ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΑΛΥΣΗ ΓΟΝΙΔΙΑΚΩΝ ΕΚΦΡΑΣΕΩΝ ΓΙΑ ΤΗ ΜΕΛΕΤΗ ΤΟΥ ΧΡΟΝΟΥ ΕΠΙΒΙΩΣΗΣ ΓΥΝΑΙΚΩΝ ΜΕ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ

Λουκία Δ. Σπινέλη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Σεπτέμβριος 2010

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- **Πολίτης Κωνσταντίνος (Επιβλέπων)**
- **Κατέρη Μαρία**
- **Κούτρας Μάρκος**

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

GENE EXPRESSION ANALYSIS FOR THE STUDY OF THE SURVIVAL TIME OF WOMEN WITH BREAST CANCER

By

Loukia D. Spineli

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master in Science in Applied Statistics

Piraeus

September 2010

Αφιερωμένο
στους γονείς μου
Δημήτρη και Σταυρούλα,
και στην αδερφή μου
Μαρία

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κο Κωνσταντίνο Πολίτη για την συνεργασία μας, την πολύτιμη βοήθειά του και την υπομονή που έδειξε καθ' όλη τη διάρκεια της συγγραφής της παρούσας εργασίας. Επίσης, εκφράζω τις ευχαριστίες μου στην κα Μαρία Κατέρη και κο Μάρκο Κούτρα για τη συμμετοχή τους στην Τριμελή Εξεταστική Επιτροπή. Ακόμα, θα ήθελα να ευχαριστήσω την Αναστασία Ελευθεράκη για τον χρόνο που διέθεσε, τις επισημάνσεις της και τη σημαντική βοήθεια που μου προσέφερε ιδιαίτερα για την συγγραφή της ενότητας Εφαρμογή στην ομαδοποίηση των εκφράσεων γονιδίων.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την υπομονή και κατανόηση που έδειξαν καθ' όλη τη διάρκεια της συγγραφής της διπλωματικής εργασίας.

Περίληψη

Χωρίς αμφιβολία η μελέτη της γονιδιακής έκφρασης αποτελεί την βάση για την κατανόηση διαφόρων ασθενειών, που προκαλούνται από γενετικές διαταραχές. Η ανακάλυψη της τεχνολογίας των μικροσυστάδων μαζί με την ανάπτυξη της πληροφορικής βοήθησαν σημαντικά τόσο στην απομόνωση αμέτρητων γονιδίων όσο και στην παρακολούθηση των ποικίλων και πολύπλοκων λειτουργιών τους στο κύτταρο. Πλέον, η μελέτη της επιβίωσης των ασθενών δεν περιορίζεται μόνο στην καταγραφή του κλινικοπαθολογικού και ιστολογικού προφίλ των ασθενών, αλλά συνδέεται στενά και με την συλλογή δεδομένων που αφορούν εκφράσεις γονιδίων, διότι επηρεάζουν σημαντικά τη διαμόρφωση αυτού του προφίλ.

Σκοπός της παρούσας εργασίας είναι η ανάλυση δεδομένων που χρησιμοποιούνται στην βιοϊατρική με στόχο τη μελέτη του χρόνου επιβίωσης σε γυναίκες με καρκίνο του μαστού. Αυτό γίνεται συνήθως με τον εντοπισμό και την ανάλυση βιολογικών καρκινικών δεικτών, βάσει των γονιδιακών εκφράσεων (gene expressions) των όγκων των ασθενών. Στην εργασία θα μελετηθούν πραγματικά δεδομένα που υπάρχουν για μια σειρά γονιδίων που μετρήθηκαν σε ιστούς ενός συνόλου ασθενών, αρχικά με την χρήση αρχικά μεθόδων ομαδοποίησης των γονιδίων, και στη συνέχεια με εξέταση του κατάλληλου μοντέλου από την ανάλυση επιβίωσης, η επίδραση που έχουν οι τιμές των γονιδιακών εκφράσεων στο χρόνο επιβίωσης των ασθενών.

Οι μέθοδοι ομαδοποίησης των γονιδίων και τα μοντέλα επιβίωσης μελετήθηκαν με την εφαρμογή του προγράμματος εντολών **R**.

Abstract

Undoubtedly the study of gene expression is the basis for understanding various diseases caused by genetic disorders. The discovery of microarray technology with the development of IT has helped significantly not only in the isolation of countless genes but also in the monitoring of the diverse and complex functions in the cell. Consequently, the study of survival of patients is not only a record of clinicopathological and histological profile of patients, but closely linked with the collection of gene expression data, because they affect significantly the shape of this profile.

The purpose of this study is to analyze the data used in biomedical research to study the survival time in women with breast cancer. This is usually done by identifying and analyzing biological tumor markers, based on the gene expressions of patients' tumor. In this study will be considered real data that exist for a number of genes measured in the tissues of a patient sample, initially appling some methods for clustering genes and then examining the influence of gene expressions in the survival time of patients using the appropriate model from the survival analysis.

The methods of gene clustering and the models of survival analysis were studied with the implementation of the programming language R.

Περιεχόμενα

Κατάλογος Πινάκων

επιζώντων την χρονική στιγμή t_j και το στρώμα g (Στρωματοποιημένος έλεγχος).

5.2.1 Καθορισμός επιπέδων δύο ψευδομεταβλητών.

5.2.2 Καθορισμός των k επιπέδων ενός παράγοντα.

6.1 Πλήθος και ποσοστό ασθενών που απεβίωσαν ή επιβίωσαν με ασθένεια για κάθε βιοδείκτη (ER-mRNA, PgR-mRNA και HER2).

6.2 Πλήθος και ποσοστό ασθενών για κάθε κατάσταση των υπό μελέτη βιοδεικτών (ER-mRNA, PgR-mRNA και HER2).

6.3a Συντελεστές παλινδρόμησης του μοντέλου **(6.3.1)** για την OS.

6.3b Συντελεστές παλινδρόμησης του μοντέλου **(6.3.1)** για την DFS.

6.4a Συντελεστές παλινδρόμησης του μοντέλου **(6.3.2)** για OS.

6.4b Συντελεστές παλινδρόμησης του μοντέλου **(6.3.2)** για DFS.

6.5a Συντελεστές παλινδρόμησης του μοντέλου **(6.3.3)** για OS.

6.5b Συντελεστές παλινδρόμησης του μοντέλου **(6.3.3)** για DFS.

6.6 Logrank έλεγχος επίδρασης του βιοδείκτη HER2 στην OS.

6.7 Logrank έλεγχος επίδρασης του βιοδείκτη HER2 στην DFS.

6.8 Logrank έλεγχος επίδρασης του βιοδείκτη PgR στην OS.

6.9 Συντελεστές παλινδρόμησης του μοντέλου **(6.4.1)** για OS.

6.10a Συντελεστές παλινδρόμησης του μοντέλου **(6.4.2)** για OS.

6.10b Συντελεστές παλινδρόμησης του μοντέλου **(6.4.2)** για DFS.

6.11 Gehan-Wilcoxon έλεγχοι σύγκρισης των δύο χημειοθεραπειών για κάθε επίπεδο του βιοδείκτη HER2 για OS.

6.12 Gehan-Wilcoxon έλεγχοι σύγκρισης των δύο χημειοθεραπειών για κάθε επίπεδο του βιοδείκτη HER2 για DFS.

6.13 Gehan-Wilcoxon έλεγχοι σύγκρισης των δύο χημειοθεραπειών για κάθε επίπεδο του βιοδείκτη PgR για OS.

6.14 Συντελεστές παλινδρόμησης του μοντέλου **(6.5.1)** για OS.

6.15 Συντελεστές παλινδρόμησης του μοντέλου **(6.5.2)** για DFS.

6.16 Πολυμεταβλητή ανάλυση παλινδρόμησης του Cox.

6.17 Ολικός έλεγχος και τοπικοί έλεγχοι των scaled Schoenfeld υπολοίπων για

Κατάλογος Σχημάτων

Εισαγωγή

Στην μελέτη επίδρασης της χημειοθεραπείας στην επιβίωση των ασθενών με καρκίνο του μαστού συνηθίζεται η παρακολούθηση της έκφρασης των βιοδεικτών ER (*Estrogen Receptor*), PgR (*Progesterone Receptor*) και HER2 (*Human Epidermal growth factor Receptor 2*), οι οποίοι βοηθούν σημαντικά στην επιλογή της κατάλληλης χημειοθεραπείας και θεραπείας καταστολής ορμονών (*hormone-suppress therapy*) για την αντιμετώπιση και την πρόληψη επανεμφάνισης καρκινικών κυττάρων στο στήθος.

