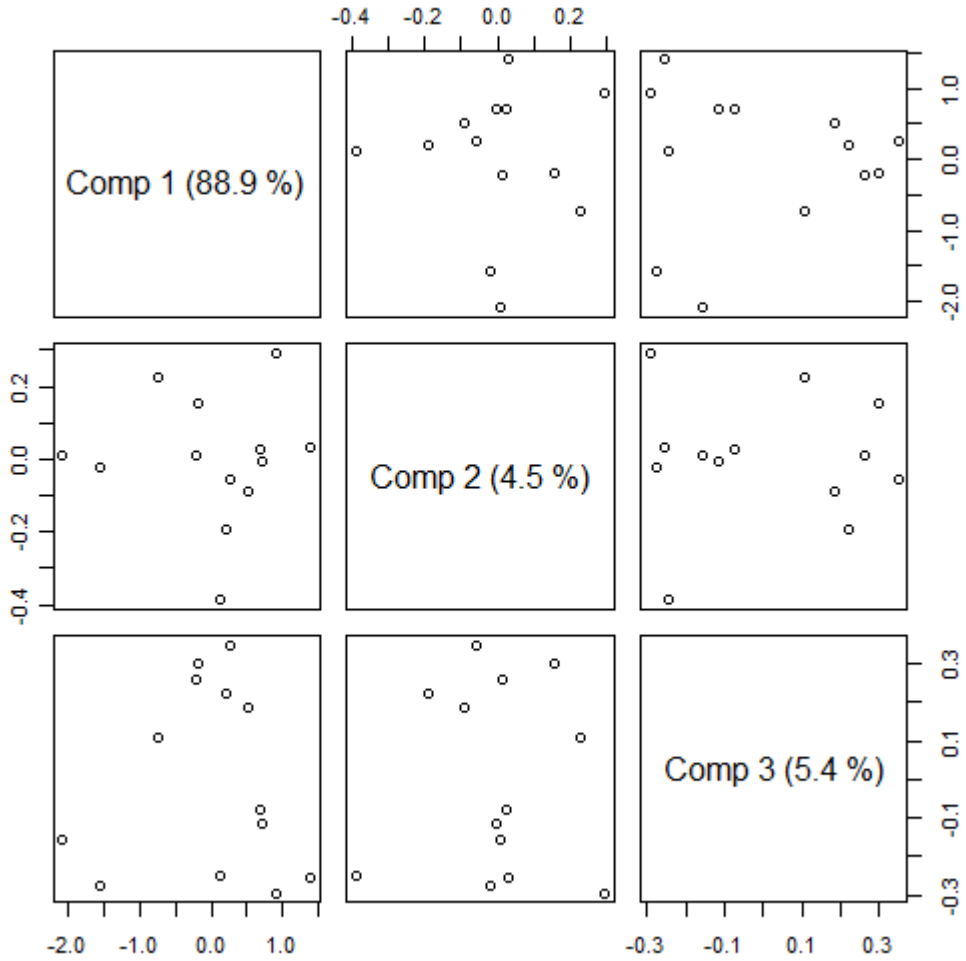
το οποίο μας δίνει τις cross-validated προβλέψεις με τις τρεις συνιστώσες που επιλέξαμε ενάντια στις τιμές που μετρήθηκαν (measured values).

Παρατηρούμε όμως, ότι τα σημεία δεν ακολουθούν σε ικανοποιητικό βαθμό την προσαρμοσμένη ευθεία, κάτι το οποίο μας υποψιάζει για ύπαρξη καμπυλότητας ή κάποιας άλλης ανωμαλίας στα δεδομένα μας.

Κάνοντας χρήση της εντολής :

plot(test1,plottype="scores",comps=1:3)

παίρνουμε τα ανά ζεύγη διαγράμματα των score τιμών για τις τρεις πρώτες κύριες συνιστώσες (7).



7. Διάγραμμα των score-τιμών

Προκειμένου να λάβουμε τις ερμηνευμένες διασπορές κάθε κύριας συνιστώσας πληκτρολογούμε την εντολή :

explvar(test1)

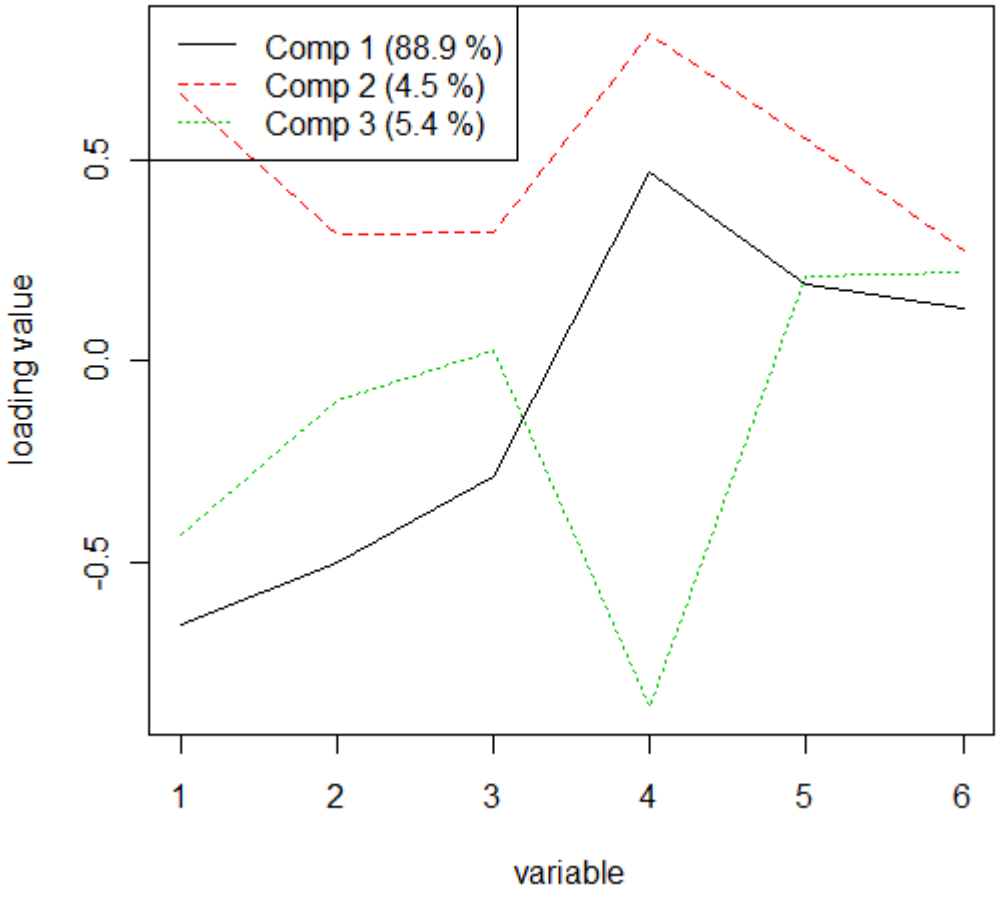
Παίρνουμε :

Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6
88.91524864	4.53463250	5.38135350	0.58893260	0.48897231	0.09086046

Δηλαδή οι τρεις πρώτες συνιστώσες είναι αυτές που ερμηνεύουν σχεδόν όλο το ποσοστό της συνολικής X διασποράς του μοντέλου μας.

Το διάγραμμα φορτίων (loadings), για τις τρεις πρώτες κύριες συνιστώσες (8), προκύπτει με την εντολή :

```
plot(test1, "loadings", comps=1:3, legendpos="topleft")
```



8. Διάγραμμα φορτίων-αριθμού Κυρίων Συνιστωσών

2η εφαρμογή

Στη 2η αυτή εφαρμογή της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων, θα χρησιμοποιήσουμε τα δεδομένα του προβλήματος " Ακετυλήνης[23] Εκτός των τριών ανεξάρτητων μεταβλητών X_1, X_2, X_3 θα χρησιμοποιήσουμε και άλλες έξι ανεξάρτητες μεταβλητές τις X_1^2, X_2^2, X_3^2, X_1X_2 , X_2X_3, X_1X_3. Οι έξι τελευταίες, χάριν συνέπειας των συμβολισμών μας θα καλούνται X_4, X_5, X_6, X_7, X_8 και X_9 αντίστοιχα.

Αρχικά, εισάγουμε τα δεδομένα μας, δηλαδή τη μεταβλητή απόκρισης Y και τις επεξηγηματικές μεταβλητές X_1-X_9 στο Minitab. Οι 9 επεξηγηματικές μεταβλητές $X_1 - X_9$ θα υποστούν centering (κεντροποίηση) αφαιρώντας από την κάθε παρατήρηση της κάθε μεταβλητής τον αντίστοιχο μέσο όρο της.

Προσαρμόζοντας το πολλαπλό γραμμικό μοντέλο Παλινδρόμησης λαμβάνουμε τα ακόλουθα :

Regression Analysis: Y versus X1; X2; X3; X4; X5; X6; X7; X8; X9

```
The regression equation is
Y = 35,8 + 0,0450 X1 + 0,493 X2 - 277 X3 - 0,0133 X4 - 19,0 X5 - 11,4 X6
    - 0,00212 X7 - 0,0327 X8 - 12193 X9
```

Predictor	Coef	SE Coef	T	P	**VIF**
Constant	35,794	1,083	33,04	0,000	
X1	0,04504	0,05556	0,81	0,449	**375,2**
X2	0,49254	0,05476	8,99	0,000	**1,8**
X3	-276,8	194,7	-1,42	0,205	**709,3**
X4	-0,013258	0,002973	-4,46	0,004	**26,9**
X5	-19,007	8,644	-2,20	0,070	**31,2**
X6	-11,436	8,354	-1,37	0,220	**6909,2**
X7	-0,002124	0,001920	-1,11	0,311	**1830,8**
X8	-0,03270	0,01163	-2,81	0,031	**3,2**
X9	-12193	7755	-1,57	0,167	**1224,5**

```
S = 0,895655   R-Sq = 99,8%   R-Sq(adj) = 99,4%
```

```
Analysis of Variance
```

Source	DF	SS	MS	F	P
Regression	9	2118,90	235,43	293,49	0,000
Residual Error	6	4,81	0,80		
Total	15	2123,71			

Source	DF	Seq SS
X1	1	1896,68
X2	1	56,30

[23] Βλ. Παράρτημα 2, Πίνακας 1, Δεδομένα από Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 328, New Jersey : John Wiley & Sons, Inc.

```
X3        1      0,46
X4        1    143,89
X5        1      0,66
X6        1     13,78
X7        1      0,65
X8        1      4,49
X9        1      1,98


Unusual Observations

Obs     X1        Y      Fit   SE Fit  Residual  St Resid
 13   -113   15,000   14,874    0,892     0,126      1,50 X
 14   -113   17,000   16,290    0,833     0,710      2,16R
 15   -113   20,500   21,927    0,594    -1,427     -2,13R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

Ομοίως με τις προηγούμενες δυο εφαρμογές, οι υψηλοί παράγοντες VIF υποδηλώνουν ύπαρξη Πολυσυγγραμμικότητας στα δεδομένα. Εισάγουμε τις επεξηγηματικές μεταβλητές καθώς και τη μεταβλητή απόκρισης στην R.

Ομοίως με το προηγούμενο παράδειγμα, δημιουργούμε τον πίνακα δεδομένων X του οποίου οι στήλες θα είναι οι 9 επεξηγηματικές και κεντροποιημένες μεταβλητές και ακολούθως τον δηλώνουμε ως pls2data.

