ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Ανάλυση Κύριων Συνιστωσών και Εφαρμογές σε Πραγματικά Σεισμολογικά Δεδομένα

Διπλωματική εργασία
της
Βασιλικής Τακτικού

Επιβλέπων: Χρήστος Κουκουβίνος
Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ
ΙΟΥΛΙΟΣ 2012

Στην οικογένειά μου

για την υποστήριξή τους

Περιεχόμενα

Περίληψη

Στην παρούσα εργασία περιγράφεται η μέθοδος της Ανάλυσης Κυρίων Συνιστωσών, μια τεχνική ανάλυσης δεδομένων με σκοπό τη δημιουργία μεταβλητών, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών, έτσι ώστε να είναι ασυσχέτιστες μεταξύ τους και να περιέχουν όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Καταλήγουμε, δηλαδή, σε ένα πιο μικρό αριθμό μεταβλητών από ότι είχαμε αρχικά, που ονομάζονται Κύριες Συνιστώσες, οι οποίες είναι ασυσχέτιστες και μπορούν να ερμηνεύσουν το μεγαλύτερο ποσοστό της διακύμανσης.

Στη συνέχεια περιγράφεται η μέθοδος της ανάλυσης κύριων συνιστωσών με επίβλεψη, που είναι παρόμοια με τη συμβατική ανάλυση κυρίων συνιστωσών με τη διαφορά ότι χρησιμοποιεί ένα υποσύνολο των μεταβλητών πρόβλεψης που έχουν επιλεγεί με βάση τη συσχέτιση τους με το εξαγόμενο αποτέλεσμα. Η τεχνική αυτή αναπτύχθηκε από τους Bair, Hastie και Tibshirani και φαίνεται να έχει καλύτερη απόδοση από τη συμβατική ανάλυση κύριων συνιστωσών.

Τέλος, πραγματοποιείται εφαρμογή σε πραγματικά σεισμολογικά δεδομένα υψηλής διάστασης με τη βοήθεια του πακέτου Clementine. Βρήκαμε τις κύριες συνιστώσες καθώς και ποιες μεταβλητές επηρεάζουν σημαντικά το μοντέλο μας.

1. Μέθοδος της Ανάλυσης Κυρίων Συνιστωσών

1.1. Εισαγωγή στο Data Mining

Το Data Mining είναι όρος που χρησιμοποιείται για να περιγράψει το σύνολο της διαδικασίας εξόρυξης γνώσης από βάσεις δεδομένων. Πιο συγκεκριμένα είναι η μη τετριμμένη εξόρυξη σιωπηρής, προηγούμενα άγνωστης, και πιθανά χρήσιμης πληροφορίας από τα δεδομένα. Αποτελεί σύγχρονη εξέλιξη και το ερευνητικό του πεδίο αποτελεί τομή μεθόδων και εργαλείων που πηγάζουν από τη στατιστική, τη μηχανική μάθηση, βάσεις και αποθήκες δεδομένων.
Οι σχέσεις που προκύπτουν από το data mining συχνά αναφέρονται ως μοντέλα ή πρότυπα (patterns) και περιλαμβάνουν γραμμικές εξισώσεις, κανόνες, συστάδες (clusters), γραφήματα, δέντρα και επαναλαμβανόμενα πρότυπα σε χρονοσειρές.
Η ιδέα του data mining είναι να κατασκευάσουμε προγράμματα ηλεκτρονικών υπολογιστών που να εξετάζουν αυτόματα τις βάσεις δεδομένων, ψάχνοντας για κανονικότητες και πρότυπα. Αν βρεθούν ισχυρά πρότυπα, αυτά είναι δυνατόν να κάνουν ακριβείς προβλέψεις για μελλοντικά δεδομένα. Βέβαια, υπάρχουν κάποια προβλήματα όπως ότι ο όγκος δεδομένων άρα και ο αριθμός πιθανών προτύπων κάποιες φορές είναι τεράστιος, πολλά πρότυπα είναι τετριμμένα και χαμηλού ενδιαφέροντος, άλλα είναι εσφαλμένα, εξαρτημένα από τυχαίες συμπτώσεις στη βάση δεδομένων που χρησιμοποιήθηκε. Επιπλέον, τα πραγματικά στοιχεία είναι ατελή ή ελαττωματικά: κάποια είναι διαστρεβλωμένα ενώ άλλα λείπουν. Οτιδήποτε ανακαλυφθεί θα ειναι ανακριβές. Επομένως αναζητούνται ανθεκτικοί αλγόριθμοι για να ανταποκρίνονται σε μη τέλεια δεδομένα και να εντοπίζουν πρότυπα και κανονικότητες στα δεδομένα αυτά.
Η εύρεση ισχυρών προτύπων, αν υπάρχουν, είναι ένα πολύ χρήσιμο εργαλείο για την ακριβή πρόβλεψη μελλοντικών δεδομένων, για τη γενίκευση από ένα δείγμα του συνόλου στο πλήρες σύνολο και για τη συμπίεση μεγάλων δεδομένων σε μικρότερα με σκοπό να γίνουν πιο κατανοητά και πιο χρήσιμα.
Για να είναι τα αποτελέσματα της εφαρμογής του data mining σε πρακτικά προβλήματα ασφαλή, δε θα πρέπει να στηρίζονται μόνο στην εφαρμογή των αλγορίθμων του σε υποδείγματα, δηλαδή στη μηχανική μάθηση (machine learning) μέσω υποδειγμάτων, αλλά να συνδυάζονται και με στατιστική ανάλυση. Μπορούμε επομένως να πούμε ότι η στατιστική ανάλυση και οι αλγόριθμοι εξόρυξης πληροφορίας από βάσεις δεδομένων αποτελούν τα δύο βασικά συστατικά του data mining για την ανάλυση δεδομένων πρακτικών προβλημάτων.
Οι αλγόριθμοι για την ανάλυση δεδομένων έχουν μελετηθεί από στατιστικούς και έχουν χρησιμοποιηθεί σε ποικίλους επιστημονικούς κλάδους εδώ και πολλά χρόνια, όμως νέοι αλγόριθμοι χρειάζεται να σχεδιαστούν για να διευθετήσουν τους περιορισμούς των υπαρχουσών τεχνικών που προκύπτουν από τους νέους τύπους δεδομένων που συλλέγονται. Η πρόσφατη πρόοδος στην τεχνολογία πληροφοριών έκανε ικανή τη συγκέντρωση τεράστιων ποσών δεδομένων στο εμπόριο και σε διάφορους επιστημονικούς κλάδους. Πολλά από αυτά τα σύνολα δεδομένων είναι

υψηλών διαστάσεων, ετερογενή, διασκορπισμένα ή χώρου-χρόνου και οι παραδοσιακές τεχνικές δε μπορούν να εφαρμοστούν σε αυτά.
Το αναδυόμενο πεδίο του data mining διορθώνει τους περιορισμούς των υπαρχουσών τεχνικών ανάλυσης δεδομένων απευθυνόμενο σε αυτούς τους νέους τύπους δεδομένων. Μέσα σε 10 χρόνια, από το τέλος του 2ου Παγκοσμίου Πολέμου, το πεδίο του data mining εξελίχθηκε ραγδαία και συνεχίζει να παράγει μεγάλο αριθμό αλγορίθμων που απευθύνονται σε περαιτέρω περιορισμούς.
Υπάρχουν πολλές προκλήσεις και απαιτήσεις που θα πρέπει να αντιμετωπίσει ο αναλυτής για την ανάπτυξη αποτελεσματικών αλγορίθμων. Μία από αυτές τις απαιτήσεις είναι η τεράστια ποικιλία των τύπων των δεδομένων και των στόχων του data mining, η οποία καθιστά αδύνατη την ύπαρξη ενός μοναδικού και ταυτόχρονα αποτελεσματικού συστήματος data mining και η οποία επιβάλλει την κατασκευή ιδιαίτερων αλγορίθμων για συγκεκριμένους τύπους δεδομένων. Μία άλλη απαίτηση που θα πρέπει να ικανοποιούν οι αλγόριθμοι είναι η αποτελεσματικότητά τους και η δυνατότητά τους να κλιμακώνονται σε μεγάλες βάσεις δεδομένων σε αποδεκτό και αναμενόμενο χρόνο. Επιπλέον, οι αλγόριθμοι θα πρέπει να διαχειρίζονται σωστά το θόρυβο και τα δεδομένα που αποτελούν εξαιρέσεις, να καταλήγουν σε αποτελέσματα που απεικονίζουν με ακρίβεια τα περιεχόμενα της βάσης δεδομένων και να παρέχουν τη δυνατότητα στον αναλυτή να εξετάσει την ανακαλυφθείσα γνώση από διαφορετικές οπτικές γωνίες και διαφορετικές μορφές, ενθαρρύνοντας την αλληλεπίδρασή τους. Τέλος, βασική απαίτηση που πρέπει να ικανοποιούν οι αλγόριθμοι του data mining είναι ότι η μη τελειότητα που παρουσιάζουν τα αποτελέσματά τους θα πρέπει να μπορεί να εκφραστεί με μέτρα αβεβαιότητας σε μορφή προσεγγιστικών ή ποσοτικών κανόνων. Έτσι οδηγούμαστε στην μελέτη της μέτρησης της ποιότητας της ανακαλυφθείσας γνώσης καθώς και του ενδιαφέροντος που παρουσιάζει και της αξιοπιστίας της, κατασκευάζοντας στατιστικά μοντέλα και εργαλεία.
Οι βάσεις δεδομένων οι οποίες χρησιμοποιούνται στα περισσότερα πρακτικά προβλήματα εξόρυξης γνώσης από δεδομένα είναι πολύ μεγάλων διαστάσεων καθώς αποτελούνται από εκατομμύρια εγγραφές και εκατοντάδες μεταβλητές (χαρακτηριστικά). Θεωρητικά, η ύπαρξη περισσότερων μεταβλητών κάνει περισσότερο αποδοτική τη διαδικασία μάθησης όμως στην πράξη η προσθήκη μη σχετικών μεταβλητών «συγχύζει» τους αλγόριθμους που εφαρμόζουμε. Η χρήση πολλών μεταβλητών για τη μοντελοποίηση μίας σχέσης με μια μεταβλητή απόκρισης μπορεί να περιπλέξει την ερμηνεία της ανάλυσης και να παραβεί την αρχή της φειδωλότητας (principle of parsimony), σύμφωνα με την οποία πρέπει να κρατήσουμε τον αριθμό των μεταβλητών σε ένα μέγεθος το οποίο να μπορούμε εύκολα να εξηγούμε. Επιπλέον, η παραμονή πάρα πολλών μεταβλητών μπορεί να οδηγήσει σε υπερπροσαρμογή όπου η γενικότητα των αποτελεσμάτων που βρίσκουμε παρεμποδίζεται επειδή τα νέα δεδομένα δε συμπεριφέρονται το ίδιο με τα δεδομένα εκπαίδευσης για όλες τις μεταβλητές.
Οι μέθοδοι μείωσης των διαστάσεων της βάσης δεδομένων χρησιμοποιούν τις συσχετίσεις μεταξύ των μεταβλητών για να μειώσουν τον αριθμό τους, να επιβεβαιώσουν ότι αυτές οι μεταβλητές είναι ανεξάρτητες και να ερμηνεύσουν τα αποτελέσματα. Δύο τέτοιες μέθοδοι είναι η ανάλυση κυρίων συνιστωσών (principal component analysis) και η παραγοντική ανάλυση (factor analysis).

1.2. Ανάλυση σε Κύριες Συνιστώσες

Η κλασσική Στατιστική με την ανάλυση κυρίων συνιστωσών έχει ως σκοπό τον προσδιορισμό των κυρίων αξόνων ενός ελλειψοειδούς που παράγεται από μια πολυδιάστατη κανονική κατανομή. Αρχικά η μέθοδος αυτή αναπτύχθηκε από τον Harold Hotelling (1933) και στη συνέχεια παρουσιάστηκε από πολλά βιβλία που ασχολούνται με την κλασσική ανάλυση πολυδιάστατων μεταβλητών (Anderson (1958), Kendall και Stuart (1968) κλπ.).
Τα τελευταία χρόνια οι αναλυτές δεδομένων έχουν αναπτύξει μια εντελώς διαφορετική άποψη για την ανάλυση κυρίων συνιστωσών, την οποία χρησιμοποιούν ως μια τεχνική για περιγραφή δεδομένων στα οποία έχουν γίνει βελτιστοποιήσεις συγκεκριμένων αλγεβρο-γεωμετρικών μεγεθών, χωρίς υποθέσεις για κατανομές ή στατιστικά μοντέλα.
Η εξέλιξη των υπολογιστών που μας παρέχουν σχετικά πακέτα, σε συνδυασμό με την εργασία του Benzecri που δημιούργησε "σχολή ανάλυσης δεδομένων" έτσι ώστε να μιλάμε για τη Γαλλική Σχολή Δεδομένων, συντέλεσαν αποφασιστικά στη διάδοση αυτής της αλγεβρο-γεωμετρικής μεθόδου ανάλυσης δεδομένων.
Το πρόβλημα που αντιμετωπίζουν συχνά οι αναλυτές δεδομένων είναι να μελετήσουν έναν πίνακα στον οποίο οι στήλες αντιπροσωπεύουν μεταβλητές και οι γραμμές είναι οι μετρήσεις των μεταβλητών στα συγκεκριμένα άτομα του δείγματος. Τις περισσότερες φορές τα δεδομένα είναι πολύ μεγάλου πλήθους και η εξαγωγή συμπερασμάτων από αυτά πάρα πολύ δύσκολη. Η ανάλυση κυρίων συνιστωσών (Principal Components Analysis) είναι μία μέθοδος μείωσης των διαστάσεων μίας βάσης δεδομένων η οποία περιγράφηκε αρχικά το 1901 από τον Karl Pearson. Μας επιτρέπει δηλαδή να συγκεντρώσουμε την πληροφορία που περιέχουν τα αρχικά μας δεδομένα σε πίνακες με λιγότερα στοιχεία και συγχρόνως παρέχει μια γεωμετρική αναπαράσταση της πληροφορίας. Η μέθοδος έχει σκοπό να δημιουργήσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε

- οι γραμμικοί αυτοί συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους, αλλά,
- να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών.

Το κέρδος από μια τέτοια διαδικασία είναι πως:

- Από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών, κάτι το οποίο για ορισμένες στατιστικές μεθόδους είναι περισσότερο χρήσιμο. Για παράδειγμα στο πρόβλημα της πολυσυγγραμικότητας στην παλινδρόμηση, αν χρησιμοποιήσουμε τις συσχετισμένες μεταβλητές οι εκτιμήσεις που θα πάρουμε δεν θα είναι συνεπείς ενώ αν χρησιμοποιούσαμε ασυσχέτιστες μεταβλητές το πρόβλημα αυτό δεν θα υπήρχε.

- Αν οι κύριες συνιστώσες που θα προκύψουν μπορούν να ερμηνεύσουν ένα μεγάλο ποσοστό της διακύμανσης τότε αυτό σημαίνει πως αντί να έχουμε p μεταβλητές όπως είχαμε αρχικά, έχουμε λιγότερες, με κόστος βέβαια ότι χάνουμε κάποιο (ελπίζουμε μικρό) ποσοστό της συνολικής μεταβλητότητας. Σε μερικές εφαρμογές αυτό είναι ζωτικής σημασίας. Για παράδειγμα σε μια τεράστια βάση δεδομένων αντί να αποθηκεύουμε όλες τις μεταβλητές μπορούμε να αποθηκεύουμε μόνο κάποιον αριθμό κυρίων συνιστωσών.

Σίγουρα χάνουμε κάποιο μέρος της πληροφορίας αλλά το κέρδος σε χώρο αλλά και ταχύτητα επεξεργασίας μπορεί να είναι τεράστιο. Από την άλλη πλευρά πολλές φορές συμβαίνει να έχουμε λίγες παρατηρήσεις αλλά πολλές μεταβλητές. Τέτοια προβλήματα για παράδειγμα εμφανίζονται στην αρχαιομετρία ένα πεδίο εφαρμογής στατιστικών μεθόδων, στην αρχαιολογία, όπου τα αντικείμενα που θέλει κάποιος να μελετήσει είναι συνήθως λίγα αλλά τα στοιχεία και οι μεταβλητές που έχει είναι πάρα πολλά. Η μείωση των διαστάσεων του προβλήματος φαντάζει η μόνη λύση για να προχωρήσει κανείς σε στατιστική επεξεργασία.

- Ένα άλλο μεγάλο πλεονέκτημα (το οποίο από την άλλη ίσως είναι και μειονέκτημα για πολλούς) είναι πως με τη μέθοδο των κυρίων συνιστωσών μπορούμε να εξετάσουμε τις συσχετίσεις ανάμεσα στις μεταβλητές και να διαπιστώσουμε πόσο οι μεταβλητές μοιάζουν ή όχι. Επίσης η μέθοδος μας επιτρέπει να αναγνωρίσουμε δίνοντας ονόματα στις καινούριες μεταβλητές (τις συνιστώσες) παρατηρώντας ποιες από τις αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές. Αυτό είναι πολύ χρήσιμο σε κάποιες επιστήμες καθώς μας επιτρέπουν να ποσοτικοποιήσουμε μη μετρήσιμες ποσότητες, όπως η αγάπη, η ευφυΐα, η ικανότητα ενός μπασκετμπολίστα, η εμπορευσιμότητα ενός προϊόντος κλπ. δηλαδή αφηρημένες έννοιες. Το γεγονός βέβαια πως τέτοιες ερμηνείες εμπεριέχουν σε μεγάλο βαθμό υποκειμενικά κριτήρια έχει οδηγήσει πολλούς στο να κατηγορούν τη μέθοδο και να μην την εμπιστεύονται.

1.2.1. Η Βασική Ιδέα της Μεθόδου των Κυρίων Συνιστωσών

Πριν ξεκινήσουμε την περιγραφή της μεθόδου των κυρίων συνιστωσών είναι χρήσιμο να δούμε κάποια πράγματα από τη γραμμική άλγεβρα τα οποία και αποτέλεσαν τη βασική ιδέα πάνω στην οποία αναπτύχθηκε η μέθοδος.
Έστω ένας τετραγωνικός συμμετρικός πίνακας **A** διαστάσεων *p x p*. Ο πίνακας αυτός μπορεί να αναπαρασταθεί ως

$$\mathbf{A} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}'$$

όπου **Λ** είναι ένας *p x p* διαγώνιος πίνακας όπου τα στοιχεία της διαγωνίου είναι οι ιδιοτιμές του πίνακα **A** , δηλαδή

$$\mathbf{\Lambda}=\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & \\ & \dots & & \\ 0 & \dots & \dots & \lambda_p \end{bmatrix}$$

και **P** ένας ορθογώνιος *p x p* πίνακας (δηλαδή ισχύει **P' P = I**) ο οποίος αποτελείται από τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών. Η παραπάνω αναπαράσταση του πίνακα Α ονομάζεται φασματική ανάλυση του πίνακα **A.**
Επομένως αφού ο πίνακας είναι ορθογώνιος θα ισχύει πως $\mathbf{P^{-1}} = \mathbf{P'}$.
Μπορεί κάποιος να δείξει με βάση τις παραπάνω ιδιότητες πως ισχύει

$$\mathbf{\Lambda} = \mathbf{P}'\,\mathbf{A}\,\mathbf{P} \qquad (1.1)$$

καθώς και

$$\mathbf{A} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}' \Leftrightarrow \mathbf{P^{-1}}\,\mathbf{A} = \mathbf{P^{-1}}\,\mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}' \Leftrightarrow$$

$$\Leftrightarrow \mathbf{P^{-1}\,\Lambda\,P = \Lambda\,P'\,P = \Lambda}$$

Αν δηλαδή ξεκινήσουμε από έναν τετραγωνικό πίνακα **Α** μπορούμε να καταλήξουμε σε έναν διαγώνιο πίνακα **Λ.**
Γιατί αυτό όμως μας είναι τόσο χρήσιμο; Αν έχουμε ένα τυχαίο διάνυσμα **X** το οποίο έχει πίνακα διακύμανσης **Σ** τότε το διάνυσμα **Y=BX** έχει πίνακα διακύμανσης **B΄ΣB.**
Αν τώρα κοιτάξουμε την σχέση (1.1) βλέπουμε πως από έναν τετραγωνικό πίνακα μπορώ να οδηγηθώ σε έναν διαγώνιο πίνακα, πολλαπλασιάζοντας με έναν κατάλληλο πίνακα **P** και άρα αν ο τετραγωνικός πίνακας είναι πίνακας διακύμανσης καταλήγουμε σε έναν διαγώνιο πίνακα διακύμανσης. Δηλαδή το τυχαίο διάνυσμα που αντιστοιχεί στον πίνακα αυτόν είναι ασυσχέτιστο. Δηλαδή αυτό που μου δίνει η φασματική ανάλυση ενός πίνακα διακύμανσης είναι πως αν πολλαπλασιάσω το αρχικό διάνυσμα με έναν κατάλληλο πίνακα μπορώ να δημιουργήσω έναν νέο διάνυσμα το οποίο να είναι ασυσχέτιστο, να έχει δηλαδή διαγώνιο πίνακα διακύμανσης.

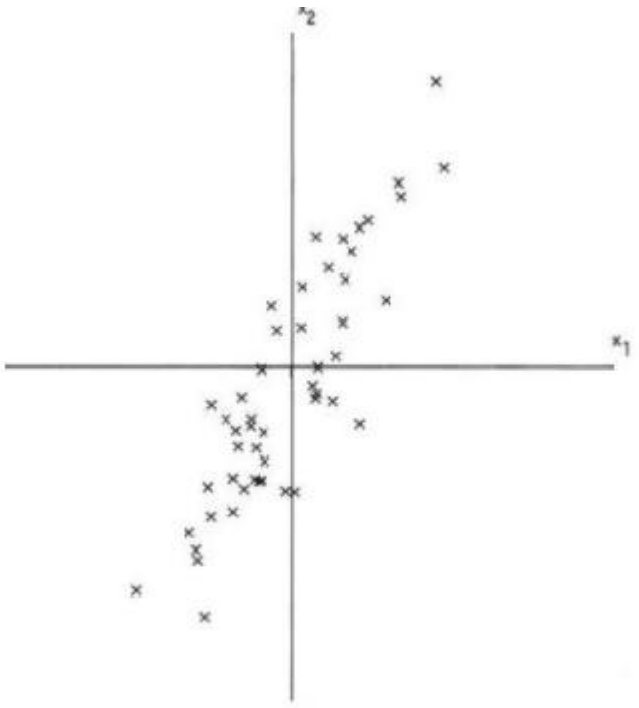
1.2.2. Εύρεση των κυριών συνιστωσών

Όπως είπαμε προηγούμενα η μέθοδος στηρίζεται στη φασματική ανάλυση ενός τετραγωνικού πίνακα. Αυτό σημαίνει πως μπορούμε να χρησιμοποιήσουμε είτε τον πίνακα διακυμάνσεων είτε τον πίνακα συσχετίσεων που είναι στην ουσία ο πίνακας διακυμάνσεων των τυποποιημένων δεδομένων.
Έστω λοιπόν πως έχουμε ένα σύνολο από διανύσματα k διαστάσεων με τη μορφή (X_1, X_2,..., X_k). Η ανάλυση κυρίων συνιστωσών μεταφέρει τα διανύσματα αυτά σε έναν άλλο χώρο, ο οποίος έχει και αυτός k διαστάσεις και είναι ο χώρος των κυρίων συνιστωσών (principal components). Μετατρέπει, δηλαδή, τα αρχικά διανύσματα στη μορφή (Y_1, Y_2,..., Y_k). Έτσι, τα νέα διανύσματα έχουν k κύριες συνιστώσες.
Οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους, και είναι υπολογισμένες με τέτοιο τρόπο, ώστε το μεγαλύτερο ποσοστό της μεταβλητότητας του δείγματος των διανυσμάτων να αντιπροσωπεύεται από όσο το δυνατό λιγότερες μεταβλητές. Πιο συγκεκριμένα, οι κύριες συνιστώσες συνηθίζεται να διατάσσονται με της εξής φθίνουσα σειρά. Η πρώτη κύρια συνιστώσα (Y_1) είναι η κύρια συνιστώσα που εκφράζει το μεγαλύτερο ποσοστό της μεταβλητότητας του δείγματος. Η δεύτερη κύρια συνιστώσα (Y_2) είναι η κύρια συνιστώσα που εκφράζει το δεύτερο μεγαλύτερο ποσοστό της μεταβλητότητας του δείγματος. Με ανάλογο τρόπο, η k -οστή κύρια συνιστώσα (Y_k) είναι η κύρια συνιστώσα η οποία εκφράζει το ελάχιστο ποσοστό της μεταβλητότητας του δείγματος. Όλες μαζί οι Y_i συνολικά εκφράζουν το 100% της μεταβλητότητας του δείγματος.
Ο λόγος για τον οποίον οι κύριες συνιστώσες κατασκευάζονται με τη λογική αυτή της φθίνουσας σειράς, είναι ότι αν ένας μικρός μόνο αριθμός $q(q<k)$ από Y_i αρκεί για να καλυφθεί ένα μεγάλο ποσοστό της μεταβλητότητας του δείγματος, τότε τα διανύσματα μπορούν να συμπιεστούν από k σε q συνιστώσες, με μικρό σφάλμα.
Κάθε μία κύρια συνιστώσα (Y_1, Y_2,..., Y_k) ορίζεται να είναι γραμμικός συνδυασμός των μεταβλητών (X_1, X_2,..., X_k) των αρχικών διανυσμάτων, δηλαδή η i-οστή κύρια συνιστώσα Y_i έχει τη μορφή:

$$Y_i = a_1X_1 + a_2X_2 + \ldots + a_kX_k.$$

Παράδειγμα 1.

Ο παραπάνω θεωρητικός ορισμός της Ανάλυσης Κύριων Συνιστωσών μπορεί να γίνει περισσότερο κατανοητός, παραθέτοντας το παρακάτω απλοποιημένο παράδειγμα. Έστω ότι υπάρχει ένα δείγμα από ένα πλήθος διανυσματικών μετρήσεων των 2 διαστάσεων. Το δείγμα αυτό απεικονίζεται στο Σχήμα 1.1. Δηλαδή, οι διανυσματικές μετρήσεις του δείγματος είναι της μορφής: (X_1, X_2) , με k =2. Εφαρμόζοντας Ανάλυση Κύριων Συνιστωσών, οι διανυσματικές μετρήσεις μετατρέπονται σε διανύσματα από κύριες συνιστώσες, και έχουν τη μορφή (Y_1, Y_2).
Ο βασικός στόχος της PCA διαδικασίας είναι να μπορούν τα διανύσματα να συμπιέζονται από 2 σε 1 διάσταση, χωρίς σημαντική αύξηση του σφάλματος. Με το παράδειγμα αυτό θα προσπαθήσουμε να διαισθανθούμε κάτω υπό ποιες συνθήκες μπορεί να επιτευχθεί αυτό, αλλά και τι σημαίνει το σφάλμα.



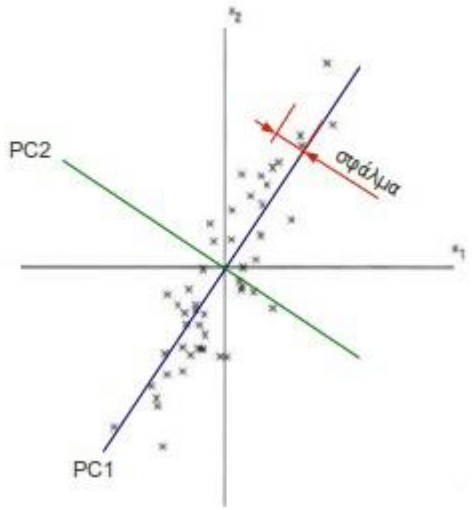
Σχήμα 1.1. Δείγμα δισδιάστατων μετρήσεων

Όπως αναφέρθηκε προηγουμένως, κάθε μία κύρια συνιστώσα είναι ένας γραμμικός συνδυασμός των πραγματικών μεταβλητών. Έτσι, μπορούμε να πούμε ότι, για κάθε μία μετασχηματισμένη σε Y_i μέτρηση του δείγματος, ισχύει ότι

$$(Y_1, Y_2) = (\alpha_1 X_1 + \alpha_2 X_2, \beta_1 X_1 + \beta_2 X_2).$$

Η κάθε μία Y_i εκφράζει έναν άξονα στο σχήμα της γραφικής παράστασης του δείγματος. Και επειδή οι Y_i είναι ανεξάρτητες και ορθογώνιες μεταξύ τους, οι δύο αυτοί άξονες θα είναι κάθετοι μεταξύ τους.
Επίσης, όπως αναφέραμε προηγουμένως, ο Y_1 θα πρέπει να εκφράζει το μεγαλύτερο δυνατό ποσοστό της μεταβλητότητας του δείγματος. Αυτό σημαίνει ότι ο άξονας του Y_1 θα πρέπει να είναι τοποθετημένος στο σχήμα με τέτοιο τρόπο ώστε να εκφράζει το μεγαλύτερο δυνατό ποσοστό της μεταβλητότητας του δείγματος.

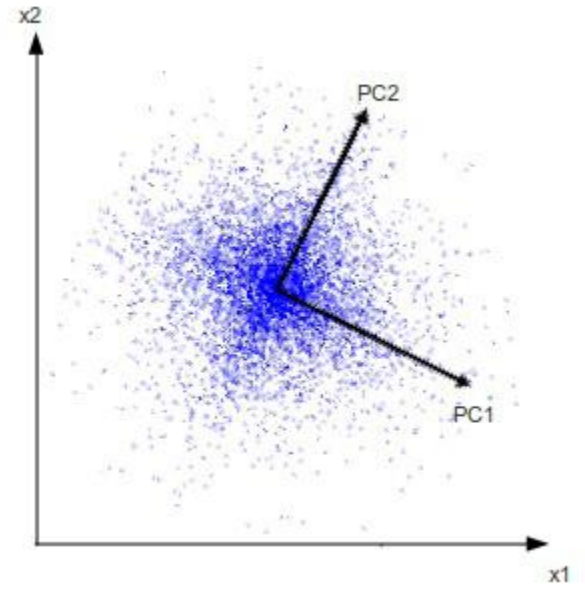
Στο Σχήμα 1.2, ο άξονας του Y_1 είναι η ευθεία κατά μήκος της οποίας οι μετρήσεις του δείγματος παρουσιάζουν τη μεγαλύτερη μεταβλητότητα.
Στο ίδιο σχήμα, ο άξονας του Y_2 είναι κάθετος στου Y_1, και κατά μήκος του η μεταβλητότητα των μετρήσεων είναι σαφώς μικρότερη.
Αν χρησιμοποιήσουμε μόνο τον Y_1, και αγνοήσουμε τον Y_2, τότε οι διανυσματικές μετρήσεις συμπιέζονται από δύο σε μία διαστάσεις. Στην περίπτωση του παραδείγματός μας, αν λάβουμε υπόψη μόνο τον άξονα του Y_1, τότε όλες οι μετρήσεις, όταν "αποσυμπιεστούν", θα βρίσκονται



Σχήμα 1.2. Εφαρμογή PCA σε συσχετισμένο δείγμα

πάνω στον άξονα του Y_1. Τότε, το σφάλμα οφείλεται στην απώλεια πληροφορίας και εκφράζεται από την απόσταση των πραγματικών
μετρήσεων από τον άξονα του Y_1, όπου βρίσκονται οι αποσυμπιεσμένες μετρήσεις. Επομένως, όσο μικρότερη είναι η μεταβλητότητα κατά μήκος του άξονα Y_2 που αγνοήθηκε, τόσο μικρότερο είναι και το σφάλμα λόγω απώλειας πληροφορίας. Στο παράδειγμά μας, οι μεταβλητές X_1 και X_2 είναι έντονα συσχετισμένες μεταξύ τους, λόγω του ότι τείνουν να έχουν μία σταθερή αναλογία. Η συσχέτιση αυτή των μεταβλητών έχει ως συνέπεια ο άξονας Y_1 να εκφράζει την σταθερή αναλογία που τείνουν να έχουν οι X_1 και X_2, τόσο, ώστε ο Y_1 να εκφράζει πολύ μεγάλο ποσοστό της μεταβλητότητας των μετρήσεων, και ο Y_2 πολύ μικρό ποσοστό.
Αν οι μετρήσεις δεν ήταν αρκετά συσχετισμένες, όπως στο Σχήμα 1.3, τότε η μεταβλητότητά τους κατά μήκος του άξονα Y_2 θα ήταν μεγάλη, και ως αποτέλεσμα, αν αγνοείτο ο Y_2, οι αποστάσεις των μετρήσεων από τον άξονα Y_1 θα ήταν μεγάλες, άρα το σφάλμα θα ήταν μεγάλο. Συνεπώς, σε αυτή την περίπτωση, θα έπρεπε να ληφθούν υπόψη και ο Y_1 και ο Y_2, και έτσι δεν θα γινόταν συμπίεση.
Από τα παραπάνω προκύπτει ένα πολύ ενδιαφέρον συμπέρασμα. Όταν οι μεταβλητές ενός δείγματος είναι συσχετισμένες μεταξύ τους, τότε αρκούν λιγότερα Y_i για να εκφράσουν ένα πολύ μεγάλο ποσοστό της μεταβλητότητας του δείγματος, και συνεπώς το δείγμα υφίσταται μεγαλύτερη συμπίεση. Αντιθέτως, αν οι μετρήσεις ενός δείγματος είναι

αρκετά ασυσχέτιστες μεταξύ τους, τότε απαιτούνται πολλοί, αν όχι όλοι, Y_i για να εκφράσουν ένα σημαντικό ποσοστό της μεταβλητότητας του δείγματος, και τότε η συμπίεση είναι πολύ μικρή.



Σχήμα 1.3. Εφαρμογή PCA σε ασυσχέτιστο δείγμα

1.3. Ερμηνεία της ΑΚΣ με Γεωμετρικούς Όρους

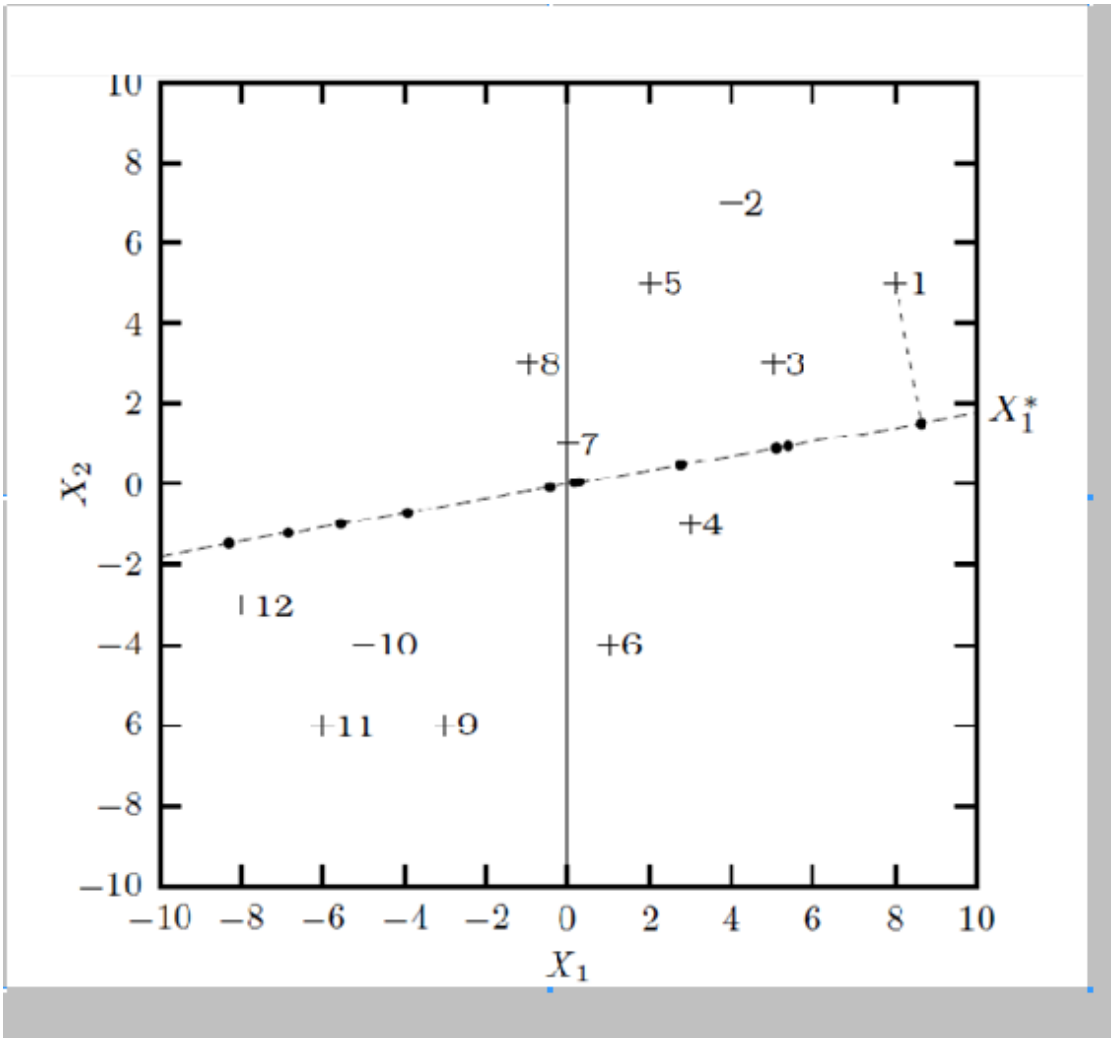
Καθορισμός Εναλλακτικών Αξόνων και Δημιουργία Νέων Μεταβλητών

Στον Πίνακα 2.1 παρουσιάζεται ένα μικρό σύνολο δεδομένων που αποτελείται από 12 παρατηρήσεις και 2 μεταβλητές οι οποίες επιπρόσθετα μετασχηματίζονται έτσι ώστε να έχουν μηδενικό μέσο όρο. Τα δεδομένα μπορούν να παρουσιαστούν και ως αποκλίσεις από το μέσο όρο, δηλαδή αφαίρεση κάθε παρατήρησης της μεταβλητής από το μέσο όρο της (mean-corrected data).
Ας υποθέσουμε ότι ο αρχικός πίνακας είναι ο P με γενικό στοιχείο p_{ij} όπου i=1,…,12 και j =1,2. Αν X ο μετασχηματισμένος πίνακας, τότε

$$x_{ij} = \frac{p_{ij} - \bar{p}_j}{s_j} \quad \text{όπου } \bar{p}_j = \sum_{i=1}^{12} \frac{p_{ij}}{12} \quad \text{και } s_j^2 = \frac{1}{12-1} \sum_{i=1}^{12} (p_{ij} - \bar{p}_j)^2.$$

Πίνακας 2.1. Αρχικές και μετασχηματισμένες μεταβλητές

Αριθμός Παρατήρησης	**Αρχική μεταβλητή p_1**	**Μετασχημα-τισμένη μεταβλητή x_1**	**Αρχική μεταβλητή p_2**	**Μετασχηματισμένη μεταβλητή x_2**
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**
Διακύμανση	**23.091**	**23.091**	**21.091**	**21.091**



Σχήμα 2.1. Γραφική παράσταση των μετασχηματισμένων μεταβλητών και προβολή των σημείων στο νέο άξονα X_1^*

Οι πίνακες συνδιακύμανσης και συσχέτισης των δύο μεταβλητών είναι αντίστοιχα

$$\mathbf{C} = \begin{bmatrix} 23.091 & 16.455 \\ 16.455 & 21.091 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1.000 & 0.746 \\ 0.746 & 1.000 \end{bmatrix}$$

Το Σχήμα 2.1 είναι μια γραφική παράσταση των μετασχηματισμένων μεταβλητών. Από τον Πίνακα 2.1 παρατηρούμε ότι οι διακυμάνσεις των μεταβλητών x_1 και x_2 είναι αντίστοιχα 23.091 και 21.091 και ότι η συνολική διακύμανση είναι 44.182 (δηλ. 23.091+21.091). Επίσης οι μεταβλητές συσχετίζονται με συντελεστή συσχέτισης r = 0.746. Τα ποσοστά της ολικής διακύμανσης που εκφράζουν οι x_1 και x_2 είναι 52.26% και 47.74% αντίστοιχα.

Έστω ένας νέος άξονας X_1^* ο οποίος σχηματίζει γωνία θ μοιρών με τον άξονα X_1. Η συντεταγμένη των σημείων ως προς το νέο άξονα X_1^* λαμβάνεται μετά από την προβολή των σημείων (παρατηρήσεων) στον άξονα X_1^*. Η νέα αυτή συντεταγμένη είναι γραμμικός συνδυασμός των συντεταγμένων κάθε σημείου σε σχέση με το ζεύγος των αριχκών αξόνων και εκφράζεται από την εξίσωση

$$x_1^* = \cos\theta * x_1 + \sin\theta * x_2 \qquad (2.1)$$

όπου x_1^* είναι η συντεταγμένη της εκάστοτε παρατήρησης σε σχέση με τον άξονα X_1^* ενώ x_1 και x_2 είναι οι συντεταγμένες της εκάστοτε παρατήρησης ως προς τους άξονες X_1 και X_2. Είναι προφανές ότι η x_1^*, η οποία είναι ένας γραμμικός συνδυασμός των αρχικών μεταβλητών, μπορεί να θεωρηθεί ως μια νέα μεταβλητή. Αν λοιπόν η κλίση του άξονα X_1^* είναι 10^o, από την εξίσωση 2.1 υπολογίζουμε τη νέα μεταβλητή x_1^*.

$$x_1^* = 0.985\, x_1 + 0.174 x_2$$

Οι τιμές της x_1^* φαίνονται στον Πίνακα 2.2 και στο Σχήμα 2.1. Για παράδειγμα, οι συντεταγμένες της πρώτης παρατήρησης είναι 8 και 5, οπότε η πρώτη τιμή της νέας μεταβλητής, που αντιστοιχεί στην προβολή του σημείου (8,5) στον άξονα X_1^*, είναι 8.747, δηλαδή,

$$0{,}985 * 8 + 0.174 * 5 = 8.747$$

Επίσης από τον Πίνακα 2.2 παρατηρούμε παρατηρούμε τα εξής:

1. Η νέα μεταβλητή έχει μηδενικό μέσο όρο.

2. Η διακύμανση της x_1^* είναι 28.659 και αντιστοιχεί στο 64.87% (28.659/44.182) της ολικής διακύμανσης των δεδομένων, δηλαδή είναι μεγαλύτερη από τις διακυμάνσεις των αρχικών μεταβλητών.

Ας υποθέσουμε ότι ο άξονας X_1^* σχηματίζει γωνία 20^o με τον άξονα X_1. Είναι φανερό πως η νέα μεταβλητή x_1^* θα έχει τώρα νέες τιμές. Αυξάνοντας διαδοχικά την κλίση του X_1^* και υπολογίζοντας την εκάστοτε x_1^* προκύπτει ο Πίνακας 2.3 ο οποίος παρουσιάζει το ποσοστό της διακύμανσης που εκφράζουν οι νέες μεταβλητές. Από το Σχήμα 2.2 παρατηρούμε ότι η αύξηση της κλίσης του X_1^* προκαλεί αύξηση της ολικής διακύμανσης που εκφράζει η x_1^*. Όμως, υπάρχει ένας και μόνο ένας νέος

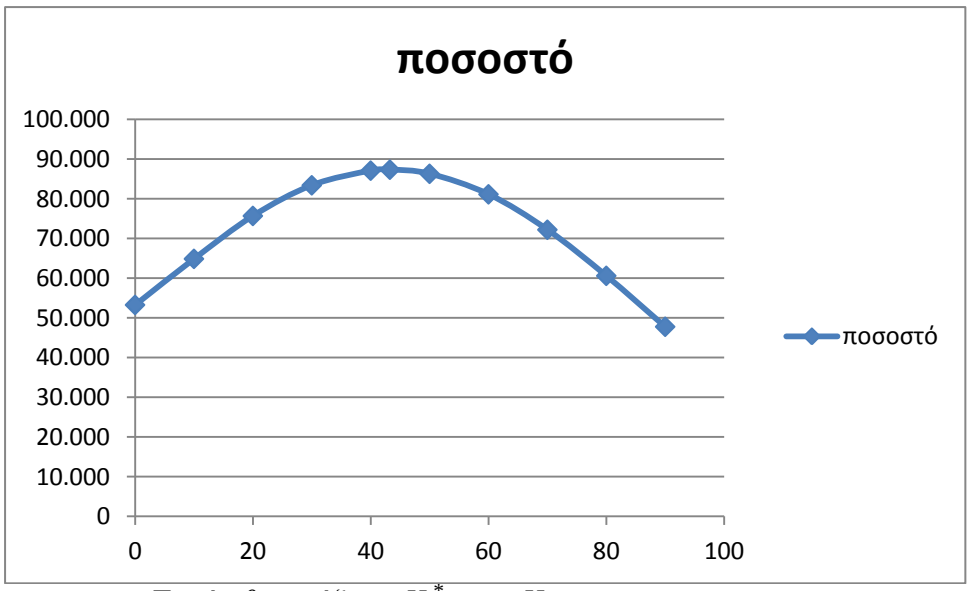
Πίνακας 2.2. Οι μετασχηματισμένες μεταβλητές και η νέα μεταβλητή x_1^* για κλίση του νέου άξονα ίση με 10^o

Αριθμός Παρατήρησης	**Μετασχηματισμένη μεταβλητή x_1**	**Μετασχηματισμένη μεταβλητή x_2**	**Μετασχηματισμένη μεταβλητή 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**
Διακύμανση	**23.091**	**21.091**	**28.659**

Πίνακας 2.3. Η διακύμανση που εκφράζουν οι εκάστοτε νέες μεταβλητές για διάφορους νέους άξονες

Γωνία θ με τον X_1	**Ολική διακύμανση**	**Διακύμανση 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



Γωνία θ του άξονα X_1^* στον X_1

Σχήμα 2.2. Ποσοστό της ολικής διακύμανσης που εκφράζει ο εκάστοτε X_1^*

άξονας του οποίου η αντίστοιχη μεταβλητή εξηγεί το μεγαλύτερο μέρος της ολικής διακύμανσης των δεδομένων. Ο άξονας αυτός έχει κλίση 43,261^o με τον άξονα X_1 και η νέα x_1^* υπολογίζεται από την εξίσωση

$$x_1^* = \cos 43.261 * x_1 + \sin 43.261 * x_2 = 0.728\, x_1 + 0.685\, x_2.$$

Ο Πίνακας 2.4 παρουσιάζει τις τιμές της νέας μεταβλητής x_1^* η οποία εκφράζει το 87.31% της ολικής διακύμανσης. Αυτό σημαίνει ότι ένα μέρος της ολικής διακύμανσης δεν εκφράζεται. Επομένως είναι δυνατό να σχεδιάσουμε ένα νέο άξονα του οποίου η αντίστοιχη νέα μεταβλητή να εξηγεί όσο το δυνατό περισσότερη διακύμανση από την υπόλοιπη διακύμανση από την υπόλοιπη διακύμανση που δεν εξηγεί η x_1^*.
Έστω λοιπόν ένας νέος άξονας X_2^* ο οποίος είναι κατακόρυφος στον X_1^*. Ως συνέπεια η γωνία θ που σχηματίζει ο X_2^* με τον X_2 ισούται με τη γωνία που σχηματίζει ο X_1^* με τον X_1, δηλαδή 43.261. Η γραμμική εξίσωση για τον υπολογισμό της νέας μεταβλητής x_2^* είναι

$$x_2^* = -\sin\theta * x_1 + \cos\theta * x_2 \quad \Leftrightarrow \quad x_2^* = -0.685\, x_1 + 0.728 x_2$$

Ο Πίνακας 2.4 παρουσιάζει επίσης τη νέα μεταβλητή x_2^* και στο Σχήμα 2.3 φαίνονται οι νέοι άξονες. Οι πίνακες συνδιακύμανσης και συσχέτισης των δύο νέων μεταβλητών είναι

$$\mathbf{C} = \begin{bmatrix} 38.576 & 0.000 \\ 0.000 & 5.606 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1.000 & 0.000 \\ 0.000 & 1.000 \end{bmatrix}$$

Από τη μελέτη του Πίνακα 2.4 και του Σχήματος 2.3 εξάγονται τα εξής συμπεράσματα:

1. Η διευθέτηση των σημείων στο δισδιάστατο χώρο δεν αλλάζει, δηλαδή τα σημεία μπορούν να παρουσιαστούν σε σχέση είτε με τους παλιούς είτε με τους νέους άξονες.