Ο βιοδείκτης ER είναι ένας υποδοχέας που ανήκει στην οικογένεια πυρηνικών ορμονικών υποδοχέων (*nuclear hormone receptors family*) και βασική λειτουργία του είναι να ρυθμίζει την έκφραση γονιδίων που μπορούν να ελέγξουν τα οιστρογόνα (*estrogens*). Αυτός ο υποδοχέας δεσμεύει τα οιστρογόνα και ενεργοποιεί τα αντίστοιχα γονίδια για την σύνθεση συγκεκριμένων πρωτεϊνών που μπορούν να επηρεάσουν τη συμπεριφορά του κυττάρου με πολλούς τρόπους.

Ο βιοδείκτης PgR είναι επίσης ένας υποδοχέας που ανήκει στην οικογένεια πυρηνικών ορμονικών υποδοχέων (*nuclear hormone receptors family*) και ο κύριος ρόλος του είναι να ρυθμίζει την έκφραση γονιδίων που μπορεί να ελέγξει η προγεστερόνη (*progesterone*). Η έκφραση του PgR παρέχει συμπληρωματικές πληροφορίες για τον ER τόσο για την πρόβλεψη της απόκρισης των όγκων σε έναν συνδυασμό θεραπείας καταστολής ορμονών και χημειοθεραπείας όσο και για την επιβίωση των ασθενών. Μερικές μελέτες έχουν αποδείξει ότι ο υποδοχέας PgR είναι ανώτερος προγνωστικός δείκτης σε σχέση με τον υποδοχέα ER λόγω της ασθενούς προγνωστικής δύναμης του τελευταίου.

Τέλος, ο βιοδείκτης HER2 είναι ένα γονίδιο που ανήκει στην οικογένεια ErbB (*epidermal growth factor receptor family*) και ρυθμίζει την ανάπτυξη, την λειτουργία και τον πολλαπλασιασμό του κυττάρου μέσω της σύνθεσης της πρωτεΐνης τυροσινική κινάση (*tyrosine kinase*). Έχει αποδειχθεί ότι η υπερέκφραση αυτού του γονιδίου (δηλαδή, η υπερπαραγωγή της τυροσινικής κινάσης) οδηγεί στην αύξηση των καρκινικών κυττάρων, τα οποία γίνονται πολύ πιο επιθετικά και κατ' επέκταση στην μη απόκριση των καρκινικών κυττάρων στον συνδυασμό χημειοθεραπείας και ορμονοθεραπείας που χορηγείται. Συνεπώς, ο βιοδείκτης HER2 έχει μεγάλη

προγνωστική δύναμη και η έκφρασή του μελετάται εντατικά στις μελέτες για τον καρκίνο του στήθους. Αξίζει να σημειώσουμε ότι ένα φυσιολογικό κύτταρο στο στήθος μπορεί να έχει περίπου 20,000 HER2, ενώ ένα καρκινικό κύτταρο στο στήθος μπορεί να έχει περίπου 1,5 εκατομμύρια HER2. Περίπου το 15% με 20% των καρκινικών κυττάρων του στήθους έχουν υπερεκφρασμένο HER2.

Οι θεραπείες καταστολής ορμονών, όπως η ταμοξιφαίνη (*tamoxifen*) βοηθούν ως πρόληψη έναντι μιας επανεμφάνισης του καρκίνου, διότι μειώνουν τα οιστρογόνα, τα οποία επηρεάζουν την έκφραση των υποδοχέων προγεστερόνης μέσω των υποδοχέων οιστρογόνων, μπλοκάροντας την είσοδό τους στους υποδοχείς ορμονών (*hormone receptors*) του κυττάρου με αποτέλεσμα τα καρκινικά κύτταρα να μην αναπτύσσονται. Οι χημειοθεραπείες, όπως η πακλιταξέλη (*paclitaxel*), είναι αναστολείς που χρησιμοποιούνται στην αντιμετώπιση του καρκίνου σταματώντας την κυτταρική διαίρεση με αποτέλεσμα να πεθαίνουν τα κύτταρα που δεν μπορούν να διαιρεθούν.

Η μελέτη της έκφρασης των βιοδεικτών ER και PgR δείχνει εάν αυτοί οι υποδοχείς επηρεάζουν τον καρκινικό όγκο. Ο καρκινικός όγκος που είναι ευαίσθητος στους υποδοχείς ορμονών δεν αυξάνεται γρήγορα, οπότε υπάρχει μεγάλη πιθανότητα να ανταποκριθεί στη χημειο-ορμονοθεραπεία (*chemohormonotherapy*) σε αντίθεση με τον καρκίνο που δεν παρουσιάζει ευαισθησία στις ορμόνες με αποτέλεσμα να μην μπορεί να ανταποκριθεί στη συγκεκριμένη χημειο-ορμονοθεραπεία. Σε αυτήν την περίπτωση επιβάλλεται να μελετήσουμε την έκφραση άλλων υποδοχέων στα καρκινικά κύτταρα, όπως του HER2 κυρίως, για να εντοπίσουμε την κατάλληλη θεραπευτική αγωγή.

Η απόκριση του καρκίνου σε μία συγκεκριμένη θεραπευτική αγωγή συμβολίζεται με ένα θετικό πρόσημο στον υπό μελέτη βιοδείκτη, ενώ η μη απόκριση του καρκίνου σε μία συγκεκριμένη θεραπευτική αγωγή συμβολίζεται με ένα αρνητικό πρόσημο. Απόκριση καρκίνου σε μία θεραπεία σημαίνει ότι μειώνεται η παραγωγή καρκινικών κυττάρων και συνεπώς μειώνεται το μέγεθος του όγκου. Σε έναν καρκινικό όγκο η εμφάνιση των ER- και PgR+, όπως επίσης και των ER+ και PgR+ είναι ένδειξη ότι ο συγκεκριμένος όγκος έχει ανταποκριθεί ικανοποιητικά στην θεραπεία. Ωστόσο, συμβαίνει το αντίθετο σ' έναν καρκινικό όγκο με ER- και PgR-. Στην περίπτωση αυτή συνηθίζεται η μελέτη του βιοδείκτη HER2. Έχει αποδειχθεί ότι ασθενείς με ER+ και HER2+ μπορούν να ωφεληθούν περισσότερο από μία θεραπεία σε σύγκριση

με τους ασθενείς με ER- και HER2+, διότι στους πρώτους τα καρκινικά κύτταρα ανταποκρίνονται πιο ικανοποιητικά στην θεραπεία αυτή.

Στην παρούσα διπλωματική εργασία εξετάζουμε την επίδραση της χημειοθεραπείας με ή χωρίς πακλιταξέλη στον συνολικό χρόνο επιβίωσης και στον χρόνο επιβίωσης χωρίς ασθένεια των ασθενών με καρκίνο στο μαστό, που υπέστησαν χειρουργική αφαίρεση του όγκου, σε συνάρτηση με τους βιοδείκτες ER, PgR και HER2. Ως χρόνος συνολικής επιβίωσης ορίζεται το χρονικό διάστημα από την παρακολούθηση της ασθενούς έως τον θάνατό της από οποιαδήποτε αιτία, ενώ ως χρόνος επιβίωσης χωρίς ασθένεια ορίζεται το χρονικό διάστημα από την παρακολούθηση της ασθενούς έως την εμφάνιση δεύτερου όγκου οπουδήποτε ή την επανεμφάνιση καρκίνου στο στήθος ή τον θάνατο λόγω καρκίνου στο στήθος ή τον θάνατο από οποιαδήποτε αιτία εκτός του καρκίνου στο στήθος, οποιοδήποτε από τα παραπάνω ενδεχόμενα συμβεί πρώτο. Η σύγκριση της χημειοθεραπείας με πακλιταξέλη έναντι της χημειοθεραπείας χωρίς πακλιταξέλη ενισχύεται λαμβάνοντας υπόψη αρκετές ιστολογικές και κλινικοπαθολογικές (*clinocopathological*) παραμέτρους, όπως το μέγεθος του όγκου σε cm (*tumor size*), την ταξινόμηση του όγκου ανάλογα με την δομή και ανάπτυξη των κυττάρων του (*tumor grade*), το πλήθος θετικών αδενωμάτων (*nodes*), την πρωτεΐνης VEGF και 37 γονίδια ομαδοποιημένα σε τρεις ομάδες.

Για αυτήν την έρευνα η Ελληνική Συνεργαζόμενη Ογκολογική Ομάδα (*Hellenic Cooperative Oncology Group, HeCOG*) πραγματοποίησε μία τυχαιποιημένη προοδευτική κλινική δοκιμή σε 595 συνολικά ασθενείς με καρκίνο στο μαστό για την λήψη είτε χημειοθεραπείας με πακλιταξέλη (E-T-CMF) είτε χημειοθεραπείας χωρίς πακλιταξέλη (E-CMF). Αυτή η κλινική δοκιμή ξεκίνησε τον Οκτώβριο του 1997 και διήρκησε 10 χρόνια. Συγκεκριμένα, ως προς το χρονοδιάγραμμα της θεραπευτικής αγωγής στην πρώτη ομάδα ασθενών χορηγείται πρώτα σε τέσσερις κύκλους *epirubicin* και στην συνέχεια σε τέσσερις κύκλους η E-CMF, ενώ στην δεύτερη ομάδα ασθενών χορηγείται πρώτα σε τρεις κύκλους *epirubicin* και στην συνέχεια σε τρεις κύκλους η E-T-CMF. Οι κύκλοι χημειοθεραπείας καθορίζονται κάθε δύο βδομάδες και στην συνέχεια χορηγείται στους ασθενείς η ορμόνη *G-CSF* (*granulocyte-colony stimulating factor*) για ανάρρωση από την χημειοθεραπεία.