```
X<-cbind(X1,X2,X3,X4,X5,X6,X7,X8,X9)
pls2data<-as.data.frame(X)
```

Ο πίνακας συσχέτισης των επεξηγηματικών μεταβλητών μέσω της εντολής cor(X) δίνεται παρακάτω :

```
           X1          X2          X3          X4          X5          X6
X1  1.00000000  0.22362776 -0.95751877 -0.12381988  0.20187257  0.4511728
X2  0.22362776  1.00000000 -0.24807978  0.04068679 -0.03120182  0.1949815
X3 -0.95751877 -0.24807978  1.00000000  0.18920174 -0.27023310 -0.6714860
X4 -0.12381988  0.04068679  0.18920174  1.00000000 -0.97139314 -0.2615720
X5  0.20187257 -0.03120182 -0.27023310 -0.97139314  1.00000000  0.3125732
X6  0.45117285  0.19498152 -0.67148598 -0.26157200  0.31257323  1.0000000
X7 -0.27074558 -0.14771083  0.50580043  0.24422032 -0.26697741 -0.9705809
X8  0.03384875  0.50795768 -0.03044338  0.39639652 -0.39027353  0.1289465
X9 -0.58653885 -0.22519502  0.77671248  0.26963623 -0.34742062 -0.9721360
           X7          X8          X9
X1 -0.2707456  0.03384875 -0.5865388
X2 -0.1477108  0.50795768 -0.2251950
X3  0.5058004 -0.03044338  0.7767125
X4  0.2442203  0.39639652  0.2696362
X5 -0.2669774 -0.39027353 -0.3474206
X6 -0.9705809  0.12894646 -0.9721360
X7  1.0000000 -0.12477883  0.8901332
```

```
X8 -0.1247788  1.00000000 -0.1580460
X9  0.8901332 -0.15804602  1.0000000
```

Στη συνέχεια πληκτρολογούμε τις γνώστες εντολές :

```
library(pls)
pls.options(plsralg="oscorespls")
```

Τώρα, προσαρμόζουμε το μοντέλο :

```
test2<-plsr(Y~X,data=pls2data,validation="LOO")
```

και με την εντολή

```
summary(test2)
```

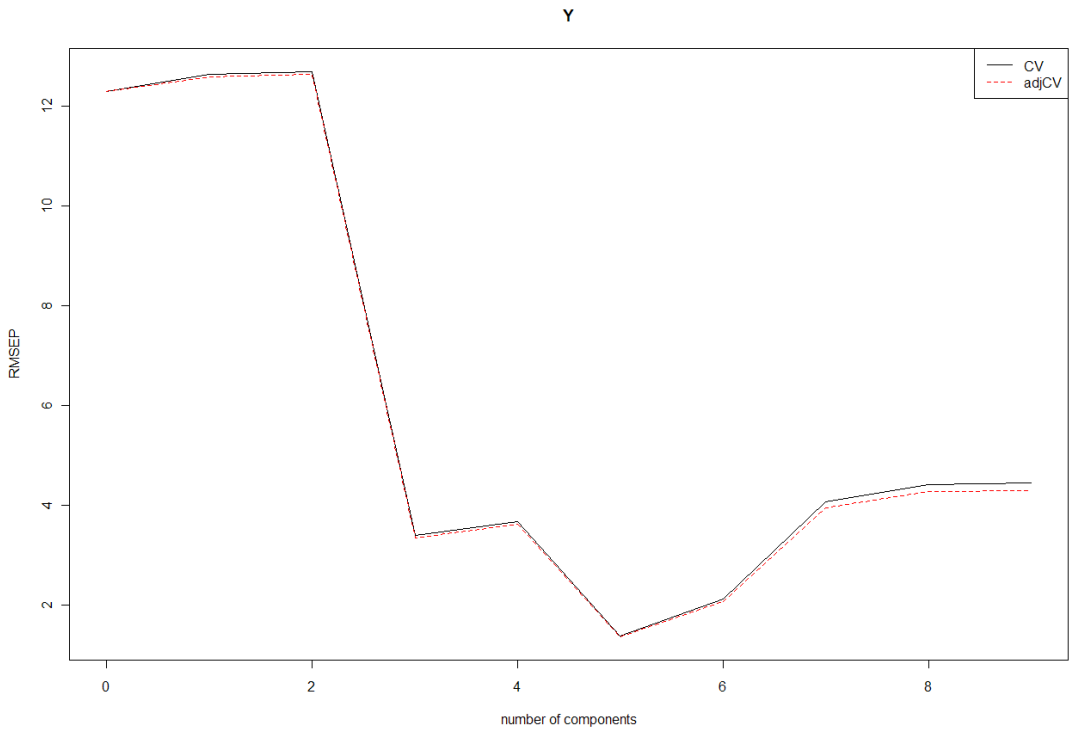
παίρνουμε :

```
Data:   X dimension: 16 9
      Y dimension: 16 1
Fit method: oscorespls
Number of components considered: 9

VALIDATION: RMSEP
Cross-validated using 16 leave-one-out segments.
     (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV         12.29    12.64    12.70    3.386    3.666    1.373    2.114
adjCV      12.29    12.59    12.64    3.349    3.626    1.356    2.068
      7 comps  8 comps  9 comps
CV      4.071    4.411    4.446
adjCV   3.949    4.279    4.291

TRAINING: % variance explained
  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X  99.396   99.97    100.0   100.00   100.00   100.00   100.00   100.00
Y   6.347   22.85     95.8    96.15    99.47    99.52    99.68    99.68
  9 comps
X  100.00
Y   99.77
```

Ομοίως με την προηγούμενη εφαρμογή της PLSR, οδηγούμαστε στα ακόλουθα συμπεράσματα. Η μέθοδος oscorespls, παρατηρούμε ότι έχει εξάγει 9 κύριες συνιστώσες, λόγω της ύπαρξης 9 επεξηγηματικών μεταβλητών. Τα validation results είναι με την RMSEP μορφή (root mean squared error of prediction). Παίρνουμε δυο εκτιμήτριες cross-validation, την κανονική CV εκτιμήτρια και τη διορθωμένη εκτιμήτρια adjCV.

Μπορούμε να κρίνουμε τα RMSEP κάνοντας το διάγραμμα τους (9), με την εντολή :