2. Η προβολή των σημείων της γραφικής στους αρχικούς άξονες είναι οι τιμές των αρχικών μεταβλητών, ενώ η προβολή των σημείων στους νέους άξονες μας δίνει τις τιμές των νέων μεταβλητών. Οι νέοι άξονες ή οι νέες μεταβλητές ονομάζονται κύριες συνιστώσες (principal components) και οι τιμές των νέων μεταβλητών ονομάζονται τιμές των κυρίων συνιστωσών (principal components scores).

3. Κάθε νέα μεταβλητή (x_1^* και x_2^*) είναι ο γραμμικός συνδυασμός των αρχικών μεταβλητών και παραμένει μετασχηματισμένη ως προς μηδενικό μέσο όρο (mean corrected).

4. Οι διακυμάνσεις των x_1^* και x_2^* είναι αντίστοιχα 38.576 και 5.606 και η συνολική διακύμανση αυτών είναι 44.182 η οποία είναι ίση με τη συνολική διακύμανση των αρχικών μεταβλητών x_1 και x_2. Ως συμπέρασμα η ολική διακύμανση των δεδομένων δεν αλλάζει, κάτι που είναι αναμενόμενο αφού παραμένει αμετάβλητη η διευθέτηση των σημείων στο χώρο.

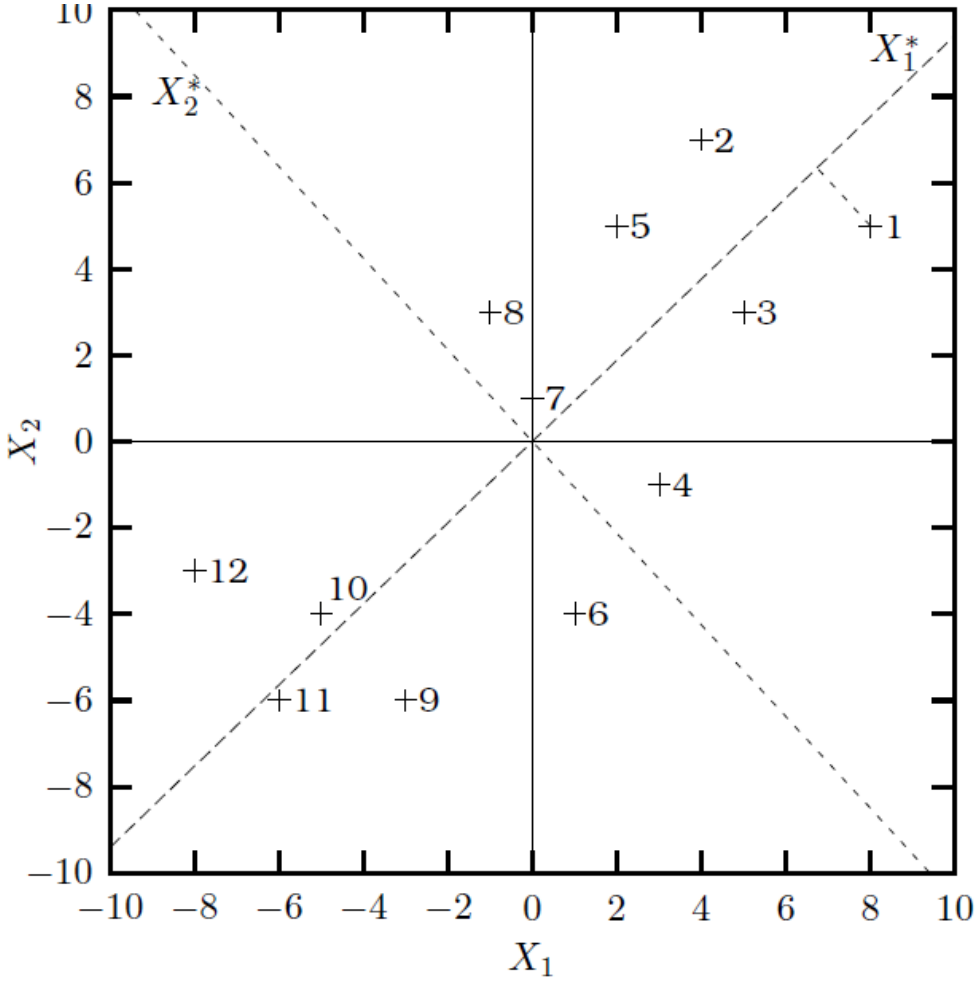
5. Τα ποσοστά της ολικής διακύμανσης που εξηγούν οι x_1^* και x_2^* είναι 87.31% και 12.69% αντίστοιχα. Έτσι η πρώτη νέα μεταβλητή x_1^* εκφράζει το μεγαλύτερο ποσοστό διακύμανσης απ'ότι οι δύο αρχικές μεταβλητές. Η δεύτερη νέα μεταβλητή x_2^* εκφράζει το ποσοστό εκείνο της ολικής διακύμανσης που δεν εκφράζεται από την x_1^*. Ωστόσο οι δύο νέες μεταβλητές μαζί εξηγούν όλη τη διακύμανση στα δεδομένα.

6. Η συσχέτιση των δύο μεταβλητών είναι μηδενική. Με άλλα λόγια οι x_1^* και x_2^* δεν συσχετίζονται.

Όσα ειπώθηκαν για την περίπτωση των δύο μεταβλητών ισχύουν και για περισσότερες. Ένα σύνολο δεδομένων που αποτελείται από p μεταβλητές μπορεί να παρουσιαστεί γραφικά σε έναν p-διάστατο χώρο σε σχέση με

Πίνακας 2.4. Οι μετασχηματισμένες αρχικές μεταβλητές και οι νέες μεταβλητές x_1^* και x_2^* για τους νέους άξονες με κλίση 43.261°

Αριθμός Παρατήρησης	**Μετασχηματισμένη μεταβλητή x_1**	**Μετασχηματισμένη μεταβλητή x_2**	**Νέα μεταβλητή x_1^***	**Νέα μεταβλητή 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**
Διακύμανση	**23.091**	**21.091**	**38.576**	**5.606**



Σχήμα 2.3. Γραφική παράσταση των μετασχηματισμένων δεδομένων και των νέων αξόνων

τους *p* αρχικούς άξονες ή τους *p* νέους άξονες. Ο πρώτος νέος άξονας X_1^* αντιστοιχεί στην πρώτη νέα μεταβλητή x_1^* η οποία εξηγεί το μέγιστο δυνατό της ολικής διακύμανσης. Στη συνέχεια κατασκευάζεται ο δεύτερος νέος άξονας κάθετα στον πρώτο, του οποίου η αντίστοιχη νέα μεταβλητή x_2^* εξηγεί το μέγιστο της διακύμανσης που δεν εξηγεί η πρώτη μεταβλητή και δεν συσχετίζεται με αυτήν. Αυτή η διαδικασία συνεχίζεται έως ότου όλοι οι *p* νέοι άξονες καθοριστούν και οι *p* νέες μεταβλητές να εκφράζουν το μέγιστο της κάθε φορά υπολοίπουσας διακύμανσης, με την προυπόθεση πάντα οι νέες μεταβλητές να μην συσχετίζονται (uncorrelated). Πρέπει επίσης να σημειωθεί ότι ο αριθμός των νέων μεταβλητών (principal components) ισούται με τον αριθμό των αρχικών μεταβλητών.

1.4. Μαθηματική Ερμηνεία (Άλγεβρα Πινάκων)

Μέχρι στιγμής επιχειρήθηκε μία διαισθητική ερμηνεία της Ανάλυσης Κύριων Συνιστωσών. Προκύπτουν όμως πολλά ερωτήματα. Ένα βασικό ερώτημα αφορά το πώς υπολογίζονται οι κύριες συνιστώσες με μαθηματικό τρόπο, αλλά και το πώς ερμηνεύεται η Ανάλυση Κυρίων Συνιστωσών από τα μαθηματικά. Στη συνέχεια θα προσπαθήσουμε να περιγράψουμε τα βήματα που οδηγούν στον μαθηματικό

υπολογισμό των κύριων συνιστωσών. Η ανάλυση θα γίνει με όσο το δυνατό πιο απλοϊκό τρόπο, αποφεύγοντας λεπτομέρειες που μπορεί να δυσκολέψουν αντί να διευκολύνουν την κατανόηση της μεθόδου.
Όπως αναφέρθηκε προηγουμένως, κάθε μία κύρια συνιστώσα (Y) είναι ένας γραμμικός συνδυασμός των αρχικών k μεταβλητών. Επίσης, όλες οι Y_i είναι ασυσχέτιστες μεταξύ τους.
Η Y_1 έχει τη μορφή

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1k}X_k$$

και θέλουμε να εκφράζει μέγιστη δυνατή διακύμανση.
Η Y_2 έχει τη μορφή

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2k}X_k$$

και θέλουμε να είναι ασυσχέτιστη με την Y_1, εκφράζοντας παράλληλα μέγιστη δυνατή διακύμανση, και το δεύτερο μεγαλύτερο ποσοστό της συνολικής διακύμανσης των δεδομένων.
Η i-στή κύρια συνιστώσα Y_i έχει τη μορφή

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ik}X_k \qquad (2.1)$$

και θέλουμε να είναι ασυσχέτιστη με τις $Y_1, Y_2, \ldots, Y_{i-1}$ εκφράζοντας παράλληλα μέγιστη δυνατή διακύμανση, και το i-στό μεγαλύτερο ποσοστό της συνολικής διακύμανσης των δεδομένων.
Με παρόμοια λογική, η k-στή κύρια συνιστώσα Y_k, έχει τη μορφή

$$Y_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kk}X_k$$

και θέλουμε να είναι ασυσχέτιστη με τις $Y_1, Y_2, \ldots, Y_{k-1}$, εκφράζοντας παράλληλα μέγιστη δυνατή μεταβλητότητα, και το ελάχιστο ποσοστό της συνολικής μεταβλητότητας του δείγματος.

Υπό μορφή πινάκων μπορεί να γραφτεί ως $\mathbf{Y} = \mathbf{AX}$ όπου $\mathbf{Y}, \mathbf{X}$ είναι διανύσματα $k \times 1$ και $\mathbf{A}$ είναι $k \times k$ πίνακας με στοιχεία

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \ldots & & \ldots & \\ a_{k1} & a_{k2} & \ldots & a_{kk} \end{bmatrix} = [\alpha_1 \quad \alpha_2 \quad \ldots. \quad \alpha_k]$$

όπου α_j είναι το διάνυσμα στήλη με στοιχεία $\alpha_j' = [\alpha_{j1} \quad \alpha_{j2} \quad \ldots. \quad \alpha_{jk}] \quad j = 1, \ldots, k$, και για να μην έχουμε προβλήματα ταυτοποίησης θέτουμε $\sum_{i=1}^{k} a_{ji}^2 = \alpha_j' \alpha_j = 1$.
Επομένως το πρόβλημα εύρεσης των κυρίων συνιστωσών είναι το πρόβλημα της εύρεσης των στοιχείων του πίνακα Α. Έχουμε όμως έναν επιπλέον περιορισμό, ότι δηλαδή οι κύριες συνιστώσες πρέπει να είναι σε φθίνουσα σειρά ως προς τη διακύμανση τους, δηλαδή η πρώτη να έχει τη μεγαλύτερη διακύμανση, η δεύτερη τη δεύτερη μεγαλύτερη και ούτω
καθεξής. Έχουμε ήδη δει (από την εξ. (1.1)) ότι τα ιδιοδιανύσματα

αποτελούν μια λύση στο πρόβλημα αν εξαιρέσουμε την τελευταία υπόθεση για τη φθίνουσα σειρά της διακύμανσης.
Ας δουλέψουμε για την πρώτη κύρια συνιστώσα $Y_1 = \alpha_1' X$. Είναι σαφές πως $Var(Y_1) = a_1' \Sigma a_1$ όπου Σ ο πίνακας διακυμάνσεων του τυχαίου διανύσματος Χ. Επομένως για να βρούμε το a_1 θα πρέπει να μεγιστοποιήσουμε την $Var(Y_1)$ με τον περιορισμό πως $\alpha_1' \alpha_1 = 1$.
Από τη μαθηματική ανάλυση, και συγκεκριμένα από τη θεωρία των ακροτάτων υπό συνθήκη, γνωρίζουμε ότι αν έχουμε μία συνάρτηση $f(x,y)$ με ακρότατο το σημείο (x_0,y_0) υπό τη δέσμευση $g(x,y)=0$,τότε ισχύει

$$\nabla f(x_0,y_0)=\lambda \nabla g(x_0,y_0) \qquad (2.4)$$

όπου λ είναι ο πολλαπλασιαστής Lagrange. Η σχέση αυτή γενικεύεται και για περισσότερες διαστάσεις.
Η παραπάνω σχέση μπορείνα εφαρμοστεί για τον υπολογισμό του μεγίστου της μεταβλητότητας της Y_1. Αυτό γίνεται αν θεωρήσουμε ότι η εύρεση του μεγίστου της συνάρτησης $Var(Y_1) = a_1' \Sigma a_1$ είναι στην πράξη υπολογισμός ακρότατου υπό τη δέσμευση $\alpha_1' \alpha_1 = 1 \Rightarrow \alpha_1' \alpha_1 - 1 = 0$.
Τότε ο υπολογισμός του μεγίστου της μεταβλητότητας της Y_1 γίνεται ως εξής

$$\nabla (a_1' \Sigma a_1) = \lambda \nabla(\alpha_1' \alpha_1 - 1) \Rightarrow \nabla (a_1' \Sigma a_1) - \lambda \nabla(\alpha_1' \alpha_1 - 1) = 0.$$

Δηλαδή θα μεγιστοποιήσουμε τη συνάρτηση

$$L(a_1) = a_1' \Sigma a_1 - \lambda(\alpha_1' \alpha_1 - 1),$$

όπου λ ο πολλαπλασιαστής Lagrange.
Χρησιμοποιώντας παραγώγους διανυσμάτων βρίσκουμε πως

$$\frac{\partial L(a_1)}{\partial a_1} = 2(\Sigma - \lambda \mathbf{I})a_1 = 0$$

και επομένως αντιστοιχεί στο να λύσουμε την εξίσωση

$$\Sigma\alpha_1 = \lambda\alpha_1 \qquad (2.5)$$

η οποία είναι η εξίσωση των ιδιοδιανυσμάτων του πίνακα Σ όπου λ είναι η ιδιοτιμή. Συνεπώς, παρατηρούμε ότι γνωρίζοντας την κατάλληλη ιδιοτιμή λ του πίνακα συνδιακύμανσης Σ μπορούμε να υπολογίσουμε το ιδιοδιάνυσμα α_1 και γνωρίζοντας το ιδιοδιάνυσμα α_1 έχουμε υπολογίσει την $1^η$ κύρια συνιστώσα

$$Y_1 = \alpha_1' X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1k}X_k.$$

Επομένως, προκύπτει το γενικευμένο συμπέρασμα. Για να υπολογιστούν και οι k Y_i η διαδικασία είναι η εξής. Βρίσκονται οι k ιδιοτιμές λ_i του πίνακα συνδιακύμανσης Σ, μετά βρίσκονται τα αντίστοιχα ιδιοδιανύσματα α_i, $i=1,\ldots,k$ και με τα ιδιοδιανύσματα αυτά υπολογίζονται οι Y_i

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ik}X_k \quad , \quad i=1,\ldots,k.$$

Δηλαδή κάθε ζεύγος ιδιοτιμής και του ιδιοδιανύσματος που τη συνοδεύει είναι λύση της εξίσωσης $\Sigma a_i = \lambda a_i$, και άρα έχουμε k δυνατές λύσεις.
Όμως προκύπτει ένα ερώτημα το οποίο δεν έχει απαντηθεί. Όπως ήδη έχουμε πει, οι Y_i ταξινομούνται σε σειρά φθίνουσα ως προς το το ποσοστό της μεταβλητότητας του δείγματος που καλύπτουν. Όπως είδαμε, κάθε Y_i αντιστοιχεί σε ένα ιδιοδιάνυσμα a_i. Όμως, το κάθε ένα ιδιοδιάνυσμα a_i σε ποια από τις k ιδιοτιμές του πίνακα Σ αντιστοιχεί ; Όταν βρεθούν οι k ιδιοτιμές λ_i του πίνακα **Σ**, δεν έχει γίνει σαφές το ποια ιδιοτιμή θα αντιστοιχιστεί σε κάθε μία Y.
Για να διερευνήσουμε ποια από τις k ιδιοτιμές του πίνακα Σ θα χρησιμοποιηθεί για τον υπολογισμό της Y_1, θα αναζητήσουμε την ιδιοτιμή η οποία δίνει το ιδιοδιάνυσμα a_1 για το οποίο η μεταβλητότητα της $Y_1 = a_1' X$ μεγιστοποιείται. Αυτό, γιατί η μεταβλητότητα του δείγματος που καλύπτει η Y_1 θέλουμε να είναι η μέγιστη δυνατή.
Η μεταβλητότητα της Y_1 γράφεται ως εξής:

$$Var(Y_1) = Var(a_1' X) = a_1' \Sigma a_1 = a_1' \lambda a_1 = \lambda\, a_1' a_1 = \lambda, \qquad (2.6)$$

αφού ισχύουν: $\Sigma a_1 = \lambda a_1$ και $a_1' a_1 = 1$.

Επομένως, η μεταβλητότητα της κύριας συνιστώσας ισούται με την αντίστοιχη ιδιοτιμή της. Αυτό σημαίνει ότι η πρώτη κύρια συνιστώσα Y_1, η μεταβλητότητα της οποίας θέλουμε να είναι η μεγαλύτερη όλων των κυρίων συνιστωσών, θα έχει μεταβλητότητα που θα ισούται με την μέγιστη ιδιοτιμή. Δηλαδή,

$$Var(Y_1) = \lambda_{max}.$$

Ομοίως, η Y_2 θέλουμε να εκφράζει το δεύτερο μεγαλύτερο ποσοστό της μεταβλητότητας του δείγματος, και συνεπώς η μεταβλητότητα της Y_2 θα ισούται με την ιδιοτιμή που έχει τη δεύτερη μεγαλύτερη τιμή.
Με παρόμοια επιχειρήματα μπορούμε να δούμε πως για όλες τις κύριες συνιστώσες τα διανύσματα a_j που χρειαζόμαστε θα αντιστοιχούν στα ιδιοδιανύσματα της j σε φθίνουσα σειρά ιδιοτιμής. Φυσικά για την εύρεση των υπόλοιπων κυρίων συνιστωσών χρειάζεται να προσθέσουμε έναν ακόμη περιορισμό: ότι οι κύριες συνιστώσες είναι ασυσχέτιστες με τις προηγούμενες τους.
Επομένως :

- Για να κατασκευάσουμε τις κύριες συνιστώσες χρειάζεται να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα Σ που χρησιμοποιούμε.

- Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμα της αντιστοιχούν στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή στη δεύτερη κύρια συνιστώσα κλπ.

- Η διακύμανση της κάθε κύριας συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί. Έτσι αν συμβολίσουμε με λ_j την j μεγαλύτερη ιδιοτιμή τότε έχουμε πως $Var(Y_j) = \lambda_j$.

- Όπως είπαμε και πριν οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και άρα ο πίνακας διακύμανσης τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές λ_j .

- Η συνολική διακύμανση των κύριων συνιστωσών θα είναι η ίδια με τη συνολική διακύμανση των αρχικών μεταβλητών εξαιτίας των ιδιοτήτων του ίχνους συμμετρικού και τετραγωνικού πίνακα. Δηλαδή θα ισχύει $tr(\Sigma) = tr(\Lambda)$ και άρα η συνολική διακύμανση διατηρείται.
- Επίσης η γενικευμένη διακύμανση των κυριών συνιστωσών είναι η ίδια με τη γενικευμένη διακύμανση των αρχικών μεταβλητών. Αυτό προκύπτει εύκολα καθώς η ορίζουσα ενός τετραγωνικού πίνακα είναι το γινόμενο των ιδιοτιμών της και άρα ισχύει

$$|\Sigma| = \prod_{i=1}^{p} \lambda_i = |\Lambda| .$$

- Η ποσότητα $\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$ μας δείχνει το ποσοστό της συνολικής διακύμανσης που εξηγεί η j συνιστώσα. Είναι ευνόητο πως αν κάποιος πάρει όλες τις συνιστώσες τότε θα διατηρήσει όλη τη διακύμανση, ενώ αν τελικά παραλείψει κάποιες συνιστώσες κάποιο ποσοστό της διακύμανσης θα χαθεί. Προφανώς συμφέρει να διατηρούμε τις πρώτες συνιστώσες που εξηγούν μεγαλύτερο κομμάτι της διακύμανσης.

Παράδειγμα 2.

Έστω ο πίνακας διακύμανσης $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$. Οι ιδιοτιμές του πίνακα είναι (σε φθίνουσα σειρά) λ_1=5.83, λ_2 =2, λ_3=0.17 καθώς και τα ιδιοδιανύσματα που τους αντιστοιχούν είναι τα

α_1'= (-0.383 0.924 0), α_2'= (0 0 1), α_3'= (0.924 0.383 0).

Χρησιμοποιώντας λοιπόν αυτά βρίσκουμε πως οι κύριες συνιστώσες είναι οι

$Y_1 = -0.383X_1 + 0.924 X_2$

$Y_2 = X_3$

$Y_3 = 0.924X_1 + 0.383 X_2$

Παρατηρούμε τα εξής:

- Προφανώς δεν υπάρχει περίπτωση μια ιδιοτιμή να είναι αρνητική αφού ένας πίνακας διακύμανσης είναι πάντα θετικά ορισμένος και άρα οι ιδιοτιμές του είναι θετικές ή μηδέν.

- Η μεταβλητή X_3 η οποία ήταν ασυσχέτιστη με τις υπόλοιπες είναι η δεύτερη κύρια συνιστώσα. Συνεπώς αν χρησιμοποιήσω ασυσχέτιστες μεταβλητές στην ανάλυση σε κύριες συνιστώσες αυτές θα ταυτιστούν με κάποια συνιστώσα και επομένως δεν κερδίζω τίποτα με το να τις χρησιμοποιήσω στην ανάλυση. Θυμηθείτε πως ένας σκοπός της ανάλυσης σε κύριες συνιστώσες ήταν να οδηγηθώ σε ασυσχέτιστες μεταβλητές. Αν ξεκινήσω από ασυσχέτιστες δεν έχει νόημα η ανάλυση.

- Αν αλλάξω τα πρόσημα στις τιμές των ιδιοδιανυσμάτων αυτά συνεχίζουν να είναι λύσεις με την έννοια ότι ικανοποιούν όλες τις συνθήκες. Επομένως οι κύριες συνιστώσες δεν είναι μοναδικές καθώς μπορώ να αλλάξω τα πρόσημα. Αυτό έχει σαν αποτέλεσμα να μην είναι ξεκάθαρη η ερμηνεία τους. Παρόλα αυτά καθώς οι τιμές των

συντελεστών δεν θα αλλάξουν σε απόλυτη τιμή μπορώ να ‘αναγνωρίσω’ κάπως τις συνιστώσες και την επίδραση των αρχικών μεταβλητών σε αυτές άσχετα με το πρόσημο.

- Οι διακυμάνσεις των τριών κυρίων συνιστωσών είναι

$$Var\,(Y_1) = 5.83$$
$$Var\,(Y_2) = 2$$
$$Var\,(Y_3) = 0.17$$

και άρα η συνολική διακύμανση είναι 8 (όπως και στα αρχικά δεδομένα). Επομένως η 1η κύρια συνιστώσα εξηγεί το 5.83/8 = 72.8% της συνολικής διακύμανσης, ενώ η 2η κύρια συνιστώσα το 25%. Και οι δύο μαζί εξηγούν το 97.8% και άρα αν αποφασίσει κάποιος να μην διατηρήσει την 3η (πιθανότατα για να περιορίσει τον αριθμό των μεταβλητών) του θα χάσει μόλις το 2.2% της πληροφορίας που είχαν τα αρχικά δεδομένα.

1.5. Αλλαγή Κλίμακας

Ένα από τα μειονεκτήματα της ανάλυσης σε κύριες συνιστώσες χρησιμοποιώντας τον πίνακα διακύμανσης είναι πως αν αλλάξει η κλίμακα μέτρησης των δεδομένων μας τότε αλλάζουν και οι κύριες συνιστώσες και η ερμηνεία τους. Για να το δούμε αυτό έστω ο πίνακας διακύμανσης που ακολουθεί και αφορά την ηλικία X_1 σε χρόνια και το βάρος X_2 σε κιλά είναι ο

$$\mathbf{\Sigma} = \begin{bmatrix} 10 & 5 \\ 5 & 3 \end{bmatrix}.$$

Αν αντί για το βάρος σε κιλά χρησιμοποιήσουμε το βάρος σε τόνους (X_2`) τότε ο πίνακας διακύμανσης γίνεται

$$\mathbf{\Sigma^*} = \begin{bmatrix} 10 & 0.005 \\ 0.005 & 0.0000003 \end{bmatrix}.$$

Οι ιδιοτιμές του $\mathbf{\Sigma}$ είναι 12.6033 και 0.3967 ενώ του $\mathbf{\Sigma^*}$ είναι 10.00000250 και 0.00000050. Είναι ξεκάθαρο πως και η όποια ερμηνεία έχει αλλάξει δραματικά. Επομένως η ανάλυση σε κύριες συνιστώσες επηρεάζεται από τις μονάδες μέτρησης των μεταβλητών.

Ένα ακόμα μειονέκτημα είναι πως αν κάποια μεταβλητή έχει πολύ μεγαλύτερη διακύμανση από τις υπόλοιπες, αυτή θα τείνει να ταυτιστεί με την πρώτη κύρια συνιστώσα. Φανταστείτε τον πίνακα

$$\mathbf{\Sigma} = \begin{bmatrix} 50 & -1 & 0.1 \\ -1 & 1 & -0.5 \\ 0.1 & -0.5 & 0.3 \end{bmatrix}.$$

Οι ιδιοτιμές του πίνακα είναι 50.0206, 1.2425 και 0.0368 ενώ οι κύριες συνιστώσες που προκύπτουν είναι οι

$$Y_1 = 0.9997X_1 - 0.0190\,X_2 + 0.0075\,X_3$$
$$Y_2 = 0.0204X_1 - 0.8840\,X_2 + 0.4669X_3$$
$$Y_3 = 0.0022X_1 + 0.4669\,X_2 + 0.8842\,X_3$$

Παρατηρείστε πως η πρώτη κύρια συνιστώσα σχεδόν ταυτίζεται με την πρώτη μεταβλητή η οποία είχε πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες.

Από τα παραπάνω εύκολα προκύπτει πως ένα τρόπος να ξεπεράσουμε τις κακές αυτές ιδιότητες της ανάλυσης σε κύριες συνιστώσες στον πίνακα διακύμανσης είναι να χρησιμοποιήσουμε τον πίνακα συσχετίσεων. Οι συσχετίσεις δεν αλλάζουν όταν αλλάξουν οι μονάδες μέτρησης ή η κλίμακα. Επίσης στην ουσία δίνουν ίδιο βάρος σε όλες τις μεταβλητές καθώς όλα τα στοιχεία της διαγωνίου είναι 1, και άρα τα προβλήματα που δημιουργούσε ο πίνακας διακύμανσης μπορούν να ξεπεραστούν.
Από την άλλη πλευρά πάντως, η γενικευμένη χρήση του πίνακα συσχετίσεων δεν ενδείκνυται καθώς η διαφορά στις διακυμάνσεις ενδέχεται να περιέχει πληροφορία πολύτιμη για το θέμα που εξετάζουμε. Ίσως δηλαδή κάποιες μεταβλητές να πρέπει να θεωρηθούν πως έχουν μεγαλύτερο βάρος εξαιτίας της και επομένως θέτοντας όλες τις μεταβλητές να έχουν το ίδιο βάρος χάνουμε χρήσιμη πληροφορία.
Κατά συνέπεια στην πράξη δεν είναι ξεκάθαρο ποιόν από τους δύο πίνακες πρέπει να χρησιμοποιούμε. Μια καλή στρατηγική είναι να αποφεύγουμε τον πίνακα διακύμανσης όταν υπάρχουν κάποιες μεταβλητές με πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες. Αν οι διακυμάνσεις διαφέρουν μεν αλλά είναι συγκρίσιμες (π.χ. αναφέρονται σε ίδιες μονάδες) τότε καλό είναι να χρησιμοποιούμε αυτή την πληροφορία. Εναλλακτικά θα μπορούσε κανείς να μετασχηματίσει τα δεδομένα του ώστε να κάνει τις διακυμάνσεις συγκρίσιμες.
Η συσχέτιση ανάμεσα στην i κύρια συνιστώσα Y_i και την j αρχική μεταβλητή X_j δίνεται από τον τύπο

$$\rho(Y_i , X_j) = \frac{a_{ij}\sqrt{\lambda_i}}{s_j^2},$$

όπου όπως και πριν a_{ij} είναι ο συντελεστής της μεταβλητής X_j στην κύρια συνιστώσα Y_i και s_j^2 είναι η διακύμανση της μεταβλητής X_j . Η συσχέτιση αυτή αποτελεί ένα μέτρο του κατά πόσο η συνιστώσα που προέκυψε σχετίζεται με τη μεταβλητή.
Μπορεί κανείς εύκολα να δει πως αν a_{ij} =0 τότε δεν υπάρχει συσχέτιση, ενώ αν a_{ij} =±1, τότε η συσχέτιση
γίνεται ±1.
Συνήθως στην πράξη δεν διατηρούμε όλες τις κύριες συνιστώσες αλλά τις πρώτες m από αυτές και αγνοούμε τις υπόλοιπες. Σε μια τέτοια περίπτωση χάνουμε πληροφορία. Μπορούμε να βρούμε το ποσοστό της διακύμανσης της αρχικής μεταβλητής X_i που εξηγούμε με τη χρήση των πρώτων m κυρίων συνιστωσών ως

$$\frac{1}{s_i^2}\sum_{j=1}^{m}\lambda_j a_{ji}^2.$$

Είναι ευνόητο πως το ποσοστό της διακύμανσης που εξηγούμε για κάθε μεταβλητή πρέπει να είναι σχετικά μεγάλο γιατί αλλιώς σημαίνει πως χάνουμε πληροφορία για τη μεταβλητή αυτή.
Ως προς την αναμενόμενη τιμή των κυρίων συνιστωσών παρατηρείστε πως

$$E(Y_i) = E(a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ik}X_k) = a_{i1}\,E(X_1) + a_{i2}\,E(X_2) + \ldots + a_{ik}\,E(X_k) = \mu_i .$$

Γενικά δηλαδή η αναμενόμενη τιμή είναι διαφορετική του 0. Για αυτό πολλές φορές κεντροποιούμε τις αρχικές μεταβλητές να έχουν μέση τιμή 0 απλά αφαιρώντας τη μέση τιμή. Αυτό δεν έχει καμιά επίδραση στη διακύμανση απλά οι προκύπτουσες κύριες συνιστώσες έχουν μέση τιμή 0.

Σε αυτό το σημείο θα πρέπει να παρατηρήσουμε τα εξής:

- Μέχρι τώρα μιλάμε γενικά για μεταβλητές και πουθενά δεν έχουμε μιλήσει για δεδομένα και τυχαία δείγματα.
- Δεν έχουμε κάνει καμιά υπόθεση για τον πληθυσμό και δεν υπάρχει κανένα μοντέλο. Η ανάλυση σε κύριες συνιστώσες είναι ένας μαθηματικός μετασχηματισμός των δεδομένων μας και τίποτα άλλο.
- Δεν υπάρχει επομένως κάποια στατιστική συμπερασματολογία μέχρι τώρα.

Ας υποθέσουμε λοιπόν πως έχουμε πια δεδομένα και συγκεκριμένα n παρατηρήσεις σε *k* μεταβλητές (X_1 , X_2,..., X_k). Από αυτά τα δεδομένα υπολογίζουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων με βάση όσα είπαμε προηγουμένως. Στη συνέχεια για τον επιλεγμένο πίνακα βρίσκουμε τις ιδιοτιμές και τα ιδιοδιανύσματα και επομένως βρίσκουμε τις κύριες συνιστώσες.
Μέχρι τώρα τα δεδομένα εισήλθαν στην ανάλυση μόνο για τον καθορισμό του πίνακα διακύμανσης (συσχέτισης) και πουθενά αλλού. Αφού μιλάμε πια για δειγματικό πίνακα διακύμανσης (συσχέτισης) αυτός εμπεριέχει κάποια μεταβλητότητα λόγω του δείγματος. Το ίδιο συμβαίνει και για τις ιδιοτιμές και τα ιδιοδιανύσματά του. Για να μπορέσει κανείς να προχωρήσει σε στατιστική συμπερασματολογία χρειάζεται να κάνει κάποιες υποθέσεις σχετικά με τον πληθυσμό από όπου προήλθε το δείγμα. Με αυτά τα προβλήματα θα ασχοληθούμε σε λίγο.

Παρατηρούμε πως αν δεν ενδιαφερόμαστε για στατιστική συμπερασματολογία η ανάλυση σε κύριες συνιστώσες είναι απλά ένας μετασχηματισμός των δεδομένων. Πώς όμως θα μετασχηματίσουμε τα δεδομένα; Για κάθε μια παρατήρηση θα δημιουργήσουμε τόσες καινούριες μεταβλητές όσες και οι κύριες συνιστώσες που αποφασίσαμε να διατηρήσουμε. Για κάθε κύρια συνιστώσα θα υπολογίσουμε την τιμή της για την παρατήρηση χρησιμοποιώντας τις αντίστοιχες τιμές των αρχικών μεταβλητών και τους συντελεστές που έχουμε βρει. Θα δούμε αναλυτικά πως γίνεται αυτό στην πράξη.

Εξάγουμε την πρώτη αλγεβρική λύση μας στην ΑΚΣ χρησιμοποιώντας γραμμική άλγεβρα. Αυτή η λύση βασίζεται σε μια σημαντική ιδιότητα της ανάλυσης ιδιοδιανυσμάτων. Το σύνολο δεδομένων είναι X, ένας $m \times n$ πίνακας, όπου m είναι ο αριθμός των τύπων μετρήσεων και n ο αριθμός των δειγμάτων. Ο στόχος συνοψίζεται ως εξής.

Βρείτε ορθοκανονικό πίνακα P όπου Y = PX τέτοια ώστε ο $C_Y \equiv \frac{1}{n-1} YY^T$ να είναι διαγωνοποιημένος. Οι γραμμές του P είναι οι κύριες συνιστώσες του X. Ξεκινάμε ξαναγράφοντας τον C_Y από την άποψη της μεταβλητότητας της επιλογής του P.

$$C_Y = \frac{1}{n-1} YY^T = \frac{1}{n-1} (PX)(PX)^T = \frac{1}{n-1} PXX^TP^T = \frac{1}{n-1} P(XX^T) P^T$$

$$= \frac{1}{n-1} PAP^T$$

Ορίσαμε ένα νέο πίνακα $A \equiv XX^T$, όπου ο Α είναι συμμετρικός. Ο οδικός μας χάρτης είναι να αναγνωρίσουμε ότι ένας συμμετρικός πίνακας (Α) διαγωνοποιείται από έναν ορθογώνιο πίνακα των ιδιοδιανυσμάτων του. Για ένα συμμετρικό πίνακα Α ισχύει

$$A = EDE^T \qquad (2.7)$$

όπου D είναι ένας διαγώνιος πίνακας και Ε ένας πίνακας των ιδιοδιανυσμάτων του Α που «τακτοποιούνται» ως στήλες.
Ο πίνακας Α έχει $r \leq m$ ορθοκανονικά ιδιοδιανύσματα όπου r είναι η τάξη του πίνακα. Η τάξη του Α είναι μικρότερη από m όταν το Α είναι εκφυλισμένος ή όλα τα δεδομένα καταλαμβάνουν έναν υπόχωρο διάστασης $r \leq m$. Διατηρώντας τον περιορισμό της ορθογωνιότητας, μπορούμε να διορθώσουμε αυτή την κατάσταση με την επιλογή (m - r) πρόσθετων ορθοκανονικών διανυσμάτων για να "γεμίσει" ο πίνακας Ε. Αυτά τα πρόσθετα διανύσματα δεν επηρεάζουν την τελική λύση, γιατί οι διακυμάνσεις που συνδέονται με τις κατευθύνσεις είναι μηδέν.
Κάνουμε ένα κόλπο. Επιλέγουμε τον πίνακα Ρ να είναι ένας πίνακας, όπου κάθε γραμμή p_i να είναι ένα ιδιοδιάνυσμα του XX^T. Με την επιλογή αυτή, $P \equiv E^T$. Αντικαθιστώντας στην εξίσωση 2.7, βρίσκουμε $A = P^TDP$. Με αυτή τη σχέση και το ότι $P^{-1} = P^T$ μπορούμε να ολοκληρώσουμε τον υπολογισμό του C_Y ως εξής

$$C_Y = \frac{1}{n-1} PAP^T = \frac{1}{n-1} P(P^TDP)P^T = \frac{1}{n-1} (PP^T)D(PP^T)$$

$$= \frac{1}{n-1} (PP^{-1})D(PP^{-1}) = \frac{1}{n-1} D$$

Είναι προφανές ότι η επιλογή του Ρ διαγωνοποιεί τον C_Y. Αυτός ήταν ο στόχος για την ΑΚΣ.

Μπορούμε να συνοψίσουμε τα αποτελέσματα της ΑΚΣ στους πίνακες Ρ και C_Y:

- Οι κύριες συνιστώσες του Χ είναι τα ιδιοδιανύσματα της XX^T ή οι γραμμές του Ρ.
- Η i-οστή διαγώνια τιμή του C_Y είναι η διασπορά του Χ κατά μήκος των p_i. Στην πράξη, ο υπολογισμός της ΑΚΣ ένος σύνολου δεδομένων Χ συνεπάγεται (1) αφαίρεση από τη μέση τιμή κάθε τύπου μέτρησης και (2) υπολογισμό των ιδιοδιανύσματα του XX^T.

1.6. Μια πιο γενική λύση: SVD

Αντλούμε άλλη μια αλγεβρική λύση για την ΑΚΣ και στη συνέχεια διαπιστώνουμε ότι η ΑΚΣ συνδέεται στενά με την αποσύνθεση μοναδικών τιμών (SVD = Singular Value Decomposition). Στην πραγματικότητα, οι δύο αυτές μέθοδοι είναι τόσο στενά συνδεδεμένες που τα ονόματα συχνά χρησιμοποιούνται αδιακρίτως. Θα δούμε ότι η αποσύνθεση μοναδικών τιμών είναι μια πιο γενική μέθοδος κατανόησης της αλλαγής βάσης. Ξεκινάμε αντλώντας γρήγορα την αποσύνθεση. Στην επόμενη ενότητα ερμηνεύουμε την αποσύνθεση και στην τελευταία συσχετίζουμε τα αποτελέσματα αυτά με την ΑΚΣ.

1.6.1. Αποσύνθεση Μοναδικών Τιμών (Singular Value Decomposition)

Έστω X ένας αυθαίρετος $n \times m$ πίνακας και X^TX ένας τάξης r, τετραγωνικός, συμμετρικός $n \times n$ πίνακας. Καθορίζουμε όλες τις ποσότητες που μας ενδιαφέρουν.

• Το $\{\hat{v}_1, \hat{v}_2, \dots, \hat{v}_r\}$είναι το σύνολο των ορθοκανονικών $m \times 1$ ιδιοδιανυσμάτων που σχετίζονται με τις ιδιοτιμές $\{\lambda_1, \lambda_2, \dots, \lambda_r\}$ για το συμμετρικό πίνακα X^TX.

$$(X^TX)\,\hat{v}_i = \lambda_i \hat{v}_i$$

• Οι $\sigma_i \equiv \sqrt{\lambda_i}$ είναι θετικές πραγματικές και καλούνται singular values.

• Το $\{\hat{u}_1, \hat{u}_2, \dots, \hat{u}_r\}$ είναι το σύνολο των ορθοκανονικών $n \times 1$ διανυσμάτων που ορίζονται από την $\hat{u}_i \equiv \frac{1}{\sigma i} X\hat{v}_i$

Εχουμε, λοιπόν, τον τελικό ορισμό που περιλαμβάνει δύο νέες ιδιότητες:

• $\hat{u}_i * \hat{u}_j = \delta_{ij}$

• $||X\hat{v}_i|| = \sigma_i$

Τώρα έχουμε όλα τα κομμάτια για την κατασκευή της αποσύνθεσης. Η "value" version της αποσύνθεσης μοναδικών τιμών είναι απλά μια επαναδιατύπωση του τρίτου ορισμού:

$$X\hat{v}_i = \sigma_i \hat{u}_i \qquad (3.1)$$

Το αποτέλεσμα αυτό λέει αρκετά. Ο X πολλαπλασιασμένος με ένα ιδιοδιάνυσμα του X^TX είναι ισούται με x φορές ένα άλλο διάνυσμα. Το σύνολο των ιδιοδιανυσμάτων $\{\hat{v}_1, \hat{v}_2, \dots, \hat{v}_r\}$ και το σύνολο των διανυσμάτων $\{\hat{u}_1, \hat{u}_2, \dots, \hat{u}_r\}$ είναι και τα δύο ορθοκανονικά σύνολα ή βάσεις στο r-διάστατο χώρο.

Μπορούμε να συνοψίσουμε αυτό το αποτέλεσμα για όλα τα διανύσματα σε ένα πολλαπλασιασμό πινάκων, ακολουθώντας την προβλεπόμενη κατασκευή στο Σχήμα 3.3. Ξεκινάμε με την κατασκευή ενός νέου διαγώνιου πίνακα Σ.

$$\mathbf{\Sigma} \equiv \begin{bmatrix} \sigma_{\tilde{1}} & & & & & 0 \\ & \ddots & & & & \\ & & \sigma_{\tilde{r}} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ 0 & & & & & 0 \end{bmatrix}$$

Σχήμα 3.1.

όπου $\sigma_{\tilde{1}} \geq \sigma_{\tilde{2}} \geq \ldots \geq \sigma_{\tilde{r}}$ είναι το ταξικά-διατεταγμένο σύνολο των μοναδικών τιμών. Κατά τον ίδιο τρόπο κατασκευάζουμε ορθογώνιους πίνακες V και U.

$$V = [\hat{v}_{\tilde{1}}, \hat{v}_{\tilde{2}}, \ldots, \hat{v}_{\tilde{m}}]$$
$$U = [\hat{u}_{\tilde{1}}, \hat{u}_{\tilde{2}}, \ldots, \hat{u}_{\tilde{n}}]$$

όπου έχουμε επισυνάψει πρόσθετα (*m*-*r*) και (*n*-*r*) ορθοκανονικά διανύσματα για να "γεμίσουμε" τους πίνακες για τους V και U. Το Σχήμα 3.3 παρέχει γραφική αναπαράσταση για το πώς όλα τα κομμάτια ταιριάζουν μεταξύ τους για να σχηματίσουν την έκδοση του πίνακα της αποσύνθεσης μοναδικών τιμών.

$$XV = U\Sigma \qquad (3.2)$$

όπου κάθε στήλη των V και U εκτελεί την εκδοχή της αποσύνθεσης (εξίσωση 3.1). Επειδή ο V είναι ορθογώνιος, μπορούμε να πολλαπλασιάσουμε και τις δύο πλευρές $V^{-1} = V^T$ για να καταλήξουμε στην τελική μορφή της αποσύνθεσης

$$X = U\Sigma V^T \quad (3.3)$$

Παρόλο που προήλθε χωρίς κίνητρο, αυτή η αποσύνθεση είναι αρκετά ισχυρή. Η εξίσωση 3.3 ορίζει ότι κάθε αυθαίρετος πίνακας X μπορεί να μετατραπεί σε έναν ορθογώνιο πίνακα, ένα διαγώνιο πίνακα και έναν άλλο ορθογώνιο πίνακα (ή μια περιστροφή, μια επέκταση και μια δεύτερη περιστροφή). Το να κατανοήσουμε την εξίσωση 3.3 είναι το αντικείμενο της επόμενης ενότητας.