Από τους 367 καρκινικούς όγκους εξάγεται mRNA χρησιμοποιώντας την αλυσίδα αντίδρασης πολυμεράσης RT-PCR (*reverse-transcriptase polymerase chain reaction*). Αυτή η «τεχνική» μας παρέχει κανονικοποιημένα (*normalised*) δεδομένα

εκφράσεων γονιδίων μεγάλου εύρους. Για όλα τα δείγματα όγκων που περιλήφθησαν στην ανάλυση τουλάχιστον το 75% των ευκαρυωτικών (*nucleated*) κυττάρων ήταν κακοήθη (*malignant*). Για την αξιολόγηση της έκφρασης των βιοδεικτών ER και PgR εφαρμόστηκε η αλυσίδα αντίδρασης πολυμεράσης kRT-PCR (*kinetic reverse-transcriptase polymerase chain reaction*). Για την διάκριση αυτών των βιοδεικτών σε θετικούς και αρνητικούς χρησιμοποιήθηκε ως «κατώφλι» (*threshold* ή *cut-off*) το 25^{o} εκατοστημόριο των τιμών των mRNAs. Παρόλο που είναι μέτρια (περίπου 30%-60%) η προγνωστική δύναμη του βιοδείκτη ER για τα οφέλη που μπορεί να προσφέρει μία ορμονοθεραπεία, η πρωτεϊνική του έκφραση θεωρείται το κριτήριο για την επιλογή των ασθενών που θα πάρουν αυτήν την θεραπεία.

Η διπλωματική εργασία αποτελείται από επτά κεφάλαια και πέντε παραρτήματα στο τέλος. Συγκεκριμένα,

- το πρώτο κεφάλαιο περιέχει βασικά στοιχεία βιολογίας, όπως την σύνθεση και την λειτουργία του DNA και RNA, ώστε στην συνέχεια να γίνει κατανοητή η περιγραφή της τεχνολογίας των μικροσυστάδων. Το πρώτο κεφάλαιο ολοκληρώνεται με τις μεθόδους κανονικοποίησης των γονιδιακών εκφράσεων και τις μεθόδους επιλογής των πιο σημαντικών από τα γονίδια που μελετώνται. Το παράρτημα Α αφορά το πρώτο κεφάλαιο και περιέχει κάποιους πίνακες ANOVA.
- το δεύτερο κεφάλαιο περιλαμβάνει τη θεωρία της ομαδοποίησης των εκφράσεων γονιδίων. Συγκεκριμένα παρουσιάζονται οι ιεραρχικές και μη ιεραρχικές μέθοδοι ομαδοποίησης των γονιδιακών εκφράσεων, καθώς επίσης και τα κατάλληλα μέτρα αξιολόγησης της ομαδοποίησης, έτσι ώστε μεταξύ των μεθόδων ομαδοποίησης που εφαρμόσαμε να επιλέξουμε εκείνη την μέθοδο που ομαδοποιεί πιο αξιόπιστα τα δεδομένα μας. Στο τέλος αυτού του κεφαλαίου αναφέρονται κάποιες μέθοδοι αντικατάστασης ελλειπουσών γονιδιακών εκφράσεων.
- το τρίτο κεφάλαιο είναι η εφαρμογή των μεθόδων ομαδοποίησης και αξιολόγησης ομαδοποίησης που περιγράφηκαν στο δεύτερο κεφάλαιο χρησιμοποιώντας το πρόγραμμα **R**. Ουσιαστικά παρουσιάζονται τα πιο σημαντικά αποτελέσματα από την ομαδοποίηση. Οι εντολές που εφαρμόστηκαν για τα αποτελέσματα του τρίτου κεφαλαίου περιγράφονται αναλυτικά στο

παράρτημα Β1. Το παράρτημα Β2 περιέχει συγκεντρωτικά τις εντολές που περιγράφηκαν στο παράρτημα Β1.

- στην αρχή του τετάρτου κεφαλαίου παρουσιάζονται βασικές εισαγωγικές έννοιες της Ανάλυσης Επιβίωσης. Στη συνέχεια μελετάται η μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης χρησιμοποιώντας Kaplan-Meier και Nelson-Aalen εκτιμητή και η σύγκριση των συναρτήσεων επιβίωσης χρησιμοποιώντας τον Log-rank ή Wilcoxon έλεγχο. Το τέταρτο κεφάλαιο ολοκληρώνεται με την μελέτη των στρωματοποιημένων ελέγχων.
- το πέμπτο κεφάλαιο περιλαμβάνει την ημι-παραμετρική εκτίμηση της συνάρτησης επιβίωσης χρησιμοποιώντας το μοντέλο αναλογικού κινδύνου του Cox.. Η παρουσίαση της συνάρτησης της μερικής πιθανοφάνειας του Cox, των ελέγχων υποθέσεων για το διάνυσμα παραμέτρων b του μοντέλου και της εκτίμησης της αναφορικής συνάρτησης κινδύνου είναι απαραίτητη για την κατανόηση του μοντέλου αναλογικού κινδύνου του Cox. Στην συνέχεια μελετάται η αξιολόγηση της υπόθεσης του αναλογικού κινδύνου τόσο γραφικά όσο και μέσω της ανάλυσης των υπολοίπων. Το παρόν κεφάλαιο ολοκληρώνεται με την περιγραφή της διαδικασίας επιλογής μεταβλητών για το μοντέλο του Cox και με την παρουσίαση μεθόδων κατάλληλων για την επεξεργασία δεδομένων υψηλής διάστασης, έτσι ώστε να διευκολυνθεί η εφαρμογή της Ανάλυσης Επιβίωσης.
- το έκτο κεφάλαιο περιέχει τα πιο σημαντικά αποτελέσματα από την εφαρμογή της Ανάλυσης Επιβίωσης χρησιμοποιώντας το πρόγραμμα **R**. Το παράρτημα Γ1 περιέχει αναλυτικά τις εντολές που χρησιμοποιήθηκαν στο πέμπτο κεφάλαιο, ενώ το παράρτημα Γ2 περιέχει συγκεντρωτικά τις εντολές του παραρτήματος Γ1.
- το έβδομο κεφάλαιο είναι ο επίλογος της διπλωματικής εργασίας, καθώς περιέχει συγκεντρωμένα όλα τα συμπεράσματα από την ομαδοποίηση των 37 γονιδίων και την ανάλυση επιβίωσης των ασθενών, σύμφωνα με τα αποτελέσματα του τρίτου και του έκτου κεφαλαίου αντίστοιχα.

Κεφάλαιο 1

Τεχνολογία Μικροσυστάδων

1.1 Εισαγωγή

Για την κατανόηση της τεχνολογίας μικροσυστάδων κρίνεται απαραίτητο να αναφερθούμε στην σύνθεση, την λειτουργία και τον ρόλο του DNA στην ανάπτυξη και λειτουργία οποιουδήποτε ζωντανού οργανισμού. Η αναγκαιότητα της αποκωδικοποίησης του DNA σε συνδυασμό με τις συνεχείς προόδους της τεχνολογίας και πληροφορικής οδήγησε στην δημιουργία των μικροσυστάδων. Στο παρόν κεφάλαιο περιγράφουμε την διαδικασία δημιουργίας εικόνων μικροσυστάδων που είναι το «προϊόν» της τεχνολογίας συστάδων, καθώς επίσης και την ερμηνεία τους. Στην συνέχεια περιγράφουμε τις πιο συνηθισμένες στην εφαρμογή μεθόδους κανονικοποίησης των ποσοτικοποιημένων αποτελεσμάτων των εικόνων, δηλαδή των γονιδιακών εκφράσεων και ολοκληρώνουμε το κεφάλαιο με την παρουσίαση και ανάλυση των πιο γνωστών μεθόδων εντοπισμού των σημαντικών γονιδίων. Ο εντοπισμός των σημαντικών γονιδίων αποτελεί το πιο κρίσιμο στάδιο στην επεξεργασία των γονιδιακών εκφράσεων πριν την εφαρμογή της κατάλληλης μεθόδου ομαδοποίησης τους.