```
plot(RMSEP(test2),legendpos="topright")
```



9. Διάγραμμα RMSEP-αριθμού Κυρίων Συνιστωσών

Λαμβάνοντας υπόψη μας τη διασπορά των X και Y που ερμηνεύει η κάθε κύρια συνιστώσα, αλλά και το παραπάνω διάγραμμα, αποφασίζουμε να διατηρήσουμε **τρεις κύριες συνιστώσες** στο μοντέλο μας. Με τρεις κύριες συνιστώσες εξηγείται το 100% της διασποράς του X και το 95.8% της διασποράς του Y, ενώ η εκτιμήτρια $CV = 3,386$.

Για να προβλέψουμε τις νέες τιμές των παρατηρήσεων της μεταβλητής απόκρισης Y του μοντέλου μας, χρησιμοποιώντας πάλι τρεις κύριες συνιστώσες, χρησιμοποιούμε την εντολή :

```
predict(test2, ncomp=3, newdata=pls2data)
```

Οι νέες αυτές τιμές είναι :

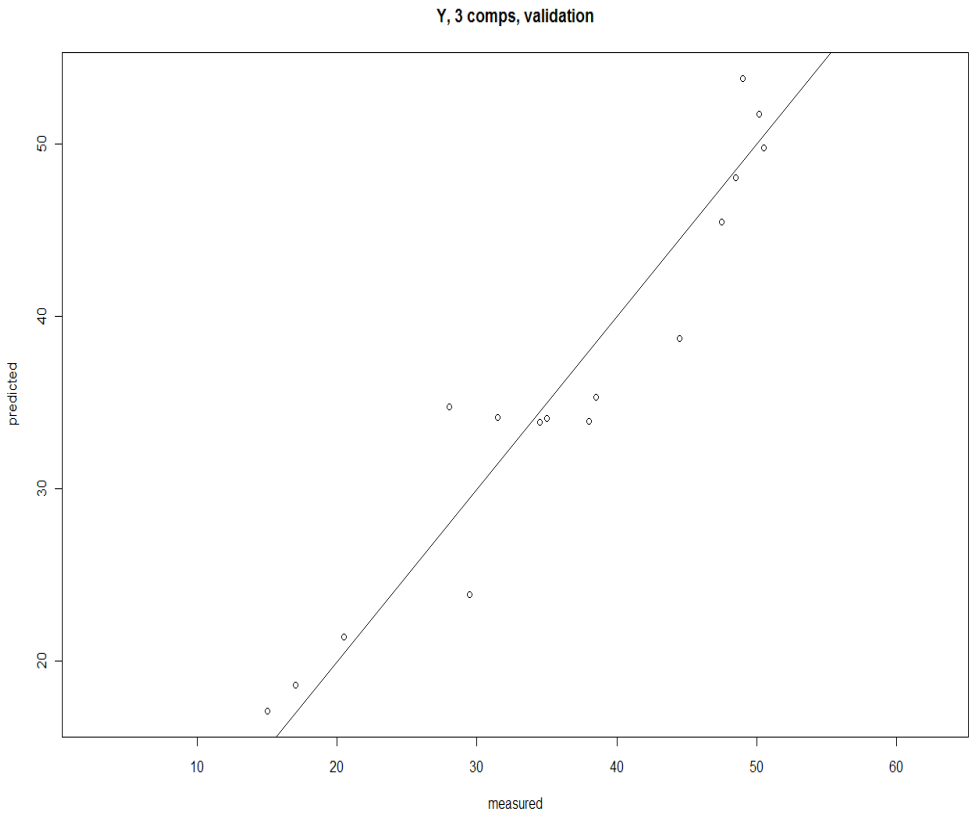
```
, , 3 comps

        Y
1  52.45382
```

2 51.36882
3 49.95641
4 48.24615
5 45.95455
6 42.30514
7 33.54387
8 33.63941
9 33.88915
10 34.14103
11 34.59636
12 35.65625
13 16.26865
14 18.08959
15 21.08688
16 26.50391

Για να πάρουμε το διάγραμμα προβλέψεων (10) έχουμε :

plot(test2,ncomp=3,asp=1,line=TRUE)

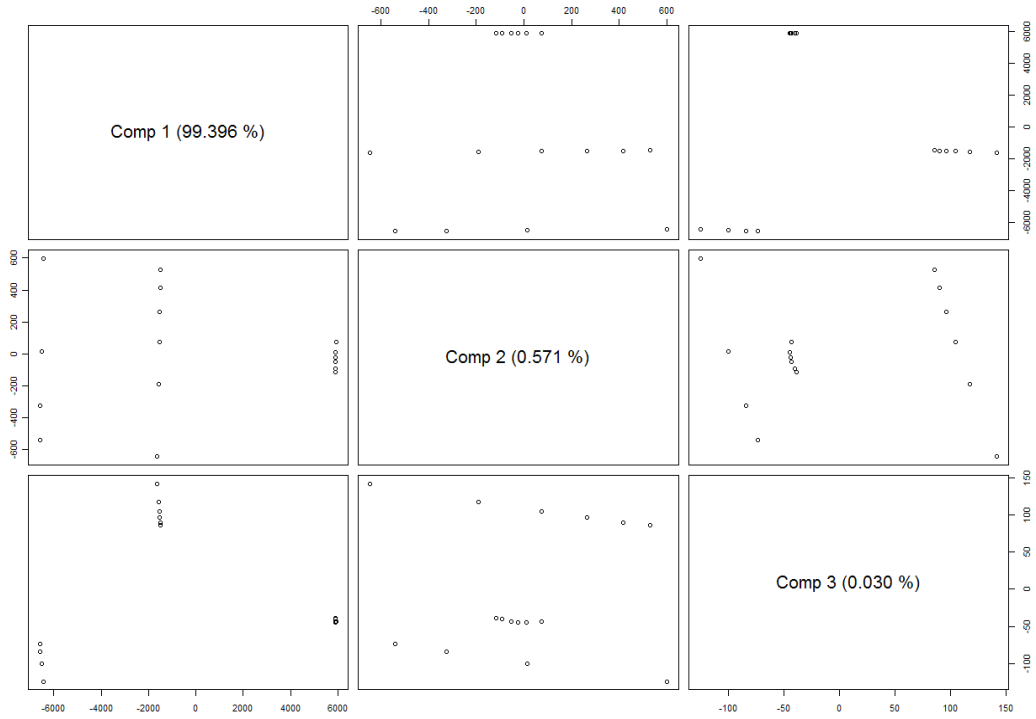
10. Διάγραμμα Προβλέψεων



και διαπιστώνουμε ότι τα σημεία ακολουθούν ικανοποιητικά την προσαρμοσμένη ευθεία, γεγονός που αποτρέπει την ύπαρξη καμπυλότητας και άλλων ανωμαλιών.

Τα ανά ζεύγη διαγράμματα των score τιμών για τις τρεις πρώτες κύριες συνιστώσες (11), τα παίρνουμε με την ακόλουθη εντολή :