1.6.2. Ερμηνεία της SVD

Η τελική μορφή της SVD (εξίσωση 3.3) είναι ένας σύντομος αλλά δύσκολος στην κατανόηση ισχυρισμός. Ας ερμηνεύσουμε ξανά την εξίσωση 3.1

$$Xa = kb$$

όπου a και b είναι διανύσματα στήλες και k μια βαθμωτή σταθερά. Το σύνολο $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m\}$ είναι ανάλογο με το a και το σύνολο $\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n\}$ είναι ανάλογο με το b. Το μοναδικό όμως είναι ότι τα $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m\}$ και $\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n\}$ είναι ορθοκανονικά σύνολα των διανυσμάτων που καλύπτουν έναν m ή n διαστάσεων χώρο, αντίστοιχα. Ειδικότερα, αόριστα μιλώντας, τα σύνολα αυτά φαίνεται να καλύπτουν όλες τις πιθανές "εισόδους" (a) και "εξόδους" (b). Μπορούμε να επισημοποιήσουμε την άποψη ότι τα $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m\}$ και $\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n\}$ καλύπτουν όλες τις πιθανές "εισόδους" και "εξόδους";
Μπορούμε να χρησιμοποιήσουμε την εξίσωση 3.3 για να κάνουμε αυτή την υπόθεση πιο ακριβή.

$$X = U\Sigma V^T$$
$$U^T X = \Sigma V^T$$
$$U^T X = Z$$

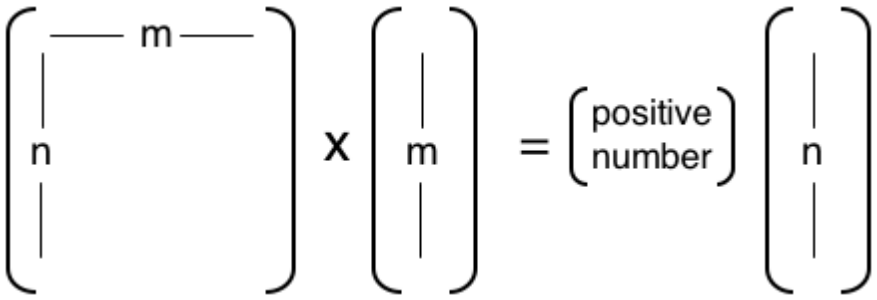
όπου έχουμε ορίσει $Z \equiv \Sigma V^T$. Οι προηγούμενες στήλες $\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n\}$ είναι τώρα γραμμές στον U^T. Συγκρίνοντας την εξίσωση αυτή με την εξίσωση $PX = Y$, το $\{\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n\}$ εκτελεί τον ίδιο ρόλο, όπως το $\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m\}$. Ως εκ τούτου, ο U^T είναι μια αλλαγή βάσης από τον X στον Z. Όπως πριν μετασχηματίζαμε διανύσματα στήλες, μπορούμε να συμπεράνουμε ότι και πάλι θα μετασχηματίσουμε διανύσματα στήλες. Το γεγονός ότι η ορθοκανονική βάση U^T (ή P) μετασχηματίζει τα διανύσματα στήλες σημαίνει ότι ο U^T είναι μια βάση που καλύπτει τις στήλες του X. Βάσεις που καλύπτουν τις στήλες ορίζονται ως ο χώρος των στηλών του X. Ο χώρος των στηλών επισημοποιεί την έννοια του τι είναι οι πιθανές "έξοδοι" κάθε πίνακα.
Υπάρχει μια περίεργη συμμετρία της αποσύνθεσης μοναδικών τιμών τέτοια ώστε να μπορούμε να καθορίσουμε μια παρόμοια ποσότητα - το χώρο των γραμμών.

$$XV = \Sigma U$$
$$(XV)^T = (\Sigma U)^T$$
$$V^T X^T = U^T \Sigma$$
$$V^T X^T = Z$$

όπου έχουμε ορίσει $Z \equiv U^T\Sigma$. Και πάλι οι γραμμές του V^T (ή οι στήλες του V) αποτελούν μια ορθοκανονική βάση για τη μετατροπή του X^T σε Z. Λόγω της μεταφοράς της στον X, προκύπτει ότι η V είναι μια ορθοκανονική βάση που καλύπτει το χώρο των γραμμών του X. Ο χώρος των γραμμών επισημοποιεί την έννοια του τι είναι δυνατές "είσοδοι" σε έναν αυθαίρετο πίνακα. Η “value” version της αποσύνθεσης μοναδικών τιμών εκφράζεται στην εξίσωση 3.1

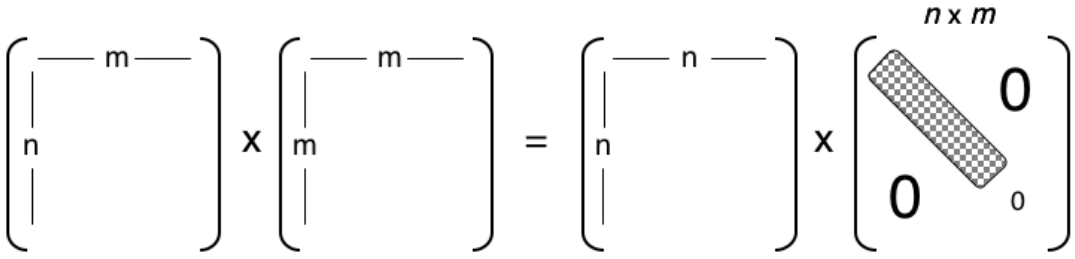
$$X\hat{v}_i = \sigma_i \hat{u}_i$$

Η μαθηματική διαίσθηση πίσω από την κατασκευή της μορφής του πίνακα είναι ότι θέλουμε να εκφράσουμε όλες τις *n* "value" version εξισώσεις με μία μόνο εξίσωση. Είναι πιο εύκολο να καταλάβουμε αυτή τη διαδικασία γραφικά. Οι πίνακες της εξίσωσης 3.1 είναι οι ακόλουθοι.



Σχήμα 3.2

Μπορούμε να κατασκευάσουμε τρεις νέους πίνακες V, U και Σ. Όλες οι μοναδικές τιμές διατάσσονται κατά σειρά $\sigma_{\tilde{1}} \geq \sigma_{\tilde{2}} \geq \ldots \geq \sigma_{\tilde{r}}$ και τα αντίστοιχα διανύσματα κατατάσσονται σε πίνακα με την ίδια σειρά κατάταξης. Κάθε ζεύγος συνδεδεμένων διανυσμάτων $\hat{v}i$ και $\hat{u}i$ στοιβάζεται στην *i*-οστή στήλη κατά μήκος των αντίστοιχων πινάκων τους. Η αντίστοιχη μοναδική τιμή σ_i τοποθετείται κατά μήκος της διαγωνίου (στην *ii*-οστή θέση) του Σ. Αυτό δημιουργεί την εξίσωση XV = UΣ, η οποία μοιάζει με το παρακάτω.



Σχήμα 3.3: Πώς να κατασκευάσουμε τη μορφή του πίνακα της SVD από τη μορφή των "τιμών".

Οι πίνακες V και U είναι $m \times m$ και $n \times n$ πίνακες αντίστοιχα, και ο Σ είναι ένας διαγώνιος πίνακας με μερικές μη μηδενικές τιμές (συμβολίζονται από τη σκακιέρα) κατά μήκος της διαγωνίου του. Η επίλυση αυτής της εξίσωσης πινάκων λύνει όλες τις εξισώσεις σε μορφή *n* "value".

1.7. SVD και PCA

Με παρόμοιους υπολογισμούς είναι προφανές ότι οι δύο μέθοδοι είναι αλληλένδετες. Επιστρέφουμε στον αρχικό $m \times n$ πίνακα δεδομένων X. Μπορούμε να ορίσουμε ένα νέο πίνακα Y ως ένα $n \times m$ πίνακα.

$$Y \equiv \frac{1}{\sqrt{n-1}} X^T$$

όπου κάθε στήλη του Y έχει μέσο μηδέν. Ο ορισμός του Y γίνεται σαφής από την ανάλυση του Y^TY.

$$Y^TY = (\frac{1}{\sqrt{n-1}} X^T)^T (\frac{1}{\sqrt{n-1}} X^T) = \frac{1}{n-1} X^{TT}X^T = \frac{1}{n-1} XX^T$$

Άρα $Y^TY = C_X$.

Από κατασκευής ο Y^TY ισοδυναμεί με τον πίνακα συνδιασποράς του X. Οι κύριες συνιστώσες του X είναι τα ιδιοδιανύσματα του C_X. Αν υπολογίσουμε την αποσύνθεση μοναδικών τιμών του Y, οι στήλες του πίνακα V περιλαμβάνουν τα ιδιοδιανύσματα του $Y^TY = C_X$. Ως εκ τούτου, οι στήλες του V οι κύριες συνιστώσες του X. Ο V καλύπτει το χώρο των γραμμών του $Y \equiv \frac{1}{\sqrt{n-1}} X^T$. Ως εκ τούτου, ο V πρέπει να καλύπτει το χώρο των στηλών του $\frac{1}{\sqrt{n-1}} X$. Μπορούμε να συμπεράνουμε ότι η εύρεση των κύριων συνιστωσών ισοδυναμεί με την εύρεση μιας ορθοκανονικής βάσης που καλύπτει το χώρο των στηλών του X.

1.8. Βήματα της Ανάλυσης Σε Κύριες Συνιστώσες

- ***Έλεγχος συσχετίσεων***

Άσχετα με το αν θα χρησιμοποιήσουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων είναι σκόπιμο να ρίξουμε μια ματιά στον πίνακα συσχετίσεων και να δούμε αν οι αρχικές μας μεταβλητές έχουν συσχετίσεις ή όχι (αυτό γίνεται κυρίως γιατί από τον πίνακα διακύμανσης δεν είναι εύκολο να δούμε την ύπαρξη συσχετίσεων). Αν δεν υπάρχουν συσχετίσεις είναι άσκοπο να συνεχίσουμε. Μεταβλητές που εμφανίζονται
ασυσχέτιστες με τις υπόλοιπες πρέπει να τις διώξουμε από την ανάλυση.
Τι εννοούμε όμως όταν λέμε να υπάρχουν συσχετίσεις; Εννοούμε πως η απόλυτη τιμή της συσχέτισης είναι μεγάλη. Αυτό δεν σημαίνει απαραίτητα πως είναι στατιστικά σημαντική, σύμφωνα με το αποτέλεσμα κάποιου ελέγχου υποθέσεων. Ακόμα και συσχετίσεις της τάξης του 0.10 τείνουν να είναι στατιστικά σημαντικές για μέτριου μεγέθους δείγματα (π.χ. 300 παρατηρήσεις). Για να είναι όμως οι συσχετίσεις ικανοποιητικές για να προχωρήσουμε σε ανάλυση σε κύριες συνιστώσες, θέλουμε να είναι της τάξης του 0.4 ή και μεγαλύτερες σε απόλυτη τιμή. Ένα μέτρο που μας επιτρέπει καλύτερα να συγκρίνουμε δύο σετ δεδομένων αλλά και να αξιολογήσουμε αν οι συσχετίσεις είναι "ενδιαφέρουσες" είναι το

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

όπου r_{ij} είναι το ij στοιχείο του πίνακα συσχετίσεων δηλαδή η συσχέτιση της X_i με τη X_j μεταβλητή. Το στατιστικό φ παίρνει τιμές κοντά στο 1 αν υπάρχουν μεγάλες συσχετίσεις, καθώς όλα τα r_{ij} πλησιάζουν σε απόλυτη τιμή τη μονάδα και άρα το άθροισμα των τετραγώνων τους είναι κοντά στο p^2 και άρα ο αριθμητής τείνει να είναι ίσος με τον παρονομαστή. Αν δεν υπάρχουν συσχετίσεις η τιμή θα είναι κοντά στο 0, καθώς μόνο τα p διαγώνια στοιχεία θα είναι 1, άρα το άθροισμα τετραγώνων θα είναι p και άρα ο αριθμητής θα μηδενιστεί. Στην πράξη τιμές πάνω από 0.4 θεωρούνται ικανοποιητικές.
Το αντίστοιχο μέτρο, στην περίπτωση που δουλεύουμε με τον πίνακα διακύμανσης, είναι το

$$\varphi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} s_{ij}^2 - \sum_{j=1}^{p} s_{jj}^2}{\sum_{i=1}^{p}\sum_{j\neq 1}^{p} s_{ii}s_{jj}}}$$

για το οποίο ισχύουν παρόμοια πράγματα.
Επομένως, ξεκινώντας την ανάλυση, θα ήταν χρήσιμο κανείς όχι απλά να δει αν οι συσχετίσεις είναι στατιστικά σημαντικά διάφορες του 0 αλλά αν είναι επαρκώς μεγάλες σε απόλυτη τιμή για να προχωρήσει.

- ***Επιλογή πίνακα που θα δουλέψουμε***

Όπως είδαμε μπορούμε να χρησιμοποιήσουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων. Μιλήσαμε προηγουμένως για το πως επιλέγουμε και με ποια κριτήρια. Πρέπει να γίνει σαφές ότι τα αποτελέσματα θα διαφέρουν ανάλογα με τον πίνακα που θα επιλέξουμε για αυτό η επιλογή είναι βασική για την αξιοποίηση των αποτελεσμάτων που θα προκύψουν.

- ***Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων***

Ανάλογα με τον πίνακα που διαλέξαμε να στηρίξουμε την ανάλυση υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα. Κρατήστε στο νου σας πως τα ιδιοδιανύσματα που δίνουν τα στατιστικά πακέτα είναι κανονικοποιημένα, δηλαδή το άθροισμα τετραγώνων τους είναι 1 και πως δεν είναι μοναδικά από την άποψη πως μπορούμε να τους αλλάξουμε πρόσημο σε όλα τα στοιχεία τους. Συνεπώς η λύση από στατιστικό πακέτο σε στατιστικό πακέτο μπορεί να διαφέρει ως προς τα πρόσημα.

- ***Απόφαση για τον αριθμό των συνιστωσών που θα κρατήσουμε***

Ίσως το πιο σημαντικό κομμάτι της ανάλυσης το οποίο δυστυχώς δεν έχει εύκολη και κοινώς αποδεκτή απάντηση. Κατ' αρχάς να διευκρινίσουμε πως επιλέγοντας λιγότερες κύριες συνιστώσες από όσες μεταβλητές είχαμε αρχικά, χάνουμε αναγκαστικά πληροφορία. Αυτό είναι το κόστος για το κέρδος μας να μειώσουμε τις διαστάσεις του προβλήματος. Συνήθως λοιπόν ενδιαφερόμαστε για κάποιον μικρότερο αριθμό συνιστωσών. Πόσες όμως; Στη βιβλιογραφία υπάρχουν πολλά κριτήρια τα οποία θα προσπαθήσουμε να περιγράψουμε. Αυτά είναι:

- **Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες.** Σύμφωνα με αυτό το κριτήριο βάζουμε κάποιο όριο (π.χ. 80%) και διαλέγουμε τόσες συνιστώσες ώστε αθροιστικά να εξηγούν μεγαλύτερο ποσοστό από το στόχο

που βάλαμε. Είναι πολύ απλό και εύκολο να το χρησιμοποιήσουμε αλλά δυστυχώς στην πράξη δεν δίνει τα καλύτερα αποτελέσματα, ιδίως αν ο στόχος είναι αρκετά υψηλός. Επίσης δεν είναι ξεκάθαρο ποιο ποσοστό της διακύμανσης πρέπει να βάλουμε σαν στόχο. Επειδή η ολική δικύμανση ισούται με το άθροισμα των χαρακτηριστικών ριζών του πίνακα συσχέτισης, ο υπολογισμός του ποσοστού της διακύμανσης που εκφράζει κάθε κύρια συνιστώσα είναι απλή υπόθεση και το ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες για τις *k* πρώτες κύριες συνιστώσες είναι,

$$P_k = \frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}$$

και αν η ανάλυση κυρίων συνιστωσών εφαρμόστηκε στον πίνακα συσχέτισης η εξίσωση είναι,

$$P_k = \frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{p}$$

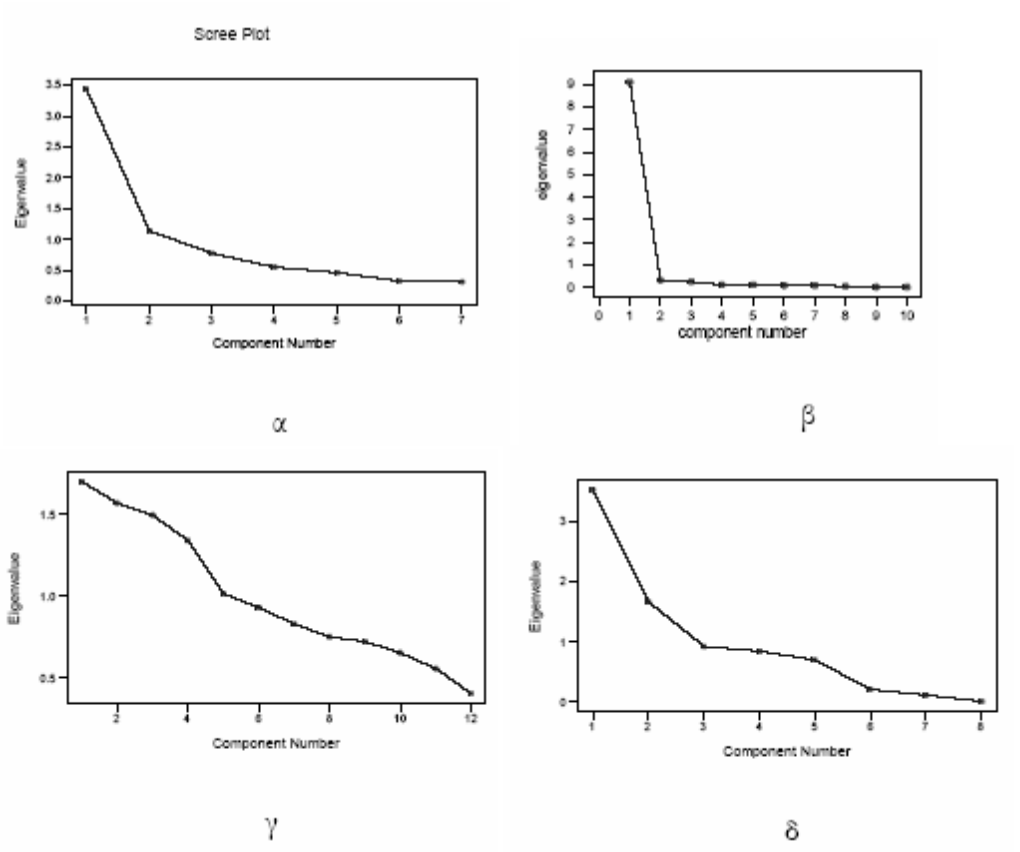
όπου *k* είναι το πλήθος των κυρίων συνιστωσών που επιλέχθηκαν και *p* είναι το σύνολο των χαρακτηριστικών ριζών του πίνακα συσχέτισης.

- **Κριτήριο του Kaiser.** Έστω λ_j οι ιδιοτιμές μας. Το κριτήριο αυτό λέει να πάρουμε τόσες ιδιοτιμές όσες είναι μεγαλύτερες από $\bar{\lambda} = \sum_{j=1}^{k} \lambda_j$, δηλαδή μεγαλύτερες από τη μέση τιμή των ιδιοτιμών. Στην περίπτωση που δουλεύουμε με πίνακα συσχετίσεων ισχύει $\bar{\lambda}$ =1 και άρα διαλέγουμε τόσες συνιστώσες όσες και οι ιδιοτιμές μεγαλύτερες της μονάδας. Το κριτήριο στηρίζεται στην εξής απλή υπόθεση. Αν οι μεταβλητές είναι ασυσχέτιστες και άρα δεν υπάρχει καμιά δομή στα δεδομένα, τότε ο πίνακας συσχετίσεων είναι ο μοναδιαίος και όλες οι ιδιοτιμές είναι ίσες με 1 (δουλεύουμε με πίνακα συσχέτισης). Επομένως κάθε ιδιοτιμή μεγαλύτερη της μονάδας δείχνει την παρουσία κάποιας δομής στα δεδομένα μας. Στην πράξη η υπόθεση αυτή είναι απλοϊκή καθώς ακόμα και αν δεν υπάρχει δομή και όλες οι ιδιοτιμές είναι 1 όταν δουλέψουμε με ένα δείγμα σίγουρα κάποιες από αυτές θα είναι μεγαλύτερες από 1 αφού το άθροισμά τους πρέπει να είναι *p*. Το κριτήριο συνήθως υπερεκτιμά τον αριθμό των συνιστωσών που χρειάζονται.

- **Ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται.** Όπως είδαμε πριν αν διατηρήσουμε *k* συνιστώσες χάνουμε κάποιο μέρος από την πληροφορία κάθε μεταβλητής και μπορούμε να βρούμε και το ποσοστό της διακύμανσης που ερμηνεύσουμε τελικά. Το κριτήριο αυτό διαλέγει τόσες συνιστώσες ώστε να ερμηνεύεται για κάθε μεταβλητή ένα υψηλό ποσοστό τουλάχιστον. Και πάλι το ποιο είναι αυτό το ποσοστό είναι υποκειμενικό. Επίσης μπορεί κάποια μεταβλητή να μην ερμηνεύεται σωστά και αυτό να οδηγήσει σε μεγάλο αριθμό συνιστωσών.

- **Scree plot.** Η γραφική παράσταση των χαρακτηριστικών ριζών (ιδιοτιμών) για κάθε κύρια συνιστώσα χρησιμοποιείται πολύ συχνά και προτάθηκε για πρώτη φορά από τον Cattell. Το scree plot (διάγραμμα διαλογής) είναι ένα γράφημα που έχει στον οριζόντιο άξονα των x τη σειρά των κυρίων

συνιστωσών και στον κάθετο άξονα των y την τιμή της κάθε ιδιοτιμής. Το κριτήριο αυτό προτείνει να πάρουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου επίπεδο, στην ουσία μέχρι να διαπιστώσουμε ότι αρχίζει να αλλάζει η κλίση. Στα scree plot που ακολουθούν Γράφημα 1) μπορεί κανείς να δει τα προβλήματα που παρουσιάζει αυτή η μέθοδος. Στο γράφημα 1β είναι ξεκάθαρο πως θα διαλέξουμε μια μόνο συνιστώσα. Στο γράφημα 1α φαίνεται να διαλέγουμε μια συνιστώσα αλλά κάποιοι θα μπορούσαν να ισχυριστούν ότι πρέπει να πάρουμε 2. Στο γράφημα 1γ τα πράγματα φαίνονται να μην είναι καθόλου καθαρά, ενώ στο 1δ φαίνεται να έχουμε 2 φορές αλλαγή κλίσης. Από αυτά τα γραφήματα γίνεται σαφές πως δεν είναι καθόλου εύκολο να χρησιμοποιήσουμε το scree plot για να επιλέξουμε αριθμό συνιστωσών. Κατ' αρχάς υπάρχει ένα υποκειμενικό κριτήριο για το που και αν αλλάζει η κλίση. Αφετέρου μερικές φορές δεν είναι καθόλου εύκολο να διακρίνει κανείς κάτι τέτοιο για αυτό το scree plot πρέπει να χρησιμοποιείται με προσοχή.



- **Παραλλαγές του Scree plot**. Μερικοί συγγραφείς με σκοπό να αποφύγουν το μειονέκτημα του scree plot ως προς την εύρεση του σημείου αλλαγής της κλίσης πρότειναν διάφορες μεθόδους για να βρει κανείς την αλλαγή κλίσης ξεκινώντας από εμπειρικές παρατηρήσεις φτάνοντας μέχρι τη χρήση γραμμικών μοντέλων. Δεν θα μπούμε σε λεπτομερή περιγραφή τέτοιων μεθόδων.

- **Η μέθοδος του σπασμένου ραβδιού (Broken Stick).** Η μέθοδος αυτή στηρίζεται στην απλή παρατήρηση πως αν πάρουμε ένα ραβδί μεγέθους 1 μονάδας και το σπάσουμε τυχαία σε p κομμάτια τότε το k μεγαλύτερο από αυτά θα έχει αναμενόμενο μήκος $g_k = \frac{1}{p} \sum_{i=k}^{p} \left(\frac{1}{i}\right)$. Επομένως συγκρίνοντας την k ιδιοτιμή με αυτή την ποσότητα μπορεί κανείς να έχει μια εικόνα για τον αν οι ιδιοτιμές προήλθαν από έναν μοναδιαίο πίνακα συσχετίσεων ή όχι. Το κριτήριο λοιπόν επιλέγει τόσες συνιστώσες όσο ισχύει $\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} > g_k$. Δεν μας ενδιαφέρει αν αργότερα ισχύσει ξανά η ανισότητα.

- **Η μέθοδος του Velicer.** Η μέθοδος στηρίζεται στους συντελεστές μερικής συσχέτισης ανάμεσα στις αρχικές μεταβλητές όταν παραλείψουμε κάποιες συνιστώσες. Αν παραλείψουμε κάποια συνιστώσα που είναι χρήσιμη θα πρέπει οι συντελεστές αυτοί να αυξηθούν απότομα και επομένως καταλαβαίνουμε πως η συνιστώσα χρειάζεται. Με τη μέθοδο αυτή αρχίζουμε να 'διώχνουμε' μια τις συνιστώσες μέχρι να βρούμε πως δεν πρέπει να διώξουμε άλλη.

- **Κανονική προσέγγιση.** Όπως θα δούμε στη συνέχεια αν μπορούμε να υποθέσουμε πως ο πληθυσμός μας ακολουθεί πολυμεταβλητή κανονική κατανομή, μπορούμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για ιδιοτιμές βασισμένοι στις δειγματικές ιδιοτιμές. Η ιδέα είναι πως δεν εμπιστευόμαστε το κριτήριο του Kaiser γιατί κάποιες ιδιοτιμές για λόγους τυχαίων κυμάνσεων μπορεί να εμφανιστούν μεγαλύτερες της μονάδας ενώ δεν είναι. Έτσι προσπαθούμε να διαχειριστούμε τη μεταβλητότητα φτιάχνοντας διαστήματα εμπιστοσύνης για τις ιδιοτιμές στηριζόμενοι στα ασυμπτωτικά αποτελέσματα της κανονικής κατανομής. Αν το 95% διάστημα εμπιστοσύνης για την i ιδιοτιμή δεν περιέχει το 1 και είναι μεγαλύτερο από αυτή την τιμή κρατάμε την αντίστοιχη κύρια συνιστώσα (υποθέτουμε πως δουλεύουμε με τον πίνακα συσχετίσεων).

- **Bootstrap.** Η μεθοδολογία bootstrap βρίσκει ολοένα και περισσότερες εφαρμογές στη στατιστική καθώς μας επιτρέπει με επαναληπτική δειγματοληψία με επανάθεση να εκτιμήσουμε ποσότητες του πληθυσμού και κυρίως τα τυπικά σφάλματα των εκτιμητριών τους. Αν η προηγούμενη μέθοδος προσπαθούσε να φτιάξει διαστήματα εμπιστοσύνης για τις ιδιοτιμές στηριζόμενη σε ασυμπτωτικά αποτελέσματα από την κανονική κατανομή, η μέθοδος bootstrap φτιάχνει τα διαστήματα χωρίς να χρειάζεται να κάνει τέτοια υπόθεση. Για να γίνει αυτό δουλεύουμε ως εξής. Παίρνουμε ένα δείγμα ίσου μεγέθους με το πραγματικό από τα δεδομένα μας κάνοντας δειγματοληψία με επανάθεση ανάμεσα στις παρατηρήσεις μας (αυτό σημαίνει πως στο δείγμα που παίρνουμε κάποια παρατήρηση μπορεί να εμφανιστεί παραπάνω από μια φορά). Στη συνέχεια για αυτό το δείγμα φτιάχνουμε τον πίνακα διακύμανσης (συσχέτισης) και βρίσκουμε τις ιδιοτιμές. Αν επαναλάβουμε τη διαδικασία πολλές φορές έχουμε σχηματίσει μια σειρά από τιμές της κατανομής των ιδιοτιμών και άρα μπορούμε να εκτιμήσουμε από αυτές τις τιμές την τυπική απόκλιση των ιδιοτιμών. Έτσι κατασκευάζουμε διαστήματα εμπιστοσύνης και ελέγχουμε αν η ιδιοτιμή που πήραμε από τα δεδομένα μας ανήκει εκεί μέσα. Είναι κατανοητό πως η μέθοδος απαιτεί μεγάλο υπολογιστικό φόρτο.

- **Cross Validation.** Η μέθοδος αυτή στηρίζεται σε επαναληπτικούς υπολογισμούς, όπου κάθε φορά αγνοούμε κάποιες τιμές των δεδομένων μας και εξετάζουμε τη συμπεριφορά των συνιστωσών προσπαθώντας να προβλέψουμε τα δεδομένα που δεν χρησιμοποιήσαμε στην ανάλυση. Επαναλαμβάνοντας τη διαδικασία αυτή πολλές φορές, έχουμε ένα σκορ που μας δείχνει αν το μοντέλο με *k* συνιστώσες δίνει καλά αποτελέσματα. Έτσι συγκρίνοντας τα αποτελέσματα για διάφορες τιμές του *k* βρίσκουμε την τιμή για την οποία τα αποτελέσματα είναι τα καλύτερα.

- **Έλεγχοι υποθέσεων.** Αν μπορούμε να υποθέσουμε κανονικότητα του πληθυσμού ο Bartlett περιέγραψε έναν έλεγχο υπόθεσης για να ελέγξουμε αν οι τελευταίες *p-k* ιδιοτιμές είναι ίσες (και επομένως δεν πρέπει να τις χρησιμοποιήσουμε). Ο έλεγχος ελέγχει τη μηδενική υπόθεση H_0 : οι τελευταίες *p-k* ιδιοτιμές είναι ίσες έναντι της H_1 : δεν είναι ίσες. Για αυτό το σκοπό χρησιμοποιεί την

$$\chi = -(n-1)\sum_{j=k+1}^{p}\ln(\lambda_j) + (n-1)(p-k)\ln\left(\frac{\sum_{j=k+1}^{p}\lambda_j}{p-k}\right),$$

η οποία ακολουθεί την κατανομή χ^2 με $\frac{1}{2}(p-k-1)(p-k+2)$ βαθμούς ελευθερίας. Παρατηρούμε ότι καθώς οι έλεγχοι θα πρέπει να γίνουν ακολουθιακά αυτό έχει σαν αποτέλεσμα το επίπεδο σημαντικότητας να διαφέρει από το *α* που χρησιμοποιούμε για κάθε έλεγχο ξεχωριστά.

- ***Εύρεση των συνιστωσών.*** Αυτό αποτελεί το πιο εύκολο ίσως κομμάτι, ιδιαίτερα στις μέρες μας που όλη τη δουλεία την κάνει ο υπολογιστής. Αρκεί να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα που επιλέξαμε για την ανάλυση, σύμφωνα με τη φασματική ανάλυση που είδαμε προηγουμένως.

- ***Ερμηνεία των συνιστωσών.*** Αυτό το κομμάτι ίσως είναι από τα πιο δύσκολα της ανάλυσης και έχει κατηγορηθεί από πολλούς συγγραφείς. Αφού λοιπόν έχουμε κατασκευάσει τις συνιστώσες πρέπει να προσπαθήσουμε να τους δώσουμε κάποια ερμηνεία, ιδιαίτερα στις πρώτες. Αυτό εξυπηρετεί τους σκοπούς της ανάλυσης καθώς ερμηνεύει τις συσχετίσεις ανάμεσα στις μεταβλητές μας αλλά και αν όλα πάνε καλά μπορούμε να ποσοτικοποιήσουμε κάποιες μη ποσοτικές μεταβλητές. Το τελευταίο είναι ιδιαίτερα χρήσιμο σε διάφορες επιστήμες όπως την ψυχολογία και το marketing. Στα πλαίσια της ερμηνευτικότητας των συνιστωσών μπορεί κανείς να καταφύγει στην περιστροφή των αξόνων, τεχνική πιο γνωστή από την παραγοντική ανάλυση που θα συζητήσουμε αργότερα. Η περιστροφή δεν είναι παρά ο πολλαπλασιασμός του πίνακα των συντελεστών που βρήκαμε με έναν ορθογώνιο πίνακα. Από τους άπειρους ορθογώνιους πίνακες μπορούμε να διαλέξουμε κάποιον με βάση κριτήρια βελτιστοποίησης, όπως για παράδειγμα κάθε συνιστώσα να έχει όσο γίνεται λιγότερες μεταβλητές με μεγάλους συντελεστές. Η περιστροφή συνήθως καταλήγει σε κάθε συνιστώσα, οι μεταβλητές να χωρίζονται πιο έντονα σε σχέση με το πρόσημο τους, δηλαδή να υπάρχουν λίγες με μεγάλες απόλυτες τιμές ενώ οι υπόλοιπες να τείνουν να έχουν συντελεστή κοντά στο μηδέν. Αυτό βοηθά να αναγνωρίζουμε πιο εύκολα τη συνιστώσα, δηλαδή στην ευκολότερη ερμηνεία της.

- ***Δημιουργία νέων μεταβλητών.*** Όπως είπαμε οι κύριες συνιστώσες είναι καινούριες μεταβλητές με κάποιες καλές ιδιότητες. Το ενδιαφέρον είναι πως μπορούμε για κάθε παρατήρηση να δημιουργήσουμε τόσες νέες μεταβλητές όσες και οι κύριες συνιστώσες που αποφασίσαμε να διατηρήσουμε, με σκοπό να χρησιμοποιήσουμε τις κύριες συνιστώσες για περαιτέρω στατιστική ανάλυση. Για να γίνει αυτό αρκεί να αντικαταστήσουμε στον τύπο της κάθε συνιστώσας τις τιμές που η παρατήρηση είχε για κάθε μεταβλητή.

1.9. Αποτελέσματα για Ανάλυση σε Κύριες Συνιστώσες από Δείγμα

Όπως είπαμε και προηγουμένως συνήθως στην ανάλυση σε συνιστώσες περιοριζόμαστε σε απλό μαθηματικό μετασχηματισμό των δεδομένων. Αν όμως μπορούμε να υποθέσουμε πως ο πληθυσμός ακολουθεί την πολυμεταβλητή κανονική κατανομή τότε προκύπτουν μερικά αποτελέσματα σχετικά με τις ποσότητες του δείγματος που μας ενδιαφέρουν. Κατ' αρχάς ξέρουμε πως ο δειγματικός πίνακας διακύμανσης δεν είναι αμερόληπτος εκτός αν έχουμε διαιρέσει κάθε στοιχείο με *n-1* και όχι με *n*. Αν **S** είναι ο μεροληπτικός πίνακας διακύμανσης τότε και οι ιδιοτιμές και τα ιδιοδιανύσματα είναι μεροληπτικά, ενώ αν έχουμε χρησιμοποιήσει τον αμερόληπτο πίνακα, έστω **S*** , τότε είναι αμερόληπτες εκτιμήτριες των αντίστοιχων ποσοτήτων του πληθυσμού. Για αυτή την ενότητα ας συμβολίσουμε με l_i τις δειγματικές ιδιοτιμές και με λ_i τις αντίστοιχες ιδιοτιμές του πληθυσμού. Κάτω από την υπόθεση της πολυμεταβλητής κανονικότητας θα ισχύει πως

$$Var(l_i) \approx \frac{2\lambda_i^2}{n}, \qquad Var(\ln l_i) = \frac{2}{n}$$

$$\text{και } \sqrt{n}(l_i - \lambda_i) \sim N(0, 2\lambda_i^2),$$

Χρησιμοποιώντας αυτά τα αποτελέσματα μπορεί κάποιος να φτιάξει προσεγγιστικά διαστήματα εμπιστοσύνης για τις ιδιοτιμές του πληθυσμού. Ένα 95% προσεγγιστικό διάστημα εμπιστοσύνης είναι το

$$\frac{l_i}{1 + 1.96\sqrt{2/n}} < \lambda_i < \frac{l_i}{1 - 1.96\sqrt{2/n}}$$

το οποίο όμως συνήθως είναι μεγαλύτερο από 95%. Μιλήσαμε προηγουμένως για τη χρήση αυτού του διαστήματος ως κριτηρίου για την επιλογή του αριθμού των συνιστωσών.

Γενικά η συνάρτηση πυκνότητας πιθανότητας των ιδιοτιμών είναι ιδιαίτερα περίπλοκη και επομένως δύσχρηστη. Επίσης μπορεί αν δειχθεί πως η συνδιακύμανση των δειγματικών ιδιοτιμών τείνει στο 0 όταν αυξάνει το μέγεθος του δείγματος, κάτι που υπονοεί πως για μεγάλα δείγματα οι ιδιοτιμές είναι ασυσχέτιστες.

Μερικές φορές είναι χρήσιμο να κάνουμε ελέγχους υποθέσεων για τα ιδιοδιανύσματα. Αυτό είναι χρήσιμο για να δούμε αν τα βάρη που δίνουμε σε κάθε μεταβλητή και επομένως η ερμηνεία στις συνιστώσες έχουν νόημα. Για παράδειγμα αν ο πίνακας συσχετίσεων περιέχει όλο θετικές συσχετίσεις η πρώτη κύρια συνιστώσα είναι ένα σταθμικός μέσος όρος. Έχουν οι μεταβλητές διαφορετικές σταθμίσεις ή είναι ένας απλός μέσος όρος; Γενικά για να ελέγξουμε αν ένα

ιδιοδιάνυσμα είναι ίσο με ένα συγκεκριμένο ιδιοδιάνυσμα c, δηλαδή για να ελέγξουμε τη μηδενική υπόθεση

$$H_0 : a_i = c \quad \text{έναντι της} \quad H_1 : a_i \neq c$$

χρησιμοποιούμε την ελεγχοσυνάρτηση

$$A=(n-1)\left\{l_i c' S^{-1} c + \frac{1}{l_i} c' S c - 2\right\} \sim \chi^2(p-1),$$

όπου **S** ο δειγματικός πίνακας διακύμανσης (συσχετίσεων) και l_i οι δειγματικές ιδιοτιμές. Η ελεγχοσυνάρτηση ακολουθεί κατανομή χ^2 με $p-1$ βαθμούς ελευθερίας. Αν τα δεδομένα προέρχονται από πολυμεταβλητή κανονική κατανομή, οι κύριες συνιστώσες έχουν μια ενδιαφέρουσα γεωμετρική ερμηνεία. Συγκεκριμένα ο μετασχηματισμός των δεδομένων σε κύριες συνιστώσες αντιστοιχεί σε μετακίνηση των αξόνων κατά τη διεύθυνση της μεγαλύτερης διακύμανσης. Έτσι οι κύριες συνιστώσες αντιστοιχούν σε αυτούς τους καινούριους άξονες.

1.10. Μερικά Χρήσιμα Αποτελέσματα

Θα παρουσιάσουμε εν συντομία μερικά ενδιαφέροντα αποτελέσματα σχετικά με την ανάλυση σε κύριες συνιστώσες και κάποιες ειδικές της περιπτώσεις.

- Αν μια μεταβλητή είναι ασυσχέτιστη με τις υπόλοιπες καλό είναι να τηναφαιρέσουμε από την ανάλυση, αφού αν παραμείνει κάποια από τις κύριες συνιστώσες θα ταυτιστεί μαζί της. Όταν δουλεύουμε με δεδομένα αυτό σημαίνει πως δεν έχει στατιστικά σημαντικές συσχετίσεις με τις υπόλοιπες και συνεπώς δεν έχει νόημα να την συμπεριλάβουμε στην ανάλυση.

- Αν δύο ιδιοτιμές προκύψουν ίδιες τότε αυτές αντιστοιχούν σε δύο όμοιες κύριες συνιστώσες κάτι που οδηγεί σε πλεονασμό. Φυσικά στην πράξη κάτι τέτοιο είναι σπάνιο. Αν λοιπόν συμβεί πρέπει να δούμε τα δεδομένα μας μήπως υπάρχει κάποιο πρόβλημα (π.χ. στήλες που επαναλαμβάνονται). Πρέπει να τονιστεί πως για δεδομένα από δείγμα έχει αποδειχτεί πως όλες οι ιδιοτιμές είναι διαφορετικές εκτός από συγκεκριμένες προβληματικές περιπτώσεις.

- Αν έχουμε μηδενικές ιδιοτιμές αυτό σημαίνει πως ο πίνακας που στηρίξαμε την ανάλυση δεν είναι πλήρους βαθμού και άρα κάποιες μεταβλητές είναι γραμμικά εξαρτημένες και πρέπει να τις διώξουμε. Στην πράξη δεν θα συναντήσουμε μηδενικές ιδιοτιμές αλλά πολύ μικρές, κοντά στο μηδέν, ιδιοτιμές. Αυτό υπονοεί ότι κάποιες μεταβλητές είναι σχεδόν γραμμικά εξαρτημένες. Αν αναλογιστεί κανείς πως τέτοιες
  ιδιοτιμές αντιστοιχούν σε συνιστώσες με σχεδόν μηδενική διακύμανση μπορούμε να τις αγνοήσουμε. Δηλαδή στην πράξη αφού δύο μεταβλητές θα παρέχουν την ίδια πληροφορία, όλη η πληροφορία θα πάει σε κάποια από τις πρώτες κύριες συνιστώσες και ότι μείνει θα πάει σε μια συνιστώσα με αμελητέα διακύμανση.

- Σε δύο διαφορετικά σετ δεδομένων μπορεί να πάρουμε τα ίδια ιδιοδιανύσματα ενώ οι ιδιοτιμές να αλλάξουν. Στην πράξη αυτό σημαίνει πως παίρνουμε τις ίδιες συνιστώσες αλλά σε κάθε περίπτωση η συνιστώσα εξηγεί άλλο ποσοστό της διακύμανσης. Συνεπώς δεν πρέπει να περιοριζόμαστε στα ιδιοδιανύσματα αλλά να κοιτάμε και τις ιδιοτιμές.

- Στη γενική περίπτωση που ο πίνακας συσχετίσεων έχει μόνο θετικά στοιχεία (όλες οι συσχετίσεις είναι θετικές) τότε η πρώτη κύρια συνιστώσα μπορεί να εκληφθεί σαν ένας σταθμικός μέσος όρος των μεταβλητών με σταθμίσεις τους αντίστοιχους συντελεστές. Επομένως σε τέτοιες περιπτώσεις μπορούμε να κατασκευάσουμε χρήσιμους δείκτες όπου οι σταθμίσεις έχουν επιλεγεί με έναν συγκεκριμένο τρόπο και όχι εμπειρικά.

- Η βασική ιδέα στην ανάλυση σε κύριες συνιστώσες είναι να γράψουμε τις συνιστώσες ως γραμμικό συνδυασμό των αρχικών μεταβλητών. Είναι εύκολο να δει κανείς πως ομοίως λύνοντας ως προς τις αρχικές μεταβλητές παίρνουμε $\mathbf{X}=\mathbf{A}'\mathbf{Y}$ επειδή ο πίνακας $\mathbf{A}$ είναι ορθογώνιος δηλαδή $\mathbf{A}'=\mathbf{A}^{-1}$. Επομένως αν έχουμε τα σκορ των συνιστωσών μπορούμε εύκολα να βρούμε τα αρχικά δεδομένα.

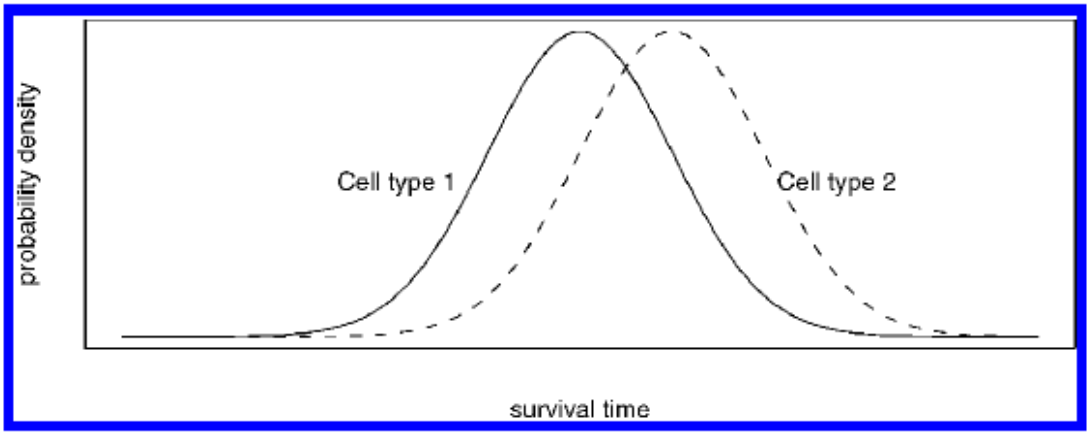
2. Πρόβλεψη από Κύριες Συνιστώσες με Επίβλεψη

2.1. Εισαγωγή στη Μέθοδο των Κυρίων Συνιστωσών με Επίβλεψη

Σε προβλήματα παλινδρόμησης, όπου ο αριθμός των μεταβλητών πρόβλεψης υπερβαίνει κατά πολύ τον αριθμό των παρατηρήσεων, οι συμβατικές τεχνικές παλινδρόμησης μπορεί να παράγουν μη ικανοποιητικά αποτελέσματα. Στην παρούσα εργασία περιγράφουμε μια τεχνική που ονομάζεται ανάλυση κύριων συνιστώσων με επίβλεψη, η οποία μπορεί να εφαρμοστεί σε τέτοιου είδους προβλήματα και αναπτύχθηκε από τους Bair et al. (2006). Η τεχνική της ανάλυσης κύριων συνιστωσών με επίβλεψη είναι παρόμοια με τη συμβατική ανάλυση κυρίων συνιστωσών με τη διαφορά ότι χρησιμοποιεί ένα υποσύνολο των μεταβλητών πρόβλεψης που έχουν επιλεγεί με βάση τη συσχέτιση τους με το εξαγόμενο αποτέλεσμα. Η τεχνική της ανάλυσης κύριων συνιστωσών με επίβλεψη μπορεί να εφαρμοστεί σε προβλήματα παλινδρόμησης και γενικευμένης παλινδρόμησης, όπως είναι η ανάλυση επιβίωσης. Συγκρίνεται ευνοϊκά με άλλες τεχνικές για τέτοιου είδους προβλήματα παλινδρόμησης, μπορεί επίσης να δώσει στοιχεία για τις επιδράσεις άλλων συμμεταβλητών, καθώς και να βοηθήσει στον εντοπισμό των πιο σημαντικών μεταβλητών πρόβλεψης. Παρέχονται, επίσης, αποτελέσματα ασυμπτωτικής θεωρίας τα οποία συμβάλλουν στη στήριξη των εμπειρικών ευρημάτων. Αυτές οι μέθοδοι θα μπορούσαν να αποτελέσουν σημαντικά εργαλεία για δεδομένα DNA μικροσυστοιχιών, καθώς επίσης μπορούν να χρησιμοποιηθούν για πιο ακριβή διάγνωση και θεραπεία του καρκίνου.