1.2 Η σημασία των μικροσυστάδων στην μελέτη της γονιδιακής έκφρασης

Από τα πιο σημαντικά επιτεύγματα της ιατρικής είναι η ανακάλυψη της γονιδιακής ακολουθίας (*genome sequence*) τόσο του ανθρώπινου γονιδιώματος όσο και αρκετών άλλων ευκαριωτικών γονιδιωμάτων. Ουσιαστικά, γονιδιακή ακολουθία είναι η μεταφορά πληροφορίας που παίρνουμε από το DNA στον ηλεκτρονικό υπολογιστή. Η κατανόηση της λειτουργίας των γονιδίων (*genomes function*) είναι στόχος ενός καινούριου ερευνητικού τομέα γνωστού ως λειτουργικότητα του γονιδιώματος (*functional genomics*). Η ανάπτυξη των τεχνολογιών για τη μελέτη της ακολουθίας

πληροφορίας του DNA (*DNA sequence*), δηλαδή ο καθορισμός των νουκλεϊκών βάσεων: αδενίνη (A), κυτοσίνη (C), γουανίνη (G) και θυμίνη (T) σ' ένα μόριο (*molecule*) του DNA, συνέβαλε στη μεγάλη επιτυχία της γονιδιακής ακολουθίας. Παρόμοιες τεχνολογίες ανακαλύπτονται συνεχώς για τη μελέτη της λειτουργικότητας του γονιδιώματος. Μεταξύ αυτών των τεχνολογιών η πιο αξιοσημείωτη είναι η τεχνολογία μικροσυστάδων του DNA (*DNA microarray technology*) με την οποία ο ερευνητής μπορεί να «φωτογραφίσει» (*snapshots*) τα επίπεδα έκφρασης όλων των γονιδίων (*gene expression levels*) σ' έναν οργανισμό. Εναλλακτικά ονόματα αυτής της τεχνολογίας είναι μικροσυστάδες του DNA (*DNA microarrays*), συστάδες DNA (*DNA arrays*), πλακίδια του DNA (*DNA chips*) και πλακίδια γονιδίων (*gene chips*) (Causton et al. 2003).

Με τις μικροσυστάδες μπορεί κανείς να αναγνωρίσει αποτελεσματικά το επίπεδο έκφρασης των γονιδίων σε διάφορους τύπους κυττάρων, να μάθει πως μεταβάλλεται το επίπεδο έκφρασής τους στα διάφορα στάδια ανάπτυξής του κυττάρου ή στα διάφορα στάδια μίας ασθένειας και να προσδιορίσει τις κυτταρικές διαδικασίες (*cellular processes*) στις οποίες συμμετέχουν. Η τεχνολογία μικροκυττάρων μας τροφοδοτεί με *terabytes* δεδομένων που αφορούν την λειτουργικότητα του γονιδιώματος και μας παρέχει ενδείξεις σχετικά με το πως αλληλεπιδρούν τα γονίδια και τα παράγωγα γονιδίων (*gene products*). Ωστόσο, η δημιουργία αυτών των δεδομένων δεν είναι εύκολη υπόθεση. Γι' αυτόν τον σκοπό έχουν αναπτυχθεί μέθοδοι και εργαλεία για την ανάλυση αυτών των πολύπλοκων και πολυάριθμων δεδομένων με τη βοήθεια της βιοπληροφορικής (*bioinformatics*) και της υπολογιστικής βιολογίας (*computational biology*) (Causton et al. 2003).

1.3 Το DNA και ο ρόλος του

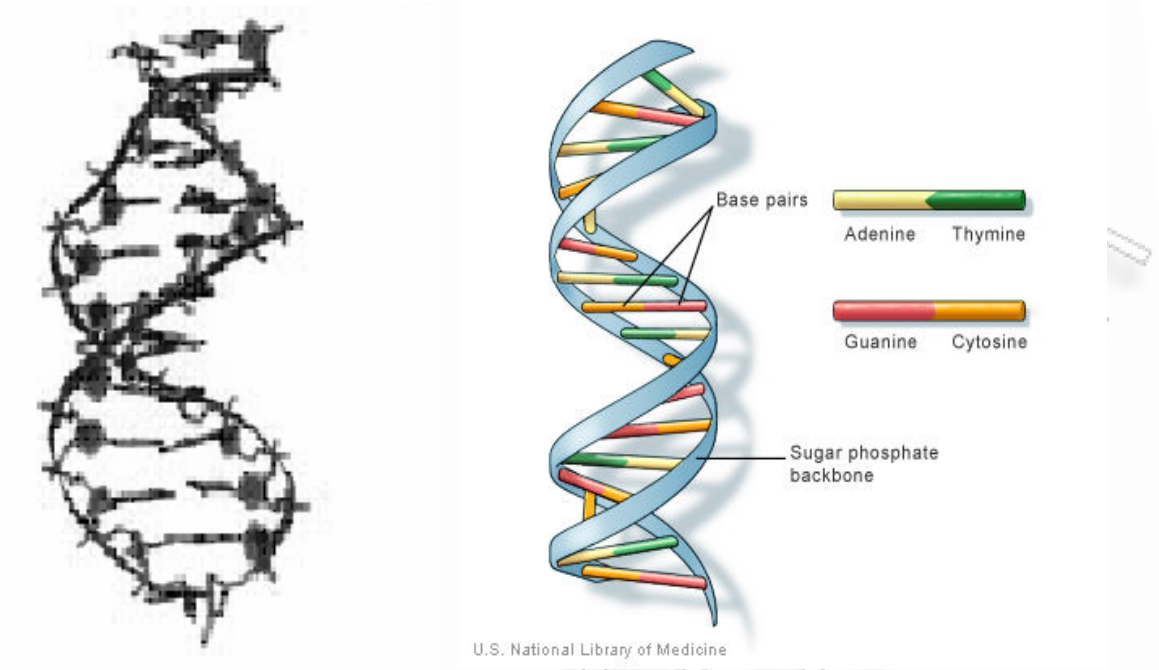
Είναι γνωστό ότι το DNA είναι ο «πυρήνας» της ζωής κάθε οργανισμού, διότι όχι μόνο περιέχει τον βιοχημικό κώδικα της κληρονομικότητάς του, αλλά επίσης παίζει σημαντικότατο ρόλο στην λειτουργία του κυττάρου, η οποία περιλαμβάνει τον μεταβολισμό, την ανάπτυξη και την παραγωγή του. Ουσιαστικά το DNA περιέχει πληροφορίες τις οποίες χρησιμοποιεί μέσω μίας πολύπλοκης διαδικασίας για την παραγωγή πρωτεϊνών, των οποίων η συμβολή είναι πολύ σημαντική σε κάθε οργανισμό, διότι συμμετέχουν σε κάθε διαδικασία που συμβαίνει εντός του κυττάρου, καταστέλλουν κάποιες βιοχημικές αντιδράσεις και επίσης προσδιορίζουν το σχήμα

του κυττάρου. Συνεπώς, επιβάλλεται να ξεκινήσει η μελέτη της κυτταρικής λειτουργίας από το DNA (Τσακανίκας, 2005).

1.3.1 Σύνθεση και λειτουργία του DNA

Για την κατανόηση των δεδομένων γονιδιακών εκφράσεων πρέπει πρώτα να αναφέρουμε το «κεντρικό δόγμα» (*central dogma*) της μοριακής βιολογίας (Causton et al. 2003). Συγκεκριμένα, το DNA (*Deoxyribonucleic acid*) είναι ένα μακρο-μόριο που αποτελείται από μία μεγάλη αλυσίδα που περιέχει ακολουθίες τυχαία επαναλαμβανόμενων μονάδων νουκλεοτιδίων (*nucleotide units*) και βρίσκεται μέσα στα χρωμοσώματα στον πυρήνα του κυττάρου. Κάθε νουκλεοτίδιο περιέχει μία φωσφατάση (*phosphatase*), μία $2' -$ δεοξυριβόζη (*deoxyribose*) και μία από τις τέσσερις νιτρογενείς βάσεις (*nitrogenous bases*) αδενίνη (A), κυτοσίνη (C), γουανίνη (G) και θυμίνη (T).

Ο πιο κοινός τύπος DNA σ' ένα κύτταρο είναι η διπλή έλικα (*double helix*), η οποία είναι αποτέλεσμα συστροφής (*twist*) σαν σε σπιράλ των δύο ατομικών «κλωστών» (*strands*) του DNA. Σε αυτόν τον διπλό έλικα κάθε βάση της μιας «κλωστής» συνδέεται με την συμπληρωματικής της βάση της άλλης «κλωστής» με τη βοήθεια δεσμών υδρογόνου (*hydrogen bonds*). Συγκεκριμένα η γουανίνη συνδέεται με την κυτοσίνη με τη βοήθεια τριών δεσμών υδρογόνου, ενώ η θυμίνη συνδέεται με την αδενίνη με τη βοήθεια δύο δεσμών υδρογόνου. Συνεπώς, οι δύο «κλωστές» του DNA χαρακτηρίζονται ως συμπληρωματικές (*complemetary).*



Εικόνα 1.3.1 Η διπλή έλικα του DNA με τις νιτρογενείς βάσεις
(Πηγή: www.genetest.org/page5.html)

1.3.2 Η αντιγραφή, η μετάφραση και η ανάστροφη αντιγραφή του DNA

Το DNA αντιγράφεται και μεταφράζεται για την σύνθεση πρωτεϊνών μέσω μίας περίπλοκης βιοχημικής διαδικασίας που συμβαίνει στο κυτταρόπλασμα, το οποίο περιλαμβάνει όλα τα στοιχεία που συνθέτουν το κύτταρο. Η αντιγραφή πραγματοποιείται στον πυρήνα του (ευκαρυωτικού) κυττάρου, όπου βρίσκεται το DNA, ενώ η μετάφραση πραγματοποιείται στα ριβοσώματα, τα οποία περιφράσσουν τον πυρήνα και είναι υπεύθυνα για την παραγωγή πρωτεΐνης χρησιμοποιώντας αμινοξέα (*amino acids*).