```
plot(test2,plottype="scores",comps=1:3)
```



11. Διάγραμμα των score-τιμών

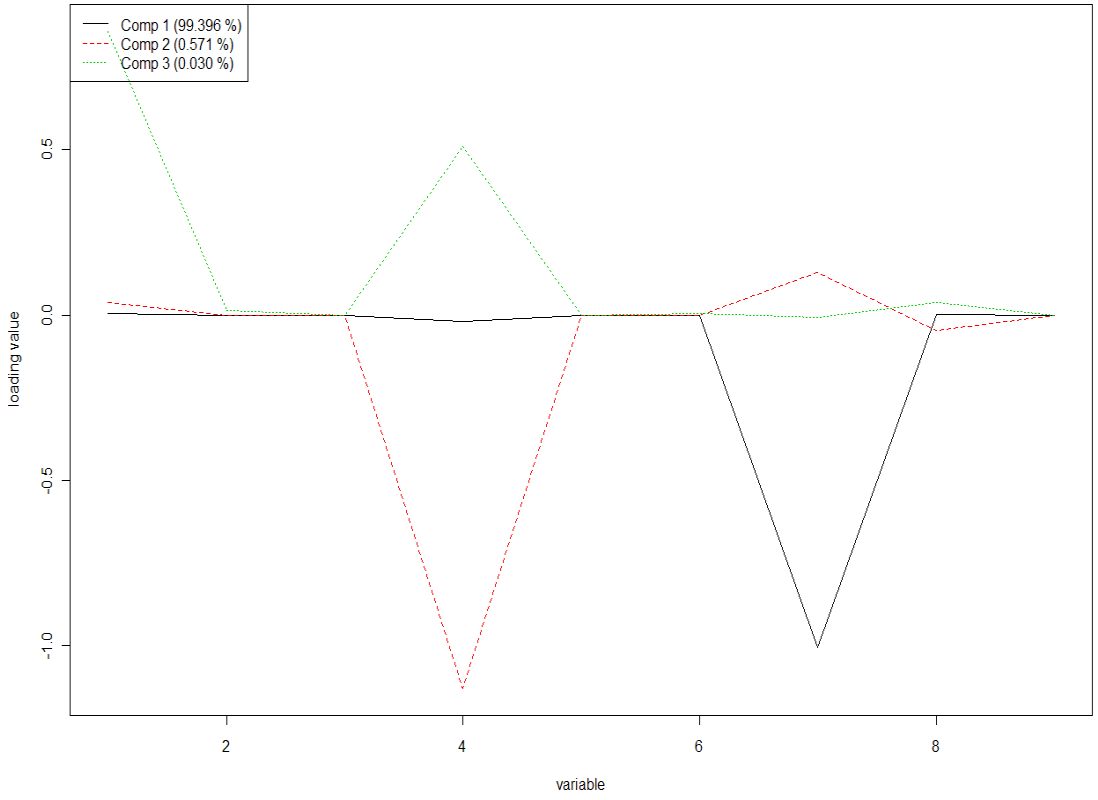
Οι ερμηνευμένες X διασπορές για κάθε κύρια συνιστώσα λαμβάνονται με την εντολή :

```
explvar(test2)
```

```
    Comp 1      Comp 2      Comp 3      Comp 4      Comp 5      Comp 6
9.939562e+01 5.709694e-01 2.954225e-02 3.754189e-03 1.145286e-04
3.807198e-07
    Comp 7      Comp 8      Comp 9
3.199030e-09 1.886417e-11 3.354137e-15
```

Για το loading διάγραμμα (12) για τις τρεις πρώτες κύριες συνιστώσες που έχουμε επιλέξει, έχουμε :