Η τεχνική της ανάλυσης κύριων συνιστωσών με επίβλεψη των Bair et al. (2006) αποτελεί μια μέθοδο πρόβλεψης μιας μεταβλητής απόκρισης Y από ένα σύνολο μεταβλητών πρόβλεψης $X_1, X_2, \ldots, X_p$, μετρούμενες σε καθένα από τα N άτομα-παρατηρήσεις. Σύμφωνα με το τυπικό σενάριο, ο αριθμός των μετρήσεων p είναι πολύ μεγαλύτερος από το N. Συγκεκριμένα στην μελέτη των Bair et al. (2006), οι μεταβλητές $X_1, X_2, \ldots, X_p$ είναι μετρήσεις της γονιδιακής έκφρασης από μικροσυστοιχίες DNA. Η μεταβλητή απόκρισης Y μπορεί να είναι μια ποσοτική μεταβλητή για την οποία θα μπορούσε να υποτεθεί ότι είναι κανονικά κατανεμημένη. Συνήθως σε μελέτες μικροσυστοιχιών, η μεταβλητή απόκρισης Y είναι ο χρόνος επιβίωσης, που υπόκειται σε λογοκρισία.

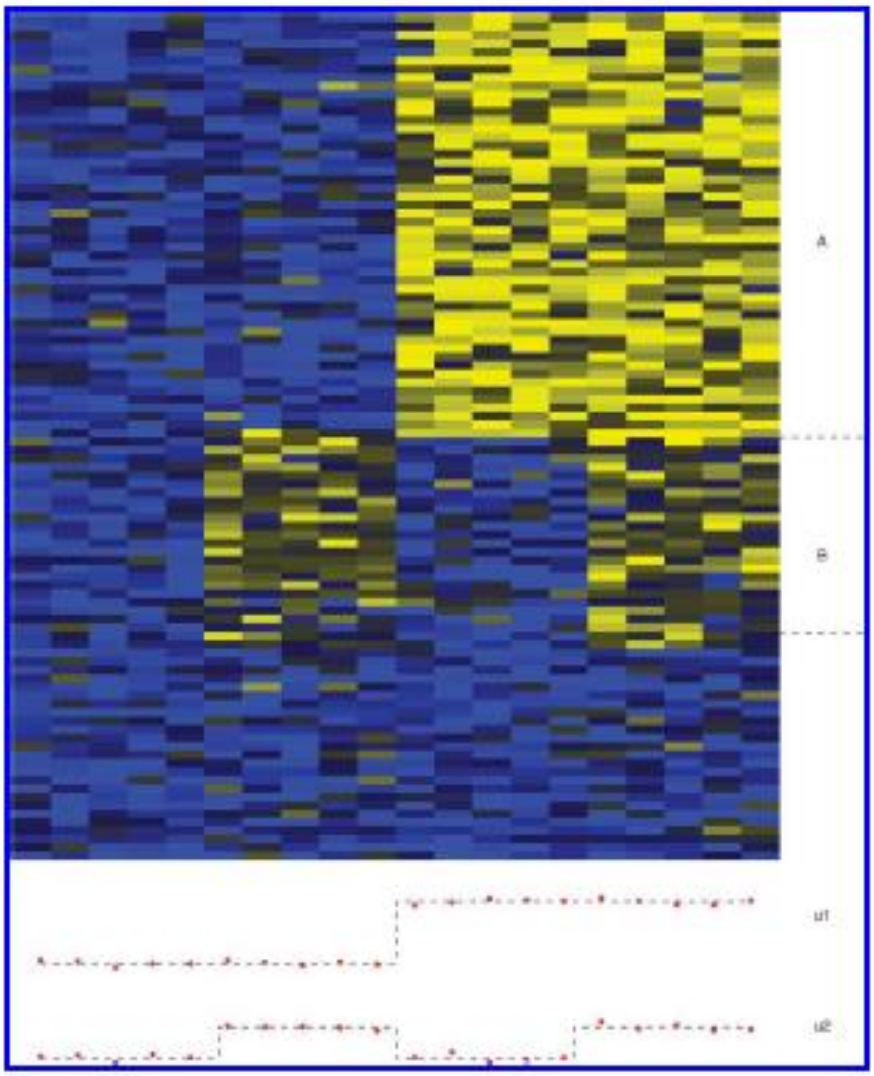
Μία προσέγγιση για την επίλυση ενός τέτοιου είδους προβλήματος θα μπορούσε να είναι μια μέθοδος πρόβλεψης με επίβλεψη. Για παράδειγμα, θα μπορούσε να χρησιμοποιηθεί μια μορφή παλινδρόμησης εφαρμόσιμη όταν $p > N$. Τα μερικά ελάχιστα τετράγωνα θα ήταν μια λογική επιλογή, καθώς και η ridge παλινδρόμηση. Ωστόσο, το Σχήμα 1 δείχνει γιατί μια ημιεπιβλέπουσα προσέγγιση μπορεί να είναι πιο αποτελεσματική.



Σχήμα 1. Βασικό Εννοιολογικό Μοντέλο. Υπάρχουν δύο τύποι κυττάρων: οι ασθενείς με τον καλό τύπο κυττάρων ζουν περισσότερο κατά μέσο όρο. Ωστόσο, υπάρχει σημαντική επικάλυψη στα δύο σύνολα των χρόνων επιβίωσης. Ως εκ τούτου, θα ήταν πολύ χρήσιμο να προσπαθήσουμε να αποκαλύψουμε τους κυτταρικούς τύπους και να τους χρησιμοποιήσουμε για να προβλέψουμε το χρόνο επιβίωσης, αντί να τον προβλέψουμε απ'ευθείας.

Έστω ότι υπάρχουν δύο τύποι κυττάρων, και ότι οι ασθενείς με τον καλό τύπο (2) ζουν περισσότερο κατά μέσο όρο. Ωστόσο, υπάρχει σημαντική επικάλυψη στα δύο σύνολα των χρόνων επιβίωσης. Ο χρόνος επιβίωσης μπορεί να θεωρηθεί ως ένας "θορυβώδης" παράγοντας διαχωρισμού του κυτταρικού τύπου. Μια πλήρως επιβλεπόμενη προσέγγιση θα έδινε το μεγαλύτερο βάρος στα γονίδια που έχουν την ισχυρότερη σχέση με την επιβίωση. Τα γονίδια αυτά είναι εν μέρει, αλλά όχι τέλεια, συσχετισμένα με τον κυτταρικό τύπο. Αν μπορούσαν να ανακαλυφθούν οι βασικότεροι κυτταρικοί τύποι των ασθενών (συχνά εκφράζονται από μεγάλο πλήθος γονιδίων που ενεργούν μαζί σε μονοπάτια), τότε θα προβλεπόταν καλύτερα η επιβίωση των ασθενών.
Στις μέρες μας μπορούμε να εξάγουμε πληροφορίες για τους σημαντικούς τύπους κυττάρων τόσο από τη σχέση μεταξύ του Y και των $X_1, X_2, \ldots, X_p$ όσο και από τη συσχέτιση μεταξύ των ίδιων των μεταβλητών πρόβλεψης. Η ανάλυση κυρίων συνιστωσών (PCA) είναι μια τυποποιημένη μέθοδος για τη μοντελοποίηση της συσχέτισης. Η παλινδρόμηση στις λίγες πρώτες κύριες συνιστώσες θα φαινόταν σαν μια φυσική προσέγγιση, αλλά αυτό δεν μπορεί να λειτουργεί πάντα καλά. Τα πλασματικά στοιχεία που δίνονται στο σχήμα 2 απεικονίζουν το πρόβλημα (αν επρόκειτο να χρησιμοποιήσουμε μόνο τη μεγαλύτερη κύρια συνιστώσα). Είναι μια heatmap απεικόνιση με κάθε γονίδιο να αντιπροσωπεύεται από μια σειρά-γραμμή και κάθε στήλη να περιέχει δεδομένα από έναν ασθενή σε μια μικροσυστοιχία. Η γονιδιακή έκφραση κωδικοποιείται από το μπλε (χαμηλή) στο κίτρινο (υψηλή). Σε αυτό το παράδειγμα, η μεγαλύτερη μεταβολή/ απόκλιση παρατηρείται στα γονίδια που σημειώνονται ως Α, με το δεύτερο σετ των 10 ασθενών να έχουν υψηλότερη έκφραση αυτών των γονιδίων από τα πρώτα 10. Το σύνολο των γονιδίων που

σημειώνονται ως Β παρουσιάζουν διαφορετική μεταβολή, με το δεύτερο και το τέταρτο μπλοκ των ασθενών να έχουν υψηλότερη έκφραση σε αυτά τα γονίδια. Το υπόλοιπο των γονιδίων δεν δείχνουν συστηματική μεταβολή. Στο κάτω μέρος της οθόνης, τα κόκκινα σημεία είναι τα πρώτα δύο μοναδικά διανύσματα u_1 και u_2 (κύριες συνιστώσες) των τιμών έκφρασης. Στις μελέτες μικροσυστοιχιών ονομάζονται μερικές φορές "ιδιογονίδια" (Alter, Brown και Botstein 2000). (Οι διακεκομμένες γραμμές αντιπροσωπεύουν τον "αληθινό" μηχανισμό ομαδοποίησης που δημιούργησε τα δεδομένα στις δύο ομάδες). Τώρα αν τα γονίδια στο Α συνδέονται στενά με το αποτέλεσμα *Y*, τότε το *Y* θα σχετίζεται σε μεγάλο βαθμό με την πρώτη κύρια συνιστώσα. Σε αυτήν την περίπτωση θα περίμενε κανείς ένα μοντέλο που χρησιμοποιεί το u_1 για να προβλέψει το *Y* να είναι πολύ αποτελεσματικό. Ωστόσο, η απόκλιση στα γονίδια Α θα μπορούσε να αντανακλά κάποια βιολογική διαδικασία που δεν σχετίζεται με το αποτέλεσμα *Y*. Στην περίπτωση αυτή, το Υ μπορεί να σχετίζεται σε μεγάλο βαθμό με το u_2 ή κάποιας υψηλότερης τάξης κύρια συνιστώσα.



Σχήμα 2. Πλασματικά Στοιχεία Μικροσυστοιχειών για Εικονογράφηση. Μια heatmap απεικόνιση με κάθε γονίδιο να αντιπροσωπεύεται από μια σειρά-γραμμή, και κάθε στήλη να δίνει τα στοιχεία από έναν ασθενή σε μια μικροσυστοιχία. Η γονιδιακή έκφραση κωδικοποιείται από μπλε (χαμηλή) μέχρι κίτρινο (υψηλή). Η μεγαλύτερη μεταβλητότητα παρατηρείται στα γονίδια που σημειώνονται με Α, με το δεύτερο σύνολο των 10 ασθενών να έχουν υψηλότερη έκφραση αυτών των γονιδίων. Το σύνολο των γονιδίων που σημειώνονται με Β δείχνουν διαφορετική μεταβλητότητα, με το δεύτερο και τέταρτο μπλοκ των ασθενών να έχουν υψηλότερη έκφραση αυτών των γονιδίων. Στο κάτω μέρος της απεικόνισης εμφανίζονται τα πρώτα δύο μοναδικά διανύσματα (κύριες συνιστώσες) του πίνακα τιμών έκφρασης (κόκκινα σημεία), καθώς και οι πραγματικές γεννήτριες ομαδοποίησης των δεδομένων (διακεκομμένες γραμμές). Αν το αποτέλεσμα είναι υψηλά συσχετισμένο με καθεμια κύρια συνιστώσα, τότε η τεχνική των κυρίων συνιστωσών με επίβλεψη θα το ανακαλύψει.

Η τεχνική των κυρίων συνιστωσών με επίβλεψη που θα περιγράψουμε στην παρούσα εργασία έχει σχεδιαστεί για να αποκαλύψει μια τέτοιου είδους δομή αυτόματα. Η τεχνική αυτή αρχικά είχε περιγραφεί σε ένα βιολογικό περιβάλλον από τους Bair και Tibshirani (2004) στο πλαίσιο μιας σχετικής μεθόδου που είναι γνωστή ως "επιβλεπόμενη ομαδοποίηση". Η βασική ιδέα της ανάλυσης κυρίων συνιστωσών με επίβλεψη είναι απλή: αντί να εκτελούμε την ανάλυση κυρίων συνιστωσών χρησιμοποιώντας όλα τα γονίδια ενός συνόλου δεδομένων, χρησιμοποιούμε μόνο εκείνα τα γονίδια με την ισχυρότερη εκτιμώμενη συσχέτιση με το Y. Στο σενάριο του σχήματος 2, αν το Y συσχετίζονταν σε μεγάλο βαθμό με τη δεύτερη κύρια συνιστώσα u_2, τότε τα γονίδια στο μπλοκ Β θα είχαν την υψηλότερη συσχέτιση με το Y. Ως εκ τούτου, θα υπολογίζαμε την πρώτη κύρια συνιστώσα χρησιμοποιώντας μόνο αυτά τα γονίδια, και με αυτόν τον τρόπο θα προέκυπτε το u_2.
Όπως δείχνει αυτό το παράδειγμα, η χρήση των κυρίων συνιστωσών βοηθά στην αποκάλυψη ομάδων γονιδίων που εκφράζονται από κοινού. Βιολογικά, μία ή περισσότερες κυτταρικές διεργασίες, συνοδευόμενες από το στέλεχος των εκφραζόμενων γονιδίων τους, καθορίζουν το αποτέλεσμα της επιβίωσης. Αυτό το ίδιο μοντέλο αποτελεί τη βάση άλλων προσεγγίσεων στην μάθηση με επίβλεψη σε μελέτες μικροσυστοιχιών, όπως το gene shaving με επίβλεψη (Hastie et al. 2000) και το tree harvesting (Hastie, Tibshirani, Botstein και Brown 2001). Η διαδικασία των κυρίων συνιστωσών με επίβλεψη μπορεί να θεωρηθεί ως ένας απλός τρόπος προσδιορισμού των ομάδων-συστάδων των σχετικών μεταβλητών πρόβλεψης με επιλογή βασισμένη σε αποτελέσματα για την αφαίρεση των άσχετων πηγών μεταβλητότητας και την εφαρμογή των κυρίων συνιστωσών για τον εντοπισμό των ομάδων των συνεκφραζόμενων γονιδίων.
Απ’ ό,τι γνωρίζουμε, οι Bair και Tibshirani (2004) ήταν οι πρώτοι που συζήτησαν την ιδέα των κυρίων συνιστωσών με επίβλεψη λεπτομερώς. Και άλλοι συγγραφείς, όμως, έχουν υποβάλει σχετικές ιδέες. Ο Ghosh (2002) εξέτασε τα γονίδια πριν από την εξαγωγή κυρίων συνιστωσών, αλλά φάνηκε να το έκανε για υπολογιστικούς λόγους.

Οι Nguyen και Rocke (2002) και οι Hi και Gui (2004) συζήτησαν τις προσεγγίσεις των μερικών ελαχίστων τετραγώνων (PLS) για την πρόβλεψη της επιβίωσης από δεδομένα μικροσυστοιχιών. Αυτή είναι μια σχετική αλλά διαφορετική μέθοδος, και τα PLS δεν αποδίδουν εξίσου καλά με την μέθοδο ανάλυσης κύριων συνιστωσών με επίβλεψη στις δοκιμές που πραγματοποιήθηκαν. Τα PLS δεν κάνουν ένα αρχικό thresholding των χαρακτηριστικών, και αυτό είναι η βασική ιδέα κλειδί της προτεινόμενης διαδικασίας των Bair et al. (2006), η οποία αποφέρει και την πολύ καλή επίδοση της μεθόδου.

Στην επόμενη ενότητα παρουσιάζεται η διαδικασία της ανάλυσης κυρίων συνιστωσών με επίβλεψη. Η ενότητα 3 δίνει μια σύντομη περίληψη της συνέπειας των αποτελεσμάτων, και η ενότητα 4 πραγματεύεται ένα μέτρο σημαντικότητας για κάθε χαρακτηριστικό ξεχωριστά και για ένα μειωμένο μοντέλο. Η ενότητα 5 δίνει ένα παράδειγμα από μια μελέτη λεμφώματος, η ενότητα 6 εξετάζει εναλλακτικές προσεγγίσεις για ημιεπιβλεπόμενη πρόβλεψη, συμπεριλαμβανομένου του "gene shaving", και η ενότητα 7 παρουσιάζει μια μελέτη προσομοίωσης η οποία συγκρίνει τις διάφορες προαναφερθέντες μεθόδους. Η ενότητα 8 συνοψίζει τα αποτελέσματα των κυρίων συνιστωσών με επίβλεψη σε κάποιες μελέτες επιβίωσης. Η ενότητα 9 δίνει λεπτομέρειες για τα θεωρητικά αποτελέσματα. Η εργασία καταλήγει σε συμπέρασμα με κάποιες γενικεύσεις, συμπεριλαμβανομένων μιας προσαρμογής συμμεταβλητών, και της χρήσης μη ταξινομημένων δεδομένων στην ενότητα 10. Στην ενότητα 11 γίνεται συζήτηση των περιορισμών της προτεινόμενης μεθόδου και παρουσιάζονται θέματα για μελλοντική εργασία.

2.2. Ανάλυση Κυρίων Συνιστωσών με Επίβλεψη

2.2.1. Περιγραφή της Μεθόδου

Υποθέτουμε ότι υπάρχουν p χαρακτηριστικά που μετρώνται σε N παρατηρήσεις (π.χ. ασθενείς). Έστω X ένας $N \times p$ πίνακας των μετρήσεων των χαρακτηριστικών (π.χ. γονιδίων), και έστω y το N-διάνυσμα των μετρήσεων του αποτελέσματος. Υποθέτουμε ότι το αποτέλεσμα είναι μια ποσοτική μεταβλητή; θα μπορούσαμε να έχουμε και άλλου τύπους αποτελεσμάτων, όπως οι λογοκριμένοι χρόνοι επιβίωσης. Παρακάτω περιγράφεται με λίγα λόγια η προτεινόμενη μέθοδος της ανάλυσης κύριων συνιστωσών με επίβλεψη:

1. Υπολογίστε τους (μονομεταβλητούς) συντελεστές κανονικής παλινδρόμησης για κάθε χαρακτηριστικό.
2. Κατασκευάστε ένα μειωμένο πίνακα δεδομένων που θα αποτελείται μόνο από τα χαρακτηριστικά των οποίων ο μονομεταβλητός συντελεστής παλινδρόμησης υπερβαίνει ένα κατώφλι (threshold) θ σε απόλυτη τιμή (το θ υπολογίζεται μέσω της διαδιακασίας διασταυρωμένης επικύρωσης (cross-validation)).
3. Υπολογίστε τις πρώτες (ή τις πρώτες λίγες) κύριες συνιστώσες του μειωμένου πίνακα δεδομένων.

4. Χρησιμοποιήστε αυτές τις κύριες συνιστώσες σε ένα μοντέλο παλινδρόμησης για να προβλεφθεί το αποτέλεσμα.

Δίνονται τώρα τις λεπτομέρειες της μεθόδου. Ας υποθέσουμε ότι οι στήλες του X (μεταβλητές) έχουν κεντραριστεί για να έχουν μέση τιμή 0. Γράφουμε την αποσύνθεση μοναδικών τιμών (SVD) του X ως

$$X = UDV^T \quad (1)$$

όπου U, D και V είναι $N \times m$, $m \times m$, και $m \times p$, και $m = \min(N - 1, p)$ είναι η τάξη του X. Εδώ ο D είναι ένας διαγώνιος πίνακας που περιέχει τις μοναδικές τιμές d_j και οι στήλες του U είναι οι κύριες συνιστώσες $u_1, u_2, \ldots, u_m$. Αυτές υποτίθεται ότι είναι διατεταγμένες, έτσι ώστε $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$.
Έστω s το p-διάνυσμα των τυποποιημένων συντελεστών παλινδρόμησης για τη μέτρηση της μονομεταβλητής επίδρασης του κάθε γονιδίου ξεχωριστά στο y,

$$s_j = \frac{x_j^T y}{\|x_j\|} \quad (2)$$

με $\|x_j\| = \sqrt{x_j^T x_j}$. Στην πραγματικότητα, μια κλίμακα εκτίμησης $\hat{\sigma}$ λείπει σε καθένα από τα s_j, αλλά επειδή είναι κοινό για όλους, μπορούμε να το παραλείψουμε. Έστω C_θ η συλλογή των δεικτών τέτοια ώστε $|s_j| > \theta$. Έχουμε δηλώσει με X_θ τον πίνακα που αποτελείται από τις στήλες του X που αντιστοιχούν στο C_θ. Η SVD του X_θ είναι

$$X_\theta = U_\theta D_\theta V_\theta^T. \quad (3)$$

Με $U_\theta = (u_{\theta,1}, u_{\theta,2}, \ldots, u_{\theta,m})$, $u_{\theta,1}$ καλείται η πρώτη κύρια συνιστώσα με επίβλεψη του X, και ούτω καθεξής. Προσαρμόζουμε τώρα ένα μοντέλο μονομεταβλητής γραμμικής παλινδρόμησης με απόκριση y και μεταβλητή πρόβλεψης $u_{\theta,1}$,

$$\hat{y}^{spc,\theta} = \bar{y} + \hat{\gamma} * u_{\theta,1}. \quad (4)$$

Σημειώνεται εδώ ότι επειδή το $u_{\theta,1}$ είναι ένα μοναδικό αριστερό διάνυσμα του X_θ, έχει μέση τιμή 0 και νόρμα μονάδα. Ως εκ τούτου $\hat{\gamma} = u_{\theta,1}^T y$, και ο σταθερός όρος είναι το $\bar{y}$, η μέση τιμή του y (εδώ επεκτείνεται ως διάνυσμα τέτοιων μέσων). Χρησιμοποιείται διασταυρωμένη επικύρωση (cross validation) του στατιστικού της λογαριθμοπιθανοφάνειας (ή της μερικής λογαριθμοπιθανοφάνειας) για να εκτιμήσουμε την καλύτερη τιμή της θ. Στα περισσότερα παραδείγματα στην εργασία των Bair et al. (2006) εξετάζονται μόνο οι πρώτες κύριες συνιστώσες με επίβλεψη. Στα παραδείγματα της ενότητας 8, επιτρέπεται η δυνατότητα χρησιμοποίησης

περισσοτέρων της μίας συνιστώσας.
Από την (3),

$$U_\theta = X_\theta V_\theta D_\theta^{-1} = X_\theta W_\theta \quad (5)$$

Έτσι, για παράδειγμα, το $u_{\theta,1}$ είναι ένας γραμμικός συνδυασμός των στηλών του X_θ : $u_{\theta,1} = X_\theta w_{\theta,1}$. Ως εκ τούτου η εκτίμηση του μοντέλου γραμμικής παλινδρόμησης μπορεί να θεωρηθεί ως μια περιορισμένη εκτίμηση γραμμικού μοντέλου χρησιμοποιώντας όλες τις μεταβλητές πρόβλεψης στο X_θ ,

$$\hat{y}^{spc,\theta} = \bar{y} + \hat{\gamma} * X_\theta w_{\theta,1} \quad (6)$$
$$= \bar{y} + X_\theta \hat{\beta}_\theta, \quad (7)$$

όπου $\hat{\beta}_\theta = \hat{\gamma} w_{\theta,1}$. Στην πραγματικότητα, "γεμίζοντας" το $w_{\theta,1}$ με μηδενικά (που αντιστοιχούν στα γονίδια που εξαιρούνται/ αποκλείονται από το C_θ), η εκτίμησή μας είναι γραμμική σε όλα τα p γονίδια.
Λαμβάνοντας υπόψη ένα διάνυσμα χαρακτηριστικών δοκιμής x*, μπορούμε να κάνουμε προβλέψεις από το μοντέλο παλινδρόμησης ως εξής:

1. Κεντράρουμε κάθε συνιστώσα του x* χρησιμοποιώντας τα μέσα που αντλούμε από τα δεδομένα εκπαίδευσης, $x_j^* \leftarrow x_j^* - \bar{x}_j$.
2. $y^* = \bar{y} + \hat{\gamma} x_\theta^{*T} w_{\theta,1} = \bar{y} + x_\theta^{*T} \hat{\beta}_\theta$,

όπου x_θ^* είναι το κατάλληλο υποδιάνυσμα του x*.
Στην περίπτωση των ασυσχέτιστων μεταβλητών πρόβλεψης, είναι εύκολο να εξακριβωθεί ότι η διαδικασία των κυρίων συνιστωσών με επίβλεψη έχει την επιθυμητή συμπεριφορά. Αποδίδει σωστά όλες τις μεταβλητές πρόβλεψης των οποίων οι τυποποιημένοι μονομεταβλητοί συντελεστές παλινδρόμησης υπερβαίνουν το θ σε απόλυτη τιμή.
Η προτεινόμενη αυτή μέθοδος των Bair et al. (2006) εφαρμόζεται,επίσης, σε γενικευμένες μορφές παλινδρόμησης, για παράδειγμα, σε δεδομένα επιβίωσης, σε προβλήματα ταξινόμησης, ή σε στοιχεία τα οποία αναλύονται τυπικά από ένα γενικευμένο γραμμικό μοντέλο. Στις περιπτώσεις αυτές, χρησιμοποιείται ένα στατιστικό σκορ στην θέση των τυποποιημένων συντελεστών παλινδρόμησης στη (2) και χρησιμοποιούμε έναν αναλογικό κίνδυνο ή την κατάλληλη γενικευμένη μορφή παλινδρόμησης στην (4).

Έστω τώρα $l_j(\beta)$ η λογαριθμοπιθανοφάνεια (ή η μερική λογαριθμοπιθανοφάνεια) η οποία συνδέει τα δεδομένα της μοναδικής μεταβλητής πρόβλεψης X_j και του αποτελέσματος y, και έστω $U_j(\beta_0) = dl/d\beta|_{\beta=\beta_0}$ και $I_j(\beta_0) = -d^2 l_j/d\beta^2|_{\beta=\beta_0}$. Τότε το στατιστικό σκορ για τη μεταβλητή πρόβλεψης j θα έχει τη μορφή

$$s_j = \frac{U_j(0)^2}{I_j(0)}. \quad (8)$$

Φυσικά, για την γκαουσιανή λογαριθμοπιθανοφάνεια, η ποσότητα αυτή είναι ισοδύναμη με τον τυποποιημένο συντελεστή παλινδρόμησης (2).
Κάποιος θα μπορούσε να εξετάσει τη διαδικασία των κυρίων συνιστωσών με επίβλεψη. Έτσι, θα βρίσκαμε χαρακτηριστικά των οποίων το εσωτερικό προϊόν με τις ισχύουσες κύριες συνιστώσες με επίβλεψη ήταν μεγαλύτερο, θα χρησιμοποιούσαμε τα χαρακτηριστικά αυτά για να υπολογίσουμε τις νέες κύριες συνιστώσες, και ούτω καθεξής. Όμως αυτή η διαδικασία θα τείνει να συγκλίνει στις συνήθεις (μη επιβλεπόμενες) κύριες συνιστώσες, γιατί δεν υπάρχει τίποτα να την κρατήσει κοντά στο αποτέλεσμα μετά το πρώτο βήμα. Μια επαναληπτική διαδικασία θα είχε νόημα μόνο εάν βασιζόταν σε κριτήριο που αφορά τόσο τη διακύμανση των χαρακτηριστικών όσο και την καλή προσαρμογή στο αποτέλεσμα. Θεωρούμε ένα τέτοιο κριτήριο στην επόμενη ενότητα, αν και τελικά δεν το ακολουθούμε (για λόγους που αναφέρονται εκεί).

2.2.2. Ένα Βασικό Μοντέλο

Το μοντέλο το οποίο χρησιμοποιήθηκε για την ανάπτυξη της μεθόδου της ανάλυσης κυρίων συνιστωσών με επίβλεψη περιγράφεται παρακάτω. Ας υποθέσουμε ότι έχουμε μια μεταβλητή απόκρισης Y που σχετίζεται με μία υποκείμενη κρυφή μεταβλητή U μέσω του γραμμικού μοντέλου,

$$Y = \beta_0 + \beta_1 U + \varepsilon. \qquad (9)$$

Επιπλέον, έχουμε μετρήσεις έκφρασης σε ένα σύνολο γονιδίων X_j με την ένδειξη $j \in P$, για τις οποίες

$$X_j = \alpha_{0j} + \alpha_{1j} U + \varepsilon_j, \qquad j \in P. \qquad (10)$$

Τα σφάλματα ε και ε_j από υπόθεση έχουν μέση τιμή 0 και είναι ανεξάρτητα από όλες τις άλλες τυχαίες μεταβλητές στα αντίστοιχα μοντέλα τους.
Έχουμε επίσης πολλά επιπρόσθετα γονίδια X_k, $k \notin P$, που είναι ανεξάρτητα του U.
Μπορούμε να σκεφτούμε το U ως ένα διακριτό ή συνεχή παράγοντα ενός τύπου κυττάρων, τον οποίο δεν μετρούμε άμεσα. Το P αντιπροσωπεύει ένα σύνολο γονιδίων που συγκροτούν ένα μονοπάτι ή μια διαδικασία που σχετίζεται με αυτό τον τύπο κυττάρων, και τα X_j είναι θορυβώδεις μετρήσεις της γονιδιακής τους έκφρασης. Θα θέλαμε να εντοπίσουμε το P, να εκτιμήσουμε το U, και ως εκ τούτου, να προσαρμόσουμε το μοντέλο πρόβλεψης (9). Αυτή είναι μια ειδική περίπτωση ενός μοντέλου λανθάνουσας/ κρυφής δομής ή ενός μοντέλου μιας συνιστώσας παραγοντικής ανάλυσης.
Ο αλγόριθμος των κυρίων συνιστωσών με επίβλεψη (SPCA) χρησιμοποιείται για την

προσαρμογή αυτού του μοντέλου:

1. Το βήμα ελέγχου υπολογίζει το σύνολο P από την $\hat{P} = C_\theta$.
2. Δεδομένου του $\hat{P}$, η SVD του X_θ εκτιμά το U στην (10) από τη μεγαλύτερη κύρια συνιστώσα $u_{\theta,1}$.
3. Τέλος, η προσαρμογή της παλινδρόμησης (4) εκτιμά την (9).

Το βήμα 1 είναι φυσικό, γιατί κατά μέσο όρο ο συντελεστής παλινδρόμησης $S_j = \mathrm{X}_j^T \mathrm{Y} / \|\mathrm{X}_j\|$ είναι μη μηδενικός μόνο αν το α_{1j} είναι μη μηδενικό (υποθέτοντας ότι τα γονίδια είναι κεντραρισμένα). Ως εκ τούτου το βήμα αυτό θα έπρεπε να επιλέγει τα γονίδια για τα οποία $j \in P$. Το βήμα 2 είναι φυσικό αν υποθέσουμε ότι τα σφάλματα ε_j έχουν μια γκαουσιανή κατανομή, με την ίδια διακύμανση/ διασπορά. Σε αυτή την περίπτωση η SVD δίνει τις μέγιστες εκτιμήσεις πιθανότητας για το μονοπαραγοντικό μοντέλο. Η παλινδρόμηση στο βήμα 3 είναι ένα προφανές τελικό βήμα. Στην πραγματικότητα, δεδομένου του P, το μοντέλο που ορίζεται από τις (9) και (10) είναι μια ειδικά δομημένη περίπτωση ενός *errors-in-variables* μοντέλου. Κάποιος θα μπορούσε να χρησιμοποιήσει ένα κριτήριο κοινής βελτιστοποίησης,

$$\min_{\beta_0,\beta_1,[\alpha_{0,j},\alpha_{1,j}],u_1,\dots,u_N} \frac{\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 u_i)^2}{\sigma_Y^2}$$

$$+ \sum_{j \in P} \frac{\sum_{i=1}^{N}(x_{ij} - a_{0j} - a_{1j}u_i)^2}{\sigma_X^2}. \quad (11)$$

Τότε είναι εύκολο να δειχθεί ότι η (11) μπορεί να λυθεί από ένα πρόβλημα επαυξημένης και σταθμισμένης SVD. Πιο αναλυτικά, σχηματίζουμε τον επαυξημένο πίνακα

$$\mathrm{X}_\alpha = (\mathrm{y} : \mathrm{X}), \qquad (12)$$

και βάζουμε το βάρος $\omega_1 = \sigma_X^2/\sigma_Y^2$ στην πρώτη στήλη και το βάρος $\omega_j = 1$ στις υπόλοιπες στήλες. Στη συνέχεια, με

$$\mathrm{v}_0 = \begin{pmatrix} \beta_0 \\ \alpha_{0j_1} \\ \vdots \\ \alpha_{0j_q} \end{pmatrix}, \qquad \mathrm{v}_1 = \begin{pmatrix} \beta_1 \\ \alpha_{1j_1} \\ \vdots \\ \alpha_{1j_q} \end{pmatrix}, \qquad (13)$$

(με $q = |P|$) η πρώτου βαθμού σταθμισμένη SVD $X_\alpha \approx 1v_0^T + uv_1^T$ λύνει το πρόβλημα της βελτιστοποίησης στην (11). Παρά το γεγονός ότι αυτή η προσέγγιση μπορεί να φαίνεται πιο σωστή από τη διαδικασία των δύο βημάτων, η SPCA έχει ένα σαφές πλεονέκτημα. Εδώ $\hat{u}_{\theta,1} = X_\theta w_{\theta,1}$, και ως εκ τούτου μπορεί να οριστεί για μελλοντικά x* δεδομένα και να χρησιμοποιειθεί για προβλέψεις. Στην errors-in-variables προσέγγιση, $\hat{u}_{EV} = X_A w_{EV}$, η οποία περιλαμβάνει το y, δεν αφήνει καμια προφανή εκτίμηση για μελλοντικά δεδομένα.

Αυτό το μοντέλο μπορεί εύκολα να επεκταθεί για να συμπεριλάβει πολλαπλές συνιστώσες $U_1, \ldots, U_m$. Ένας τρόπος για να γίνει αυτό είναι να υποθέσουμε ότι

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m U_m + \varepsilon \qquad (14)$$

και

$$X_j = \alpha_{0j} + \sum_{m=1}^{M} a_{1jm} U_m + \varepsilon_j, \quad j \in P. \qquad (15)$$

Η προσαρμογή αυτού του μοντέλου προχωρά όπως και πριν, μόνο που τώρα εξάγουμε το M αντί μια κύρια συνιστώσα από το X_θ.

Προσομοιώθηκαν δεδομένα από ένα σενάριο όπως αυτό του Σχήματος 2. Χρησιμοποιήθηκαν 1.000 γονίδια και 40 δείγματα, όλα με βάση το μοντέλο σφάλματος να είναι γκαουσιανό με μοναδιαία διακύμανση. Στη συνέχεια ορίστηκαν τα μέσα διανύσματα μ_1 και μ_2 ως εξής. Χωρίζουμε τα δείγματα σε διαδοχικά μπλοκ των 10, που υποδεικνύονται από τα σύνολα (a, b, c, d). Τότε

$$\mu_{1i} = \begin{cases} -2 & \text{αν } i \in a \cup b \\ +2 & \text{διαφορετικά} \end{cases} \qquad (16)$$

και

$$\mu_{2i} = \begin{cases} -1 & \text{αν } i \in a \cup c \\ +1 & \text{διαφορετικά} \end{cases} \qquad (17)$$

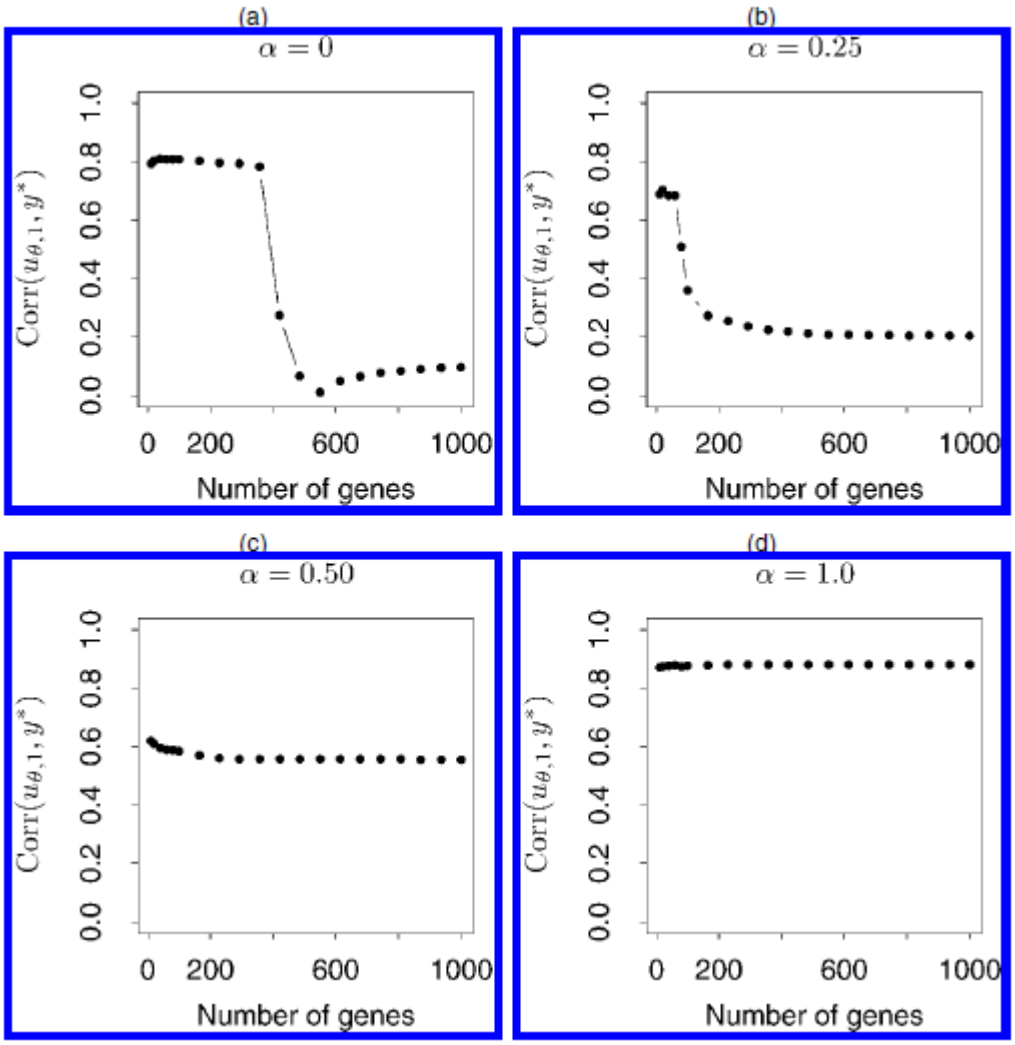
Τα πρώτα 200 γονίδια έχουν μέση δομή μ_1,

$$x_{ij} = \mu_{1i} + \varepsilon_{ij}, \quad j = 1, \ldots, 200, \; i = 1, \ldots, 40. \qquad (18)$$

Τα επόμενα 50 γονίδια έχουν μέση δομή μ_2,

$x_{ij} = \mu_{2i} + \varepsilon_{ij}, \quad j = 201,\ldots, 250, i = 1,\ldots, 40.$ (19)

Σε όλες τις περιπτώσεις ισχύει ότι $\varepsilon_{ij} \sim N(0, 1)$, το οποίο ισχύει και για τα υπόλοιπα 750 γονίδια. Τέλος, το αποτέλεσμα παράγεται ως $y_i = \alpha * \mu_{1i} + (1-\alpha) * \mu_{2i} + \varepsilon_i$, όπου το ε_i είναι N (0, 1). Οι πρώτες δύο κύριες συνιστώσες του X είναι περίπου μ_1 και μ_2 (βλ. Σχήμα. 2).
Δοκιμάστηκαν διάφορες τιμές του $\alpha \in [0, 1]$, όπως φαίνεται στο Σχήμα 3. Απεικονίζεται η συσχέτιση της μεταβλητής πρόβλεψης των κυρίων συνιστωσών με επίβλεψη με μια ανεξάρτητη (σύνολο ελέγχου) πραγματοποίηση του y αφού το θ κατά τη διαδικασία επιλογής $|s_j| > \theta$ ποικίλει. Ο αριθμός των γονιδίων που επιβιώνουν από τον έλεγχο απεικονίζεται στον οριζόντιο άξονα. Η τελείως δεξιά άκρη του διαγράμματος αντιπροσωπεύει την παλινδρόμηση των τυποποιημένων κυρίων συνιστωσών. Όταν $\alpha = 0$, το αποτέλεσμα συσχετίζεται με τη δεύτερη κύρια συνιστώσα, η κύρια συνιστώσα με επίβλεψη βελτιώνεται στην παλινδρόμηση κυρίων συνιστωσών. Όταν το α φτάνει το 0,5, αυτό το πλεονέκτημα παύει να υπάρχει, αλλά η κύρια συνιστώσα με επίβλεψη δεν είναι χειρότερη από την παλινδρόμηση κυρίων συνιστωσών.



Σχήμα 3. Συσχέτιση Μεταξύ της Πρώτης Κύριας Συνιστώσας με Επίβλεψη $u_{\theta,1}$ και μιας Έκβασης Εξέτασης y που, όταν το Βάρος α που δόθηκε στην Πρώτη Κύρια Συνιστώσα στην Παραγωγή των Δεδομένων ποικίλλει. Ο αριθμός των γονιδίων που χρησιμοποιούνται από τη διαδικασία εμφανίζεται στον οριζόντιο άξονα, σε κάθε πίνακα. Η απότομη εναλλαγή (a) και (b) αντιστοιχεί στο σημείο στο οποίο η σειρά των κυρίων συνιστωσών αντιστρέφεται.

2.3. Συνέπεια των Κυρίων Συνιστωσών με Επίβλεψη

Η παλινδρόμηση των τυποποιημένων κυρίων συνιστωσών δεν είναι συνεπής όσο το μέγεθος του δείγματος και ο αριθμός των χαρακτηριστικών αυξάνονται, ενώ οι κύριες συνιστώσες με επίβλεψη είναι συνεπείς κάτω από κατάλληλες υποθέσεις (βλ. ενότητα 9).

Θεωρούμε ένα μοντέλο λανθάνουσας/κρυφής μεταβλητής της μορφής (9) και (10) για δεδομένα με N δείγματα και p χαρακτηριστικά. Δηλώνουμε τον πλήρη $N \times p$ πίνακα χαρακτηριστικών X, και το $N \times P_1$ μπλοκ του X με X_1, ώστε να αντιστοιχούν με τα

χαρακτηριστικά $j \in P$. Υποθέτουμε ότι όταν $N \to \infty$, $p/N \to \gamma \in (0, \infty)$ και $p_1/N \to 0$ "γρήγορα". Τα p και p_1 μπορούν να καθοριστούν ή να προσεγγίζουν το ∞. Δεδομένης αυτής της ρύθμισης, αποδεικνύονται τα παρακάτω:

• Έστω $\tilde{U}$ η κορυφαία κύρια συνιστώσα του X και έστω $\tilde{\beta}$ ο συντελεστής παλινδρόμησης του Y στο $\tilde{U}$. Τότε $\tilde{U}$ δεν είναι γενικά συνεπές για το U, και επίσης το $\tilde{\beta}$ δεν είναι γενικά συνεπές για το β (το U είναι μια τυχαία μεταβλητή). Ας υποθέσουμε ότι μας δίνεται το X_1. Στη συνέχεια, αν το $\hat{U}$ είναι η κορυφαία κύρια συνιστώσα του X_1 και το $\hat{\beta}$ είναι ο συντελεστής παλινδρόμησης του Y στο $\hat{U}$, τότε είναι και τα δύο συνεπή.
• Αν το X_1 δεν δίνεται αλλά υπολογίζεται μέσω thresholding μονοπαραγοντικών χαρακτηριστικών (όπως στη διαδικασία των κυρίων συνιστωσών με επίβλεψη), τότε τα αντίστοιχα $\hat{U}$ και $\hat{\beta}$ είναι συνεπή.

Ανάλογα αποτελέσματα προκύπτουν, επίσης, για το μοντέλο αναλογικών κινδύνων του Cox.