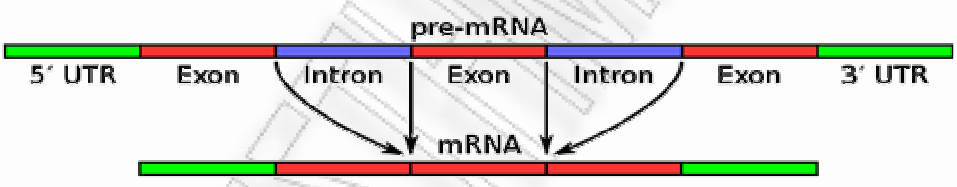
Τμήμα του DNA αντιγράφεται σε RNA με την βοήθεια ενζύμων που ονομάζονται RNA-πολυμεράσες (*RNA-polymerases*), τα οποία προσθέτουν ένα RNA νουκλεοτίδιο τη φορά σε μορφή ακολουθίας, έτσι ώστε να σχηματιστεί η «κλωστή» του RNA . Το RNA είναι ένας σημαντικός τύπος μορίου, το οποίο αποτελείται από μία μεγάλη αλυσίδα από ακολουθίες νουκλεοτιδίων. Το RNA μοιάζει αρκετά με το DNA, ωστόσο διαφέρει από αυτό στα παρακάτω σημεία:

1. τα νουκλεοτίδια του RNA περιέχουν ριβόζη (*ribose*), ενώ τα νουκλεοτίδια του DNA περιέχουν δεοξυριβόζη, η οποία είναι ένας τύπος ριβόζης που στερείται ένα άτομο οξυγόνου,

2. συνήθως το RNA είναι μία μονή «κλωστή» (*single-stranded*), ενώ το DNA είναι ένας διπλός έλικας (*double-stranded*),
3. το RNA έχει τη (νιτρογενή) βάση ουρακίλη (U), η οποία απουσιάζει στο DNA, ενώ το DNA έχει την βάση θυμίνη, η οποία απουσιάζει από το RNA.

Να σημειώσουμε ότι υπάρχουν διάφοροι τύποι των RNAs ανάλογα με την λειτουργία τους στο κύτταρο. Τα RNAs που συμμετέχουν στην σύνθεση πρωτεΐνης διακρίνονται στα mRNAs (*messenger RNAs*), rRNAs (*ribosomal RNAs*) και tRNAs (*transfer RNAs*).

Κατά την έναρξη της αντιγραφής (*transcription*) του DNA, το πρώτο RNA που δημιουργείται είναι το pre-mRNA (*precursor mRNA*). Το pre-mRNA περιέχει τα extrons, που είναι κωδικοποιημένες πληροφορίες που αφορούν την σύνθεση πρωτεΐνης και τα introns, που είναι μη κωδικοποιημένες πληροφορίες (προς το παρόν οι βιολόγοι δεν γνωρίζουν την χρήση τους). Τα introns αφαιρούνται από το pre-mRNA, οπότε τα extrons ενώνονται σχηματίζοντας το mRNA. Η διαδικασία αυτή ονομάζεται συγκόλληση (*splicing*) των extrons (εικόνα 1.3.2).

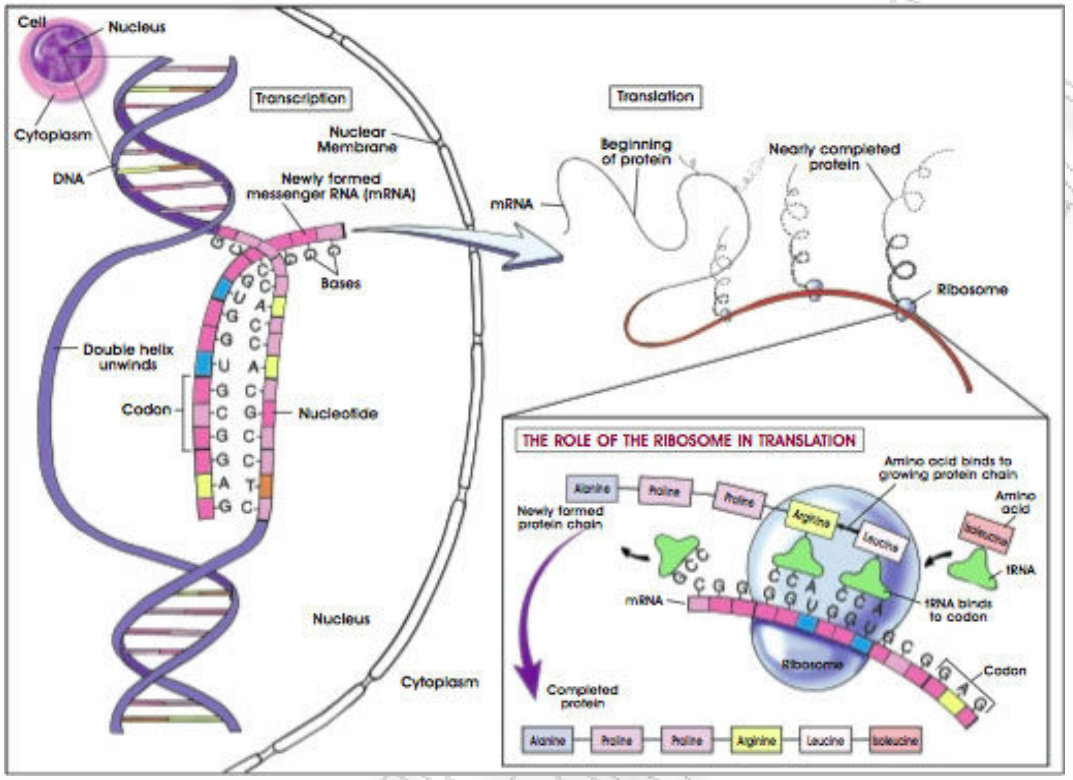


Εικόνα 1.3.2 Η διαδικασία «συγκόλλησης» των extrons μετά την αφαίρεση των introns **(Πηγή: www.en.wikipedia.org/wiki/Intron)**

Το mRNA μεταφέρει αποκλειστικά πληροφορίες για την σύνθεση πρωτεΐνης. Ουσιαστικά, αυτές οι πληροφορίες είναι τα codons που περιέχει, δηλαδή διαδοχικές τριάδες νουκλεοτιδίων. Κάθε codon θεωρείται ως «κωδικός» του αμινοξέος που θα συμμετέχει στην πρωτεϊνική σύνθεση. Γνωρίζουμε ότι κάθε mRNA περιέχει μία ακολουθία νουκλεοτιδίων, οπότε θα περιέχει μία ακολουθία από codons, δηλαδή «κωδικούς» των αντίστοιχων αμινοξέων.

Μόλις το mRNA εξαχθεί από τον πυρήνα, θα δεσμευτεί από τουλάχιστον ένα ριβόσωμα σε οποιαδήποτε στιγμή. Στην συνέχεια με τη βοήθεια των tRNAs οι πληροφορίες του mRNA μεταφράζονται σε πρωτεΐνη. Το tRNA είναι μία μικρή αλυσίδα RNA, η οποία περιέχει περίπου 80 νουκλεοτίδια. Κάθε tRNA μεταφέρει ένα

συγκεκριμένο αμινοξύ σε μία συνεχώς αναπτυσσόμενη (*growing*) πολυπεπτιδική αλυσίδα (*polypeptide chain*), η οποία βρίσκεται μέσα στο ριβόσωμα. Επιπλέον, το



Εικόνα 1.3.3 Η αντιγραφή, η μεταφορά και η μετάφραση του mRNA για την σύνθεση πρωτεΐνης

(Πηγή: www.stemcells.nih.gov/info/scireport/appendixa.asp)

tRNA περιέχει ένα anticodon, το οποίο συνδέεται με το συμπληρωματικό του codon που βρίσκεται στο mRNA. Για να πραγματοποιηθεί η μετάφραση (*translation*), το ριβόσωμα βοηθάει να συνδεθούν τα anticodons των tRNAs με τα codons του mRNA, τα οποία θεωρούνται συμπληρωματικά των πρώτων. Έπειτα το ριβόσωμα παίρνει το αμινοξύ που περιέχει κάθε ένα από τα παραπάνω tRNAs και σχηματίζει μία γραμμική αλυσίδα αμινοξέων, η οποία αντιστοιχεί σε μία πρωτεΐνη. Δηλαδή, τα αμινοξέα που συνθέτουν την αλυσίδα αυτή καθορίζουν ποια πρωτεΐνη θα συντεθεί, οπότε διαφορετικές αλυσίδες αμινοξέων αντιστοιχούν στην σύνθεση διαφορετικών πρωτεϊνών.