```
plot(test2, "loadings", comps=1:3, legendpos="topleft")
```



12. Διάγραμμα φορτίων-αριθμού Κυρίων Συνιστωσών

ΣΥΜΠΕΡΑΣΜΑΤΑ

Έχοντας περιγράψει ενδελεχώς τις δυο μεθόδους Παλινδρόμησης (PCR και PLSR) και ακολούθως παραθέσει τρεις πρακτικές εφαρμογές καταλήγουμε στα παρακάτω συμπεράσματα. Και οι δυο μέθοδοι ερμηνεύουν σχεδόν όλη τη διασπορά των αρχικών μας δεδομένων μέσω ενός μικρού αριθμού κυρίων συνιστωσών τις οποίες ο αλγόριθμος που κάθε φορά χρησιμοποιείται εξάγει. Στην περίπτωση μας και οι τρεις εφαρμογές που παρατέθηκαν χρησιμοποίησαν τρεις κύριες συνιστώσες για τη βέλτιστη ερμηνεία του μοντέλου. Παρατηρούμε ότι στην πλειονότητα των περιπτώσεων οι δυο μέθοδοι χρησιμοποιούν τον ίδιο αριθμό κυρίων συνιστωσών, χωρίς βέβαια να αποκλείεται το αντίθετο. Στην επιλογή του αριθμού των κυρίων συνιστωσών που τελικώς διατηρήσαμε στο τελικό μοντέλο, καθοριστικής σημασίας ήταν τα διαγράμματα που σε κάθε εφαρμογή χρησιμοποιήθηκαν.

Αξίζει να τονιστεί ότι σε αντίθεση με την PCR, η μέθοδος PLSR συνηθίζεται να βρίσκει εφαρμογή σε περιπτώσεις που οι μεταβλητές απόκρισης είναι περισσότερες από μία. Οφείλεται στην ικανότητά της να μοντελοποιεί και να αναλύει πολλές Y μεταβλητές ταυτόχρονα, κάτι που μας δίνει μια καλύτερη εικόνα απ' ότι όταν έχουμε ένα ξεχωριστό μοντέλο για κάθε μια Y μεταβλητή. Εάν όμως οι Y μεταβλητές είναι συσχετισμένες μεταξύ τους συνίσταται η ταυτόχρονη ανάλυση τους στο ίδιο μοντέλο.

ΠΑΡΑΡΤΗΜΑ 1ο (ΠΙΝΑΚΑΣ ΔΕΔΟΜΕΝΩΝ ΘΕΩΡΙΑΣ)

Δεδομένα από Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 314, New Jersey : John Wiley & Sons, Inc. Παραπομπή στη σελίδα 27, υποκεφάλαιο 2.2. Περιπτώσεις εμφάνισης της Πολυσυγγραμμικότητας.

Πίνακας 1. Δεδομένα προβλήματος " χρόνος παράδοσης "

Y	X_1	X_2
16,68	7	560
11,50	3	220
12,03	3	340
14,88	4	80
13,75	6	150
18,11	7	330
8,00	2	110
17,83	7	210
79,24	30	1460
21,50	5	605
40,33	16	688
21,00	10	215
13,50	4	255
19,75	6	462
24,00	9	448
29,00	10	776
15,35	6	200
19,00	7	132
9,50	3	36
35,10	17	770
17,90	10	140
52,32	26	810
18,75	9	450
19,83	8	635
10.75	4	150

Τα δεδομένα του προβλήματος “χρόνος παράδοσης“ περιγράφουν τον απαιτούμενο χρόνο ενός ανεφοδιαστή μηχανημάτων αυτόματης πώλησης προκειμένου να γεμίζει συνέχεια τα μηχανήματα πηγαίνοντας απ’ το ένα στο άλλο με τα πόδια. Η μεταβλητή απόκρισης Y περιγράφει τον χρόνο που κάθε φορά απαιτείται σε δευτερόλεπτα, ενώ οι μεταβλητές X_1 και X_2 περιγράφουν τον αριθμό κουτιών που πρέπει να προσθέσει ο υπάλληλος και την απόσταση μεταξύ των μηχανημάτων σε πόδια (ft) αντίστοιχα.

ΠΑΡΑΡΤΗΜΑ 2ο (ΠΙΝΑΚΕΣ ΔΕΔΟΜΕΝΩΝ ΠΡΑΚΤΙΚΩΝ ΕΦΑΡΜΟΓΩΝ)

Δεδομένα από Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 333, New Jersey : John Wiley & Sons, Inc. Παραπομπή στη σελίδα 31, υποκεφάλαιο 2.3. Διάγνωση της Πολυσυγγραμμικότητας.

Δεδομένα από Montgomery et al.,(2006). Introduction to Linear Regression Analysis, 4th ed, p. 328, New Jersey : John Wiley & Sons, Inc. Παραπομπή στη σελίδα 112, υποκεφάλαιο 5.2. Εφαρμογή της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων.

Πίνακας 1. Δεδομένα προβλήματος " Ακετυλήνης "

Y	X_1	X_2	X_3
49	1300	7,5	0,012
50,2	1300	9	0,012
50,5	1300	11	0,0115
48,5	1300	13,5	0,013
47,5	1300	17	0,0135
44,5	1300	23	0,012
28	1200	5,3	0,04
31,5	1200	7,5	0,038
34,5	1200	11	0,032
35	1200	13,5	0,026
38	1200	17	0,034
38,5	1200	23	0,041
15	1100	5,3	0,084
17	1100	7,5	0,098
20,5	1100	11	0,092
29,5	1100	17	0,086

Τα δεδομένα του προβλήματος " Ακετυλήνης ", περιγράφουν μέσω της μεταβλητής απόκρισης το επί τοις εκατό ποσοστό (%) του επτανίου που μετατράπηκε σε ακετυλήνη. Οι εξαρτημένες μεταβλητές X_1, X_2, X_3 περιγράφουν τη θερμοκρασία του αντιδραστήρα σε βαθμούς Κελσίου, το επί τοις εκατό ποσοστό (%) του υδρογόνου που περιέχεται στο επτάνιο και το χρόνο επαφής σε δευτερόλεπτα αντίστοιχα.

Δεδομένα από Fearn T., (1983). A Misure of Ridge Regression in the Calibration of a Near Infrared Reflectance Instrument, *Journal of the Royal Statistical Society*, vol. 32, (1), p. 74. Παραπομπή στη σελίδα 96, υποκεφάλαιο 5.1. Εφαρμογή της Παλινδρόμησης Κυρίων Συνιστωσών.

Πίνακας 2. Δεδομένα προβλήματος " Περιεκτικότητα σε πρωτείνη "

Y	X_1	X_2	X_3	X_4	X_5	X_6
9,23	468	123	246	374	386	-11
8,01	458	112	236	368	383	-15
10,95	457	118	240	359	353	-16
11,67	450	115	236	352	340	-15
10,41	464	119	243	366	371	-16