2.4. Σκορ Σημαντικότητας με μια Μειωμένη Μεταβλητή Πρόβλεψης

Έχοντας βρει τη μεταβλητή πρόβλεψης $u_{\theta,1}$, και πώς αξιολογούνται οι συνεισφορές των p ξεχωριστών χαρακτηριστικών; Ορίζεται το σκορ σημαντικότητας (*importance score*) ως το εσωτερικό γινόμενο μεταξύ κάθε χαρακτηριστικού και του $u_{\theta,1}$,

$$\text{imp}_j = \langle x_j, \quad u_{\theta,1} \rangle. \qquad (20)$$

Χαρακτηριστικά j με μεγάλες τιμές του $|\text{imp}_j|$ συμβάλουν περισσότερο στην πρόβλεψη του y. Αν τα χαρακτηριστικά είναι τυποποιημένα, τότε αυτό είναι απλά η συσχέτιση μεταξύ κάθε γονιδίου και της κύριας συνιστώσας με επίβλεψη.
Σε ορισμένες εφαρμογές, θα θέλαμε να έχουμε ένα μοντέλο που χρησιμοποιεί μόνο ένα μικρό αριθμό χαρακτηριστικών. Για παράδειγμα, μια μεταβλητή πρόβλεψης που απαιτεί μετρήσεις έκφρασης για μερικές χιλιάδες γονίδια δεν είναι πιθανό να είναι χρήσιμη σε καθημερινές κλινικές ρυθμίσεις. Οι μικροσυστοιχίες είναι υπερβολικά ακριβές και περίπλοκες για καθημερινή χρήση, και απλούστερες δοκιμασίες, όπως η ανάποδη αντιγραφή-αλυσιδωτή αντίδραση πολυμεράσης μπορούν να μετρήσουν μόνο 50 ή 100 γονίδια τη φορά. Επιπλέον, η απομόνωση ενός μικρότερου συνόλου γονιδίων θα μπορούσε να βοηθήσει την κατανόηση της ασθένειας από βιολογικής πλευράς.
Υπάρχουν διάφοροι τρόποι για να αποκτήσουμε μια σειρά από μειωμένα μοντέλα. Ο αλγόριθμος LAR (Efron, Hastie, Johnstone και Tibshirani 2004) παρέχει μια βολική μέθοδο για τον υπολογισμό των lasso λύσεων. Ένα μειονέκτημα αυτής της προσέγγισης είναι ότι η σειρά των μοντέλων θα περιλαμβάνει συνήθως διαφορετικά σύνολα χαρακτηριστικών, τα οποία μπορεί να είναι δύσκολο για έναν επιστήμονα να αφομοιώσει.
Εδώ παίρνουμε μια απλούστερη προσέγγιση. Ορίζουμε

$$\hat{u}_{red} = \sum_{|imp_j|>\gamma} l_j \times x_j , \qquad (21)$$

όπου $l_j = u_{\theta,1}^T x_j / d_1$ είναι το loading για το j-οστό χαρακτηριστικό και d_1 είναι η πρώτη μοναδική τιμή από την SVD (3). Αυτή η μεταβλητή πρόβλεψης κρατά μόνο τα χαρακτηριστικά με σκορ σημαντικότητας γ ή μεγαλύτερα, και σταθμίζει αυτά τα χαρακτηριστικά από τα loadings τους.
Θα μπορούσε κανείς να υπολογίσει τα σκορ σημαντικότητας, και την αντίστοιχη μειωμένη μεταβλητή πρόβλεψης, για όλα τα χαρακτηριστικά (όχι μόνο αυτά που χρησιμοποιήθηκαν στον υπολογισμό των κυρίων συνιστωσών με επίβλεψη). Για παράδειγμα, θα μπορούσε να υπάρχει ένα χαρακτηριστικό όχι στο πρώτο σύνολο που έχει υψηλότερο εσωτερικό γινόμενο με την κύρια συνιστώσα με επίβλεψη από ό,τι ένα χαρακτηριστικό στο πρώτο σύνολο. Ωστόσο, το ενδιαφέρον συγκεντρώνεται στα χαρακτηριστικά του πρώτου συνόλου για δυο λόγους.
Με την προσέγγιση αυτή, η τιμή $\gamma = 0$ δίνει την αρχική μεταβλητή πρόβλεψης των κυρίων συνιστωσών με επίβλεψη, διευκολύνοντας τη σύγκριση μεταξύ των ολόκληρων και των μειωμένων μοντέλων.
Δεύτερον, επιτρέποντας στο μειωμένο μοντέλο να χρησιμοποιεί χαρακτηριστικά που είναι έξω από το πρώτο σύνολο οδηγεί σε μια επαναλαμβανόμενη έκδοση της διαδικασίας κατά την οποία ξαναυπολογίζεται η κύρια συνιστώσα με επίβλεψη χρησιμοποιώντας γονίδια με το υψηλότερο σκορ σημαντικότητας, υπολογίζονται τα νέα αποτελέσματα και γίνεται επανάληψη. Ωστόσο, η διαδικασία αυτή θα συγκλίνει τυπικά σε μια συνηθισμένη ανάλυση πρώτης κύριας συνιστώσας (δηλαδή χωρίς επίβλεψη). Ως εκ τούτου, δεν εξετάστηκε αυτή η επαναλαμβανόμενη έκδοση, και περιορίστηκε η προσοχή σε γονίδια που περνούν το αρχικό threshold.

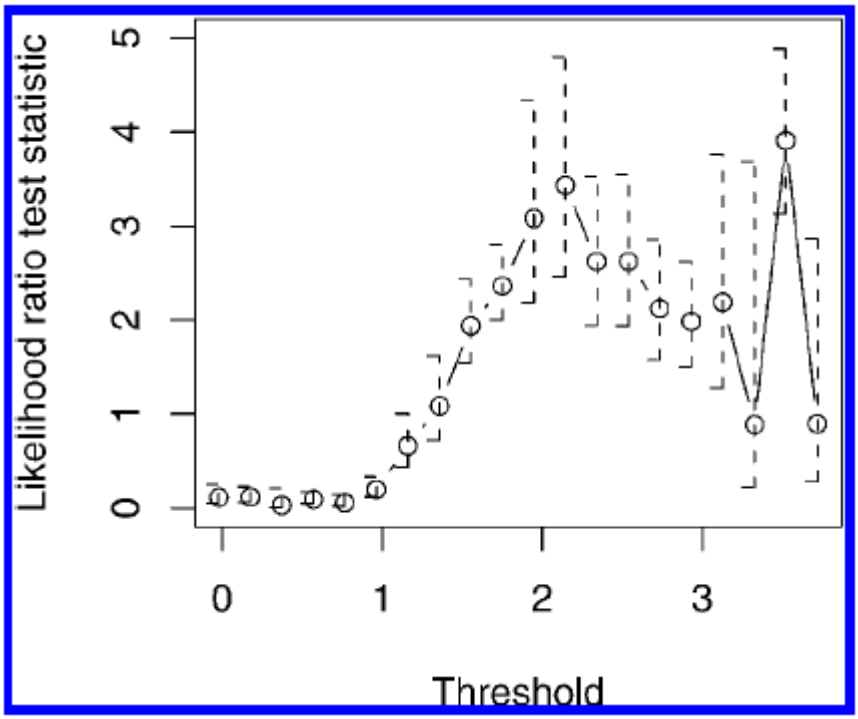
2.5. Παράδειγμα: Επιβίωση των Ασθενών από Λέμφωμα

Αυτό το σύνολο δεδομένων, από τους Rosenwald et al. (2002), αποτελείται από 240 δείγματα από ασθενείς με διάχυτο λέμφωμα μεγάλων Β-κυττάρων (DLBCL = diffuse large B-cell lymphoma), με μετρήσεις γονιδιακής έκφρασης για 7.399 γονίδια. Το αποτέλεσμα ήταν ο χρόνος επιβίωσης, είτε όπως παρατηρείται είτε σωστά λογοκριμένος. Χωρίστηκαν τυχαία τα δείγματα σε ένα σύνολο εκπαίδευσης μεγέθους 160 και σε ένα σύνολο εξέτασης μεγέθους 80. Τα αποτελέσματα διαφόρων διαδικασιών δίνονται στον Πίνακα 1. Χρησιμοποιήθηκαν τα γονίδια με τα κορυφαία 25 αποτελέσματα Cox (με όριο (cutoff) 3,53) στον υπολογισμό της πρώτης κύριας συνιστώσας με επίβλεψη. Παρόλο που τα PLS (περιγράφονται στην ενότητα 6) παρέχουν μια ισχυρή μεταβλητή πρόβλεψης επιβίωσης, οι κύριες συνιστώσες με επίβλεψη είναι ακόμη πιο ισχυρές.

Πίνακας 1. Δεδομένα Λεμφώματος: Αποτελέσματα του Συνόλου Εξέτασης για τις Διάφορες Μεθόδους

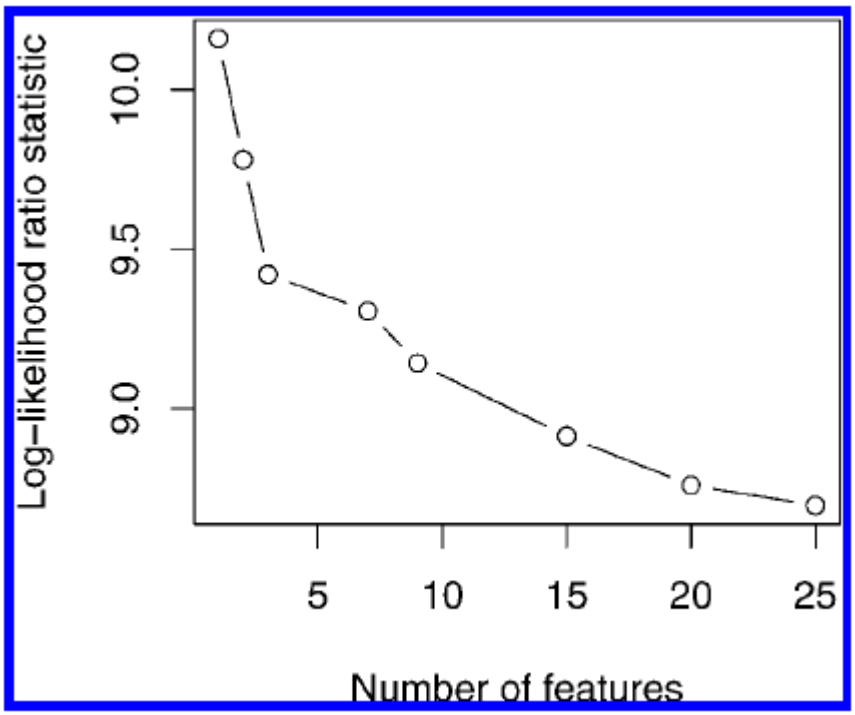
Μέθοδος	Z-score	P value
Πρώτη κύρια συνιστώσα	-1.04	0.2940
Μερικά ελάχιστα τετράγωνα	2.51	0.0112
Πρώτη κύρια συνιστώσα με επίβλεψη	-2.93	0.0045

Το Σχήμα 4 δείχνει την καμπύλη διασταυρωμένης επικύρωσης για την εκτίμηση του καλύτερου threshold. Κάθε μοντέλο έχει εκπαιδευτεί, και στη συνέχεια ο λόγος της λογαριθμοπιθανοφάνειας (LR = likelihood ratio) της στατιστικής δοκιμής υπολογίζεται στα δεδομένα που έμειναν εκτός. Για να υπάρχουν επαρκή στοιχεία στα δεδομένα που έμειναν εκτός ώστε να υπολογιστεί ένα σημαντικό στατιστικό LR, χρησιμοποιούμε δύο φορές διασταυρωμένη επικύρωση (αντί για την πιο τυπική που είναι πέντε ή δέκα φορές). Αυτή η διαδικασία επαναλαμβάνεται πέντε φορές και παίρνουμε το μέσο όρο των αποτελεσμάτων. Στα πειράματα τα οποία πραγματοποιήθηκαν, αυτή η μέθοδος παράγει μια λογική εκτίμηση του καλύτερου threshold, αλλά συχνά υποτιμά το στατιστικό σύνολο δοκιμών LR (επειδή τα σύνολα εκπαίδευσης και επικύρωσης είναι τα μισά των πραγματικών μεγεθών). Αυτή είναι η υπόθεση εδώ, όπου το διασταυρωμένα-επικυρωμένο στατιστικό LR είναι απλά σημαντικό, αλλά το στατιστικό σύνολο δοκιμών LR είναι πολύ σημαντικό.



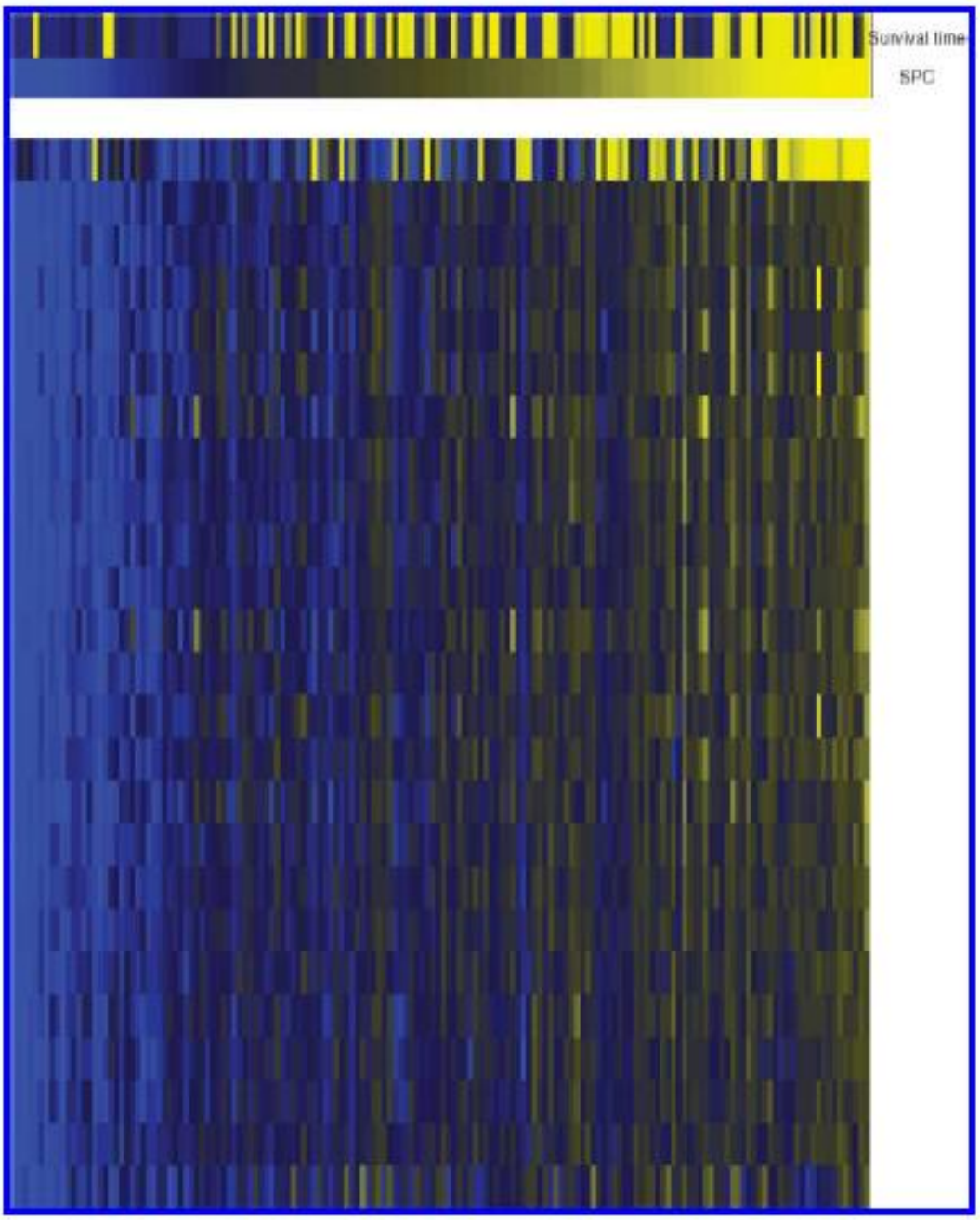
Σχήμα 4. Δεδομένα Λεμφώματος: Καμπύλη Διασταυρωμένης Επικύρωσης για τον Υπολογισμό του Καλύτερου Threshold

Το παράδειγμα αυτό δείχνει επίσης ότι η διαδικασία μπορεί να είναι ευαίσθητη στην τιμή του threshold. Αν αντί για αυτό επιλεγεί ένα threshold 2 (μια λογική επιλογή σύμφωνα με το Σχήμα 4), τότε επιλέγονται 865 γονίδια. Η συσχέτιση της προκύπτουσας κύριας συνιστώσας με επίβλεψη με αυτή που διαπιστώθηκε με 25 γονίδια είναι μόνο 0.5 περίπου. Η μεταβλητή πρόβλεψης της κύριας συνιστώσας με επίβλεψη δίνει μια *p*-τιμή 0,02 στο σύνολο εξέτασης. Αυτή είναι σημαντική, αλλά όχι τόσο ισχυρή όσο αυτή από τη μεταβλητή πρόβλεψης των 25 γονιδίων.
Το Σχήμα 5 δείχνει το στατιστικό λογαριθμοπιθανοφάνειας του συνόλου εξέτασης (test set log-likelihood ratio statistic) που λαμβάνεται από την προσαρμογή μοντέλων παλινδρόμησης διαφόρων μεγεθών στην έξοδο της παλινδρόμησης των κυρίων συνιστωσών με επίβλεψη. Παρατηρείται ότι αν χρησιμοποιούνται τα κορυφαία λίγα γονίδια, τότε δεν υπάρχει καμία απώλεια στη δύναμη πρόβλεψης.



Σχήμα 5. Δεδομένα Λεμφώματος: Στατιστικό Λογαριθμοπιθανοφάνειας του Συνόλου Εξέτασης που Λαμβάνεται από την Προσέγγιση της Μειωμένης Μεταβλητής Πρόβλεψης

Το Σχήμα 6 δείχνει τα κορυφαία 25 γονίδια και τα loadings τους. Λεπτομέρειες δίνονται στη λεζάντα του σχήματος.



Σχήμα 6. Δεδομένα Λεμφώματος: Heatmap Απεικόνιση των Κορυφαίων 25 Γονιδίων. Οι δύο πρώτες σειρές του σχήματος δείχνουν τους χρόνους επιβίωσης που παρατηρήθηκαν και την πρώτη κύρια συνιστώσα (SPC) $u_{\theta,1}$. Για χρόνους επιβίωσης T λογοκριμένους σε χρόνο c, δείχνουμε το $\hat{E}(T|T \geq c)$ που βασίζεται στον εκτιμητή Kaplan-Meier. Όλες οι στήλες έχουν ταξινομηθεί με αυξανόμενη τιμή του $u_{\theta,1}$. Στα δεξιά του heatmap απεικονίζονατι τα loadings $w_{\theta,1}$ (βλέπε (6)). Τα γονίδια (σειρές) ταξινομούνται με μειούμενη τιμή του loading τους. Όλα τα γονίδια αλλά και το τελευταίο έχουν θετικά loadings.

2.6. Κάποιες Εναλλακτικές Προσεγγίσεις

Σε αυτή την ενότητα θα συζητηθούν κάποιες εναλλακτικές προσεγγίσεις σε αυτό πρόβλημα, κάποιες κλασσικές και κάποιες που αντικατοπτρίζουν άλλες προσεγγίσεις που έχουν διερευνηθεί στη διεθνή βιβλιογραφία.

2.6.1. Ridge Παλινδρόμηση

Η ridge παλινδρόμηση είναι μια κλασσική διαδικασία παλινδρόμησης όταν υπάρχουν πολλές συσχετισμένες μεταβλητές πρόβλεψης, και θα μπορούσε λογικά να εφαρμοστεί στο τωρινό περιβάλλον. Η ridge παλινδρόμηση προσαρμόζει το μοντέλο της πλήρως γραμμικής παλινδρόμησης, αλλά διαχειρίζεται το μεγάλο αριθμό των μεταβλητών πρόβλεψης σε αυτές τις ρυθμίσεις γονιδιώματος με κανονικοποίηση (Hastie και Tibshirani 2003). Η ridge παλινδρόμηση λύνει την

$$\min_{\beta} \|\mathrm{y} - \beta_0 - \mathrm{X}\beta\|^2 + \lambda\|\beta\|^2, \qquad (22)$$

όπου ο δεύτερος όρος συρρικνώνει τους συντελεστές στο 0. Η παράμετρος κανονικοποίησης λ ελέγχει το ποσό της συρρίκνωσης, και ακόμη και για τα πιο μικρά λ > 0, η λύση ορίζεται είναι μοναδική. Μπορεί επίσης να δειχθεί ότι αυτή η μορφή κανονικοποίησης συρρικνώνει τους συντελεστές των ισχυρά συσχετισμένων μεταβλητών πρόβλεψης το ένα προς το άλλο, το οποίο αποτελεί μια πολύ ελκυστική ιδιότητα. Χρησιμοποιώντας την εκπροσώπηση μοναδικών τιμών (1), οι προσαρμοσμένες τιμές από την ridge παλινδρόμηση έχουν τη μορφή

$$\begin{aligned}\hat{\mathrm{y}}^{\mathrm{RR}} &= \bar{\mathrm{y}} + \mathrm{X}(\mathrm{X}^T\mathrm{X} + \lambda\mathrm{I})^{-1}\mathrm{X}\mathrm{y} \\ &= \bar{\mathrm{y}} + \sum_{j=1}^{m} \mathrm{u}_j \frac{d_j^2}{d_j^2+\lambda}\, \mathrm{u}_j^T\mathrm{y}. \qquad (23)\end{aligned}$$

Η ridge παλινδρόμηση είναι σαν μια ομαλή έκδοση της παλινδρόμησης των κυρίων συνιστωσών. Αντί να διατηρεί τις πρώτες *k* κύριες συνιστώσες και να απορρίπτει τις υπόλοιπες, σταθμίζει τις διαδοχικές συνιστώσες με έναν παράγοντα που μειώνεται με τη μείωση της ιδιοτιμής d_j^2. Η ridge παλινδρόμηση είναι μια γραμμική μέθοδος. Δηλαδή, το $\hat{\mathrm{y}}^{\mathrm{RR}}$ είναι μια γραμμική συνάρτηση του y. Αντίθετα, η SPCA δεν είναι γραμμική, λόγω του αρχικού βήματος επιλογής γονιδίων.

2.6.2. Lasso

Η Lasso είναι μια παραλλαγή της ridge παλινδρόμησης που λύνει την

$$\min_{\beta} \|\mathrm{y} - \beta_0 - \mathrm{X}\beta\|^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \qquad (24)$$

όπου ο δεύτερος όρος συρρικνώνει τους συντελεστές στο 0. Η απόλυτη τιμή της συνάρτησης ποινής έχει την ελκυστική ιδιότητα ότι μπορεί να συρρικνώνει κάποιους συντελεστές ακριβώς στο 0. Ο υπολογισμός του Lasso είναι πιο δύσκολος από ό,τι ο υπολογισμός της ridge παλινδρόμησης. Το πρόβλημα (24) είναι μια κυρτή

βελτιστοποίηση, η οποία μπορεί να είναι πολύ δύσκολη αν ο αριθμός των *p* χαρακτηριστικών είναι μεγάλος. Ο αλγόριθμος της παλινδρόμησης με τη λιγότερη γωνία (LARS = least-angle regression) (Efron et al. 2004) παρέχει μια αποτελεσματική μεθόδο για τον υπολογισμό του lasso, εκμεταλλευόμενος το γεγονός ότι όσο αλλάζει το λ, τα προφίλ των εκτιμήσεων είναι γραμμικά κατά τμήματα. Για άλλα μοντέλα βασισμένα σε πιθανότητα, όπως το μοντέλο του Cox, η Ευκλείδεια απόσταση στην (22) αντικαθίσταται από την (αρνητική) λογαριθμοπιθανοφάνεια ή τη μερική λογαριιθμοπιθανοφάνεια. Τα προφίλ των συντελεστών δεν είναι γραμμικά κατά τμήματα, οπότε η προσέγγιση LARS δεν μπορεί να εφαρμοστεί.

Όταν το *p* είναι μεγαλύτερο από το μέγεθος του δείγματος *N*, ο αριθμός των μη μηδενικών συντελεστών σε μία lasso λύση είναι το πολύ *N* (για οποιοδήποτε λ). Παρόλο που οι αραιές/σποραδικές λύσεις είναι γενικά ελκυστικές, αυτές οι λύσεις μπορεί να είναι πολύ αραιές, γιατί, για παράδειγμα, για δεδομένα μικροσυστοιχειών θα επέτρεπαν μόνο *N* γονίδια να εμφανιστούν σε ένα συγκεκριμένο μοντέλο. Παρακάτω παρουσιάζονται διάφορες προσεγγίσεις για τις κύριες συνιστώσες με επίβλεψη που τροποποιούν το κριτήριο βελτιστοποίησης πίσω από την ανάλυση κυρίων συνιστωσών με έναν εποπτικό τρόπο.

2.6.3. Μερικά Ελάχιστα Τετράγωνα

Τα PLS είναι μια τέτοια προσέγγιση, με μακρά ιστορία (Hastie, Tibshirani και Friedman 2001). Τα PLS λειτουργούν ως εξής:

1. Κάθε μία από τις μεταβλητές τυποποιείται ώστε να έχει μέση τιμή 0 και νόρμα μονάδα, και υπολογίζονται οι συντελεστές μονοπαραγοντικής παλινδρόμησης $w = X^T y$.
2. Ορίζεται $u_{PLS} = Xw$, και το χρησιμοποιείται σε ένα μοντέλο γραμμικής παλινδρόμησης με το y.

Παρόλο που τα PLS πηγαίνουν να βρουν επόμενες ορθογώνιες συνιστώσες, μία συνιστώσα αρκεί. Τα PLS χρησιμοποιούν ρητά το y στην εκτίμηση της κρυφής μεταβλητής τους. Μπορεί να αποδειχθεί ότι το (κανονικοποιημένο) w στα PLS λύνει το

$$\max_{\|w\|=1} \text{corr}^2(y, Xw)\, \text{var}(Xw), \qquad (25)$$

ένας συμβιβασμός μεταξύ της παλινδρόμησης και της PCA.
Συμπερασματικά ο όρος διακύμανσης κυριαρχεί, και ως εκ τούτου ότι τα PLS σε γενικές γραμμές θα ήταν παρόμοια με την παλινδρόμηση κυρίων συνιστωσών. Αυτό μπορεί να ιδωθεί στο πλαίσιο του θεωρημένου μοντέλου (9)-(10). Οι αναμενόμενες τιμές του συντελεστή παλινδρόμησης w_j είναι

$$E(w_j) = \beta_1 \sum_j \frac{a_j}{a_j^2 + \sigma_j^2}. \qquad (26)$$

Τώρα, αν $\sigma_j^2 = 0$, τότε η κατεύθυνση των PLS $\sum_j w_j x_{ij}$ μειώνεται σε $\beta_1 \sum_j x_{ij}/a_j$. Όμως στην περίπτωση αυτή, ο κρυφός παράγοντας *U* ισούται με $\sum_j X_j/a_j$, έτσι ώστε οι δύο λύσεις να συμφωνούν (όπως προσδοκάται).

Ως εκ τούτου, αφού απομονωθεί το μπλοκ των σημαντικών χαρακτηριστικών, κάνοντας παλινδρόμηση κυρίων συνιστωσών ή PLS είναι πιθανό να προκύψουν παρόμοια αποτελέσματα. Το κύριο πλεονέκτημα των κυρίων συνιστωσών με επίβλεψη επί της τυπικής PLS διαδικασίας είναι η χρήση του threshold ώστε να εκτιμηθούν τα σημαντικά χαρακτηριστικά. Τα PLS διατηρούν όλα τα χαρακτηριστικά και μπορεί να πλήττονται από το θόρυβο στα ασήμαντα χαρακτηριστικά.

2.6.4. Μικτό Κριτήριο Διασποράς – Συνδιασποράς

Η μεγαλύτερη κύρια συνιστώσα είναι ο κανονικοποιημένος γραμμικός συνδυασμός z = Xv των γονιδίων με τη μεγαλύτερη διασπορά δείγματος. Ένας άλλος τρόπος επίβλεψης θα ήταν η εύρεση ενός γραμμικού συνδυασμό z = Xv που να έχει και μεγάλη διακύμανση και μια μεγάλη (ορθογώνια) συνδιακύμανση με το y, και να οδηγεί στο κριτήριο συμβιβασμού

$$\max_{\|v\|=1}(1-a)\mathrm{var}(z) + a\mathrm{cov}(z,y)^2, \quad \text{έτσι ώστε } z = Xv. \qquad (27)$$

Αυτό ισοδυναμεί με το

$$\max_{\|v\|=1}(1-a)v^T X^T X v + a\, v^T X^T y y^T X v. \qquad (28)$$

Αν το y κανονικοποιείται στη μοναδιαία νόρμα, τότε ο δεύτερος όρος στην (28) είναι ένα άθροισμα τετραγώνων παλινδρόμησης (παλινδρομώντας το z στο y) και ερμηνεύεται ως: "η διακύμανση του z εξηγείται από το y". Η λύση v μπορεί να υπολογιστεί αποτελεματικά ως το πρώτο δεξί μοναδικό διάνυσμα του επαυξημένου $(N+1) \times p$ πίνακα,

$$X_\alpha = \begin{pmatrix} (1-\alpha)^{1/2}X \\ \alpha^{1/2} y^T X \end{pmatrix}. \qquad (29)$$

Αλλάζοντας την τιμή της παραμέτρου α, ελέγχουμε την ποσότητα της επίβλεψης. Παρά το γεγονός ότι το μικτό κριτήριο μπορεί να καθοδηγήσει την ακολουθία των ιδιοδιανυσμάτων, όλα τα γονίδια έχουν μη μηδενικά loadings πράγμα που προσθέτει πολλή διακύμανση στη λύση.

2.6.5. Επιβλεπόμενο Gene Shaving

Οι Hastie et al. (2000) πρότειναν το "gene shaving" ως μία μέθοδο για την ομαδοποίηση των γονιδίων. Ο πρωταρχικός στόχος της μεθόδου τους ήταν να βρουν μικρές ομάδες με υψηλά συσχετισμένα γονίδια, των οποίων ο μέσος όρος παρουσίασε έντονη διακύμανση στα δείγματα. Αυτό το πέτυχαν μέσα από μια επαναληπτική διαδικασία, η οποία υπολόγιζε κατ' επανάληψη τη μεγαλύτερη κύρια συνιστώσα ενός υποσυνόλου των γονιδίων, αλλά μετά από κάθε επανάληψη απομάκρυνε ένα μέρος

από τα γονίδια με μικρά φορτία. Η διαδικασία αυτή παράγει μια ακολουθία ένθετων υποσυνόλων ομάδων γονιδίων, με διαδοχικά ισχυρότερη pairwise συσχέτιση και διακύμανση της μεγαλύτερης κύριας συνιστώσας.
Πρότειναν επίσης μια επιβλεπόμενη έκδοση του gene shaving, που χρησιμοποιεί ακριβώς ένα μικτό κριτήριο της μορφής (28). Αν και η μέθοδος αυτή έχει δύο παραμέτρους ρύθμισης, το α και το μέγεθος του υποσυνόλου, εδώ το α καθορίζεται στην ενδιάμεση τιμή 0,5 και η προσοχή εστιάζεται στο μέγεθος του υποσυνόλου. Όπως και στην SPCA, για κάθε υποσύνολο η μεγαλύτερη κύρια συνιστώσα χρησιμοποιείται για να εκφράσει τα γονίδιά του.
Αυτή η μέθοδος είναι παρόμοια με την SPCA. Παράγει κύριες συνιστώσες υποσυνόλων γονιδίων, όπου η επιλογή του υποσυνόλου γίνεται υπό επίβλεψη. Τα πειράματα τα οποία πραγματοποιήθηκαν στην επόμενη ενότητα δείχνουν ότι το gene shaving μπορεί να παρουσιάσει πολύ παρόμοιες επιδόσεις με την SPCA, η τελευταία έχοντας το πλεονεκτήμα ότι είναι απλούστερο να καθοριστεί και έχει μόνο μία ρυθμιστική παράμετρο να επιλεγεί.

2.6.6. Ένα Άλλο Μικτό Κριτήριο

Η μεγαλύτερη κανονικοποιημένη κύρια συνιστώσα u_1 είναι το μεγαλύτερο ιδιοδιάνυσμα του XX^T. Αυτό προκύπτει εύκολα από την SVD (1) και ως εκ τούτου $XX^T = UD^2 U^T$. Διαισθητικά, επειδή

$$u_1^T XX^T u_1 = \sum_{j=1}^{p} \langle u_1, x_j \rangle^2 , \qquad (30)$$

αναζητείται το διάνυσμα u_1 που βρίσκεται πλησιέστερα στο μέσο όρο καθενός από τα x_j. Μια φυσική τροποποίηση με επίβλεψη είναι να διαταραχθεί αυτό το κριτήριο με τρόπο που να ενθαρρύνει το κύριο ιδιοδιάνυσμα να ευθυγραμμιστεί με το y,

$$\max_{u_1, \|u_1\|=1} (1 - a) \sum_{j=1}^{p} \langle u_1, x_j \rangle^2 + a \langle u_1, y \rangle^2. \qquad (31)$$

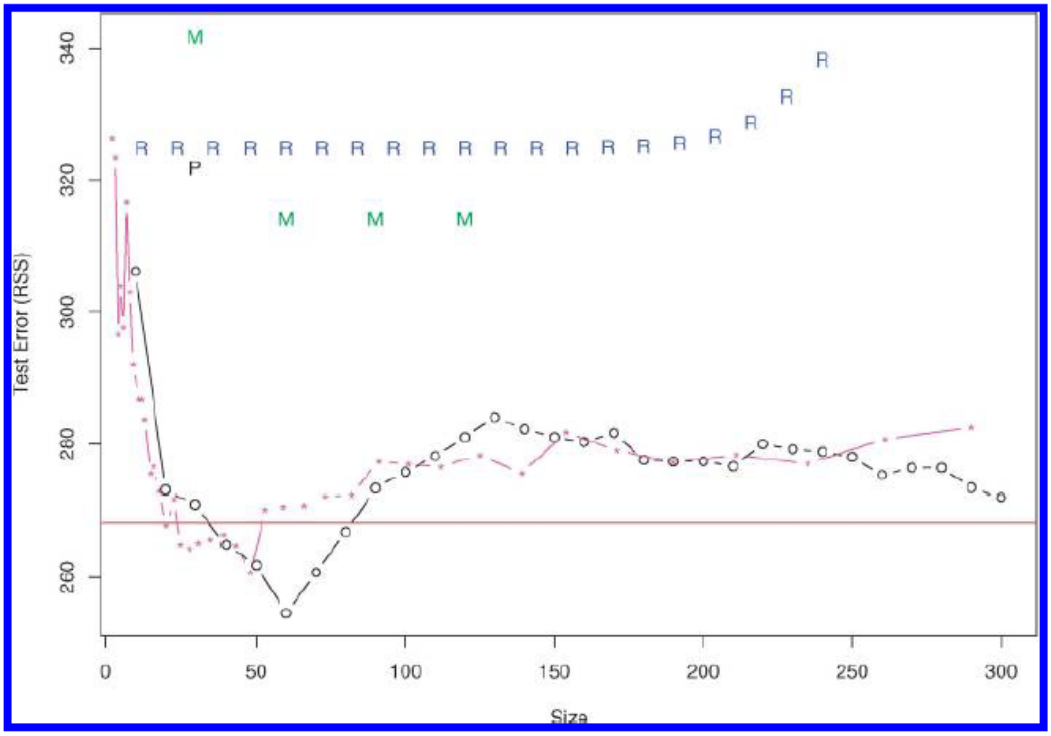
Η επίλυση της (31) ισοδυναμεί με την εύρεση του μεγαλύτερου ιδιοδιανύσματος της

$$C(y ; \alpha) = (1 - \alpha) + XX^T + \alpha yy^T. \qquad (32)$$

Αντίστοιχα, θα μπορούσε κανείς να φτιάξει έναν επαυξημένο πίνακα X_α με το y στην $(p + 1)$-οστή στήλη. Αν δοθούν βάρη α σε αυτή τη σειρά και $(1 - \alpha)$ στις p πρώτες σειρές, τότε η σταθμισμένη SVD του X_α ισοδυναμεί με μια ιδιοαποσύνθεση της (31). Αυτή είναι ακριβώς η κατάσταση που περιγράφεται στο errors-in-variables μοντέλο (11) - (13) στην ενότητα 2.2. Όπως αναφέρεται εκεί, η εκτίμηση u_1 περιλαμβάνει το y καθώς και το x_j, και έτσι δεν μπορεί να χρησιμοποιηθεί απευθείας με τα δεδομένα της εξέτασης.

2.6.7. Συζήτηση των Μεθόδων

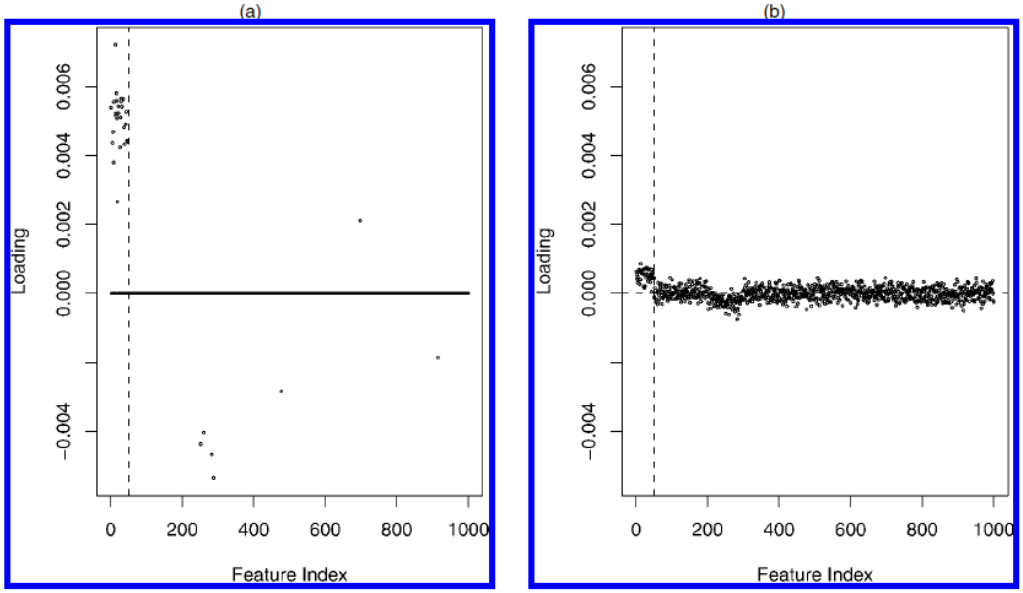
Το Σχήμα 7 δείχνει τις μεθόδους που συζητήθηκαν νωρίτερα σε ένα παράδειγμα προσομοίωσης με $N = 100$ δείγματα και $p = 5.000$ χαρακτηριστικά. Τα δεδομένα που παράγονται σύμφωνα με το μοντέλο κρυφής μεταβλητής (35), όπου υπάρχουν τέσσερις κυρίαρχες κύριες συνιστώσες, και η μία που σχετίζεται με την απόκριση κατατάσσεται με τον αριθμό 3 (όταν εκτιμάται από τα δεδομένα). Οι μέθοδοι προσδιορίζονται στη λεζάντα του σχήματος. Το σημείο *M* αντιστοιχεί σε παλινδρόμηση κυρίων συνιστωσών χρησιμοποιώντας τη μεγαλύτερη κύρια συνιστώσα. Η SPCA και το "shaving" αποδίδουν πολύ καλύτερα από ό,τι οι άλλες μέθοδοι.



Σχήμα 7. Ένα Παράδειγμα Προσομοίωσης που Αποδεικνύει την Τυπική Συμπεριφορά των Διαφορετικών Μεθόδων. Τα δεδομένα παράγονται σύμφωνα με το μοντέλο (35) που περιγράφεται στην επόμενη ενότητα, με N = 100 και p = 5.000. Η ridge παλινδρόμηση, τα PLS, και το μικτό κριτήριο όλα υποφέρουν από τις πολύ υψηλές διαστάσεις. Αν και δεν φαίνεται, η παράμετρος κανονικοποίησης λ για τα ridge σημεία αυξάνεται προς τα δεξιά, όπως και το α για το μικτό κριτήριο, η leftmost τιμή να είναι 0. Τόσο το shaving όσο και η SPCA καταχωρούνται ανάλογα με το μέγεθος του υποσυνόλου. Η γραμμή με την ένδειξη "αλήθεια" χρησιμοποιεί το γνωστό γραμμικό συνδυασμό 50 χαρακτηριστικών όπως η μεταβλητή πρόβλεψης της παλινδρόμησης.

Στο Σχήμα 8 δίνεται μια ιδέα για το τι συμβαίνει. Βλέπουμε τα πρώτα 1.000 από τα 5.000 loadings χαρακτηριστικών για δύο από τις μεθόδους όπως παρουσιάζονται στο Σχήμα 7 (επιλέγονται στα καλύτερα σημεία λύσης). Και οι δύο μέθοδοι προσδιόρισαν

σωστά τη σημαντική συνιστώσα (αυτή που σχετίζεται με το y και αφορά τα πρώτα 50 χαρακτηριστικά). Σε μια κανονική SVD του X, η σημαντική αυτή συνιστώσα κυριαρχήθηκε από δύο άλλες συνιστώσες. Πιο αναλυτικά, τα δεδομένα εκπαίδευσης από το μοντέλο (35) έχουν τέσσερις ενσωματωμένες συνιστώσες, με μοναδικές τιμές 99,9, 88,3, 80,9, και 80,5. Εμπειρικά, διαπιστώθηκε ότι η συνιστώσα τρία ταυτίζεται με το μηχανισμό αντίδρασης, αλλά η μοναδική της τιμή είναι ακριβώς πάνω από το επίπεδο του θορύβου (η πέμπτη μοναδική τιμή ήταν 79,2). Ωστόσο, το μικτό κριτήριο φέρνει επίσης θορυβώδους συντελεστές, κάπως μικρότερους για όλες τις άλλες μεταβλητές, ενώ αντίθετα η SPCA θέτει τα περισσότερα από τα άλλα loadings στο 0. Οι συντελεστές για το shaving δείχνουν ένα πολύ παρόμοιο σχέδιο με την SPCA, ενώ εκείνοι για την ridge παλινδρόμηση και τα PLS είναι πολύ παρόμοιοι με το μικτό κριτήριο και δεν φαίνονται εδώ.



Σχήμα 8. Loadings Χαρακτηριστικών ***w*** *για την SPCA (a) και το Μικτό Κριτήριο (28) (b). Τα πρώτα 1.000 από 5.000 παρουσιάζονται, στο «καλύτερο» σημείο λύσης. Η κάθετη γραμμή δείχνει ότι τα πρώτα 50 μεταβλητές παρήξαν την απόκριση. Ενώ και οι δύο αυτές μέθοδοι ήταν σε θέση να ξεπεράσουν τις δύο πρώτες κυρίαρχες κύριες συνιστώσες (που δεν είχαν σχέση με το* ***y****), η SPCA είναι σε θέση να αγνοήσει τις περισσότερες από τις μεταβλητές, και το μικτό κριτήριο δίνει σε όλες βάρος (αν και περισσότερο βάρος στις πρώτες 50).*

Η μέθοδος του shaving παίρνει κατά καιρούς την τελείως λάθος συνιστώσα. Η SPCA τείνει να είναι πιο αξιόπιστη και είναι απλούστερο να καθορίσει, και ως εκ τούτου είναι η μέθοδος που επιλέγεται. Οι προσομοιώσεις στην επόμενη ενότητα υποστηρίζουν αυτή την επιλογή.

2.7. Μελέτες Προσομοίωσης

Πραγματοποιήθηκαν τρεις μελέτες προσομοίωσης για να συγκριθεί η απόδοση των μεθόδων που εξετάζονται. Παρακάτω περιγράφονται οι δύο πρώτες μελέτες , και η τρίτη αργότερα. Κάθε προσομοιωμένο σύνολο δεδομένων X αποτελείται από 5.000 "γονίδια" (γραμμές) και 100 "ασθενείς" (στήλες). Έστω ότι το x_{ij} χαρακτηρίζει το "επίπεδο έκφρασης" του i-οστού γονιδίου και του j-οστού ασθενή. Στην πρώτη μελέτη παράχθηκαν τα δεδομένα ως εξής:

$$x_{ij} = \begin{cases} 3 + \varepsilon_{ij} & \text{αν } i \leq 50,\ j \leq 50 \\ 4 + \varepsilon_{ij} & \text{αν } i \leq 50,\ j > 50 \\ 3.5 + \varepsilon_{ij} & \text{αν } i > 51 \end{cases}, \qquad (33)$$

όπου τα ε_{ij} είναι ανεξάρτητες κανονικές τυχαίες μεταβλητές με μέση τιμή 0 και διασπορά 1. Επίσης, έστω

$$y_j = \frac{\sum_{i=1}^{50} x_{ij}}{25} + \varepsilon_j, \qquad (34)$$

όπου τα ε_j είναι ανεξάρτητες κανονικές τυχαίες μεταβλητές με μέση τιμή 0 και τυπική απόκλιση 1,5.

Η προσομοίωση αυτή σχεδιάστηκε ώστε να υπάρχουν δύο "υποκατηγορίες" όγκου. Οι ασθενείς 1-50 ανήκουν στην κατηγορία όγκου 1, και έχουν ελαφρώς χαμηλότερα μέσα επίπεδα έκφρασης απ' ό,τι οι ασθενείς με όγκο κατηγορίας 2. Επιπλέον, επειδή το y είναι ανάλογο με το άθροισμα του επιπέδου έκφρασης των πρώτων 50 γονιδίων, το y είναι ελαφρώς χαμηλότερο για τους ασθενείς με όγκο κατηγορίας 1. Τα υπόλοιπα 4.950 γονίδια δεν έχουν σχέση με το y.
Εφαρμόστηκαν οκτώ μεθόδοι σε αυτό το προσομοιωμένο σύνολο δεδομένων: παλινδρόμηση κυρίων συνιστωσών, παλινδρόμηση κυρίων συνιστωσών χρησιμοποιώντας μόνο την πρώτη κύρια συνιστώσα, PLS (μία κατεύθυνση), ridge παλινδρόμηση, lasso, κύριες συνιστώσες με επίβλεψη, μικτή διακύμανση-συνδιακύμανση, και gene shaving. Εκπαιδεύτηκε καθένα από αυτά τα μοντέλα χρησιμοποιώντας ένα προσομοιωμένο σύνολο δεδομένων που παράγεται όπως περιγράφηκε νωρίτερα. Επιλέχθηκε η βέλτιστη τιμή των παραμέτρων ρύθμισης για κάθε μέθοδο χρησιμοποιώντας 10-fold διασταυρωμένη επικύρωση. Στη συνέχεια, χρησιμοποιήθηκε η ίδια διαδικασία για να δημιουργήθει ένα ανεξάρτητο σύνολο δεδομένων εξέτασης και χρησιμοποιήθηκαν τα μοντέλα που κατασκευάστηκαν για την πρόβλεψη του y στο σύνολο δεδομένων εξέτασης.
Η διαδικασία αυτή επαναλήφθηκε 10 φορές και προέκυψε ο μέσος όρος των

αποτελεσμάτων. Ο Πίνακας 2 δίνει τα σφάλματα που παράγονται από το κάθε μοντέλο.

Πίνακας 2. Αποτελέσματα της Μελέτης Προσομοίωσης βασισμένα στα "Εύκολα" Προσομοιωμένα Δεδομένα

Μέθοδος	**CV error**	**Test error**
PCR	$293.4_{(17.21)}$	$217.6_{(10.87)}$
PCR-1	$316.8_{(20.52)}$	$239.4_{(11.94)}$
PLS	$291.6_{(13.11)}$	$218.2_{(12.03)}$
Ridge παλινδρόμηση	$298.0_{(14.72)}$	$224.2_{(12.35)}$
Lasso	$264.0_{(13.06)}$	$221.9_{(12.72)}$
Κύριες συνιστώσες με επίβλεψη	$233.2_{(11.23)}$	$176.4_{(10.14)}$
Μικτή διακύμανση-συνδιακύμανση	$316.7_{(19.52)}$	$238.7_{(10.24)}$
Gene shaving	$223.0_{(8.48)}$	$172.5_{(9.25)}$

ΣΗΜΕΙΩΣΗ: Κάθε εγγραφή στον πίνακα αντιπροσωπεύει το τετραγωνικό σφάλμα των προβλέψεων του συνόλου εξέτασης κατά μέσο όρο πάνω από 10 προσομοιώσεων. Το τυπικό σφάλμα της κάθε εκτίμησης σφάλματος είναι σε παρένθεση. Οι μέθοδοι πρόβλεψης είναι: παλινδρόμηση κυρίων συνιστωσών (PCR), PCR που περιορίζεται στη χρήση μόνο μίας κύριας συνιστώσας (PCR-1), μερικά ελάχιστα τετράγωνα (PLS), ridge παλινδρόμηση, lasso, κύριες συνιστώσες με επίβλεψη, μικτή διακύμανση-συνδιακύμανση, και gene shaving.