Η διαδικασία της μετάφρασης τερματίζεται με την επέμβαση των rRNAs. Δηλαδή, αυτός ο τύπος RNA λειτουργεί ως καταλύτης της πρωτεϊνικής σύνθεσης. Το rRNA συνδέεται με μία πρωτεΐνη και δημιουργείται ένα ριβόσωμα. Συνεπώς, το ριβόσωμα

αποτελεί μία νουκλεοπρωτεΐνη (*nucleoprotein*). Το ριβόσωμα περιέχει τέσσερα διαφορετικά rRNAs, εκ των οποίων τα τρία συντίθενται στον πυρηνίσκο (*nucleolus*) του κυττάρου, ενώ το τέταρτο συντίθεται αλλού. Να σημειώσουμε ότι το DNA και το περιβάλλον καθορίζουν ποιες πρωτεΐνες θα παραχθούν, σε ποια ποσότητα και πότε (Τσακανίκας, 2005).

Μία άλλη διαδικασία που μπορεί να πραγματοποιηθεί στον πυρήνα του κυττάρου είναι η ανάστροφη αντιγραφή (*reverse transcription*). Η διαδικασία αυτή πραγματοποιείται με την επέμβαση των βιολόγων στο mRNA που αντιγράφηκε, πριν αρχίσει η διαδικασία της μετάφρασης του, και είναι μία γρήγορη μέθοδος να απομονώσουμε τα γονίδια που θέλουμε να μελετήσουμε. Με την βοήθεια του ενζύμου γνωστού ως ανάστροφη μεταγραφάση (*reverse transcriptase*) και των νιτρογενών βάσεων του DNA, σχηματίζεται μία συμπληρωματική «κλωστή» του πρωταρχικού mRNA με το οποίο μπορεί να ενωθεί, διότι κάθε βάση του πρωταρχικού mRNA συνδέεται με την συμπληρωματικής της στο συμπληρωματικό αντίγραφο του mRNA. Η συμπληρωματική «κλωστή» που σχηματίζεται ονομάζεται συμπληρωματικό DNA (*complementary DNA*) ή cDNA, διότι αποτελεί μία συμπληρωματική «κλωστή» του DNA που μπορεί να σχηματίσει ένα υβρίδιο διπλής έλικας DNA αν ενωθεί με μία άλλη συμπληρωματική «κλωστή. Ένα cDNA έχει συνήθως μήκος 500 έως 5,000 βάσεων (McLachlan et al. 2004). Τα cDNAs παίζουν σημαντικό ρόλο στα πειράματα μικροσυστάδων, διότι συμμετέχουν σε μία διαδικασία δημιουργίας υβριδίου DNA, γνωστή ως υβριδισμός (*hybridization*). Με τον όρο υβριδισμό περιγράφεται η σύνδεση μίας συμπληρωματικής μονής «κλωστής» DNA με μία άλλη συμπληρωματική μονή «κλωστή» για τον σχηματισμό ενός υβριδίου της διπλής έλικας DNA. Ολόκληρη η τεχνολογία των μικροσυστάδων στηρίζεται στον υβριδισμό, όπως θα δούμε στα επόμενα κεφάλαια, διότι αποτελεί τη βάση για την πραγματοποίηση πειραμάτων που αφορούν εκφράσεις γονιδίων (McLachlan et al. 2004).

1.4 Οι μικροσυστάδες και η μέτρηση της γονιδιακής έκφρασης

Μία μικροσυστάδα (*microarray*) είναι ένα εργαλείο που χρησιμοποιείται τα τελευταία 10 χρόνια στην μοριακή βιολογία και αποτελείται από πολλούς πίνακες (*arrays*), οι οποίοι περιέχουν χιλιάδες μικροσκοπικές κουκίδες (*spots*) με τμήματα

ακολουθιών του DNA. Τα χαρακτηριστικά αυτών των πινάκων περιγράφονται παρακάτω (McLachlan et al. 2004):

- Αυτοί οι πίνακες είναι κατασκευασμένοι από νάιλον ή γυαλί επιστρωμένο με σιλικόνη έτσι ώστε να απομακρύνεται το νερό και να κολλάει το DNA πάνω στην κουκίδα.
- Συνήθως είναι τετράγωνοι διαστάσεων $0.5\text{cm} \times 0.5\text{cm}$ ή ορθογώνιοι διαστάσεων $2.5\text{cm} \times 7.5\text{cm}$.
- Κάθε κουκίδα αντιπροσωπεύει ένα γονίδιο και είναι γνωστή με το όνομα *feature*. Το μέγεθος κάθε κουκίδας κυμαίνεται από 2μm έως 5μm. Οι κουκίδες περιέχουν τα *probes*, τα οποία μπορεί να είναι τμήματα της διπλής έλικας του DNA, cDNAs ή ολιγονουκλεοτίδια (*oligonucleotides*), δηλαδή μικρά πολυμερή νουκλεϊκών οξέων με 20 ως 80 το πολύ ζεύγη βάσεων. Σε κάθε κουκίδα είναι δυνατόν να υπάρχουν ένα ή περισσότερα *probes* (εικόνα 1.4.1).

Το γονίδιο αποτελεί τμήμα του DNA και περιέχει πληροφορίες για έναν συγκεκριμένο τύπο πρωτεΐνης ή για μία αλυσίδα RNA. Συνεπώς, το DNA περιέχει πολλά γονίδια με αποτέλεσμα ο διπλός έλικάς του να έχει πολύ μεγάλο μήκος. Ο ρόλος του γονιδίου είναι να χρησιμοποιεί αυτές τις πληροφορίες, ώστε να συμβάλει στη δημιουργία και την ανάπτυξη των κυττάρων ενός οργανισμού και να μεταφέρει γενετικά χαρακτηριστικά (*genetic traits*) στους απογόνους. Στην μελέτη των γονιδίων μεγαλύτερο ενδιαφέρον έχουν οι εκφράσεις τους. Η έκφραση γονιδίου (*gene expression*) είναι μία διαδικασία κατά την οποία οι πληροφορίες που περιέχει το γονίδιο χρησιμοποιούνται για τη δημιουργία ενός προϊόντος γονιδίου (*gene product*), το οποίο μπορεί να είναι πρωτεΐνη ή ένα λειτουργικό RNA (είναι προϊόν του rRNA γονιδίου ή του tRNA γονιδίου, όπου δεν περιέχουν πληροφορίες για την σύνθεση πρωτεΐνης σε αντίθεση με το mRNA). Η έκφραση των γονιδίων είναι κωδικοποιημένη στο DNA και αρχίζει με την αντιγραφή του γονιδίου σε RNA και ολοκληρώνεται με την μετάφραση του RNA και συνεπώς την σύνθεση πρωτεΐνης. Το γονίδιο που ρυθμίζει (*regulate*) την παραγωγή μεγάλης ποσότητας πρωτεΐνης θεωρείται ότι υπέρ-εκφράζεται (*over-expressed* ή *up-regulated*), ενώ το γονίδιο που ρυθμίζει την παραγωγή μικρής ποσότητας πρωτεΐνης θεωρείται ότι υπό-εκφράζεται (*under-expressed* ή *down-regulated*).