9,51	499	147	273	404	433	5
8,67	463	119	242	370	377	-12
7,75	462	115	238	370	353	-13
8,05	488	134	258	393	377	-5
11.39	483	141	264	384	398	-2
9,95	463	120	243	367	378	-13
8,25	456	111	233	365	365	-15
10,57	512	161	288	415	443	12
10,23	518	167	293	421	450	19
11,87	552	197	324	448	467	32
8,09	497	146	271	407	451	11
12,55	592	229	360	484	524	51
8,38	501	150	274	406	407	11
9,64	483	137	260	385	374	-3
11,35	491	147	269	389	391	1
9,70	463	121	242	366	353	-13
10,75	507	159	285	410	445	13
10,75	474	132	255	376	383	-7
11,47	496	152	276	396	404	6

Τα δεδομένα του προβλήματος " περιεκτικότητας σε πρωτεΐνη ", Fearn (1983), αποτελούν τα αποτελέσματα ενός πειράματος που εκτελείται για τη βαθμονόμηση ενός σχεδόν υπέρυθρου οργάνου ανάκλασης NIR, για τη μέτρηση των πρωτεϊνών στα δείγματα σπορίου εδάφους. Η μεταβλητή απόκρισης Y εκφράζει την επί τοις εκατό (%) περιεκτικότητα της πρωτεΐνης, ενώ οι επεξηγηματικές μεταβλητές $X_1 - X_6$ εκφράζουν τις μετρήσεις της ανάκλασης της ακτινοβολίας απ' τα δείγματα σπορίου σε έξι διαφορετικά μήκη κύματος στο φάσμα $1680 - 2310$ nm. Οι παραπάνω μετρήσεις είναι όλες λογαριθμημένες.

Δεδομένα από Pietrogrande et al., (1989). Principal component analysis in structure-retention and retention-activity studies of benzodiazepines. *Chemometrics and Intelligent Laboratory Systems*, vol.5, p. 258. Παραπομπή στη σελίδα 104, υποκεφάλαιο 5.2. Εφαρμογή της Παλινδρόμησης Μερικών Ελαχίστων Τετραγώνων.

Πίνακας 3. Δεδομένα προβλήματος " Χρωματογράφου "

X_1	X_2	X_3	X_4	X_5	X_6	Y_1	Y_2
2,90	2,19	1,49	0,58	-0,76	-0,41	-0,39	2,50
3,17	2,67	1,62	0,11	-0,82	-0,52	-1,58	1,63
3,20	2,69	1,55	-0,31	-0,96	-0,33	-1,13	2,33
3,25	2,78	1,78	-0,56	-0,99	-0,55	-1,18	1,25
3,26	2,77	1,83	-0,53	-0,91	-0,45	-0,71	2,05
3,16	2,71	1,66	0,10	-0,80	-0,51	-1,58	1,19
3,26	2,74	1,68	0,62	-0,71	-0,39	-0,43	3,14
3,29	2,96	1,67	-0,35	-1,19	-0,71	-2,79	1,63
3,59	3,12	1,97	-0,62	-0,93	-0,56	-1,15	1,79
3,68	3,16	1,93	-0,54	-0,82	-0,50	-0,39	2,60
4,17	3,46	2,12	-0,56	-0,97	-0,55	-0,64	2,41
4,77	3,72	2,29	-0,82	-1,37	-0,80	-2,14	0,71
5,04	4,04	2,44	-1,14	-1,40	-0,86	-3,57	1,44

Ο παραπάνω πίνακας περιέχει τα χρωματογραφικά δεδομένα για μια σειρά δεκατριών βενζοδιαζεπινών σε έξι διαφορετικές αντεστραμμένες και κανονικές φάσεις συστημάτων, τα οποία εξήχθησαν από το άρθρο των Pietrogrande, Dondi , Borea και Bighi (1989). Μέσω των δεδομένων αυτών, θα περιγραφεί η σχέση μεταξύ της μοριακής δομής των βενζοδιαζεπινών και της χρωματογραφικής τους διατήρησης ή της βιολογικής τους δραστηριότητας.

Οι μεταβλητές απόκρισης είναι οι Y_1 και Y_2 οι οποίες καθορίζουν τη βιολογική δραστηριότητα. Συγκεκριμένα η μεταβλητή Y_1 εκφράζει τη συγγένεια δέσμευσης των υποδοχέων ενώ η μεταβλητή Y_2 ερμηνεύει μια ψυχοφαρμακολογική δοκιμή που έχει πραγματοποιηθεί. Οι πρώτες τρεις επεξηγηματικές μεταβλητές αφορούν συστήματα αντεστραμμένων φάσεων (reversed phase systems) με τη μεταβλητή $X_1 = C_{18}$, την $X_2 = Ph$ και την $X_3 = CN - R$. Οι υπόλοιπες τρεις επεξηγηματικές μεταβλητές αφορούν συστήματα κανονικών φάσεων (normal phase systems) με τη μεταβλητή $X_4 = NH_2$, τη μεταβλητή $X_5 = CN - N$ και την $X_6 = Si$.

ΕΥΡΕΤΗΡΙΟ ΟΡΩΝ ΚΑΙ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

- **CENTERING** : ΚΕΝΤΡΟΠΟΙΗΣΗ
- **SCALING** : ΤΥΠΟΠΟΙΗΣΗ
- **LEAST SQUARES METHOD** : ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ
- **CORRELATION** : ΣΥΣΧΕΤΙΣΗ
- **REGRESSOR** : ΕΠΕΞΗΓΗΜΑΤΙΚΗ ΜΕΤΑΒΛΗΤΗ
- **VIF** : ΠΑΡΑΓΟΝΤΑΣ ΕΛΕΓΧΟΥ ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑΣ
- **RIDGE REGRESSION** : ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΟΡΥΦΟΓΡΑΜΜΗΣ
- **CONDITION INDICES** : ΔΕΙΚΤΕΣ ΚΑΤΑΣΤΑΣΗΣ
- **OVERFITTED MODEL** : ΥΠΕΡΟΡΙΣΜΕΝΟ ΜΟΝΤΕΛΟ
- **BIASING PARAMETER** : ΠΑΡΑΜΕΤΡΟΣ ΜΕΡΟΛΗΨΙΑΣ
- **PCR** : ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ
- **PLSR** : ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ
- **PLSC** : ΣΥΜΜΕΤΡΙΚΗ ΜΕΘΟΔΟΣ ΜΕΡΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ
- **OLS** : ΚΛΑΣΙΚΗ ΜΕΘΟΔΟΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ
- **Ε.Μ.Π** : ΕΚΤΙΜΗΤΡΙΑ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ
- **ANOVA** : ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ
- **MULTIPLE LINEAR REGRESSION** : ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ
- **SVD** : ΜΕΘΟΔΟΣ ΙΔΙΟΑΝΑΛΥΣΗΣ
- **EIGENVECTOR** : ΙΔΙΟΔΙΑΝΥΣΜΑ
- **EIGENVALUE** : ΙΔΙΟΤΙΜΗ
- **COVARIANCE MATRIX** : ΜΗΤΡΑ ΣΥΝΔΙΑΣΠΟΡΑΣ
- **LOADING MATRIX** : ΜΗΤΡΑ ΦΟΡΤΙΟΥ
- **SCORES MATRIX** : ΜΗΤΡΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ
- **SINGULAR VALUES** : ΙΔΙΑΖΟΥΣΕΣ ΤΙΜΕΣ
- **SINGULAR VECTORS** : ΙΔΙΑΖΟΝΤΑ ΔΙΑΝΥΣΜΑΤΑ
- **LATENT VARIABLES** : ΛΑΝΘΑΝΟΥΣΕΣ ΜΕΤΑΒΛΗΤΕΣ
- **WEIGHTS** : ΒΑΡΗ
- **RESIDUALS** : ΥΠΟΛΟΙΠΑ
- **PREDICTORS** : ΠΡΟΒΛΕΠΟΥΣΕΣ
- **CROSS VALIDATION** : ΤΕΧΝΙΚΗ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ
- **PRESS** : ΣΤΑΤΙΣΤΙΚΟ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ
- **FITTING** : ΠΡΟΣΑΡΜΟΓΗ
- **VALIDATION** : ΕΠΙΚΥΡΩΣΗ
- **JACK-KNIFING TECHNIQUE** : ΣΤΑΤΙΣΤΙΚΗ ΤΕΧΝΙΚΗ ΑΞΙΟΛΟΓΗΣΗΣ ΔΙΑΣΠΟΡΑΣ ΕΚΤΙΜΗΤΡΙΑΣ
- **MINITAB** : ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ
- **R** : ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΞΕΝΟΓΛΩΣΣΗ

Abdi H. and Williams L.J., (2010). Principal component analysis, Wiley Interdisciplinary Review: *Computational Statistics*, vol 2, pp. 433-459.

Alun T., (2009). Bootstrapping, jackknifing and cross validation. Reusing your data, week 08, " Statistic for Biomedical Informatics ", lesson, Department of Biomedical Informatics, University of Utah.

Catell R.B., (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, vol 1, (2), pp. 245-276.

Cliff N., (1988). The eigenvalues-greater-than-one-rule and the reliability of components. *Psychological Bulletin*, vol 103, pp. 276-279.

Dayal B.S. and MacGregor J.F., (1997). Improved PLS Algorithms. *Journal of Chemometrics,* vol 11, (1), pp. 73-85.

De Jong S., (1993). SIMPLS : An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems,* vol 18, pp. 251-263.

Draper N.R. and Smith H., (1998). *Applied Regression Analysis,* 3rd ed., New Jersey : John Wiley & Sons, Inc.

Farrar D.E. and Glauber R.R., (1967). Multicollinearity in Regression Analysis : The problem revisited, *Review of Econometrics and Statistics,* vol 49, pp. 92-107.

Fearn T., (1983). A Misure of Ridge Regression in the Calibration of a Near Infrared Reflectance Instrument, *Journal of the Royal Statistical Society*, vol 32, (1), pp. 73-79.

Hastie T., Tibshirani R. and Friedman J., (2001). *The Elements of Statistical Learning,* New York : Springer-Verlag.

Hoerl A.E. and Kennard R.W., (1970). Ridge Regression: Biased Estimation for the Non Orthogonal Problems, *Technometrics*, vol 12, (1), pp. 55-67.

Krishnan A., Williams L.J, McIntosh A.R. and Abdi H., (2010). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review, *Neuroimage*, pp. 1-21.

Lachenbruch P.A. and Mickey M.R., (1968). Estimation of Error Rates in Discriminant Analysis, *Technometrics*, vol 10, (1), pp. 1-11.

Marquardt D.W., (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, *Technometrics,* vol 12, pp. 591-612.

Martens H. and Martens M., (2000). Modified jack-knife estimation of parameter uncertainty in bilinear modeling (PLSR). *Food Qual. Preference*, vol 11, pp. 5-16.

Martens H. and Naes T., (1989). *Multivariate Calibration,* Chichester : Wiley.

Mevik B.H. and Wehrens R., (2007). The pls Package: Principal Component and Partial Least Squares Regression in R, *Journal of Statistical Software*, vol 18, (2), pp. 1-24.

Mevik B.H. and Cederkvist H.R., (2004). Mean Squared Error of Prediction (MSEP) Estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). *Journal of Chemometrics*, vol 18, (9), pp. 422-429.

Montgomery D.C., Peck E.A. and Vining G.G., (2006). *Introduction to Linear Regression Analysis*, 4th ed., New Jersey : John Wiley & Sons, Inc.

Pietrogrande M.C., Dondi F., Borea P.A. and Bighi C., (1989). Principal analysis in structure-retention and retention-activity studies of benzodiazepines, *Chemometrics and Intelligent Laboratory Systems*, vol 5, pp. 257-262.

Sharma S., (1996). Principal Component Analysis, *Applied Multivariate Techniques,* pp. 58-79, New York : John Wiley & Sons, Inc.

Wold S., Sjostrom M. and Eriksson L., (2001). Pls-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, vol 58, pp. 109-130.

ΕΛΛΗΝΟΓΛΩΣΣΗ

Οικονόμου Π. και Καρώνη Χ., (2010). «*Στατιστικά μοντέλα παλινδρόμησης».* Αθήνα : Εκδόσεις Συμεών.

ΗΛΕΚΤΡΟΝΙΚΗ

http://cran.r-project.org/web/packages/pls/