Βλέπουμε ότι το gene shaving και οι κύριες συνιστώσες με επίβλεψη γενικά παράγουν μικρότερη διασταυρωμένη επικύρωση και δοκιμαστικά σφάλματα από οποιεσδήποτε από τις άλλες μεθόδους, με το πρώτο να έχει ένα μικρό πλεονέκτημα. Η παλινδρόμηση κυρίων συνιστωσών και τα PLS έδωσαν συγκρίσιμα αποτελέσματα (αν και η παλινδρόμηση κυρίων συνιστωσών απέδωσε ελαφρώς χειρότερα όταν περιορίστηκε σε μία συνιστώσα).
Στη συνέχεια, δημιουργήθηκε ένα νέο προσομοιωμένο σύνολο δεδομένων με τα παρακάτω χαρακτηριστικά.

$$x_{ij} = \begin{cases} 3 + \varepsilon_{ij} & \text{αν } i \leq 50,\ j \leq 50 \\ 4 + \varepsilon_{ij} & \text{αν } i \leq 50,\ j > 50 \\ 3.5 + 1.5\, I\left(u_{1j} < 0.4\right) + \varepsilon_{ij} & \text{αν } 51 \leq i \leq 100 \\ 3.5 + 0.5\, I\left(u_{2j} < 0.7\right) + \varepsilon_{ij} & \text{αν } 101 \leq i \leq 200 \\ 3.5 - 1.5\, I\left(u_{3j} < 0.3\right) + \varepsilon_{ij} & \text{αν } 201 \leq i \leq 300 \\ 3.5 + \varepsilon_{ij} & \text{αν } i > 301 \end{cases} \qquad (35)$$

Εδώ οι u_{ij} είναι ομοιόμορφες τυχαίες μεταβλητές στο (0,1) και το $I(x)$ είναι μια δείκτρια συνάρτηση. Για παράδειγμα, για καθένα από τα γονίδια 51-100, μια

μοναδική τιμή u_{1j} δημιουργείται για το δείγμα j. Αν αυτή η τιμή είναι μεγαλύτερη από 0,4, τότε όλα τα γονίδια σε αυτό το μπλοκ αυξάνονται κατά 1,5. Το κίνητρο για αυτή την προσομοίωση είναι ότι υπάρχουν και άλλες ομάδες γονιδίων με παρόμοιες μορφές έκφρασης που δεν έχουν σχέση με το y. Αυτό είναι πιθανό να συμβεί σε πραγματικά δεδομένα μικροσυστοιχιών, επειδή υπάρχουν μονοπάτια γονιδίων (που έχουν πιθανώς παρόμοιες μορφές έκφρασης) που δεν σχετίζονται με το y. Τα Σχήματα 7 και 8 απεικονίζουν μερικές από τις μεθόδους που εφαρμόζονται για την κατανόηση αυτού του μοντέλου.
Το πείραμα που περιγράφηκε νωρίτερα επαναλήφθηκε χρησιμοποιώντας την (35) για να δημιουργήθουν τα σύνολα δεδομένων αντί της (33). Τα αποτελέσματα δίνονται στον Πίνακα 3. Οι περισσότερες από τις μεθόδους είχαν χειρότερες επιδόσεις σε αυτό το "δύσκολο" πείραμα. Για άλλη μια φορά, το gene shaving και οι κύριες συνιστώσες με επίβλεψη παρήξαν μικρότερα σφάλματα από οποιεσδήποτε από τις ανταγωνιστικές μεθόδους. Το gene shaving δείχνει πολύ μεγαλύτερη μεταβλητότητα από ό,τι οι κύριες συνιστώσες με επίβλεψη σε αυτή την περίπτωση.

Πίνακας 3. Αποτελέσματα της Μελέτης Προσομοίωσης βασισμένα στα "Δύσκολα" Προσομοιωμένα Δεδομένα

Μέθοδος	**CV error**	**Test error**
PCR	$302.4_{(17.48)}$	$327.6_{(14.49)}$
PCR-1	$325.6_{(20.05)}$	$354.6_{(14.99)}$
PLS	$299.6_{(17.10)}$	$321.8_{(16.12)}$
Ridge παλινδρόμηση	$301.0_{(18.47)}$	$328.0_{(16.38)}$
Lasso	$286.9_{(16.92)}$	$322.8_{(21.24)}$
Κύριες συνιστώσες με επίβλεψη	$242.3_{(15.38)}$	$268.9_{(10.47)}$
Μικτή διακύμανση-συνδιακύμανση	$322.5_{(19.64)}$	$349.8_{(16.02)}$
Gene shaving	$234.0_{(12.46)}$	$276.6_{(13.43)}$

ΣΗΜΕΙΩΣΗ: Κάθε εγγραφή στον πίνακα αντιπροσωπεύει το τετραγωνικό σφάλμα των προβλέψεων του συνόλου εξέτασης κατά μέσο όρο πάνω από 10 προσομοιώσεων. Το τυπικό σφάλμα της κάθε εκτίμησης σφάλματος είναι σε παρένθεση. Οι μέθοδοι πρόβλεψης είναι οι ίδιες όπως στον Πίνακα 2.

Μια τρίτη μελέτη προσομοίωσης ήταν αρκετά διαφορετική από τις πρώτες δύο. Χρησιμοποιήθηκαν σύνολα δεδομένων εκπαίδευσης και εξέτασης από τους Rosenwald et al. (2002), έτσι ώστε να αποκτήθουν γονίδια με "πραγματικής ζωής" συσχέτιση. Με τον καθορισμό των δεδομένων έκφρασης, δημιουργήθηκαν ανεξάρτητοι τυποποιημένοι γκαουσιανοί συντελεστές θ_j, και τελικά ένα ποσοτικό αποτέλεσμα $y_i = \sum_{j=1}^{p} x_{ij}\theta_j + \sigma Z$, με Z την τυποποιημένη γκαουσιανή. Με σ = 3, περίπου το 30% της διακύμανσης στο αποτέλεσμα εξηγήθηκε από το αληθινό μοντέλο. Πολλαπλά σύνολα δεδομένων παράχθηκαν με αυτό τον τρόπο, με τα δεδομένα έκφρασης να μένουν σταθερά.
Η ridge παλινδρόμηση είναι η εκτίμηση του Bayes σε αυτό το σκηνικό, οπότε θα

περίμενε κανείς να έχει τις καλύτερες επιδόσεις. Στον Πίνακα 4 παρουσιάζονται τα αποτελέσματα. Η ridge παλινδρόμηση είναι η καλύτερη, ακολουθούμενη στο σφάλμα διασταυρωμένης επικύρωσης από τα PLS και στα σφάλματα εξέτασης από το lasso. Οι άλλες μέθοδοι είναι σημαντικά χειρότερες. Στον πίνακα 5 παρουσιάζεται ο μέσος αριθμός των γονιδίων που χρησιμοποιούνται από τις κύριες συνιστώσες με επίβλεψη και το lasso στις τρεις μελέτες προσομοίωσης. Το lasso χρησιμοποιεί λιγότερα γονίδια από ό,τι οι κύριες συνιστώσες με επίβλεψη σε κάθε περίπτωση. Ωστόσο, στις δύο πρώτες μελέτες προσομοίωσης, ο αριθμός που έχει επιλεγεί από τις κύριες συνιστώσες με επίβλεψη είναι πιο κοντά στον πραγματικό αριθμό (50). Επιπλέον, αν υπάρχουν N δείγματα και το N είναι μικρότερο από το συνολικό αριθμό των χαρακτηριστικών p, τότε το lasso δεν μπορεί ποτέ να επιλέξει περισσότερα από N χαρακτηριστικά. Αυτό θα μπορούσε να είναι πολύ περιοριστικό, γιατί δεν υπάρχει λόγος σε γενικές γραμμές για να πρέπει ο πραγματικός αριθμός των σημαντικών γονιδίων να είναι μικρότερος από N.

Πίνακας 4. Τρίτη Μελέτη Προσομοίωσης: Γκαουσιανή prior για Αληθινούς Συντελεστές

Μέθοδος	**CV error/ 1000**	**Test error/ 1000**
PCR	$399.423_{(16.617)}$	$194.489_{(16.298)}$
PCR-1	$559.708_{(29.637)}$	$283.356_{(24.320)}$
PLS	$322.513_{(11.142)}$	$203.375_{(16.978)}$
Ridge παλινδρόμηση	$304.215_{(9.858)}$	$132.251_{(55.45)}$
Lasso	$356.886_{(15.281)}$	$169.266_{(10.217)}$
Κύριες συνιστώσες με επίβλεψη	$417.972_{(16.485)}$	$203.374_{(16.978)}$
Μικτή συνδιακύμανση (y)	$418.250_{(10.975)}$	$202.293_{(16.805)}$
Μικτή συνδιακύμανση (ŷ)	$551.924_{26.251}$	$286.255_{(23.149)}$
Gene shaving	$402.876_{(11.897)}$	$197.000_{(17.040)}$

Πίνακας 5. Μέσος Αριθμός Γονιδίων (και τυπική απόκλιση) για Επιβλέπουσες Κύριες Συνιστώσες και Lasso σε Καθεμία από τις 3 Μελέτες Επιβίωσης

Μέθοδος	**Προσομοίωση 1**	**Προσομοίωση 2**	**Προσομοίωση 3**
Κύριες συνιστώσες με επίβλεψη	$44.5_{(9.4)}$	$54.4_{(10.9)}$	$95.7_{(16.4)}$
Lasso	$32.8_{(8.7)}$	$23.1_{(6.2)}$	$42.9_{(5.5)}$

2.8. Εφαρμογή σε Διάφορες Μελέτες Επιβίωσης

Εδώ συγκρίνονται διάφορες μεθόδοι για την άσκηση ανάλυσης επιβίωσης σε πραγματικά σύνολα δεδομένων DNA μικροσυστοιχιών. (Μερικά από αυτά τα αποτελέσματα αναφέρθηκαν επίσης από τους Bair και Tibshirani το 2004). Εφαρμόστηκαν οι μεθόδοι για τέσσερα διαφορετικά σύνολα δεδομένων. Πρώτα, εξετάστηκε ένα σύνολο δεδομένων μικροσυστοιχιών που αποτελείται από ασθενείς με διάχυτο λέμφωμα μεγάλων Β-κυττάρων (Rosenwald et al. 2002). Υπάρχουν 7.399 γονίδια, 160 εκπαιδευόμενοι ασθενείς, και 80 ασθενείς υπό εξέταση σε αυτό το σύνολο δεδομένων. Έπειτα, θεωρήθηκε ένα σύνολο δεδομένων καρκίνου του μαστού με 4.751 γονίδια και 97 ασθενείς. Χωρίστηκε αυτό το σύνολο δεδομένων σε ένα σύνολο εκπαίδευσης 44 ασθενών και σε ένα σύνολο εξέτασης 53 ασθενών.
Στη συνέχεια, εξετάστηκε ένα σύνολο δεδομένων καρκίνου του πνεύμονα με 7.129 γονίδια και 86 ασθενείς, το οποίο χωρίστηκε σε ένα σύνολο εκπαίδευσης 43 ασθενών και σε ένα σύνολο εξέτασης 43 ασθενών. Τέλος, θεωρήθηκε ένα σύνολο δεδομένων από ασθενείς με οξεία μυελογενή λευχαιμία, που αποτελείται από 6.283 γονίδια και 116 ασθενείς. Αυτό το σύνολο δεδομένων χωρίστηκε σε ένα σύνολο εκπαίδευσης 59 ασθενών και σε ένα σύνολο εξέτασης 53 ασθενών.
Εκτός από τις κύριες συνιστώσες με επίβλεψη, εξετάστηκαν οι ακόλουθες μέθοδοι: παλινδρόμηση κυρίων συνιστωσών, μερικά ελάχιστα τετράγωνα, lasso, και δύο άλλες μέθοδοι που ονομάζουμε "median cut" και "ομαδοποίηση του Cox", τα οποία περιγράφονται από τους Bair και Tibshirani (2004). Και οι δύο αυτές τελευταίες μέθοδοι μετατρέπουν το πρόβλημα σε ένα δύο τάξεων πρόβλημα ταξινόμησης και στη συνέχεια εφαρμόζουν τον πλησιέστερο συρρικνωμένο ταξινομητή κέντρου βάρους των Tibshirani, Hastie, Narasimhan, και Chu (2001). Η μέθοδος median cut χωρίζει τους ασθενείς σε στρώματα υψηλού ή χαμηλού κινδύνου, ανάλογα με το αν θα επιβιώσουν μετά το μέσο χρόνο επιβίωσης. Η μέθοδος της "ομαδοποίησης του Cox" είναι σαν τις κύριες συνιστώσες με επίβλεψη και χρησιμοποιεί ομαδοποίηση δύο μέσων που εφαρμόζεται στα γονίδια με τα υψηλότερα αποτελέσματα Cox.
Για τα PLS, την ridge παλινδρόμηση, και το lasso, κατέστη δυνατή η χρησιμοποίηση περισσότερων από μίας συνιστώσας, και ο αριθμός αυτός επιλέχθηκε με διασταυρωμένη επικύρωση. Τα αποτελέσματα δίνονται στον Πίνακα 6. Συνολικά, οι κύριες συνιστώσες με επίβλεψη αποδίδουν καλύτερα από ό,τι οι ανταγωνιστικές μέθοδοι. Ωστόσο, στο παράδειγμα DLBCL, το lasso αποδίδει καλύτερα. Αυτό δεν αποτελεί έκπληξη, επειδή η χρήση του lasso ως μετα-επεξεργαστή για κύριες συνιστώσες με επίβλεψη έδειξε ότι μόνο λίγα γονίδια απαιτούνται για καλή πρόβλεψη σε αυτό το παράδειγμα.

Πίνακας 6. Σύγκριση των Διαφορετικών Μεθόδων σε 4 Διαφορετικά Σύνολα Δεδομένων από Μελέτες για Καρκίνο

		(a)DLBCL	
Μέθοδος	R^2	*p value*	*NC*
(1) Median cut	0.05	0.047	
(2) Ομαδοποίηση του Cox	0.08	0.006	
(3) SPCA	0.11	0.003	2
(4) Παλινδρόμηση κυρίων συνιστωσών	0.01	0.024	2
(5) PLS	0.10	0.004	3
(6) Lasso	0.16	0.0002	NA

		(b)Καρκίνος του μαστού	
Μέθοδος	R^2	*p value*	*NC*
(1) Median cut	0.13	0.0042	
(2) Ομαδοποίηση του Cox	0.21	0.0001	
(3) SPCA	0.27	2.1×10^{-5}	1
(4) Παλινδρόμηση κυρίων συνιστωσών	0.22	0.0003	3
(5) PLS	0.18	0.0003	1
(6) Lasso	0.14	0.001	NA

		(c)Καρκίνος του πνεύμονα	
Μέθοδος	R^2	*p value*	*NC*
(1) Median cut	0.15	0.0016	
(2) Ομαδοποίηση του Cox	0.07	0.0499	
(3) SPCA	0.36	1.5×10^{-7}	3
(4) Παλινδρόμηση κυρίων συνιστωσών	0.11	0.0156	1
(5) PLS	0.18	0.0044	1
(6) Lasso	0.26	0.0001	NA

		(d)AML	
Μέθοδος	R^2	*p value*	*NC*
(1) Median cut	0.07	0.0487	
(2) Ομαδοποίηση του Cox	0.08	0.0309	
(3) SPCA	0.16	0.0013	3
(4) Παλινδρόμηση κυρίων συνιστωσών	0.08	0.0376	1
(5) PLS	0.07	0.0489	1
(6) Lasso	0.05	0.0899	NA

ΣΗΜΕΙΩΣΗ: Οι μέθοδοι είναι (1) ορισμός δειγμάτων σε ομάδα "χαμηλού κινδύνου" ή "υψηλού κινδύνου" με βάση το μέσο χρόνο επιβίωσής τους, (2) χρήση ομαδοποίησης δύο μέσων με βάση τα γονίδια με τα μεγαλύτερα αποτελέσματα Cox, (3) μέθοδος κυρίων συνιστωσών με επίβλεψη, (4) παλινδρόμηση κυρίων συνιστωσών, (5) παλινδρόμηση μερικών ελαχίστων τετραγώνων, και (6) lasso. Ο πίνακας παραθέτει το R^2 (ποσοστό της λογαριθμοπιθανοφάνειας που εξηγείται), τις p τιμές για τις προβλέψεις του συνόλου εξέτασης, καθώς και τον αριθμό των συνιστωσών που χρησιμοποιούνται.

2.9. Θεωρητικά Αποτελέσματα

Στην ενότητα αυτή δίνουμε στοιχεία για την ανάλυση κυρίων συνιστωσών με επίβλεψη στην γκαουσιανή παλινδρόμηση. Αποτελέσματα συνέπειας για δεδομένα επιβίωσης συζητούνται στην απόδειξη του Θεωρήματος 2.

2.9.1. Ανάλυση κυρίων συνιστωσών με επίβλεψη στην γκαουσιανή παλινδρόμηση

Έστω ότι οι γραμμές του X είναι ανεξάρτητες και ίδια κατανεμημένες. Στη συνέχεια, μπορεί να διατυπωθεί ένα μοντέλο πληθυσμού ως εξής. Δηλώνοντας τις γραμμές με X_i^T (i = 1, . . . , N), προκύπτει το μοντέλο

$$X_i \sim \mathrm{N}_p(\mu, \Sigma),$$

όπου Σ ($p \times p$) είναι ο πίνακας συνδιακύμανσης. Χωρίς βλάβη της γενικότητας, υποθέτουμε ότι μ = 0, επειδή μπορεί να εκτιμηθεί με αρκετή ακρίβεια από τα δεδομένα.
Έστω ότι ο X είναι χωρισμένος ως X = (X_1, X_2), όπου ο X_1 είναι $N \times p_1$ και ο X_2 είναι $N \times p_2$ με $p_1 + p_2 = p$. Έστω, επίσης, ότι η αντίστοιχη διαμέριση του Σ δίνεται από τον τύπο

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}. \qquad (36)$$

Επιπλέον, θεωρείται ότι ο Σ_1 ($p_1 \times p_1$) μπορεί να αναπαρασταθεί ως

$$\Sigma_1 = \sum_{k=1}^{M} \lambda_k \theta_k \theta_k^T + \sigma^2 \mathrm{I}, \qquad (37)$$

όπου θ_k (k = 1, ... , M) είναι αμοιβαία ορθοκανονικά ιδιοδιανύσματα και οι ιδιοτιμές $\lambda_1 \geq \cdots \geq \lambda_M > 0$. Εδώ το $\sigma^2 > 0$ αντιπροσωπεύει τη συμβολή του (ισότροπου)

"background θορύβου" που είναι άσχετος με τις αλληλεπιδράσεις μεταξύ των γονιδίων. Αυτό το μοντέλο μπορεί να περιγραφεί ως ένα μοντέλο συνδιακύμανσης για εκφράσεις γονιδίου που είναι μια *M*-τάξης διατάραξη της ταυτότητας.
Εδώ $1 \leq M \leq p_1 - 1$.
Αντίστοιχα, μπορούν να εκφραστούν οι μεταβλητές πρόβλεψης μέσω του ακόλουθου μοντέλου παραγοντικής ανάλυσης. Έστω *P* το σύνολο των γονιδίων που αποτελούν τις στήλες του πίνακα X_1. Τότε, $|P| = p_1$ και οι

$$X_{ij} = \sum_{k=1}^{M} \sqrt{\lambda_k}\, \theta_{jk}\eta_{ik} + \sigma w_{ij}, \qquad j \in P, \qquad (38)$$

αντιπροσωπεύουν τις μετρήσεις έκφρασης για τα γονίδια στο σύνολο *P* της *i*-οστής σειράς (επαναληπτική), $i = 1,\ldots, N$ [βλ. (10) στην ενότητα 2.2]. Εδώ τα $\eta_{ik} \sim$ N(0,1) και είναι ανεξάρτητα από τα $w_{ij} \sim$ N(0,1).
Η κύρια υπόθεση είναι ότι ο X_1 είναι ο πίνακας που περιέχει όλες τις στήλες των οποίων οι μεταβολές σχετίζονται με τις μεταβολές στο y. Πρώτον, έστω ότι η διαδικασία επιλογής είναι τέτοια ώστε να επιλέγει το X_1 με πιθανότητα που τείνει στο 1 όταν $N \to \infty$. Στην ενότητα 10.4 λαμβάνεται το πιο ρεαλιστικό σενάριο στο οποίο εκτιμάται ο υπόχωρος αυτός από δεδομένα. Οι βασικές παραδοχές που αφορούν τον πίνακα Σ_1 δίνονται από τους όρους A1-A2 ή, πιο γενικά, από τους όρους A1΄ και A2΄. Στην ενότητα 10.2 αποδεικνύεται ότι οι προϋποθέσεις αυτές είναι επαρκείς για τη συνέπεια των συνήθων/κανονικών, βασισμένων στην PCA εκτιμητών του θ_k και λ_k, $k = 1,\ldots, M$, όταν εκτελείται μια τέτοια PCA στο δειγματικό πίνακα συνδιακύμανσης του X_1. Από αυτό προκύπτει ότι μπορούν να εκτιμηθούν με συνέπεια οι παράμετροι στο μοντέλο παλινδρόμησης κυρίων συνιστωσών για την απόκριση y που περιγράφεται μέσω της (40), (βλ. ενότητα 10.2 για λεπτομέρειες).

A1. Οι "σημαντικές" ιδιοτιμές του Σ_1 ικανοποιούν (προϋπόθεση αναγνωρισιμότητας για ιδιοδιανύσματα) την

$$\lambda_1 > \cdots > \lambda_M > 0,$$

και το *M* είναι ένας σταθερός θετικός ακέραιος.
A2. $p_1 \to \infty$ όσο το *N* αυξάνεται προς το άπειρο με τέτοιο τρόπο ώστε $p_1 / N \to 0$.

Είναι πιθανό ότι η διακύμανση του θορύβου σ^2 και οι "σημαντικές" ιδιοτιμές λ_k επίσης ποικίλλουν ανάλογα με το *N*. Με αυτή τη ρύθμιση, για να εξασφαλιστεί η συνέπεια, θα πρέπει να αντικαταστηθούν οι όροι A1 και A2 με τους παρακάτω:

A1΄. Οι ιδιοτιμές είναι τέτοιες ώστε $\lambda_k / \lambda_1 \to \rho_k$ για $k = 1,\ldots, M$ με $1 = \rho_1 > \rho_2 > \cdots > \rho_M > 0$ και $\lambda_1 \to c > 0$ όσο $N \to \infty$. Επιπλέον, $\sigma^2 \to \sigma_0^2 \in [0, \infty)$ όσο $N \to \infty$.

A2′. Το p_1 ποικίλλει ανάλογα με το N με τέτοιο τρόπο ώστε $\sigma^2 p_1 / (N\lambda_1) \to 0$ καθώς $N \to \infty$.

Ο όρος A1′ είναι μια ασυμπτωτική προυπόθεση αναγνωρισιμότητας για τα ιδιοδιανύσματα $\theta_1, \ldots, \theta_M$. Αυτό ισχύει επειδή αν $\rho_k = \rho_{k+1}$ για κάποια $k = 1, \ldots, M-1$, τότε για μεγάλα N, και για κάθε 2×2 ορθογώνιο πίνακα C, οι στήλες του πίνακα $C[\theta_k : \theta_{k+1}]$ είναι περίπου τα ιδιοδιανύσματα του Σ_1 που αντιστοιχούν στις ιδιοτιμές λ_k και λ_{k+1}. Αυτό θα υπονοούσε ένα πολύ ιδιαίτερο είδος ασυνέπειας στις εκτιμήσεις των θ_k και θ_{k+1}, ακόμα κι αν είναι δυνατό να εκτιμηθεί ο αντίστοιχος ιδιοχώρος συνεπώς. Για να αποφευχθούν οι τεχνικές λεπτομέρειες που σχετίζονται με αυτή την κατάσταση, θα γίνει περιορισμός στον όρο A1′. Η προυπόθεση $\lambda_1 \to c > 0$, σε συνδυασμό με το πρώτο μέρος του όρου A1′, σημαίνει ότι όλες οι M ιδιοτιμές λ_k συγκλίνουν σε θετικά όρια.

Παρατήρηση 1. Οι όροι A1′ και A2′ παρέχουν τη δυνατότητα $\lambda_1/\sigma^2 \to \infty$ και p_1 / N να συγκλίνει σε ένα θετικό όριο. Λαμβάνεται το μοντέλο (38). Έστω ότι $M = 1$. Στην περίπτωση αυτή, αν $\sqrt{\lambda_1}\theta_{j1}$ είναι περίπου της ίδιας τάξης μεγέθους για κάθε $j \in P$, τότε $\lambda_1 \sim p_1$ για μεγάλα p_1.
Ακόμη και αν είναι διαφορετικά, είναι λογικό να υπάρχει η πεποίθηση ότι το "signal-to-noise ratio" λ_1/σ^2 πηγαίνει στο ∞ όταν $p_1 \to \infty$, επειδή η παρουσία μεγαλύτερου αριθμού γονιδίων που συνδέονται με ένα κοινό κρυφό παράγοντα δίνει μεγαλύτερη ποσότητα πληροφοριών.

Έστω ότι η SVD του X_1 δίνεται από την

$$X_1 = UDV^T, \quad \text{όπου}$$

$$U \text{ είναι } N \times M, D \text{ είναι } m \times m \text{ και } V \text{ είναι } p_1 \times m, \qquad (39)$$

$$\text{με } m = \min(N, p_1).$$

Εδώ N είναι ο αριθμός των παρατηρήσεων (ασθενείς) και p_1 είναι η διάσταση (αριθμός γονιδίων). Έστω $u_1, \ldots, u_m$ οι στήλες του U και έστω $v_1, \ldots, v_m$ οι στήλες του V. Για προφανείς λόγους, $\hat{\theta}_k = v_k$, $k = 1, \ldots, M$. Επίσης, ορίζονται τα διαγώνια στοιχεία του D με $d_1 > \cdots > d_m$.
Το μοντέλο για την απόκριση είναι

$$y = \beta_0 \frac{1}{\sqrt{N}} 1 + \sum_{k=1}^{K} \beta_k \frac{1}{\sqrt{N}} \eta_k + Z, \qquad (40)$$

όπου $K \leq M$, 1 είναι το διάνυσμα με 1 σε κάθε συντεταγμένη, και $Z \sim N_N(0, \frac{\tau 2}{N} I)$ είναι ανεξαρτήτο του X για κάποια $\tau \in [0, \infty)$.
Μπορεί να φαίνεται από την (40) ότι οι παράμετροι που σχετίζονται με την κατανομή

του μεγέθους του δείγματος εξαρτώνται από το μέγεθος του δείγματος *N*, αλλά στην πραγματικότητα το μοντέλο (40) είναι ένα ακριβές ανάλογο του μοντέλου για την απάντηση που δόθηκε από τις (9) και (10). Αυτό μπορεί να γίνει αντιληπτό διαιρώντας την (9) με $\sqrt{N}$, παίρνοντας $\alpha_{1jm} = \theta_{jm}$στην (10) και θέτοντας $U_m = \eta_m$ για $m = 1, \ldots, M$.

Παρατήρηση 2. Το μοντέλο θα μπορούσε να είχε περιγραφεί από την άποψη παρόμοιων ποσοτήτων για ολόκληρο το σύνολο δεδομένων, δηλαδή, το X (αντίστοιχα το Σ). Υπάρχουν δύο προβλήματα που σχετίζονται με αυτή τη διατύπωση. Πρώτον, δεν είναι καθόλου πιθανό ότι όλη η συστηματική μεταβολή στις εκφράσεις των γονιδίων σχετίζεται με τη μεταβολή στην απόκριση. Έτσι, ακόμη και αν το μοντέλο (36)-(37) είναι αληθινό, δεν υπάρχει καμία εγγύηση ότι οι μεγαλύτερες *K* ιδιοτιμές του Σ είναι και οι μεγαλύτερες *K* ιδιοτιμές του Σ_1. Αυτό θα έχει ως αποτέλεσμα την προσθήκη ψευδών (δηλαδή που δεν σχετίζονατι με την απόκριση y) συστατικών στο μοντέλο.

Η δεύτερη δυσκολία αφορά την ακρίβεια των εκτιμήσεων. Επειδή συνήθως το *p* είναι πολύ μεγάλο (στην πραγματικότητα πολύ μεγαλύτερο από, ή τουλάχιστον σε σύγκριση με, το μέγεθος του δείγματος *N*), δεν πρόκειται σχεδόν ποτέ να ικανοποιείται η υπόθεση A2΄ (με το p_1 να αντικαθίσταται από το *p*). Όμως η υπόθεση για το p_1 είναι λογική, γιατί μόνο λίγα γονίδια αναμένεται να συνδέονται με ένα συγκεκριμένο τύπο ασθένειας. Η παραβίαση αυτών των όρων έχει ως αποτέλεσμα μια ασυνέπεια στις εκτιμήσεις του θ_k (βλ. την επόμενη ενότητα για λεπτομέρειες). Έτσι, η διαδικασία επιλογής των γονιδίων πριν από την εκτέλεση της PCA παλινδρόμησης δεν είναι μόνο λογική, αλλά και ουσιαστικά απαραίτητη.

2.9.2. Αποτελέσματα για την Εκτίμηση των θ_k και λ_k

Για να συζητηθεί η συνέπεια των ιδιοδιανυσμάτων θ_k, θεωρείται η ποσότητα $\text{dist}(\hat{\theta}_k, \theta_k)$, όπου dist είναι ένα μέτρο απόστασης μεταξύ δύο διανυσμάτων στη p_1-διάστατη μοναδιαία σφαίρα. Επιλέγεται είτε dist(a, b) = ∠ (a, b) (δηλαδή, η γωνία μεταξύ a και b) ή $\text{dist}(a, b) = \|a - \text{sign}(a^T b)b\|_2$ για $a, b \in S^{p_1}$.
Έστω, αρχικά, ότι εκτελείται PCA σε ολόκληρο το σύνολο δεδομένων X και εκτιμάται το θ_k από το $\tilde{\theta}_k$, τον περιορισμό του *k*-οστού δεξιού μοναδικού διανύσματος του X στις συντεταγμένες που αντιστοιχούν στο σύνολο X_1. Στη συνέχεια, το ακόλουθο αποτέλεσμα ισχυρίζεται ότι αν το *p* είναι πολύ μεγάλο, τότε μπορεί να μην υπάρχει συνέπεια.

Θεώρημα 1 (Lu 2002). Έστω ότι η (38) και η προυπόθεση A1 θεωρούν (και υποθέτουν ότι τα σ^2 και λ_k είναι σταθερά) ότι $p/N \to \gamma \in (0, \infty)$ όταν $N \to \infty$. Τότε

$\text{dist}(\tilde{\theta}_k, \theta_k)$ δεν τείνει στο 0 με πιθανότητα όταν $N \to \infty$,

δηλαδή, η συνήθης βασισμένη στην PCA εκτίμηση του θ_k είναι ασυνεπής.

Σύμφωνα με τους ίδιους όρους όπως στο θεώρημα 1, οι ιδιοτιμές του δείγματος είναι επίσης ασυνεπείς εκτιμήσεις για τις ιδιοτιμές των πληθυσμών. Οι Baik και Silverstein (2004) βρήκαν σχεδόν σίγουρα όρια των ιδιοτιμών του δείγματος σε παρόμοια δομή βάσει ελάχιστων υποθέσεων κατανομής.
Από τώρα και έπειτα γίνεται επεξεργασία αποκλειστικά της αποσύνθεσης μοναδικών τιμών του X_1. Η βασισμένη στην PCA εκτίμηση της *k*-οστής μεγαλύτερης ιδιοτιμής του Σ_1 συμβολίζεται με $\hat{l}_k$, $k = 1, 2, \ldots, m$. Ισχύει ότι $\hat{l}_k = \frac{1}{N} d_k 2$. Η αντίστοιχη ποσότητα του πληθυσμού είναι $l_k := \lambda_k + \sigma^2$.
Ένας φυσικός εκτιμητής του λ_k είναι το $\hat{\lambda}_k = \max\{\hat{l}_k - \sigma^2, 0\}$ αν το σ^2 είναι γνωστό. Αν το σ^2 είναι άγνωστο, τότε μπορεί να εκτιμηθεί με διάφορες στρατηγικές. Μία προσέγγιση είναι να χρησιμοποιηθεί ο μέσος των διαγώνιων στοιχείων του $\frac{1}{N} X_1^T X_1$ ως μία (συνήθως μεροληπτική) εκτίμηση του σ^2 και στη συνέχεια να οριστεί $\hat{\lambda}_k = \max\{\hat{l}_k - \hat{\sigma}^2, 0\}$.
Τώρα αποδεικνύουμε τη συνέπεια για την PCA να περιορίζεται στον πίνακα X_1. Δεν δίνεατι πλήρη απόδειξη αυτού του αποτελέσματος, διότι είναι μακριά και τεχνικής φύσεως. Όμως στην απόδειξη του Θεωρήματος 2 δίνουμε μια περίληψη της απόδειξης για την περίπτωση $p_1 / N \to 0$ και τα $\{\lambda_k\}_{k=1}^M$ και σ^2 είναι σταθερά. Οι λεπτομέρειες έχουν δοθεί από τον Paul (2005).

Θεώρημα 2. Έστω $\text{dist}(a, b) = \|a - \text{sign}(a^T b)b\|_2$. Έστω, επίσης, $h(x) := \frac{x^2}{1+x}$ και $g(x,y) := \frac{(x-y)^2}{xy}$. Ας υποτεθεί ότι η (38) ισχύει και ότι το σύνολο *P* επιλέγεται με πιθανότητα που τείνει στο 1 όταν $N \to \infty$.

Απόδειξη του Θεωρήματος 2

Όπως ήδη αναφέρθηκε, έχει αποδειχθεί το αποτέλεσμα υπό τις προυποθέσεις A1 και A2. Για να αποδειχθεί το Θεώρημα 2, χρειάζεται το ακόλουθο λήμμα για τα ιδιοδιανύσματα ενός συμμετρικού πίνακα.

Λήμμα A.1. Για κάποιο $m \in \mathbb{N}$, έστω A και B δύο συμμετρικοί $m \times m$ πίνακες. Έστω ότι οι ιδιοτιμές του πίνακα A συμβολίζονται με $\lambda_1(A) \geq \cdots \geq \lambda_m(A)$. Θέτουμε $\lambda_0(A) = \infty$ και $\lambda_{m+1}(A) = -\infty$. Για κάθε $r \in \{1, \ldots, m\}$, αν $\lambda_r(A)$ είναι μια ιδιοτιμή πολλαπλότητας 1, δηλαδή, $\lambda_{r-1}(A) > \lambda_r(A) > \lambda_{r+1}(A)$, συμβολίζοντας στη συνέχεια με p_r το ιδιοδιάνυσμα που σχετίζεται με την *r*-οστή μεγαλύτερη ιδιοτιμή,

$$p_r(A + B) - \text{sign}(p_r(A + B)^T p_r(A))\, p_r(A)$$

$$= -H_r(A)B\, p_r(A) + R_r, \qquad \text{(A.1)}$$

όπου $H_r(A) := \sum_{s \neq r} (\lambda_s(A) - \lambda_r(A))^{-1} P_{E_s}(A)$ και $P_{E_s}(A)$ συμβολίζει τον πίνακα προβολής πάνω στον ιδιοχώρο E_s που αντιστοιχεί στην ιδιοτιμή $\lambda_s(A)$, (ενδεχομένως πολυδιάστατη). Επιπλέον, ο όρος R_r που απομένει μπορεί να οριοθετείται από

$$\|\mathrm{R}_r\| \leq \begin{cases} \|\mathrm{H}_r(\mathrm{A})\mathrm{B}\lambda_r(\mathrm{A})\| \left[\frac{2\Delta_r(1+\Delta_r)}{1-2\Delta_r(1+\Delta_r)} + \frac{\|\mathrm{H}_r(\mathrm{A})\mathrm{B}\lambda_r(\mathrm{A})\|}{\left(1-2\Delta_r(1+\Delta_r)\right)^2}\right] \\ \qquad\qquad \text{αν } \Delta_r < \frac{\sqrt{5}-1}{2} \\ 10\Delta_r^2 \quad \text{πάντα} \end{cases} \tag{A.2}$$

όπου

$$\Delta_r = \frac{\|B\|_2}{\min_{1\leq s\neq r\leq m}|\lambda_s(\mathrm{A})-\lambda_r(\mathrm{A})|}\,. \tag{A.3}$$

Απόδειξη. Αυτό προκύπτει από βελτίωση του επιχειρήματος που δίνεται στην απόδειξη του λήμματος A.1 των Kneip και Utikal (2001).

Το Λήμμα A.1 δίνει μια πρώτης τάξης επέκταση του ιδιοδιανύσματος ενός διαταραγμένου πίνακα. Τώρα μπορούμε να πάρουμε ως πίνακα A τον πίνακα Σ_1, τον πίνακα συνδιακύμανσης του $\{\mathrm{X}_j : j \in \mathrm{P}\}$, και τότε μπορούμε να πάρουμε τον B να είναι η διαφορά S_1 - Σ_1, όπου $\mathrm{S}_1 = \frac{1}{N}\mathrm{X}_1^T\mathrm{X}_1$. Παρατηρούμε ότι

$$\mathrm{H}_r(\Sigma_1) = \sum_{1\leq s\neq r\leq M} \frac{1}{\lambda_s - \lambda_r}\theta_s\,\theta_s^T - \frac{1}{\lambda_r}\left(1 - \sum_{s=1}^{M}\theta_s\theta_s^T\right)$$

και

$$\mathrm{p}_r(\Sigma_1) = \theta_r.$$

Από το Λήμμα A.1, χρειαζόμαστε μόνο πιθανολογικά όρια για τις ποσότητες $\|\mathrm{H}_r(\mathrm{A})\mathrm{B}p_r(\mathrm{A})\|$ και $\|\mathrm{S}_1 - \Sigma_1\|$. Η πρώτη περιλαμβάνει κάποιο μεγάλη σε διάρκεια, αλλά απλό υπολογισμό, και για τη δεύτερη χρειαζόμαστε ένα φράγμα για τον όρο $\left\|\frac{1}{N}\mathrm{W}^T\mathrm{W} - 1.\right\|$. Γι 'αυτό, χρησιμοποιείται το παρακάτω λήμμα, η απόδειξη του οποίου χρησιμοποιεί ανισότητες με μεγάλη απόκλιση για τετραγωνικές μορφές των Γκαουσιανών τυχαίων μεταβλητών.

Λήμμα A.2. Υποθέτουμε ότι $n, L \to \infty$ τέτοια ώστε $L/n \to 0$. Έστω ότι ο Z συμβολίζει έναν $L\times n$ πίνακα με ανεξάρτητες και ίδια κατανεμημένες N(0,1) εισόδους. Συμβολίζουμε με l_1 και l_L τη μεγαλύτερη και τη μικρότερη ιδιοτιμή του $\frac{1}{n}\mathrm{Z}\mathrm{Z}^T$ αντίστοιχα. Έχουμε

$$\mathrm{P}\left(l_1 - 1 > 2\left(\sqrt{\log(n/L)} + \pi\right)\sqrt{\frac{L}{n}}\right) \leq \Delta_{nL}^{-1}(L/n)^{(L/2)(1+o(1))}\left(1 + o(1)\right)$$

και

$$\mathrm{P}\left(l_L - 1 > -2\left(\sqrt{\log(n/L)} + \pi\right)\sqrt{\frac{L}{n}}\right) \le (1 + \Delta_{nL}^{-1})(L/n)^{(L/2)(1+o(1))}\left(1 + o(1)\right),$$

όπου $\Delta_{nL} = \sqrt{2L}\sqrt{\log(n/L)}$.

Απόδειξη. Για $\mathrm{a} \in \mathrm{S}^L$ (L-διάστατη μοναδιαία σφαίρα), ορίζεται $g(\mathrm{a}, \mathrm{Z}) = \frac{1}{n}\mathrm{a}^T\mathrm{Z}\mathrm{Z}^T\mathrm{a}$. Ως συνάρτηση του a, η $g(\mathrm{a}, \mathrm{Z})$ είναι 1-Lipschitz με σταθερά Lipschitz $2\left\|\frac{1}{n}\mathrm{Z}\mathrm{Z}^T\right\| = 2l_1$. Αυτό ισχύει γιατί $g(\mathrm{a}, \mathrm{Z}) - g(\mathrm{b}, \mathrm{Z}) = (\mathrm{a} - b)^T\ \frac{1}{n}\mathrm{Z}\mathrm{Z}^T(\mathrm{a} + \mathrm{b})$. Έστω F_δ ένα ελάχιστο κάλυμμα της σφαίρας S^L από μπάλες ακτίνας $\delta < 1$. Στη συνέχεια ένα απλό επιχείρημα δείχνει ότι

$$\left(\frac{1}{\delta}\right)^{L-1} \le |F_\delta| \le 2\left(\frac{\pi}{\delta}\right)^{L-1}. \qquad \text{(A.4)}$$

Από ορισμό

$$l_1 = \max_{\mathrm{a}\in\mathrm{S}^L} g(\mathrm{a}, \mathrm{Z}) \text{ και } l_L = \min_{\mathrm{a}\in\mathrm{S}^L} g(\mathrm{a}, \mathrm{Z}).$$

Ως εκ τούτου, από το κάλυμμα του S^L από σφαίρες ακτίνας δ κεντραρισμένες σε σημεία στο F_δ και το όριο Lipschitz στην g, προκύπτει ότι

$$\max_{\mathrm{a}\in F_\delta} g(\mathrm{a}, \mathrm{Z}) \le l_1 \le \max_{\mathrm{a}\in F_\delta} g(\mathrm{a}, \mathrm{Z}) + 2\delta l_1 \qquad \text{και}$$

$$\min_{\mathrm{a}\in F_\delta} g(\mathrm{a}, \mathrm{Z}) - 2\delta l_1 \le l_L \le \min_{\mathrm{a}\in F_\delta} g(\mathrm{a}, \mathrm{Z}). \qquad \text{(A.5)}$$

Τώρα χρησιμοποιούμε το γεγονός ότι αν $\mathrm{a} \in \mathrm{S}^L$ και οι είσοδοι του $L\times n$ πίνακα Z είναι ανεξάρτητες και ίδια κατανεμημένες N(0,1) μεταβλητές, τότε ο $\mathrm{Z}^T\mathrm{a}$ έχει ανεξάρτητες και ίδια κατανεμημένες N(0,1) εισόδους και έτσι $g(\mathrm{a}, \mathrm{Z}) \sim \chi^2_{(n)}/n$. Τέλος, δίνονται κάποιες ανισότητες μεγάλης απόκλισης για χ^2 τυχαίες μεταβλητές.

$$\mathrm{P}\left(\chi^2_{(n)} > n(1 + \varepsilon)\right) \le e^{-3n\varepsilon^2/16}, \quad 0 < \varepsilon < \frac{1}{2}, \qquad \text{(A.6)}$$

$$\mathrm{P}\left(\chi^2_{(n)} > n(1 - \varepsilon)\right) \le e^{-n\varepsilon^2/4}, \quad 0 < \varepsilon < 1\,, \qquad \text{(A.7)}$$

και

$$\mathrm{P}\left(\chi^2_{(n)} > n(1 + \varepsilon)\right) \le \frac{\sqrt{2}}{\varepsilon\sqrt{n}} e^{-n\varepsilon^2/4}, \quad 0 < \varepsilon < n^{1/16},\ \ n \ge 16. \qquad \text{(A.8)}$$

Έστω $0 < t < 1$ και $0 < \delta < \frac{t}{2(1+t)}$. Τότε από την πρώτη ανισότητα στην (A.5), για $n \ge 16$,

$$\mathrm{P}(l_1 - 1 > t) \le \mathrm{P}\left(\max_{\mathrm{a}\in F_\delta} g(\mathrm{a}, \mathrm{Z})\,(1 - 2\delta)^{-1} - 1 > t\right)$$

$$= \mathrm{P}\left(\max_{\mathrm{a}\in F_\delta}[g(\mathrm{a}, \mathrm{Z}) - 1] > t(1 - 2\delta) - 2\delta\right)$$

$$\leq |F_\delta| \mathrm{P}\left(\frac{\chi^2_{(n)}}{n} - 1 > t(1-2\delta) - 2\delta\right)$$

$$\leq 2\left(\frac{\pi}{\delta}\right)^{N-1} \frac{\sqrt{2}}{\sqrt{n}(t(1-2\delta)-2\delta)}$$

$$\times \exp\left[-\frac{n}{4}(t(1-2\delta)-2\delta)^2\right],$$

από τις (A.4) και (A.8). Τώρα επιλέγουμε $\delta := \delta_n = \pi\sqrt{L/n}$ και $t := t_n = \left(2\sqrt{\log(n/L)} + 2\pi\right)\sqrt{\frac{L}{n}}$, τα οποία ικανοποιούν τους περιορισμούς για n αρκετά μεγάλο. Τότε

$$t(1-2\delta) - 2\delta = 2\sqrt{\log\left(\frac{n}{L}\right)}\sqrt{\frac{L}{n}}\left(1 - 2\pi\sqrt{\frac{L}{n}}\right) - 4\pi^2\frac{L}{n} = 2\sqrt{\log\left(\frac{n}{L}\right)}\sqrt{\frac{L}{n}}(1-\varepsilon_n),$$

όπου $\varepsilon_n = 2\pi\sqrt{L/n}\left(1 + \pi\left(\log\left(\frac{n}{L}\right)\right)^{-1/2}\right)$. Επειδή $\varepsilon_n = o(1)$ όταν $n \to \infty$,

$$\mathrm{P}\left(l_1 - 1 > \left(2\sqrt{\log\left(\frac{n}{L}\right)} + 2\pi\right)\sqrt{\frac{N}{n}}\right)$$

$$\leq \frac{\sqrt{2}}{2\sqrt{L}\sqrt{\log(n/L)}\,(1-\varepsilon_n)}\left(\frac{n}{L}\right)^{(L-1)/2}$$

$$\times \exp\left[-L\log\left(\frac{n}{L}\right)(1-\varepsilon_n)^2\right]$$

$$= \frac{1}{\sqrt{2L}\sqrt{\log(n/L)}\left(1+o(1)\right)}\left(\frac{L}{n}\right)^{(L/2)\left(1+o(1)\right)} \quad \text{όταν } n \to \infty \quad \text{(A.9)}$$

Στη συνέχεια, χρησιμοποιώντας τη δεύτερη ανισότητα στην (A.5), για $t = t_n$ και $\delta = \delta_n$ όπως επιλέχθηκε νωρίτερα,

$$\mathrm{P}(l_L - 1 < -t)$$

$$\leq \mathrm{P}\left(\min_{\mathrm{a}\in F_\delta} g(\mathrm{a},\mathrm{Z}) - 1 < -t + 2\delta l_1\right)$$

$$\leq \mathrm{P}\left(\min_{\mathrm{a}\in F_\delta}[\,g(\mathrm{a},\mathrm{Z}) - 1] < -t + 2\delta(1+t)\right) + \mathrm{P}(l_1 - 1 > t)$$

$$\leq |F_\delta|\,\mathrm{P}\left(\frac{\chi^2_{(n)}}{n} - 1 < -(t(1-2\delta) - 2\delta)\right) + \mathrm{P}(l_1 - 1 > t).$$

Επιπλέον, χρησιμοποιώντας τις (A.4) και (A.7) ακριβώς όπως πριν για να φράξουμε τον πρώτο όρο στη δεξιά πλευρά και μετά την (A.9), παίρνουμε, όταν $n \to \infty$,

$$\mathrm{P}\left(l_L - 1 < \left(2\sqrt{\log\left(\frac{n}{L}\right)} + 2\pi\right)\sqrt{\frac{L}{n}}\right)$$

$$\leq \left(1 + \frac{1}{\sqrt{2L}\sqrt{\log(n/L)}(1+o(1))}\right)\left(\frac{L}{n}\right)^{(L/2)(1+o(1))}. \qquad \text{(A.10)}$$

• Έστω ότι οι όροι A1΄ και A2΄ ισχύουν. Στη συνέχεια, για $1 \leq k \leq M$,

$$\mathrm{Edist}^2\left(\hat{\theta}_k, \theta_k\right) \leq \left[\frac{p_1}{Nh(\lambda_k/\sigma^2)} + \frac{1}{N}\sum_{k\neq k'}\frac{1}{g(\lambda_k+\sigma^2, \lambda_{k'}+\sigma^2)}\right] \times \left(1 + o(1)\right). \qquad (41)$$

Αν, επιπλέον, $\lambda_1/\sigma^2 \to \infty$, τότε $\hat{l}_k = \lambda_k\left(1 + o_P(1)\right)$.