Με την χρήση των μικροσυστάδων μπορούμε να ποσοτικοποιήσουμε τις γονιδιακές εκφράσεις για να μπορούμε να συγκρίνουμε γονιδιακές εκφράσεις

φυσιολογικών και μη φυσιολογικών κυττάρων για να εντοπίσουμε τα γονίδια που σχετίζονται με την εμφάνιση κάποιων ασθενειών ή να μελετήσουμε τις γονιδιακές εκφράσεις κάτω από διάφορες συνθήκες περιβάλλοντος (για παράδειγμα, η θερμοκρασία και η θεραπευτική αγωγή) (Τσακανίκας, 2005). Η διαδικασία που ακολουθείται στις μικροσυστάδες για την ποσοτικοποίηση των εκφράσεων γονιδίων παρουσιάζεται επιγραμματικά παρακάτω:

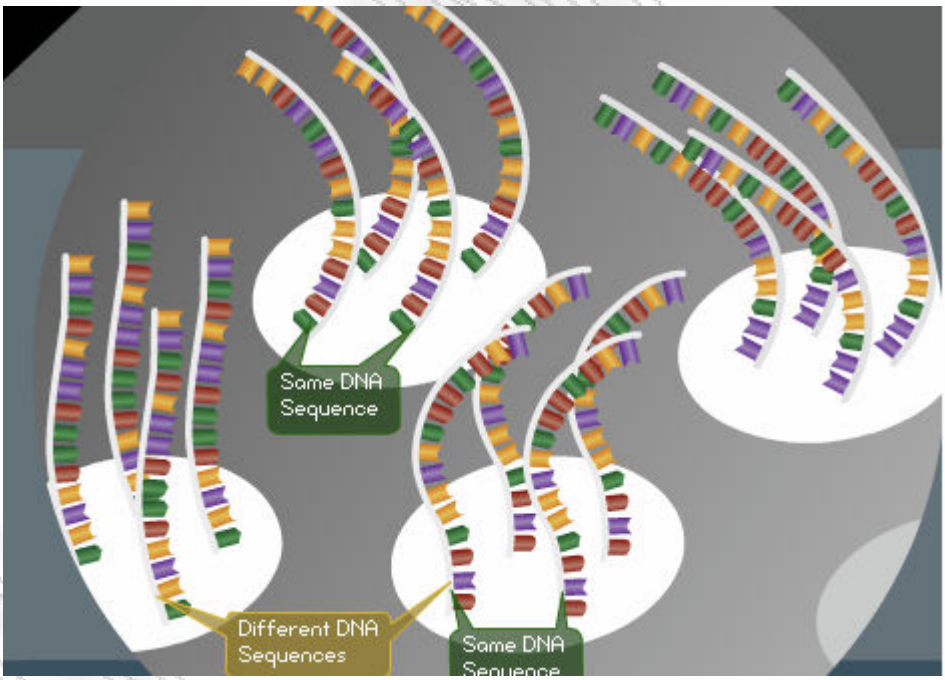
1. Τοποθέτηση του DNA στους πίνακες των μικροσυστάδων
2. Αντιγραφή του mRNA και εξαγωγή του cDNA από τα υπό μελέτη κύτταρα
3. Υβριδισμός
4. Σάρωση των πινάκων και επεξεργασία εικόνας
5. Κανονικοποίηση των εντάσεων του υβριδισμού
6. Επιλογή των σημαντικών γονιδίων

1.4.1 Ο ορισμός και ο ρόλος των *probes* και των «στόχων» στον υβριδισμό

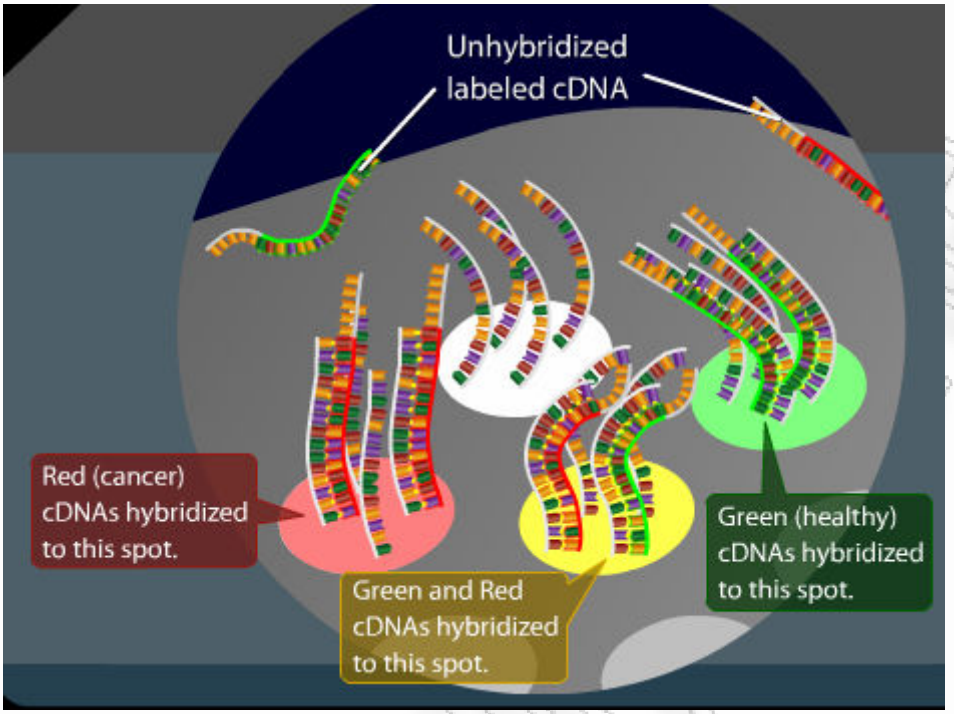
Στα *probes* είναι γνωστή η ακολουθία των βάσεων, δηλαδή η σειρά που εμφανίζονται οι βάσεις στα τμήματα του DNA που έχουμε απομονώσει και προσκολλήσει στις κουκίδες. Σε κάθε κουκίδα είναι δυνατόν να υπάρχουν ένα ή περισσότερα *probes* με την ίδια ακολουθία βάσεων, ενώ μεταξύ των κουκίδων η ακολουθία βάσεων στα *probes* συνήθως διαφέρει (εικόνα 1.4.1). Κάθε *probe* υφίσταται μετουσίωση (d*enaturation*), δηλαδή υπόκειται σε υψηλές θερμοκρασίες ή εκτίθεται σε υδροξείδιο του νατρίου (*sodium hydroxide*) με αποτέλεσμα να σπάνε οι δεσμοί υδρογόνου που συνδέουν τις συμπληρωματικές βάσεις των «κλωστών» (*strands*) του DNA και να διαιρείται η διπλή έλικα DNA σε δύο «κλωστές» DNA. Στην συνέχεια σημαδεύουμε (*tag* ή *label*) κάθε *probe* με κάποια μοριακή σήμανση (*molecular marker*), η οποία μπορεί να είναι είτε ραδιενεργά (*radioactive*) είτε φθορίζοντα (*fluorescent*) μόρια, τα οποία αποτελούν συστατικά ενός μορίου και μπορούν να κάνουν φθορίζον (*fluorescent*) το στοιχείο που θα περιβάλλουν, διότι απορροφούν ενέργεια συγκεκριμένου μήκους κύματος (*wavelength*) και επανεκπέμπουν την ενέργεια σε διαφορετικό μήκος κύματος, ώστε να μπορούμε να διακρίνουμε τον υβριδισμό ενός *probe* με τον συμπληρωματικό του «στόχο» μέσω της επεξεργασίας εικόνας και να μελετήσουμε την ένταση αυτού του υβριδισμού.

Τα *probes* χρησιμοποιούνται στην διαδικασία του υβριδισμού, όπου συμμετέχουν τα cDNAs, τα οποία δημιουργήσαμε μέσω της ανάστροφης αντιγραφής σε κάποια

υπό μελέτη κύτταρα. Για τα συγκεκριμένα cDNAs δεν γνωρίζουμε την ακολουθία των βάσεων που περιέχουν, γι' αυτό τα θεωρούμε ως «στόχους» (*targets*). Οι «στόχοι» σημαδεύονται, επίσης, με μία μοριακή σήμανση κάποιου χρώματος (συνήθως με κόκκινο ή πράσινο). Στην συνέχεια, οι στόχοι τοποθετούνται στον πίνακα με τη βοήθεια κάποιας τεχνολογίας έκχυσης μελάνης και πραγματοποιείται ο υβριδισμός, όπου ένας ή περισσότεροι «στόχοι» ενώνονται με το συμπληρωματικό τους *probe* σχηματίζοντας ένα υβρίδιο διπλής έλικας DNA (εικόνα 1.4.2). Ένα *probe* μπορεί να ενωθεί πλήρως ή μερικώς με τον συμπληρωματικό του «στόχο», δηλαδή οι βάσεις του *probe* ενώνονται αντίστοιχα με όλες ή με μερικές βάσεις του συμπληρωματικού cDNA. Είναι δυνατόν κάποια *probes* να μην υποστούν υβριδισμό, όταν δεν υπάρχουν «στόχοι» με βάσεις συμπληρωματικές των βάσεων που περιέχουν, ώστε να σχηματιστεί ένα υβρίδιο διπλής έλικας DNA. Μόλις ολοκληρωθεί ο υβριδισμός απομακρύνονται (*wash off*) από τον πίνακα τα cDNAs που δεν συμμετείχαν στον υβριδισμό.