• Αν το σ^2 και τα λ_k είναι σταθερά και οι προϋποθέσεις A1 και A2 ισχύουν, τότε ισχύει και η (41) και $\hat{l}_k \xrightarrow{P} l_k = \lambda_k + \sigma^2$ όταν $N \to \infty$.

2.9.3. Εκτίμηση του $\boldsymbol{\beta_k}$

Σε αυτή την ενότητα θα συζητηθεί η εκτίμηση των παραμέτρων β_k, $k = 1, \ldots, K$. Για να γίνει η ανάλυση απλούστερη, το σ^2 και τα λ_k θεωρούνται σταθερά και οι συνθήκες A1 και A2 υποτίθεται ότι ισχύουν. Το μοντέλο για τη μεταβλητή απόκρισης είναι $\mathrm{y} = \widehat{\beta}_0 + \sum_{k=1}^{K}\hat{\beta}_k\,\hat{\mu}_k$. Έστω ότι ή το σ^2 είναι γνωστό ή μια συνεπής εκτίμηση είναι διαθέσιμη. Μετά ορίζεται $\hat{\lambda}_k = \max\{\hat{l}_k - \sigma^2, 0\}$. Έστω u_k όπως πριν και ορίζεται το $\tilde{\mathrm{u}}_k$ ως $\frac{1}{\sqrt{\hat{\lambda}_k}}\frac{1}{\sqrt{N}}\mathrm{X}_1\mathrm{v}_k$ αν $\hat{\lambda}_k > 0$, και όπως κάθε άλλο μοναδιαίο διάνυσμα [π.χ. $(1,0, \ldots ,0)^T$] διαφορετικά. Ορίζεται μια εκτίμηση του β_k (για $1 \leq k \leq K$) ως $\tilde{\beta}_k = \tilde{\mathrm{u}}_k^T\mathrm{y}$. Η επίδοσή του μπορεί να συγκριθεί με μια άλλη εκτίμηση $\hat{\beta}_k = \mathrm{u}_k^T\mathrm{y}$ με το u_k όπως πριν. Επίσης, ορίζεται $\hat{\beta}_0 = \tilde{\beta}_0 = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\mathrm{y}_j$.

Παρατηρείται ότι

$$\mathrm{u}_k = \frac{1}{d_k}\mathrm{X}_1\mathrm{v}_k = \left(\hat{l}_k\right)^{-1/2}\frac{1}{\sqrt{N}}\mathrm{X}_1\hat{\theta}_k = \left(\hat{l}_k\right)^{-1/2}\left[\sum_{l=1}^{M}\sqrt{\lambda_l}\left(\theta_l^T\hat{\theta}_k\right)\frac{1}{\sqrt{N}}\eta_l + \frac{\sigma}{\sqrt{N}}\mathrm{W}\hat{\theta}_k\right],$$

όπου W είναι ένας $N \times p_1$ πίνακας του οποίου οι γραμμές είναι w_i^T (i =1, … , N). Τότε, επειδή $\hat{\theta}_k = \theta_k + \varepsilon_k$ ($\hat{\theta}_k^T\theta_k > 0$), όπου $\|\varepsilon_k\|_2 = O_P\left(\sqrt{p_1/N}\right)$,

$$\mathrm{u}_k = \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_k+\sigma^2}}\frac{1}{\sqrt{N}}\eta_k\left(1 + o_P(1)\right) + \frac{\sigma}{\sqrt{\lambda_k+\sigma^2}}\frac{1}{\sqrt{N}}\mathrm{W}\theta_k\left(1 + o_P(1)\right) + \delta_k\,, \qquad (42)$$

όπου $\|\delta_k\|_2 = O_P\left(\sqrt{p_1/N}\right)$. Για να αποδειχθεί ο τελευταίος αυτός ισχυρισμός, χρειάζεται να χρησιμοποιηθεί μόνο το θεώρημα 2 σε συνδυασμό με το γεγονός ότι $\left\|\frac{1}{N}\mathrm{W}^T\mathrm{W}\right\|_2 = 1 + o_P(1)$, γιατί

$$2\left\|\frac{1}{\sqrt{N}}\mathrm{W}\varepsilon_k\right\|_2 \le 2\left\|\frac{1}{N}\mathrm{W}^T\mathrm{W}\right\|_2 \|\varepsilon_k\|_2 = O_P\left(\frac{p_1}{N}\right)$$

και

$$|\varepsilon_k^T\theta_l| \le \|\varepsilon_k\|_2\|\theta_l\|_2 = O_P\left(\frac{p_1}{N}\right) \text{ για } 1 \le l \ne k \le M,$$

και, τελικά, $\|\eta_l\|_2 = \sqrt{N}\big(1 + o_P(1)\big)$ για κάθε $l = 1, \dots, M$.
Από αυτό προκύπτει ότι

$$\tilde{\mathrm{u}}_k = \tfrac{1}{\sqrt{N}}\eta_k\big(1 + o_P(1)\big) + \tfrac{\sigma}{\sqrt{\lambda_k}}\tfrac{1}{\sqrt{N}}\mathrm{W}\theta_k\big(1 + o_P(1)\big) + \tilde{\delta}_k\,, \qquad (43)$$

όπου $\left\|\tilde{\delta}_k\right\|_2 = O_P\big(\sqrt{p_1/N}\big)$. Τα διανύσματα $\{\mathrm{W}\theta_k : k = 1, \dots, M\}$ είναι ανεξάρτητα από την $\mathrm{N}_N(0, \mathrm{I})$ και από τα $\{\eta_k : k = 1, \dots, M\}$, επειδή τα θ_k είναι αμοιβαία ορθοκανονικά. Για να επιτευχθεί η συνέπεια του $\tilde{\beta}_k$, $k = 1, \dots, K$, προκύπτει (από την (43)) ότι

$$\tilde{\beta}_k = \beta_0\frac{1}{\sqrt{N}}\tilde{\mathrm{u}}_k^T 1 + \sum_{l=1}^{K}\beta_l\frac{1}{N}\left[\left(\eta_k + \frac{\sigma}{\sqrt{\lambda_k}}\mathrm{W}\theta_k\right)\big(1 + o_P(1)\big) + \sqrt{N}\tilde{\delta}_k\right]^T\eta_l + \tilde{\mathrm{u}}_k^T\mathrm{Z}$$

$$= \beta_0\left(O_P\left(\frac{1}{\sqrt{N}}\right) + o_P(1)\right) + \beta_k\left(1 + o_P(1) + \tilde{\delta}_k^T\frac{1}{\sqrt{N}}\eta_k\right)$$

$$+ \sum_{l\ne k}\beta_l\left(O_P\left(\frac{1}{\sqrt{N}}\right) + \tilde{\delta}_k^T\frac{1}{\sqrt{N}}\eta_l\right) + O_P\left(\frac{1}{\sqrt{N}}\right)$$

$$= \beta_k\big(1 + o_P(1)\big),$$

επειδή $\frac{1}{N}\eta_k^T\eta_l = O_P\big(1/\sqrt{N}\big)$ αν $k \ne l$ και $\frac{1}{N}\eta_l^T\mathrm{W}\theta_k = O_P\big(1/\sqrt{N}\big)$ για όλα τα k, l, (από ανεξαρτησία), $\left\|\tilde{\delta}_k\right\|_2 = o_P(1)$, και $\tilde{\mathrm{u}}_k^T\mathrm{Z} = \|\tilde{\mathrm{u}}_k\|_2\,\langle\frac{\tilde{\mathrm{u}}_k}{\|\tilde{\mathrm{u}}_k\|_2}, \mathrm{Z}\rangle$. Ο δεύτερος όρος στο τελευταίο αποτέλεσμα είναι μια $\mathrm{N}(0,\frac{\tau^2}{N})$ τυχαία μεταβλητή, και ο πρώτος όρος είναι $\sqrt{(\lambda_k + \sigma^2)/\lambda_k}\,\big(1 + o_P(1)\big)$ από την (43).
Είναι εύκολο να επαληθεύτεί ότι $\hat{\beta}_0 = \beta_0\big(1 + o_P(1)\big)$. Όμως, από την προηγούμενη ανάλυση, φαίνεται ξακάθαρα ότι ο εκτιμητής $\hat{\beta}_k = \mathrm{u}_k^T\mathrm{y}$, για $1 \le k \le K$, δεν είναι γενικά συνεπής. Στην πραγματικότητα, $\hat{\beta}_k = \sqrt{\frac{\lambda_k}{\lambda_k+\sigma^2}}\beta_k\big(1 + o_P(1)\big)$ όταν τα λ_k και το σ^2 είναι σταθερά. Ωστόσο, όπως φαίνεται στην Παρατήρηση 1, είναι λογικό να υποθέσουμε ότι $\lambda_1/\sigma^2 \to \infty$ όταν $p_1, N \to \infty$. Αυτό θα διασφαλίσει (μέσω του πρώτου μέρους του Θεωρήματος 2) ότι ο παράγοντας $\sqrt{\lambda_k/\hat{l}_k} \to 1$ με πιθανότητα όταν $N \to \infty$ και ισχύουν οι προυποθέσεις A1΄ και A2΄. Επομένως, $\hat{\beta}_k = \beta_k\big(1 + o_P(1)\big)$, για $1 \le k \le K$. Αυτό επικυρώνει κατά ένα τρόπο τον ισχυρισμό ότι το να υπάρχουν

περισσότερα γονίδια (δηλαδή, μεγαλύτερα του p_1) που να σχετίζονται με την απόκριση δίνει καλύτερη πρόβλεψη.

2.9.4. Συνέπεια του Σχεδίου Επιλογής Συντεταγμένων: Μοντέλο Παλινδρόμησης

Στην ενότητα αυτή περιγράφονται κάποιες καταστάσεις κάτω από τις οποίες η SPCA θα επιλέγει με συνέπεια το σύνολο *P* των συντεταγμένων (γονίδια) των οποίων η μεταβλητότητα σχετίζεται με αυτή της απόκρισης μέσω του μοντέλου που δίνεται από τις (36), (38) και (40). Όλες αυτές οι εργασίες γίνονται κάτω από την υπόθεση ότι $p_1 = O(N^{\alpha})$ για κάποια $\alpha \in (0,1)$ και log*p* = log*N*. Η δεύτερη υπόθεση ένα ευρύ φάσμα πιθανών καταστάσεων. Το βασικό σημείο που τονίζεται είναι ότι για να μπορεί να ανακτηθεί το σύνολο *P* των μεταβλητών πρόβλεψης που σχετίζονται με την απόκριση, μπορεί να χρειαστούν ορισμένες προϋποθέσεις αναγνωρισιμότητας για το σύνολο αυτό. Η μέθοδος μπορεί να λειτουργήσει κάτω από γενικότερες συνθήκες, αλλά εδώ η προσοχή θα περιοριστεί στις περιπτώσεις που είναι αναλυτικά τιθασεύσιμες και σχετικά απλό να ερμηνευτούν.
Πρώτον, παρατηρείται ότι το διάνυσμα των μονομεταβλητών αποτελεσμάτων μπορεί να γραφεί ως $\mathrm{s} = \mathrm{H}_{\mathrm{X}}^{-1}\mathrm{X}^{T}\mathrm{y}$, όπου $\mathrm{H}_{\mathrm{X}} = \mathrm{diag}\,(\|\mathrm{x}_1\|, \ldots, \|\mathrm{x}_p\|)$. Επειδή οι γραμμές του X_2 είναι ανεξάρτητες της $\mathrm{N}_{p_2}(0, \Sigma_2)$ τυχαίες μεταβλητές ανεξάρτητες του X_1, το μη κανονικοποιημένο διάνυσμα του αποτελέσματος $\tilde{\mathrm{s}} := \mathrm{X}^{T}\mathrm{y}$ μπορεί να εκφραστεί, με βάση την (38), στη μορφή

$$\tilde{\mathrm{s}} = \begin{bmatrix} \left(\sum_{k=1}^{M} \sqrt{\lambda_k}\, \theta_k \eta_k^{T} + \sigma \mathrm{W}^{T}\right)\mathrm{y} \\ \Sigma_2^{1/2}\mathrm{C}\mathrm{y} \end{bmatrix}, \qquad (44)$$

όπου C είναι ένας $p_2 \times N$ πίνακας του οποίου οι καταχωρήσεις είναι ανεξάρτητες και ίδια κατανεμημένες N(0, 1) ανεξάρτητες από τα X_1 και Z (και ως εκ τούτου και το y). Παρατηρούμε ότι το W^{T} είναι ανεξάρτητο του y. Για εκθεσιακούς σκοπούς, προτιμάται το $\tilde{\mathrm{s}}$ έναντι του s.
Αυτό δείχνει ότι αν θεωρηθεί το *j*-οστό στοιχείο του $\tilde{\mathrm{s}}$ για $j \in P$, τότε

$$\frac{1}{\sqrt{N}}\tilde{s}_j = \frac{1}{N}\left(\sum_{k=1}^{M} \sqrt{\lambda_k}\theta_{jk}\eta_k^{T}\right) \times \left(\beta_0 1 + \sum_{k'=1}^{K} \beta_{k'}\eta_{k'} + \sqrt{N}\mathrm{Z}\right) + \frac{\sigma}{\sqrt{N}}(\mathrm{W}^{T}\mathrm{y})_j$$

$$= \beta_0 \sum_{k=1}^{M} \sqrt{\lambda_k}\theta_{jk} O_P\left(\frac{1}{\sqrt{N}}\right) + \sum_{k=1}^{K} \beta_k \sqrt{\lambda_k}\theta_{jk}\left(1 + O_P\left(\frac{1}{\sqrt{N}}\right)\right)$$

$$+\sum_{k=1}^{M}\sqrt{\lambda_k}\theta_{jk}\sum_{k'\neq k}^{K}\beta_{k'}\,O_P\left(\frac{1}{\sqrt{N}}\right)$$

$$+\sigma\left(\sum_{k=0}^{K}\beta_k\right)O_P\left(\frac{1}{\sqrt{N}}\right)$$

$$=\sum_{k=1}^{K}\beta_k\sqrt{\lambda_k}\theta_{jk}+O_P\left(\tfrac{1}{\sqrt{N}}\right).$$

Όμως από την άλλη πλευρά, αν $j \notin P$, τότε, υποθέτοντας ότι η $\|\Sigma_2\|_2$ είναι άνω φραγμένη,

$$\frac{1}{\sqrt{N}}\tilde{s}_j=\frac{1}{\sqrt{N}}\left(\Sigma_2^{1/2}\mathrm{Cy}\right)_j=\left(\Sigma_2^{1/2}\right)_j^T\frac{1}{\sqrt{N}}\mathrm{Cy}=O_P\left(\frac{1}{\sqrt{N}}\right).$$

Έτσι, για να είναι το "σήμα" $\zeta_j^K:=\sum_{k=1}^{K}\beta_k\sqrt{\lambda_k}\theta_{jk}$ ανιχνεύσιμο, πρέπει να είναι $\gg 1/\sqrt{N}$. Η μεγάλη απόκλιση στα φράγματα δείχνει ότι μπορούν να ανακτηθούν με επαρκή ακρίβεια μόνο οι συντεταγμένες j για τις οποίες $\left|\zeta_j^K\right|\geq c_0\sqrt{\log N/N}$ για κάποια σταθερά $c_0 > 0$ (η οποία εξαρτάται από το σ, τα λ_k, τα β_k και τη $\|\Sigma_2\|_2$). Δυνητικά, πολλά ζ_j^K μπορεί να είναι μικρότερα από αυτά, και ως εκ τούτου, οι συντεταγμένες αυτές δεν θα επιλεγούν με μεγάλη πιθανότητα. Αν το όριο (threshold) γίνει πολύ μικρό, τότε θα περιλαμβάνει πολλές "εσφαλμένες" συντεταγμένες (πχ. αυτές με $j \notin P$), η οποίες μπορεί να προκαλέσουν προβλήματα στην εκτίμηση με διάφορους τρόπους που έχουν ήδη συζητηθεί.

Αν $K = 1$, τότε η j-οστή συνιστώσα του διανύσματος σήματος ζ^K είναι ανάλογη με το $\sqrt{\lambda_1}\theta_{j1}$. Έτσι το σύστημα θα επιλέξει μόνο αυτές τις συντεταγμένες j για τις οποίες το $\sqrt{\lambda_1}\left|\theta_{j1}\right|$είναι μεγάλο. Αυτό μπορεί να μην εξαντλήσει το σύνολο $\{1, \ldots, p_1\}$, αλλά όσον αφορά τη συνεπή εκτίμηση των θ_1 και λ_1 είναι αρκετό. Έτσι, όταν $K = 1$, το σύστημα επιλογής συντεταγμένων είναι συνεπές.

Στην περίπτωση όπου $K > 1$, μπορεί να υπάρξει κάποιο πρόβλημα. Αυτό επειδή η μέθοδος που περιγράφηκε βασίζεται σε μία σταθερή γραμμική συνάρτηση του διανύσματος $\mathrm{t}_j=\left(\sqrt{\lambda_1}\theta_{j1},\ldots,\sqrt{\lambda_k}\theta_{jk}\right)$, το ζ_j^K. Έτσι, ακόμη και αν μία τουλάχιστον είσοδος του t_j είναι αρκετά μεγάλη, η συντεταγμένη j μπορεί να χαθεί. Με άλλα λόγια, όταν $K > 1$, σε γενικές γραμμές δεν υπάρχει καμία εγγύηση ότι το σύστημα επιλογής συντεταγμένων είναι συνεπές. Μια πιο προσεκτική ματιά δείχνει ότι σε γενικές γραμμές δεν υπάρχουν επαρκείς περιορισμοί αναγνωρισιμότητας για το σύνολο των μεταβλητών πρόβλεψης P. Ένας τρόπος να επιβληθεί αυτός ο περιορισμός είναι να υποτεθεί ότι το $\left|\zeta_j^K\right|$ είναι πάνω από ένα όριο της μορφής

$c_0\sqrt{\log N/N}$ όποτε το $\|\mathrm{t}_j\|$ είναι πάνω από ένα όριο $c_1\sqrt{\log N/N}$ για ορισμένες σταθερές c_0, $c_1 > 0$. Ακόμα κι αν αυτή η προυπόθεση δεν μπορεί να ικανοποιηθεί ακριβώς, αποδεικνύεται ότι χρειάζεται μόνο ο εξής, κάπως ασθενέστερος περιορισμός:

A3. Το σύνολο των μεταβλητών P που καθορίζει το X_1 είναι τέτοιο ώστε αν το $P_{N,\beta}$ δηλώνει το σύνολο όλων των $j \in P$ με

$$\left|\langle t_j, \frac{\beta}{\|\beta\|}\rangle\right| \geq c_0\sqrt{\frac{\log N}{N}} \qquad (45)$$

για κάποια σταθερά $c_0 > 0$ ανεξάρτητη του β, τότε

$$\sum_{j\in P\setminus P_{N,\beta}} \frac{1}{\lambda_1}\|\mathrm{t}_j\|^2 \to 0 \quad \text{όταν } N \to \infty.$$

Ο A3 είναι ένας περιορισμός για ολόκληρο το μοντέλο, όχι μόνο για την κατανομή των μεταβλητών πρόβλεψης X. Η φυσική σημασία του περιορισμού αυτού έχει ως εξής: οι μεταβλητές πρόβλεψης της κλάσης P που έχουν σημαντική διακύμανση επειδή οι πρώτες K συνιστώσες στο μοντέλο (38) είναι υψηλά συσχετισμένες με την απόκριση. Αυτό συμβαίνει επειδή $\sigma^2 + \|\mathrm{t}_j\|^2 + \sum_{k=K+1}^{M} \lambda_k\, \theta_{jk}^2$ είναι η διακύμανση της j-οστής μεταβλητής πρόβλεψης, και $\frac{1}{\sqrt{N}}\zeta_j^K$ είναι η συνδιακύμανση των X_j και y. Σημαίνει επίσης ότι οι συντεταγμένες που μπορεί να απορρίφθηκαν έχουν αμελητέα συμβολή στη συνολική μεταβλητότητα που σχετίζεται με τις πρώτες K συνιστώσες της μεταβολής. Αυτή η προϋπόθεση πληρούται αυτομάτως όταν $K = 1$.
Ένας διαφορετικός τρόπος να επιβληθεί η αναγνωρισιμότητα στο σύνολο των μεταβλητών πρόβλεψης P είναι να επιβληθεί ένας περιορισμός στην παράμετρο $\beta = (\beta_1, \ldots, \beta_K)$. Με τον τρόπο αυτό απαιτείται ότι ένας συγκεκριμένος $K \times K$ πίνακας $\mathrm{H}(\beta)$, οι είσοδοι του οποίου είναι πολυώνυμα στα β_K, έχει μικρό αριθμό όρων. Ο H ορίζεται ως εξής. Η πρώτη στήλη του $\mathrm{H}(\beta)$ είναι το ίδιο το β. Μπορεί να χρησιμοποιηθεί αυτός ο περιορισμός για να διασφαλιστεί η επιλογή όλων των μεγάλων συντεταγμένων ακόμη και όταν $K \gg 1$. Το σύστημα επιλογής θα μπορούσε να γενικευτεί ως εξής.
Για τους ακεραίους $r = 1, 2, \ldots, K$, το σύνολο J_r ορίζεται να είναι το σύνολο των συντεταγμένων j τέτοιο ώστε $\left|s_j^{(r)}\right| > a_j^{(r)}$ όπου το $a_j^{(r)}$ είναι ένα όριο της τάξης $\sqrt{\log N}$ και $s_j^{(r)}$είναι η j-οστή συντεταγμένη του $\frac{1}{\sqrt{N}}\left(\mathrm{X}^T \mathrm{y}^{(r)}\right)$ όπου η l-οστή συντεταγμένη του $\mathrm{y}^{(r)}$ είναι $\left(\sqrt{N}\mathrm{y}_l\right)^{2r-1}$. Ειδικότερα, $\mathrm{y}^{(1)} = \sqrt{N}\mathrm{y}$, έτσι ώστε να είναι $s^{(1)} = s$, όπως καθορίστηκε προηγουμένως. Τέλος, λαμβάνεται η ένωση $J := \cup_{r=1}^{K} J_r$ και το J να είναι η τελική επιλογή. Μια ανάλυση του συστήματος αυτού δείχνει ότι για $j \in P$, $\frac{1}{\sqrt{N}} s_j^{(r)} = \mathrm{t}_j^T \mathrm{H}_r(\beta) + O_P\left(1/\sqrt{N}\right)$, όπου $\mathrm{H}_r(\beta)$ είναι η r-οστή

στήλη του H. Στη συνέχεια, από τον περιορισμό στον πίνακα H(β), για κάθε $j \in P$, ισχύει $j \notin J$ αν και μόνο αν το $\|t_j\|$ είναι «μικρό» (που σημαίνει μικρότερο από από ένα ορισμένο όριο της μορφής $c_2\sqrt{\log N/N}$) για κάποια σταθερά c_2.

Παρατήρηση 3. Από τις δύο μεθόδους για την επιβολή των περιορισμών της αναγνωρισιμότητας, η δεύτερη είναι ομολογουμένως αρκετά ειδική και δεν έχει σημαντική γενίκευση πέρα από τη ρύθμιση της παλινδρόμησης. Ωστόσο, ο πρώτος περιορισμός μπορεί συχνά να ικανοποιηθεί στην πράξη, επειδή ένα μέρος της μεταβλητότητας στις μεταβλητές πρόβλεψης μπορεί να συνδέεται άμεσα με τη μεταβλητότητα στην απόκριση. Αυτό είναι πιθανό να αληθεύει αν, για παράδειγμα, υπάρχει μια αιτιώδης σχέση.

2.10. Μερικά Πρακτικά Ζητήματα και Γενικεύσεις

Εδώ αναφέρονται μερικοί τρόποι με τους οποίους οι κύριες συνιστώσες με επίβλεψη μπορούν να εφαρμοστούν στην πράξη.

Κοινή Προσαρμογή με άλλες Συμμεταβλητές. Τυπικά, μπορεί να υπάρχουν συμμεταβλητές που μετρώνται σε κάθε μία από τις περιπτώσεις, και μπορεί να έχει ενδιαφέρον να προσαρμοστούν για αυτές. Για παράδειγμα, σε μελέτες επιβίωσης της γονιδιακής έκφρασης, εκτός από τις μεταβλητές πρόβλεψης $X_1, X_2, \ldots, X_p$, θα μπορούσαν να υπάρχουν συμμεταβλητές $z = (z_1, z_2, \ldots, z_k)$, όπως το στάδιο του όγκου και ο τύπος του όγκου. Μπορεί να υπάρχει ενδιαφέρον στην εξεύρεση μεταβλητών πρόβλεψης γονιδιακής έκφρασης που λειτουργούν ανεξάρτητα από το στάδιο και τον όγκο. Δηλαδή, έχοντας προσαρμοστεί για τους παράγοντες αυτούς, η μεταβλητή πρόβλεψης γονιδιακής έκφρασης παραμένει στενά συνδεδεμένη με την επιβίωση.
Για να συγκρίνει κανείς τη μεταβλητή πρόβλεψης της επιβλεπόμενης κύριας συνιστώσας με τις ανταγωνιστικές μεταβλητές πρόβλεψης, μπορεί απλά να τα τοποθετήσει μαζί σε ένα μοντέλο για το σύνολο εξέτασης. Στο παράδειγμα του λεμφώματος, ο Διεθνής Προγνωστικός Δείκτης (IPI = International Prognostic Index) (χαμηλός, μεσαίος ή υψηλός) είναι μια κλινική μεταβλητή πρόβλεψης της επιβίωσης που χρησιμοποιείται ευρέως. Η μεταβλητή πρόβλεψης της κύριας συνιστώσας με επίβλεψη και το IPI τοποθετούνται στο σύνολο εξέτασης και μετά καθορίζονται οι p τιμές όταν καθένα από αυτά αφαιρείται ξεχωριστά από το κοινό μοντέλο. Αυτές ήταν 0.001 για την κύρια συνιστώσα με επίβλεψη και 0.05 για το IPI. Έτσι, η επίδραση της μεταβλητής πρόβλεψης των κυρίων συνιστωσών με επίβλεψη είναι ισχυρά ανεξάρτητη από το IPI, ενώ το IPI είναι μόνο μετρίως ανεξάρτητο από τη μεταβλητή πρόβλεψης των κυρίων συνιστωσών με επίβλεψη.
Η μεταβλητή πρόβλεψης των κυρίων συνιστωσών με επίβλεψη θα μπορούσε να ψάξει για μεταβολή που να είναι ανεξάρτητη των ανταγωνιστικών μεταβλητών πρόβλεψης. Για να γίνει αυτό, πραγματοποιείται μια γραμμική παλινδρόμηση κάθε

γονιδίου στις ανταγωνιστικές μεταβλητές πρόβλεψης, αντικαθιστώντας τις μετρήσεις κάθε γονιδίου από ό,τι έχει μείνει από τη διαδικασία αυτή. Στη συνέχεια εφαρμόζεται η διαδικασία των κυρίων συνιστωσών με επίβλεψη στον υπολειπόμενο πίνακα. Αυτή η διαδικασία «αποσυσχετίζει» τη γονιδιακή έκφραση και τις ανταγωνιστικές μεταβλητές πρόβλεψης και εξαναγκάζει τις κύριες συνιστώσες να είναι ορθογώνιες με τις ανταγωνιστικές μεταβλητές πρόβλεψης. Η ίδια προσέγγιση μπορεί να χρησιμοποιηθεί με άλλες μεθόδους, όπως τα PLS.

Χρήση δειγμάτων χωρίς επισήμανση. Σε ορισμένες ρυθμίσεις, διαθέτουμε και δεδομένα "*με επισήμανση*" (π.χ., προφίλ γονιδιακής έκφρασης με μετρημένους χρόνους επιβίωσης) και χωρίς (απλά προφίλ γονιδιακής έκφρασης). Στην πραγματικότητα, μπορεί κανείς να έχει πολλά μη επισημασμένα δείγματα και μόνο λίγα επισημασμένα, λόγω της απόκτησης πληροφοριών για το αποτέλεσμα μπορεί να είναι πιο δύσκολο. Σε αυτό το πλαίσιο ίσως να είναι χρήσιμο να χρησιμοποιηθούν τα μη επισημασμένα δεδομένα με κάποιο τρόπο, επειδή περιέχουν πληροφορίες για τη συσχέτιση μεταξύ των χαρακτηριστικών. Λόγω της απλής μορφής της μεταβλητής πρόβλεψης των κυρίων συνιστωσών με επίβλεψη, υπάρχει ένας εύκολος τρόπος να γίνει αυτό. Έστω ότι οι πίνακες χαρακτηριστικών για τα επισημασμένα και για τα μη επισημασμένα δεδομένα είναι οι X^L και X^U. Στο πρώτο βήμα, χρησιμοποιείται μόνο τον X^L (και το αποτέλεσμα) ώστε να επιλεγούν τα χαρακτηριστικά. Στη συνέχεια, χρησιμοποιείται όλο το σύνολο των χαρακτηριστικών (X^L, X^U) για τον υπολογισμό των κυρίων συνιστωσών. Οι προστιθέμενες πληροφορίες που παρέχονται από τα μη επισημασμένα δείγματα μπορούν να βελτιώσουν την ακρίβεια των κυρίων συνιστωσών με επίβλεψη.

Εφαρμογή σε άλλους τύπους δεδομένων. Η ιδέα των κυρίων συνιστωσών με επίβλεψη μπορεί να εφαρμοστεί σε άλλους τύπους μέτρων αποτελέσματος, όπως τα αποτελέσματα ταξινόμησης. Στην περίπτωση αυτή, θα μπορούσαν να επιλεγούν τα χαρακτηριστικά που έχουν τη μεγαλύτερη between-class to within-class μεταβολή, και μετά θα υπολογίζονταν οι κύριες συνιστώσες των επιλεγμένων δεδομένων. Στη συνέχεια, η κύρια συνιστώσα θα έμπαινε σε μια πολλαπλή λογιστική παλινδρόμηση για να προβλεφθεί η ετικέτα της κλάσης. Αν και η διαδικασία αυτή φαίνεται πολλά υποσχόμενη, δεν έχουν ακόμη βρεθεί παραδείγματα όπου διορθώνονται μέθοδοι, όπως η προσέγγιση του πλησιέστερου συρρικνωμένου κέντρου βάρους (Tibshirani et al. 2001). Η εξήγηση μπορεί να βρίσκεται στο soft- thresholding που είναι συνυφασμένο με το πλησιέστερο συρρικνωμένο κέντρο βάρους. Αυτό μπορεί να έχει την ίδια θετική επίδραση όπως το threshold στις κύριες συνιστώσες με επίβλεψη.

2.11. Συζήτηση και Περιορισμοί

Οι κύριες συνιστώσες με επίβλεψη αντιπροσωπεύουν ένα πολλά υποσχόμενο εργαλείο για την πρόβλεψη στην παλινδρόμηση και για γενικευμένα προβλήματα παλινδρόμησης. Είναι μια απλή ιδέα που έχει δοκιμαστεί πιθανώς πολλές φορές στην πράξη. Εδώ έχει διερευνηθεί η εφαρμογή της σε μελέτες γονιδιακής έκφρασης.

Η παλινδρόμηση είναι ένα σημαντικό και δύσκολο πρόβλημα στη στατιστική. Είναι ιδιαίτερα δύσκολη όταν ο αριθμός των χαρακτηριστικών *p* υπερβαίνει κατά πολύ τον αριθμό των παρατηρήσεων *N*. Υπερπροσαρμογή μπορεί να συμβεί ακόμα και με μετρίως πολύπλοκα μοντέλα, και ο προσδιορισμός των σημαντικών χαρακτηριστικών είναι πολύ επικίνδυνος λόγω του μεγάλου αριθμού των χαρακτηριστικών, πολλά από τα οποία συχνά συσχετίζονται σε μεγάλο βαθμό. Παρά τη δυσκολία στον εντοπισμό των σημαντικών χαρακτηριστικών, αυτό αποτελεί υψηλή προτεραιότητα για τους βιολόγους στις μελέτες της γονιδιακής έκφρασης.
Οι κύριες συνιστώσες με επίβλεψη προσεγγίζουν αυτό το δύσκολο πρόβλημα μέσω μιας ημιεπιβλεπόμενης στρατηγικής, ψάχνοντας τη δομή στα δεδομένα που θα τα ευθυγραμμίζει με το αποτέλεσμα. Μόνο στη συνέχεια της διαδικασίας θα γίνει προσπάθεια να μειωθεί το σύνολο των χαρακτηριστικών σε μια πολύ μικρότερη λίστα (μέσω των σκορ σημαντικότητάς του). Μια κρίσιμη πρακτική πτυχή αυτού του αποτελέσματος σημασίας είναι το γεγονός ότι παρέχει μια σταθερή διάταξη των χαρακτηριστικών. Έτσι, ξεκινώντας με μια λίστα από 200 χαρακτηριστικά αναζητείται ένα υπομοντέλο που να περιέχει μόλις 20 χαρακτηριστικά. Το κατασκευασμένο μοντέλο αποτελείται από τα 20 χαρακτηριστικά μεταξύ των 200 με τα μεγαλύτερα αποτελέσματα σημασίας. Αντίθετα, χρησιμοποιώντας μια μέθοδο όπως το lasso, θα μπορούσαν να συμβούν τα ακόλουθα: παραδίδεται ένα μοντέλο με 200 χαρακτηριστικά, και στη συνέχεια ζητάται ένα μικρότερο μοντέλο, το οποίο να περιέχει μόνο 20 χαρακτηριστικά. Για να επιτευχθεί αυτό αλλάζει το φράγμα του lasso οπότε αποκτάται ένα νέο μοντέλο που περιέχει 20 χαρακτηριστικά, μερικά από τα οποία ήταν, ή και κανένα, στην αρχική λίστα των 200! Αυτό φαίνεται σαν μια ικανοποιητική προσέγγιση στην επιλογή μοντέλου σε αυτή τη ρύθμιση.
Παρά τις ενθαρρυντικές επιδόσεις των κυρίων συνιστωσών με επίβλεψη, το πρόβλημα της μεγάλων διαστάσεων παλινδρόμησης είναι πολύ δύσκολο και θα πρέπει να προσεγγίζεται με προσοχή. Υπάρχουν πολλά θέματα που χρειάζονται περαιτέρω ανάπτυξη και προσεκτική μελέτη. Πολλά από αυτά επισημάνθηκαν από τους συντάκτες και τους referees. Κάποια αναφέρονται παρακάτω:

- Η ικανότητα διασταυρωμένης επικύρωσης για να επιλεγεί το "σωστό" σύνολο γονιδίων δεν έχει αποδειχθεί θεωρητικά. Στην πράξη, φαίνεται να αποδίδει ικανοποιητικά, αλλά μπορεί μερικές φορές να παρουσιάσει μεγάλη μεταβλητότητα, ιδιαίτερα όταν τα μεγέθη του δείγματος είναι μικρά. Όταν ο αριθμός των κυρίων συνιστωσών K είναι > 1, ο όρος που απαιτείται για να εξασφαλιστεί η εκλογή των σωστών μεταβλητών είναι πολύ δύσκολο να επαληθευτεί στην πράξη. Θα ήταν χρήσιμο να διερευνηθούν και άλλες προσεγγίσεις για πολλαπλές συνιστώσες. Με μεγάλους αριθμούς υψηλά συσχετισμένων χαρακτηριστικών, είναι σημαντικό να γίνει κατανοητό πότε μπορεί και πότε δεν μπορεί να γίνει απομόνωση των σημαντικών βασικών χαρακτηριστικών.

- Το μοντέλο κρυφής μεταβλητής που χρησιμοποιείται σε αυτή την εργασία είναι μια λογική αφετηρία, αλλά δεν μπορεί να είναι ρεαλιστικό στην πράξη. Θα μπορούσε κάποιος να έχει μια κατάσταση στην οποία η απόκριση είναι οριακά ανεξάρτητη από

τις ενεργές μεταβλητές πρόβλεψης, ακόμη και από κοινού εξαρτημένη από αυτά. Μια άλλη κατάσταση θα είχε όλες τις μεταβλητές πρόβλεψης οριακά εξαρτημένες από την απόκριση, ενώ ένα σύνολο είναι ανεξάρτητο από την απάντηση που δόθηκε στο υπόλοιπο των μεταβλητών πρόβλεψης. Σε αυτές τις περιπτώσεις, η διαδικασία των κυρίων συνιστωσών με επίβλεψη θα αποτύγχανε.

- Το μοντέλο απόκρισης που μελετάται σε αυτή την εργασία είναι ένα απλό γραμμικό (ή γενικευμένα γραμμικό) μοντέλο. Θα ήταν χρήσιμο να εξεταστεί κατά πόσον οι κύριες συνιστώσες με επίβλεψη αποδίδουν καλά όταν η απόκριση είναι μια πιο σύνθετη λειτουργία των κρυμμένων παραγόντων. Κάποιος θα μπορούσε να χρησιμοποιήσει threshold γραμμικής συσχέτισης (όπως περιγράφεται στο άρθρο), αλλά στη συνέχεια να χρησιμοποιήσει μια spline βάση (αντί για γραμμική βάση) στο μοντέλο απόκρισης. Στην πράξη, έχουμε διαπιστώσει ότι μια φυσική κυβική spline βάση με δύο ή τρεις κόμβους μπορεί να συλλάβει απλές μη γραμμικότητες στη λειτουργία της απόκρισης.

- Μια βασική πτυχή της μεθόδου είναι η προεπιλογή των χαρακτηριστικών σύμφωνα με τη συσχέτισή τους με το αποτέλεσμα. Αυτό ελαχιστοποιεί την επίδραση ενός μεγαλύτερου αριθμού θορυβωδών χαρακτηριστικών στο μοντέλο πρόβλεψης. Είναι πιθανό ότι αυτή η προεπιλογή μπορεί να χρησιμοποιηθεί αποτελεσματικά με άλλες μεθόδους παλινδρόμησης, όπως τα μερικά ελάχιστα τετράγωνα και η ridge παλινδρόμηση. Η εργασία είναι επικεντρωμένη στις κύριες συνιστώσες με επίβλεψη λόγω της εντυπωσιακής τους απλότητας.

- Απαιτείται περαιτέρω εργασία στη ρύθμιση του μοντέλου του Cox, επειδή τα αποτελέσματα εκεί δεν είναι ακόμη αυστηρά.

- Οι κύριες συνιστώσες με επίβλεψη είναι ελκυστικές λόγω της απλότητάς τους. Ωστόσο, όπως αναφέρθηκε νωρίτερα, άλλες μέθοδοι, όπως τα μερικά ελάχιστα τετράγωνα και το lasso, θα μπορούσαν να εφαρμοστούν μετά το thresholding των γονιδίων. Αυτές θα μπορούσαν επίσης να αποδώσουν καλά και αξίζει να διερευνηθούν. Επιπλέον, υπάρχουν άλλες στενά συνδεδεμένες μέθοδοι που θα πρέπει να μελετηθούν και να συγκριθούν με τις κύριες συνιστώσες με επίβλεψη. Αυτές περιλαμβάνουν τις προσεγγίσεις επαρκούς μείωσης διαστάσεων που χρησιμοποιούνται από τους Chiaromonte, Cook και Li (2002) και τον Cook (2004). Το άρθρο αυτό αντιμετωπίζει την επιλογή μεταβλητής πρόβλεψης. Σχετικές εφαρμογές σε δεδομένα μικροσυστοιχειών περιλαμβάνουν αυτές των Antoniadis, Lambert-Lacroix και Leblanc (2003), και Bura και Pfeiffer (2003).

  Η εφαρμογή των κυρίων συνιστωσών με επίβλεψη σε ιατρικό περιβάλλον συζητείται στους Zhao et al. (2006)

3. Εφαρμογή της PCA σε Πραγματικά Σεισμολογικά Δεδομένα

Εισαγωγή στο Πρόβλημα

Πρόκειται για μία βάση δεδομένων με 10.333 (n) εγγραφές και 11 επεξηγηματικές μεταβλητές.

Οι 11 αυτές επεξηγηματικές μεταβλητές που μελετήθηκαν είναι οι εξής :

X_1	**Χρόνια (years)**
X_2	**Νομός (1-54)**
X_3	**Γεωγραφικό μήκος (longitude)**
X_4	**Γεωγραφικό πλάτος (latitude)**
X_5	**Ένταση (1-12)**
X_6	**Απόσταση από το σεισμό (km)**
X_7	**Hyper distance (degrees)**
X_8	**Αζιμούθιο (degrees)**
X_9	**Επίκεντρο, στον άξονα x (τεταγμένη)**
X_{10}	**Επίκεντρο, στον άξονα y (τετμημένη)**
X_{11}	**Βάθος (0-700 km)**

Και η μεταβλητή απόκρισης Y – **Magnitude** με τιμές **0** όταν το μέγεθος του σεισμού είναι **<6.5** και **1** όταν είναι **>6.5**

Εφαρμογή Clementine

ΒΗΜΑΤΑ :

1. Ανοίγουμε το πρόγραμμα Clementine από το Windows Start menu
2. Φορτώνουμε το αρχείο με τις 10.333 εγγραφές και τις 11 επεξηγηματικές μεταβλητές .
3. Τοποθετούμε στο stream canvas ένα type node με σκοπό να διαβαστούν οι τύποι των τιμών των πεδίων. Έτσι, καθορίζεται ο τύπος των δεδομένων (type) για κάθε πεδίο και επιπρόσθετα καθορίζεται η κατεύθυνση (direction) που επιδεικνύει το ρόλο που παίζει κάθε πεδίο στην μοντελοποίση.

Πιο αναλυτικά :

Ο τύπος πληροφορίας για κάθε πεδίο πρέπει να τεθεί πριν τα πεδία χρησιμοποιηθούν στα διάφορα modeling nodes. Το Clementine διακρίνει τους εξής τύπους δεδομένων:

- Εύρος (range)
  Το range χρησιμοποιείται για να περιγράψει συνεχείς αριθμητικές τιμές, ένα σύνολο ή μία κλίμακα 0-100 ή 0.75-1.25. Μία τιμή range μπορεί να είναι ακέραιος, πραγματικός αριθμός ή ημερομηνία/ώρα.
- Διακριτοποίηση (discrete)
  Το discrete χρησιμοποιείται για να περιγράψει αλφαριθμητικές τιμές όταν ένας ακριβής αριθμός διαφορετικών τιμών είναι άγνωστος (π.χ 1,5,8)
- Δίτιμη παράμετρος-λογική παράμετρος τύπου Boolean (flag)
  Το flag χρησιμοποιείται από δεδομένα με δύο μόνο τιμές yes/no ή 0/1 ή 1/2.
- Σύνολο (set)
  Το set χρησιμοποιείται για να περιγράψει δεδομένα με πολλαπλές διακεκριμένες τιμές όπου η καθεμιά αντιμετωπίζεται σαν μονάδα ενός συνόλου, ή διακεκριμένες κατηγορίες όπως small/medium/large.
- Ανένταχτος τύπος (typeless)
  Το typeless χρησιμοποιείται για δεδομένα που δεν εντάσσονται σε καμία από τις παραπάνω κατηγορίες ή για δεδομένα τύπου set με πάρα πολλές διακεκριμένες τιμές. Η επιλογή του τύπου typeless ορίζεται αυτόματα το πεδίο του direction σε none, δηλαδή το πεδίο που δεν μπορεί να χρησιμοποιηθεί σε μοντέλα.

Τα δεδομένα εντάσσονται αρχικά σε μία από τις παραπάνω κατηγορίες με το που εισέρχονται στο σύστημα. Για παράδειγμα, ο discrete τύπος δίνεται προσωρινά σε κατηγορικές μεταβλητές μέχρι να μπορεί να προσδιορισθεί αν πρόκειται για set ή flag τύπο και ο τύπος range δίνεται σε όλες τις αριθμητικές μεταβλητές.

Η τιμή του direction ενός πεδίου σχετίζεται μόνο με τη μοντελοποίηση. Υπάρχουν τέσσερις δυνατές κατευθύνσεις :

- IN: το πεδίο χρησιμοποιείται σαν input, δηλαδή είναι μία τιμή που θα βοηθήσει στην πρόβλεψη
- OUT: το πεδίο χρησιμοποιείται σαν output-στόχος της τεχνικής μοντελοποίησης. Είναι, δηλαδή, το πεδίο που θα προβλέψουμε.
- BOTH: το πεδίο επιτρέπεται να είναι και input και output σε κανόνα συσχέτισης (association rule). Όλες οι άλλες τεχνικές μοντελοποίησης αγνοούν αυτό το πεδίο.
- NONE: το πεδίο δεν χρησιμοποιείται στην μοντελοποίηση

Στο δικό μας dialog box του type node οι τύποι των 11 επεξηγηματικών μεταβλητών ήταν τύπου Range. Η απόκριση y=magnitude αποτελεί τον στόχο πρόβλεψης, δηλαδή ο στόχος μας στην παρούσα μελέτη ορίζεται να είναι με άλλα λόγια το μέγεθος (σφοδρότητα) του σεισμού. Η y λοιπόν είναι τύπου Flag εφόσον είναι δίτιμη με τιμές y=0 (<6.5) και y=1 (>6.5).

Στην δική μας εφαρμογή όσον αφορά την τιμή του direction έχουμε direction input για τις 11 επεξηγηματικές μεταβλητές που το In δείχνει ότι θα χρησιμοποιηθούν σαν

μεταβλητές πρόβλεψης και direction output για το y=magnitude που το Out δείχνει ότι πρόκειται για το πεδίο που θέλουμε να προβλέψουμε. Οι επεξηγηματικές μεταβλητές χωρίζονται σε συνεχείς (range) και κατηγορικές (ordinal). Κατηγορικές είναι οι X_2 και X_5, ενώ συνεχείς είναι όλες οι υπόλοιπες.

4. Τοποθετούμε στο stream canvas ένα Partition node με σκοπό να χωρίσουμε τα δεδομένα σε δεδομένα εκπαίδευσης (training set) και δεδομένα ελέγχου-εξέτασης (test dataset).

   Πιο αναλυτικά :

Σε ένα τυπικό πρόβλημα του data mining, έχουμε ένα σύνολο δεδομένων εκπαίδευσης (training set) στο οποίο γνωρίζουμε την τιμή του αποτελέσματος και τις τιμές των χαρακτηριστικών που μας ενδιαφέρουν, και προσπαθούμε με βάση αυτά τα δεδομένα να κατασκευάσουμε ένα μοντέλο πρόβλεψης. Το μοντέλο αυτό θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης (test set), στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης. Στην περίπτωση τώρα όπου ο αλγόριθμος που εφαρμόζουμε στηρίζεται σε κατασκευή και εκτίμηση μοντέλου, τα δεδομένα διαχωρίζονται σε δύο υποσύνολα: 1) τα δεδομένα εκπαίδευσης (training data) τα οποία χρησιμοποιούνται για την προσαρμογή του μοντέλου και 2) τα δεδομένα ελέγχου (test data) που χρησιμοποιούνται για τον υπολογισμό της γενικευμένης τιμής σφάλματος του τελικά επιλεγμένου μοντέλου. Καθένα από αυτά τα σύνολα θα πρέπει να επιλεγεί ανεξάρτητα. Αυτός είναι ένας τρόπος να μεγιστοποιήσουμε τα δεδομένα που παράγουν το μοντέλο με το οποίο θα ασχοληθούμε στην πραγματικότητα. Αυτό που είναι σημαντικό, είναι ότι το ποσοστό σφάλματος δεν καθορίζεται με βάση κανένα από αυτά τα δεδομένα.
Γενικά, μπορούμε να πούμε ότι η ποιότητα του μοντέλου είναι ανάλογη του όγκου των διαθέσιμων δεδομένων, αν και συχνά βαίνει φθίνουσα όταν ο όγκος του συνόλου εκπαίδευσης υπερβαίνει κάποιο όριο. Επίσης, και η αξιοπιστία της εκτίμησης του σφάλματος είναι ανάλογη του όγκου των δεδομένων ελέγχου. Τα προβλήματα αρχίζουν όταν δεν υπάρχει επαρκής όγκος δεδομένων και επομένως περιορίζεται το ποσό των δεδομένων που μπορεί να χρησιμοποιηθεί ως σύνολο εκπαίδευσης και σύνολο ελέγχου. Σε τέτοια σύνολα δεδομένων ένα μέρος των δεδομένων χρησιμοποιείται για τον έλεγχο και το υπόλοιπο για την εκπαίδευση. Αυτή η διαδικασία ονομάζεται διαδικασία παρακράτησης (holdout procedure) και το δίλημμα που προκύπτει τώρα είναι πώς διαχωρίσουμε το αρχικό σύνολο έτσι ώστε και τα δύο σύνολα να είναι μεγάλα.
Είναι δύσκολο να δώσουμε ένα γενικό κανόνα σχετικά με το πώς επιλέγεται ο αριθμός των παρατηρήσεων που καταχωρείται σε καθένα από αυτά τα τρία σύνολα, καθώς εξαρτάται από το ποσοστό θορύβου στα δεδομένα και το μέγεθος του δείγματος εκπαίδευσης. Μία τυπική διάκριση που χρησιμοποιείται είναι το 75% στο σύνολο εκπαίδευσης και από 25% στα σύνολα ελέγχου αντίστοιχα.

Εφαρμόζουμε λοιπόν τον παραπάνω διάκριση στις δικές μας 10.333 εγγραφές και προκύπτουν:

- Training set : $\approx 75\% \times 10.333 = 7.749,75$ υποδείγματα
- Test set : $\approx 25\% \times 10.333 = 2.583,25$ υποδείγματα

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	Partition	$F-Factor-1	$F-Factor-2
1	0....	1995....	1....	23.738	38.037	3....	153....	160....	95.0...	22....	38.1...	47....	1_Training	-0.226	0.933
2	0....	1995....	2....	21.396	38.636	3....	75.0...	88.0...	314....	22....	38.1...	47....	1_Training	0.580	-1.308
3	0....	1995....	2....	21.434	38.656	3....	74.0...	88.0...	317....	22....	38.1...	47....	1_Training	0.595	-1.334
4	0....	1995....	2....	21.318	38.428	3....	67.0...	81.0...	295....	22....	38.1...	47....	2_Testing	0.385	-1.256
5	0....	1995....	2....	20.874	38.891	3....	127....	136....	308....	22....	38.1...	47....	1_Training	0.983	-0.760
6	0....	1995....	2....	21.441	38.563	3....	66.0...	81.0...	311....	22....	38.1...	47....	1_Training	0.488	-1.369
7	0....	1995....	3....	22.002	37.040	3....	125....	133....	180....	22....	38.1...	47....	1_Training	0.101	0.080
8	0....	1995....	3....	21.727	37.049	3....	126....	134....	192....	22....	38.1...	47....	1_Training	0.191	0.009
9	0....	1995....	3....	21.706	36.964	3....	135....	143....	192....	22....	38.1...	47....	1_Training	0.271	0.100
10	0....	1995....	3....	21.613	37.128	3....	120....	129....	197....	22....	38.1...	47....	1_Training	0.176	-0.081
11	0....	1995....	4....	22.362	38.532	4....	52.0...	70.0...	39.0...	22....	38.1...	47....	1_Training	-1.448	0.344
12	0....	1995....	4....	22.554	38.380	4....	54.0...	72.0...	64.0...	22....	38.1...	47....	1_Training	-1.263	0.195
13	0....	1995....	5....	22.859	36.696	3....	180....	186....	154....	22....	38.1...	47....	2_Testing	0.406	0.804
14	0....	1995....	5....	22.429	37.053	3....	129....	137....	162....	22....	38.1...	47....	1_Training	0.016	0.243
15	0....	1995....	3....	21.936	37.100	3....	118....	127....	183....	22....	38.1...	47....	2_Testing	0.064	-0.006
16	0....	1995....	3....	21.617	37.051	3....	128....	136....	196....	22....	38.1...	47....	2_Testing	0.235	0.002
17	0....	1995....	3....	21.673	37.239	3....	106....	116....	196....	22....	38.1...	47....	1_Training	0.049	-0.211
18	0....	1995....	3....	21.972	37.214	3....	105....	115....	181....	22....	38.1...	47....	1_Training	-0.061	-0.119
19	0....	1995....	6....	22.754	37.975	3....	69.0...	84.0...	106....	22....	38.1...	47....	1_Training	-0.861	0.049
20	0....	1995....	6....	22.500	38.121	3....	44.0...	64.0...	95.0...	22....	38.1...	47....	1_Training	-1.134	-0.105
21	0....	1995....	6....	22.934	37.903	3....	87.0...	99.0...	108....	22....	38.1...	47....	1_Training	-0.701	0.203
22	0....	1995....	6....	22.936	37.931	3....	86.0...	98.0...	106....	22....	38.1...	47....	2_Testing	-0.723	0.206
23	0....	1995....	6....	22.933	37.828	3....	90.0...	101....	113....	22....	38.1...	47....	1_Training	-0.645	0.194
24	0....	1995....	6....	23.052	37.790	3....	101....	111....	113....	22....	38.1...	47....	1_Training	-0.552	0.301
25	0....	1995....	2....	21.385	37.841	3....	65.0...	80.0...	238....	22....	38.1...	47....	1_Training	-0.007	-0.880
26	0....	1995....	2....	21.247	37.845	3....	75.0...	88.0...	243....	22....	38.1...	47....	1_Training	0.106	-0.822
27	0....	1995....	2....	21.215	37.924	3....	74.0...	87.0...	250....	22....	38.1...	47....	2_Testing	0.144	-0.880
28	0....	1995....	2....	21.559	37.855	3....	51.0...	70.0...	230....	22....	38.1...	47....	2_Testing	-0.167	-0.948
29	0....	1995....	2....	21.624	37.595	3....	71.0...	85.0...	209....	22....	38.1...	47....	1_Training	-0.153	-0.628
30	0....	1995....	2....	21.838	37.399	3....	86.0...	98.0...	190....	22....	38.1...	47....	1_Training	-0.157	-0.358
31	0....	1995....	1....	21.758	38.094	4....	22.0...	52.0...	252....	22....	38.1...	47....	1_Training	-0.229	-1.337
32	0....	1995....	1....	21.679	38.045	3....	31.0...	56.0...	247....	22....	38.1...	47....	1_Training	-0.205	-1.237
33	0....	1995....	1....	21.570	37.942	3....	45.0...	65.0...	239....	22....	38.1...	47....	1_Training	-0.157	-1.067

5. Τοποθετούμε στο stream canvas ένα Feature selection node  το οποίο συνδέουμε στο Partition node με σκοπό να επιλεγούν για επίπεδο σημαντικότητας α=0.05 οι σημαντικές μεταβλητές.

   Πιο αναλυτικά :

Στο δικό μας πρόβλημα εξόρυξης δεδομένων, όπως συμβαίνει και στην πλειοψηφία των προβλημάτων εξόρυξης δεδομένων, εμπεριέχονται εκατοντάδες πεδία-μεταβλητές τα οποία είναι πιθανόν να χρησιμοποιηθούν με σκοπό την πρόβλεψη. Σαν αποτέλεσμα, χρειάζεται να ξοδευτεί αρκετός χρόνος και προσπάθεια για να εξεταστεί ποια από αυτά τα πεδία πρέπει να συμπεριληφθούν στο μοντέλο. Για να μειώσουμε στο ελάχιστο τις πιθανές επιλογές, ο αλγόριθμος της επιλογής των χαρακτηριστικών (Feature Selection Algorithm) μπορεί να χρησιμοποιηθεί για να προσδιορίσει τα πεδία εκείνα τα οποία είναι πιο σημαντικά για τη δεδομένη ανάλυση.
Η επιλογή χαρακτηριστικών αποτελείται από τρία βήματα :

- **Screening (Κρισάρισμα)**
  Σε αυτό το βήμα απομακρύνονται οι μη σημαντικές και προβληματικές μεταβλητές πρόβλεψης καθώς και εγγραφές, όπως στην περίπτωση που

έχουμε μεταβλητές με πολλές ελλειπούσες τιμές ή μεταβλητές με πολύ μεγάλη ή πολύ μικρή διακύμανση για να τις καθιστά χρήσιμες.

- **Ranking (Στοίχιση)**
  Σε αυτό το βήμα ξεχωρίζονται οι εναπομείναντες μεταβλητές πρόβλεψης και καθορίζονται ranks βασισμένα στην σημαντικότητα.

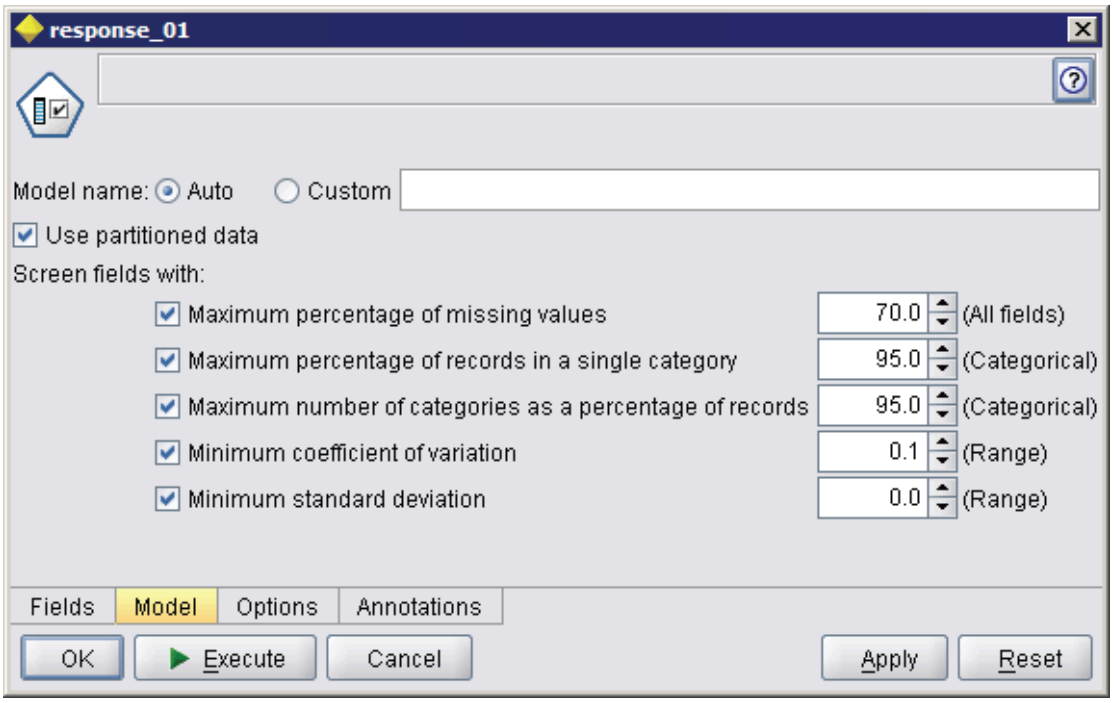
- **Selecting (Επιλογή)**
  Σε αυτό το βήμα αναγνωρίζεται το υποσύνολο των χαρακτηριστικών που θα χρησιμοποιηθεί στα μοντέλα που ακολουθούν κρατώντας μόνο τις πιο σημαντικές μεταβλητές πρόβλεψης και φιλτράροντας ή αποκλείοντας όλες τις υπόλοιπες.

Τα πλεονεκτήματα από την επιλογή χαρακτηριστικών είναι ότι η διαδικασία της μοντελοποίησης απλοποιείται και φυσικά γίνεται ταχύτερη. Μειώνοντας τον αριθμό των πεδίων που χρησιμοποιούνται στο μοντέλο μειώνεται ο χρόνος αξιολόγησης του μοντέλου και επιπρόσθετα αποκτούμε απλούστερα, ακριβέστερα μοντέλα τα οποία μπορούν πολύ πιο εύκολα να εξηγηθούν .

Model tab-Options tab

Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα model tab το οποίο περιλαμβάνει βασικές επιλογές για το μοντέλο καθώς και ρυθμίσεις που επιτρέπουν την εύρεση κριτηρίων για το κρισάρισμα των μεταβλητών πρόβλεψης.

Feature Selection Model tab



Model name: (auto) το όνομα του μοντέλου παράγεται αυτόματα.
Use partitioned data: (v) το τικάρουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι τα δεδομένα μόνο από το training set χρησιμοποιούνται για την κατασκευή του μοντέλου.

Τα πεδία κρισάρονται με την βοήθεια των παρακάτω κτριτηρίων :

- **Μέγιστο ποσοστό ελλειπουσών τιμών**
  Κρισάρει τα πεδία με μεγάλο αριθμό ελλειπουσών τιμών, που προσφέρουν ελάχιστη πληροφορία πρόβλεψης.
- **Μέγιστο ποσοστό εγγγραφών σε μια απλή κατηγορία**
  Κρισάρει τα πεδία τα οποία έχουν πάρα πολλές εγγραφές που να ανήκουν στην ίδια κατηγορία συγκριτικά με τον συνολικό αριθμό των εγγραφών.
- **Μέγιστος αριθμός των κατηγοριών ως ποσοστό των εγγραφών**
  Κρισάρει τα πεδία με πολλές κατηγορίες συγκριτικά με τον συνολικό αριθμό των εγγραφών, δηλαδή εάν ένα μεγάλο ποσοστό των κατηγοριών περιέχει μόνο μία περίπτωση, το πεδίο δεν μπορεί παρά να χρησιμοποιηθεί ελάχιστα.
- **Έλαχιστος συντελεστής μεταβολής**
  Κρισάρει τα πεδία με συντελεστή διακύμανσης μικρότερο ή ίσο από το καθορισμένο ελάχιστο όριο. Εάν η τιμή είναι κοντά στο 0, δεν υπάρχει μεγάλη μεταβλητότητα στις τιμές της μεταβλητής.
- **Ελάχιστη τυπική απόκλιση**
  Κρισάρει τα πεδία με τυπική απόκλιση μικρότερη ή ίση από το καθορισμένο ελάχιστο όριο.

Οι εγγραφές οι οποίες έχουν ελλειπούσες τιμές για το πεδίο στόχου ή ελλειπούσες τιμές για όλες τις μεταβλητές πρόβλεψης, αποκλείονται αυτόματα από όλους τους υπολογισμούς μέσα στο rankings.

Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα options tab το οποίο μας επιτρέπει να καθορίσουμε τις default (εξ'ορισμού) ρυθμίσεις για την επιλογή ή τον αποκλεισμό των πεδίων πρόβλεψης του μοντέλου.
Σε αυτό το βήμα θεωρούμε μία μεταβλητή πρόβλεψης τη φορά για να εξετάσουμε πόσο καλά κάθε μεταβλητή πρόβλεψης ξεχωριστά προβλέπει τη μεταβλητή στόχο. Οι μεταβλητές πρόβλεψης ιεραρχούνται σύμφωνα με το κριτήριο που καθορίζεται από τον πειραματιστή.
Η τιμή σημαντικότητας κάθε μεταβλητής ή διαφορετικά ένα μέτρο το οποίο χρησιμοποιείται για να βάλει σε σειρά τα πεδία ή τα αποτελέσματα σε ποσοστιαία κλίμακα ορίζεται ως $(1-p)$ όπου p είναι η τιμή p value του κατάλληλου στατιστικού τεστ που εξετάζει την σχέση μεταξύ της υποψήφιας μεταβλητής πρόβλεψης και της μεταβλητής στόχο.
Στην δική μας εφαρμογή χρησιμοποιήσαμε τιμή p value βασισμένη στο στατιστικό του Pearson, το Pearson chi-square το οποίο εξετάζει την ανεξαρτησία του στόχου και

της μεταβλητής πρόβλεψης χωρίς να δείχνει τη δύναμη ή την κατεύθυνση οποιασδήποτε υπάρχουσας σχέσης.

Αναλυτικά :

Το Pearson chi-square είναι ένα τεστ ανεξαρτησίας μεταξύ X και Y το οποίο περιλαμβάνει τη διαφορά μεταξύ των παρατηρούμενων και των αναμενόμενων συχνοτήτων. Τα αναμενόμενα κελιά συχνότητας κάτω από την μηδενική υπόθεση υπολογίζονται-εκτιμώνται από τον τύπο

$$\hat{N}_{ij} = N_{i.}N_{.j} / N$$

Κάτω από την μηδενική υπόθεση, το Pearson chi-square συγκλίνει ασυμπτωτικά σε μία κατανομή chi-square $\chi_d{}^2$ με $d = (I-1)(J-1)$ βαθμούς ελευθερίας.

Το p value το οποίο βασίζεται στο Pearson chi-square X^2 υπολογίζεται ως

p value $=\text{Prob}\,(x_d{}^2 \succ X^2)$ όπου το $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(N_{ij} - \hat{N}_{ij}\right)^2 / \hat{N}_{ij}$.

Όπου το X η μεταβλητή πρόβλεψης με την θεώρηση ότι έχουμε I κατηγορίες

Y η μεταβλητή στόχος με J κατηγορίες

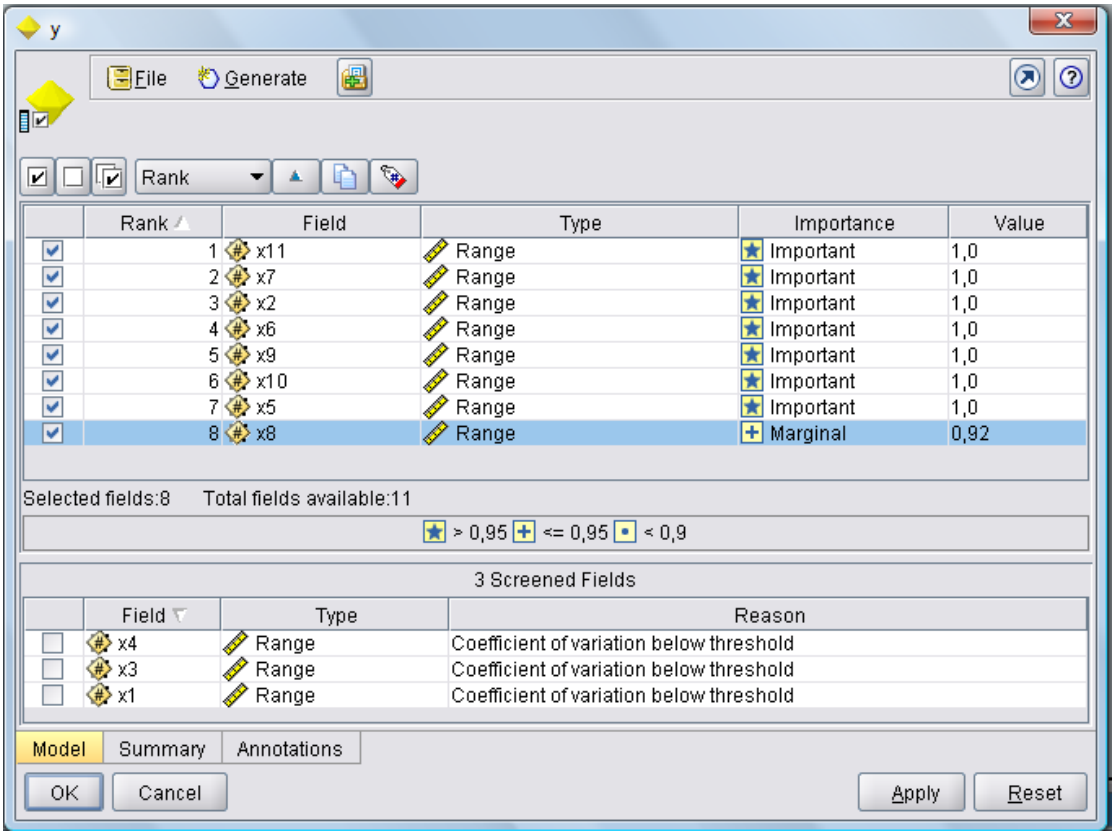
N ο συνολικός αριθμός των περιπτώσεων

N_{ij} ο αριθμός των περιπτώσεων όπου $X = i$ και $Y = j$

$N_{i.}$ ο αριθμός των περιπτώσεων όπου $X = i$ και $N_{i.} = \sum_{j=1}^{J} N_{ij}$

$N_{.j}$ ο αριθμός των περιπτώσεων όπου $Y = j$ και $N_{.j} = \sum_{i=1}^{I} N_{ij}$

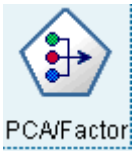
Feature Selection Options tab



Πραγματοποιήσαμε ξανά το βήμα αυτό βασιζόμενοι στο Likelihood-ratio chi-square, το οποίο είναι παρόμοιο με το Pearson chi-square, αλλά εξετάζει επιπλέον την ανεξαρτησία μεταξύ στόχου και μεταβλητής πρόβλεψης. Τα αποτελεσματα τα οποία προέκυψαν από το στατιστικό Likelihood-ratio chi-square συνέπιπταν με αυτά του Pearson chi-square στατιστικού.

Μετά από όλη αυτή τη διαδικασία επιλογής μεταβλητών παρατηρούμε ότι οι επεξηγηματικές μου μεταβλητές από 11 που ήταν αρχικά μειώνονται στις 7.
Για επίπεδο σημαντικότητας α=0.05 οι σημαντικές 7 αυτές μεταβλητές με *p* value =0.0 παρουσιάζονται παρακάτω:

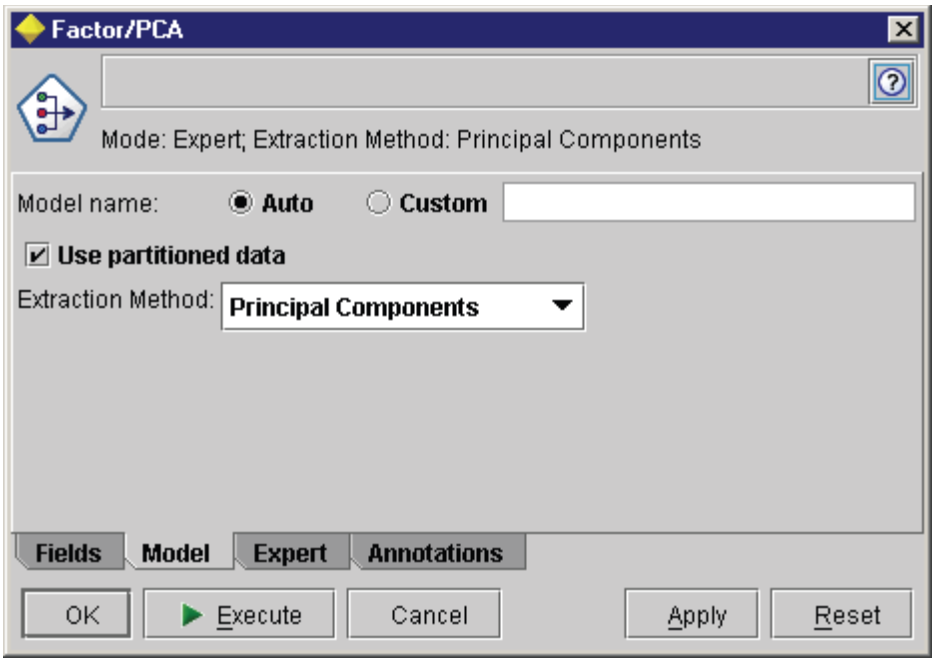
X_2	**Νομός (1-54)**
X_5	**Ένταση (1-12)**
X_6	**Απόσταση από το σεισμό (km)**
X_7	**Hyper distance (degrees)**
X_9	**Επίκεντρο, στον άξονα x (τεταγμένη)**
X_{10}	**Επίκεντρο, στον άξονα y (τετμημένη)**
X_{11}	**Βάθος (0-700 km)**

1. Τοποθετούμε στο stream canvas ένα PCA/Factor node 

το οποίο συνδέουμε με το Partition node έτσι ώστε να μειώσουμε την πολυπλοκότητα των δεδομένων μας.

Model tab/Options tab

PCA/Factor Model tab

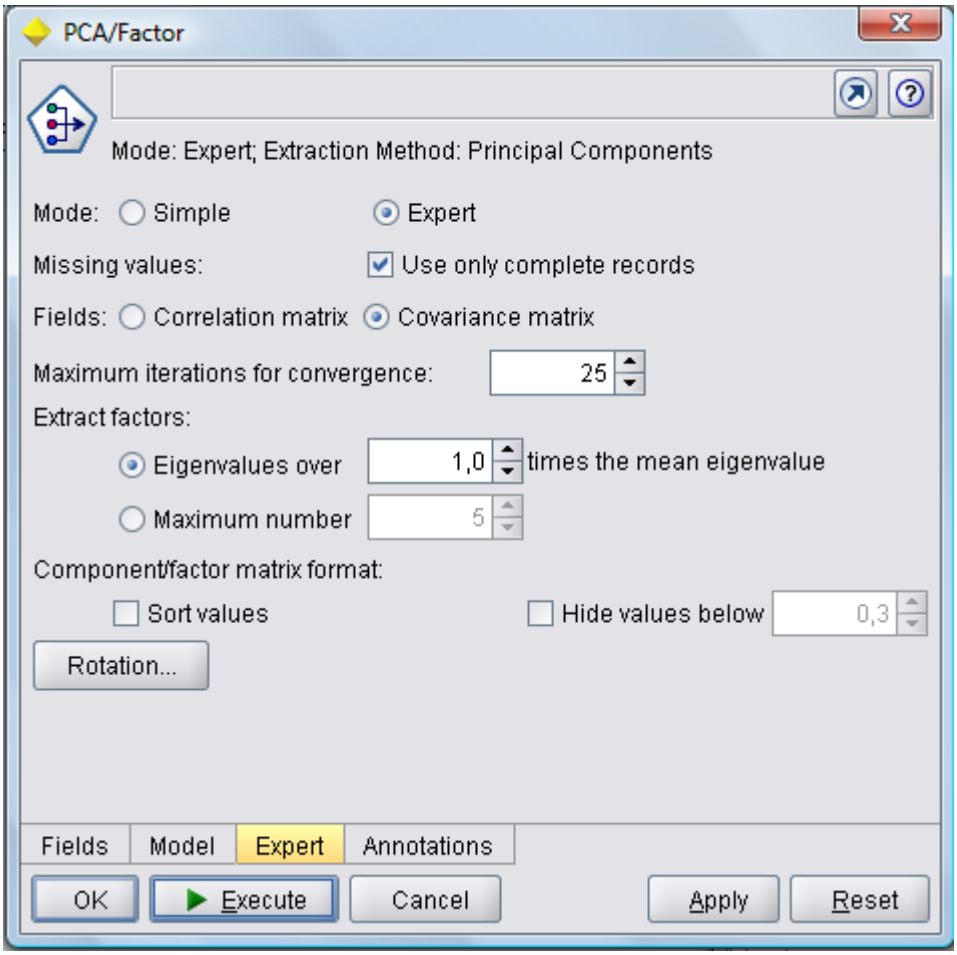


Model name: (auto) το όνομα του μοντέλου παράγεται αυτόματα.
Use partitioned data: (v)

Στη συγκεκριμένη εφαρμογή, για να μειώσουμε τον αριθμό των δεδομένων, θα χρησιμοποιήσουμε ανάλυση κυρίων συνιστωσών (PCA) η οποία βρίσκει συνιστώσες που συνοψίζουν τα πεδία εφαρμογής.

Με βάση τις γνώσεις μας στην PCA προτιμάμε την επιλογή Expert αντί της Simple ώστε να τελειοποιήσουμε τη διαδικασία εκπαίδευσης.

PCA/Factor Expert tab

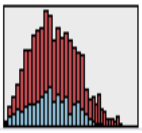
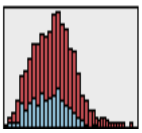


Missing values: δεν έχουμε ελλειπούσες τιμές οπότε τικάρουμε το κουτί
Fields: χρησιμοποιούμε τον πίνακα συνδιακύμανσης των πεδίων εισαγωγής στην εκτίμηση του μοντέλου
Maximum iterations for convergence: προσδιορίζουμε το μέγιστο αριθμό επαναλήψεων για την εκτίμηση του μοντέλου
Extract factors: **Eigenvalues over**: το μοντέλο θα κρατήσει τις συνιστώσες των οποίων οι τιμές των ιδιοτιμών είναι μεγαλύτερες από τη μέση ιδιοτιμή
Component matrix format: **Hide values below**: Τα αποτελέσματα με τιμές κάτω από το καθορισμένο threshold δεν θα συμπεριλαμβάνονται στον εξαγόμενο πίνακα και έτσι προκύπτει ευκολότερα το πρότυπο πίνακα (pattern)
Rotation: εδώ δεν κάνουμε περιστροφή

PCA/Factor Model Nugget

Ένα PCA/Factor model nugget αντιπροσωπεύει το μοντέλο παραγοντικής ανάλυσης ή ανάλυσης κυρίων συνιστωσών (PCA) που δημιουργήθηκε από ένα PCA/Factor node. Περιέχουν όλες τις πληροφορίες που συλλαμβάνονται από το μοντέλο εκπαίδευσης,

καθώς και πληροφορίες σχετικά με την επίδοση και τα χαρακτηριστικά του μοντέλου. Όταν εκτελούμε ένα stream που περιέχει την εξίσωση του μοντέλου παραγόντων (factor equation model), το node προσθέτει ένα νέο πεδίο για κάθε παράγοντα ή συνιστώσα του μοντέλου. Τα νέα ονόματα των πεδίων προέρχονται από το όνομα του μοντέλου με πρόθεμα $F- και και κατάληξη –n, όπου n είναι ο αριθμός του παράγοντα ή της συνιστώσας. Αν, για παράδειγμα, το μοντέλο μας ονομάζεται Factor και περιέχει δύο παράγοντες, τα νέα πεδία θα ονομάζονταν $F-Factor-1 και $F-Factor-2.

$F-Factor...		Range	-2.096	4.210	0.009	1.002	0.466	--	10333
$F-Factor...		Range	-2.359	5.027	0.003	0.998	0.566	--	10333

Το Model tab για ένα Factor model nugget εμφανίζει την εξίσωση αποτελέσματος του παράγοντα, για κάθε παράγοντα. Τα αποτελέσματα των παραγόντων ή των συνιστωσών υπολογίζονται πολλαπλασιάζοντας κάθε τιμή του πεδίου εισαγωγής με το συντελεστή του και προσθέτοντας τα αποτελέσματα.

Equation For Factor-1

-0,0003104 * x1 +
0,000178 * x2 +
-0,000006354 * x3 +
0,0000009887 * x4 +
-0,00003652 * x5 +
0,004406 * x6 +
0,004467 * x7 +
0,006715 * x8 +
0,0001272 * x9 +
-0,000102 * x10 +
0,0007681 * x11 +
+ -1,668

Equation For Factor-2

-0,0001728 * x1 +
0,0003154 * x2 +
0,0000547 * x3 +
-0,000004137 * x4 +
-0,00006644 * x5 +
0,005106 * x6 +
0,005035 * x7 +
-0,006761 * x8 +
0,00003392 * x9 +
-0,00003521 * x10 +
0,0003822 * x11 +
+ 0,3148

Στη συνέχεια δίνονται ορισμένες λεπτομερείς πληροφορίες πάνω στο εκτιμώμενο μοντέλο και στην απόδοσή του.

Communalities

	Raw		Rescaled	
	Initial	**Extraction**	**Initial**	**Extraction**
x1	216,296	16,413	1,000	,076
x2	191,593	14,845	1,000	,077
x3	1,361	,321	1,000	,236
x4	1,562	,002	1,000	,001
x5	2,310	,649	1,000	,281
x6	5494,006	5420,852	1,000	,987
x7	5457,422	5420,029	1,000	,993
x8	11028,850	11027,846	1,000	1,000
x9	17,891	2,348	1,000	,131
x10	18,476	1,564	1,000	,085
x11	714,592	96,661	1,000	,135
Extraction Method: Principal Component Analysis.				

Communalities: εμφανίζουν το ποσοστό της διακύμανσης κάθε πεδίου που εξηγείται από τους παράγοντες ή τις συνιστώσες. Το *Initial* δίνει τα αρχικά communalities με το πλήρες σύνολο των παραγόντων (το μοντέλο ξεκινά με τόσους παράγοντες όσα και τα πεδία εισαγωγής (input fields), και το *Extraction* δίνει τα communalities που βασίζονται στο σύνολο των παραγόντων που διατηρούνται.

Total Variance Explained

	Component	Initial Eigenvalues(a)			Extraction Sums of Squared Loadings		
		Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
Raw	1	11736,447	50,710	50,710	11736,447	50,710	50,710
	2	10265,084	44,352	95,062	10265,084	44,352	95,062
	3	715,730	3,092	98,155			
	4	191,973	,829	98,984			
	5	169,651	,733	99,717			
	6	44,775	,193	99,911			
	7	16,153	,070	99,980			
	8	1,709	,007	99,988			
	9	1,357	,006	99,994			
	10	1,208	,005	99,999			
	11	,271	,001	100,000			
Rescaled	1	11736,447	50,710	50,710	2,037	18,520	18,520
	2	10265,084	44,352	95,062	1,965	17,864	36,384
	3	715,730	3,092	98,155			
	4	191,973	,829	98,984			
	5	169,651	,733	99,717			
	6	44,775	,193	99,911			
	7	16,153	,070	99,980			
	8	1,709	,007	99,988			
	9	1,357	,006	99,994			
	10	1,208	,005	99,999			
	11	,271	,001	100,000			

Extraction Method: Principal Component Analysis.

a. When analyzing a covariance matrix, the initial eigenvalues are the same across the raw and rescaled solution.

Ολική διακύμανση που εξηγείται (Total variance explained): εμφανίζει τη συνολική διακύμανση που εξηγείται από τους παράγοντες του μοντέλου. Οι αρχικές ιδιοτιμές (*Initial Eigenvalues*) δείχνουν τη διακύμανση που εξηγείται από το πλήρες σύνολο των αρχικών παραγόντων. Τα ποσά εξαγωγής τετραγωνικών loadings (*Extraction Sums of Squared Loadings*) δείχνουν τη διακύμανση που εξηγείται από τους παράγοντες που διατηρούνται στο μοντέλο. Τα ποσά περιστροφής τετραγωνικών loadings (*Rotation Sums of Squared Loadings*) δείχνουν τη διακύμανση που εξηγείται από τους παράγοντες που έχουν υποστεί περιστροφή. Για λοξές περιστροφές, τα ποσά περιστροφής τετραγωνικών loadings δείχνουν μόνο τα ποσά των τετραγωνικών loadings, και δεν δείχνουν ποσοστά διακύμανσης.

Component Matrix(a)

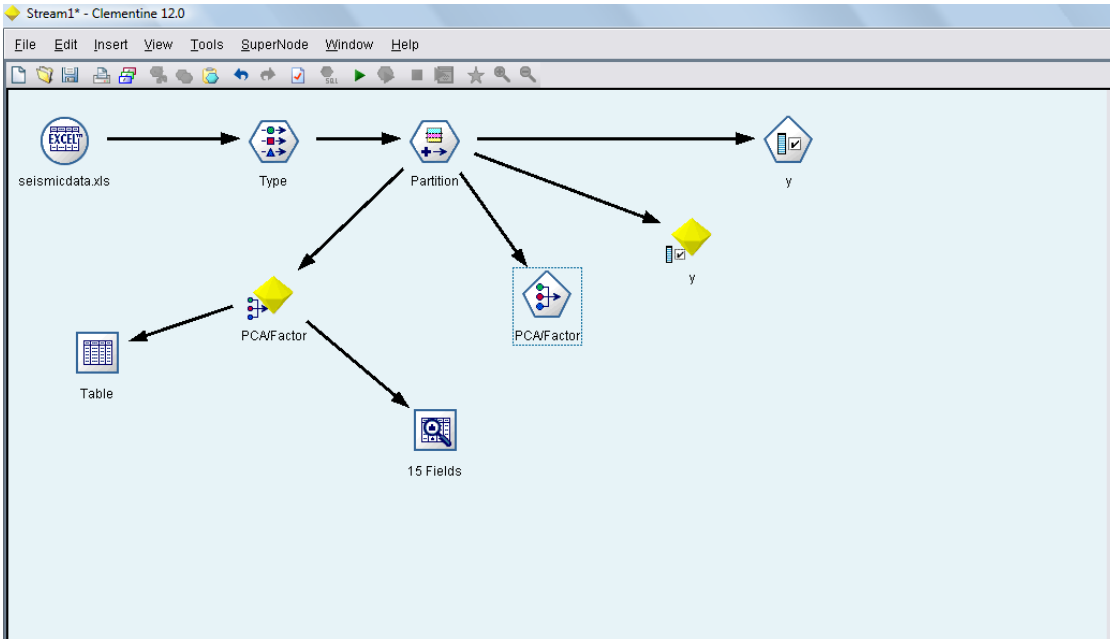
	Raw		Rescaled	
	Component		Component	
	1	2	1	2
x1	-3,642	-1,774	-,248	-,121
x2	2,089	3,238	,151	,234
x3	-,075	,562	-,064	,481
x4	,012	-,042	,009	-,034
x5	-,429	-,682	-,282	-,449
x6	51,712	52,409	,698	,707
x7	52,429	51,684	,710	,700
x8	78,808	-69,405	,750	-,661
x9	1,492	,348	,353	,082
x10	-1,197	-,361	-,279	-,084
x11	9,015	3,923	,337	,147

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Παραγοντικός πίνακας ή πίνακας συνιστωσών (Factor (or component) matrix): δείχνει τις συσχετίσεις μεταξύ των πεδίων εισαγωγής και των παραγόντων που δεν έχουν υποστεί περιστροφή.

Η τελική εικόνα του stream canvas μετά την εφαρμογή των αλγορίθμων με τη βοήθεια του Clementine είναι:


Stream1* - Clementine 12.0
File Edit Insert View Tools SuperNode Window Help
seismicdata.xls
Type
Partition
y
y
PCA/Factor
PCA/Factor
Table
15 Fields

Βιβλιογραφία

1. Alter, O., Brown, P., and Botstein, D. (2000), "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proceedings of the National Academy of Sciences USA*, 97, 10101–10106.

2. Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003), "Effective Dimension Reduction Methods for Tumor Classification Using Gene Expression Data," *Bioinformatics*, 19, 563–570.

3. Baik, J., and Silverstein, J. W. (2004), "Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models," *arXiv:math.ST*.

4. Bair, E., Hastie, T., Paul, D., Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association.*

5. Bair, E., and Tibshirani, R. (2004), "Semi-Supervised Methods to Predict Patient Survival From Gene Expression Data," *PLoS Biology*, 2, 511–522.

6. Bura, E., and Pfeiffer, R. M. (2003), "Graphical Methods for Class Prediction Using Dimension Reduction Techniques on DNA Microarray Data," *Bioinformatics*, 19, 1252–1258.

7. Chiaromonte, F., Cook, R., and Li, B. (2002), "Sufficient Dimension Reduction in RegressionsWith Categorical Predictors," *The Annals of Statistics*, 30, 475–497.

8. Cook, R. (2004), "Testing Predictor Contributions in Sufficient Dimension Reduction," *The Annals of Statistics*, 32, 1062–1092.

9. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

10. Ghosh, D. (2002), "Singular Value Decomposition Regression Models for Classification of Tumors From Microarray Experiments," *Pacific Symposium on Biocomputing*, 7, 18–29.

11. Hastie, T., and Tibshirani, R. (2003), "Efficient Quadratic Regularization for Expression Arrays," technical report, Stanford University.

12. Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001), "Supervised Harvesting of Expression Trees," *Genome Biology*, 2, 1–12.

13. Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Botstein, D., and Brown, P. (2000), "Identifying Distinct Sets of Genes With Similar Expression Patterns via 'Gene Shaving'," *Genome Biology*, 1, 1–21.

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

15. Hi, H., and Gui, J. (2004), "Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data," *Bioinformatics,* 5, 1208–1215.

16. Johnstone, I., and Lu, A. Y. (2006), "Sparse Principal Components Analysis," *Journal of the American Statistical Association.*

17. Kneip, A., and Utikal, K. J. (2001), "Inference for Density Families Using Functional Principal Component Analysis," *Journal of the American Statistical Association,* 96, 519–542.

18. Lu, A. Y. (2002), "Sparse Principal Components Analysis for Functional Data," technical report, Stanford University..

19. Nguyen, D., and Rocke, D. (2002), "Partial Least Squares Proportional Hazard Regression for Application to DNA Microarrays," *Bioinformatics,* 18, 1625–1632.

20. Paul, D. (2005), "Nonparametric Estimation or Parametric Components," Ph.D. thesis, Stanford University.

21. Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., and Staudt, L. M. (2002), "The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large B-Cell Lymphoma," *The New England Journal of Medicine,* 346, 1937–1947.

22. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2001), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Science,* 99, 6567–6572.

23. Zhao, H., Ljungberg, B., Grankvist, K., Rasmuson, T., Tibshirani, R., and Brooks, J. (2006), "Gene Expression Profiling Predicts Survival in Conventional Renal Cell Carcinoma," *PLoS Medicine,* 3, 1–10.

24. Clementine 12.0 Modelling Nodes, pp.44-48 and 216-224.

25. Shlens, J. (2005), "A Tutorial on Principal Components Analysis," Systems Neurobiology Laboratory, *Salk Insitute for Biological Studies* La Jolla, CA 92037 and *Institute for Nonlinear Science,* University of California, San Diego La Jolla, CA 92093-0402.

26. Καρλής, Δ. (2001), "Στοιχεία Πολυμεταβλητής Ανάλυσης Δεδομένων," Πανεπιστήμιο Αιγαίου pp. 39-59.

27. Παπαγρηγορίου, Ν. (2001), "Εφαρμογή της Στατιστικής Ανάλυσης των Κυρίων Συνιστωσών σε πειραματικά δεδομένα που καταγράφηκαν μετά από ωσμωτική αφυδάτωση ακτινιδίων," Τεχνολογικό Εκπαιδευτικό Ίδρυμα Θεσσαλονίκης pp. 29-42.

28. Μάσσου, Ε. (2008), "Αλγόριθμοι Εξόρυξης Πληροφορίας και Στατιστική Ανάλυση Δεδομένων," Εθνικό Μετσόβιο Πολυτεχνείο pp. 1-4.