Εικόνα 1.4.1 Τα probes κάθε κουκίδας πριν τον υβριδισμό.
(Πηγή: www.learn.genetics.utah.edu/content/labs/microarray/)



Εικόνα 1.4.2 Τα *probes* κάθε κουκίδας μετά τον υβριδισμό. Σε κάποιες κουκίδες έχουν σχηματιστεί υβρίδια διπλή έλικας DNA, ενώ σε κάποιες κουκίδες αυτό δεν ήταν εφικτό. Τα cDNAs που δεν υπέστησαν υβριδισμό απομακρύνονται από τον πίνακα.
(Πηγή: www.learn.genetics.utah.edu/content/labs/microarray/)

1.4.2 Μέτρηση της έντασης του υβριδισμού

Μετά τον υβριδισμό το επόμενο βήμα είμαι η χρήση ενός σαρωτή (*scanner*) λέιζερ, ο οποίος θα «διαβάσει» τους φθορίζοντες «στόχους» και θα μετρήσει την ένταση του υβριδισμού. Πολλοί εκτυπωτές εκπέμπουν λέιζερ συγκεκριμένου μήκος κύματος πάνω στους «στόχους» που έχουν υποστεί υβριδισμό από τα συμπληρωματικά τους *probes*, έτσι ώστε να διεγείρουν τα φθορίζοντα μόρια που έχουν προσκολληθεί στους «στόχους». Έπειτα, αυτά τα φθορίζοντα μόρια απορροφούν τα φωτόνια που προέρχονται από το λέιζερ και τα εκπέμπουν προς οποιαδήποτε κατεύθυνση. Μόλις ένα κλάσμα αυτών των φωτονίων συγκεντρώνεται από έναν ανιχνευτή (*detector*), ο οποίος μετράει και καταγράφει την ακτινοβολία αυτών των φωτονίων και στη συνέχεια την μετατρέπει σε ηλεκτρικό ρεύμα (*electrical signal*). Επειδή χρησιμοποιούμε συνήθως δύο διαφορετικές φθορίζουσες αποχρώσεις (*fluorescent dyes*), τότε ο πίνακας σαρώνεται σε δύο διαφορετικά μήκη κύματος,

οπότε δημιουργούνται δύο 16-bit εικόνες, μία για κάθε χρώμα φθορισμού (Τσακανίκας, 2005). Συνήθως, οι φθορίζουσες αποχρώσεις που χρησιμοποιούνται είναι η Cyanine 3 ή Cy3 (πράσινο) και η Cyanine 5 ή Cy5 (κόκκινο), διότι μπορούμε να ξεχωρίσουμε εύκολα τις φασματικές εκπομπές τους με όσο το δυνατόν μικρότερο σφάλμα (McLachlan et al. 2004).

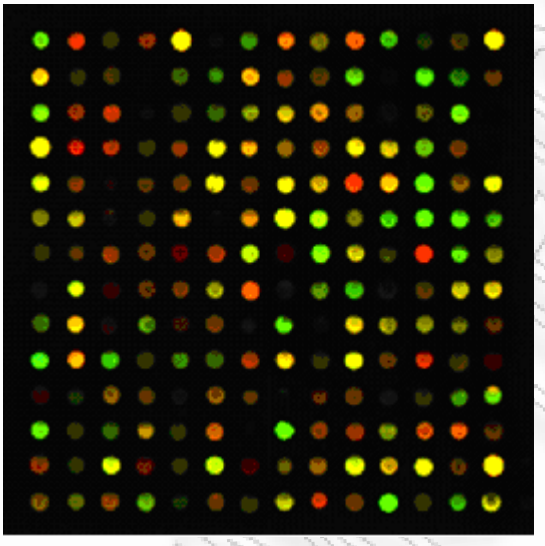
1.4.3 Επεξεργασία και ανάλυση της εικόνας μικροσυστάδας

Οι επεξεργαστές εικόνων (*imagers*) εφαρμόζουν πάνω στον πίνακα ένα πλέγμα (*grid*), για να εντοπίσουν την ένταση του ρεύματος που καταγράφηκε σε κάθε κουκίδα, ώστε να διευκολυνθούν στην μελέτη του πίνακα. Ουσιαστικά αναθέτουν συντεταγμένες σε κάθε κουκίδα. Ο ανιχνευτής θα καταγράψει το φως που προέρχεται μόνο από κουκίδες με ζεύγη συμπληρωματικών βάσεων που έχουν σημαδευτεί με φθορίζουσες αποχρώσεις. Όσο πιο μεγάλο είναι το πλήθος αυτών των ζευγών τόσο πιο μεγάλη είναι η ένταση του υβριδισμού στις αντίστοιχες κουκίδες. Ο τεχνικός του εργαστηρίου (*lab technician*) ρυθμίζει τον σαρωτή για κάθε μικροσυστάδα που επεξεργάζεται έτσι ώστε να προλάβει ενδεχόμενες διακυμάνσεις στο μήκος κύματος και την τάση του λέιζερ, όπως επίσης και τυχόν θόρυβο φωτονίων, ανάκλαση του λέιζερ ή φθορισμού του φόντου των κουκίδων που μπορεί να επηρεάσουν το τελικό σήμα που παράγεται από τον σαρωτή και επομένως την εικόνα της μικροσυστάδας που δημιουργείται (McLachlan et al. 2004).

Μετά την σάρωση του πίνακα εμφανίζεται μία ψηφιακή εικόνα του σε μία οθόνη, όπου δείχνει τις εντάσεις φθορισμού με την μορφή χρωματισμένων εικονοστοιχείων (*pixels)*, όπου κάθε εικονοστοιχείο αντιστοιχεί σε μία κουκίδα του πίνακα (εικόνα 1.4.3). Οι εντάσεις των αποχρώσεων εμφανίζονται στην εικόνα. Αν θεωρήσουμε ότι για τα *probes* χρησιμοποιήσαμε πράσινη φθορίζουσα απόχρωση, ενώ για τους «στόχους» χρησιμοποιήσαμε κόκκινη φθορίζουσα απόχρωση, τότε τα χρώματα των εικονοστοιχείων ερμηνεύονται ως εξής (McLachlan et al. 2004):

- Το κόκκινο χρώμα σημαίνει ότι σε αυτήν την κουκίδα υπήρχε μεγάλος αριθμός cDNAs που πραγματοποίησαν υβριδισμό με τα *probes*.
- Το πράσινο χρώμα σημαίνει ότι σε αυτήν την κουκίδα υπήρχε μικρός αριθμός cDNAs που πραγματοποίησαν υβριδισμό με τα *probes*.

- Το κίτρινο χρώμα σημαίνει ότι σε αυτήν την κουκίδα υπάρχει ισοδυναμία στο πλήθος των cDNAs που πραγματοποίησαν υβριδισμό με τα *probes*.
- Το μαύρο χρώμα σημαίνει ότι τα *probes* αυτής της κουκίδας δεν πραγματοποίησαν υβριδισμό με κανένα cDNA.



Εικόνα 1.4.3 Η τελική εικόνα που προκύπτει από τον συνδυασμό της εικόνας με τις πράσινες αποχρώσεις και της εικόνας με τις κόκκινες αποχρώσεις.
(Πηγή: www.imbb.forth.gr/people/poirazi/researchEP.html)

Παρατήρηση

Ανάλογα με τον σκοπό που πραγματοποιείται ένα πείραμα σε μικροσυστάδες υπάρχουν τρεις βασικές τεχνολογίες μικροσυστάδων (Τσακανίκας, 2005):

1. Οι cDNA μικροσυστάδες, όπου μετράται η ποσότητα του mRNA ως δείκτης έκφρασης του επιπέδου του γονιδίου.
2. Οι SNP μικροσυστάδες (*Single Nucleotide Polymorphism microarrays*), όπου χρησιμοποιείται για την ανίχνευση πολυμορφισμών ή μεταλλάξεων σε έναν πληθυσμό.
3. Ο Συγκριτικός Γονιδιακός Υβριδισμός (*Comparative Genomic Hybridization, CGH*), όπου ερευνώνται οι αλλαγές στον αριθμό αντιγραφής κάποιου συγκεκριμένου γονιδίου που σχετίζεται με κάποια ασθένεια.

1.4.4 Ανάλυση cDNA μικροσυστάδων

Η πιο γνωστή τεχνολογία μικροσυστάδων περιλαμβάνει την ανάλυση των cDNA μικροσυστάδων, με τις οποίες συγκρίνουμε τις γονιδιακές εκφράσεις σε δύο διαφορετικά δείγματα κυττάρων, όπως για παράδειγμα κύτταρα ίδιου τύπου πριν και μετά την έκθεσή τους σε κάποια αλλαγή συνθηκών που προκαλεί μετάλλαξη των κυττάρων (υγιή και μη υγιή κύτταρα αντίστοιχα). Για να μεγιστοποιήσουμε τον αριθμό των γονιδίων που μπορούμε να αναπαραστήσουμε σε κάθε πίνακα της μικροσυστάδας χρειαζόμαστε αρκετές εκατοντάδες ζευγάρια νουκλεϊκών βάσεων (*probes*) (Τσακανίκας, 2005). Ένα πείραμα με μία cDNA μικροσυστάδα περιγράφεται ως εξής (εικόνα 1.4.4):

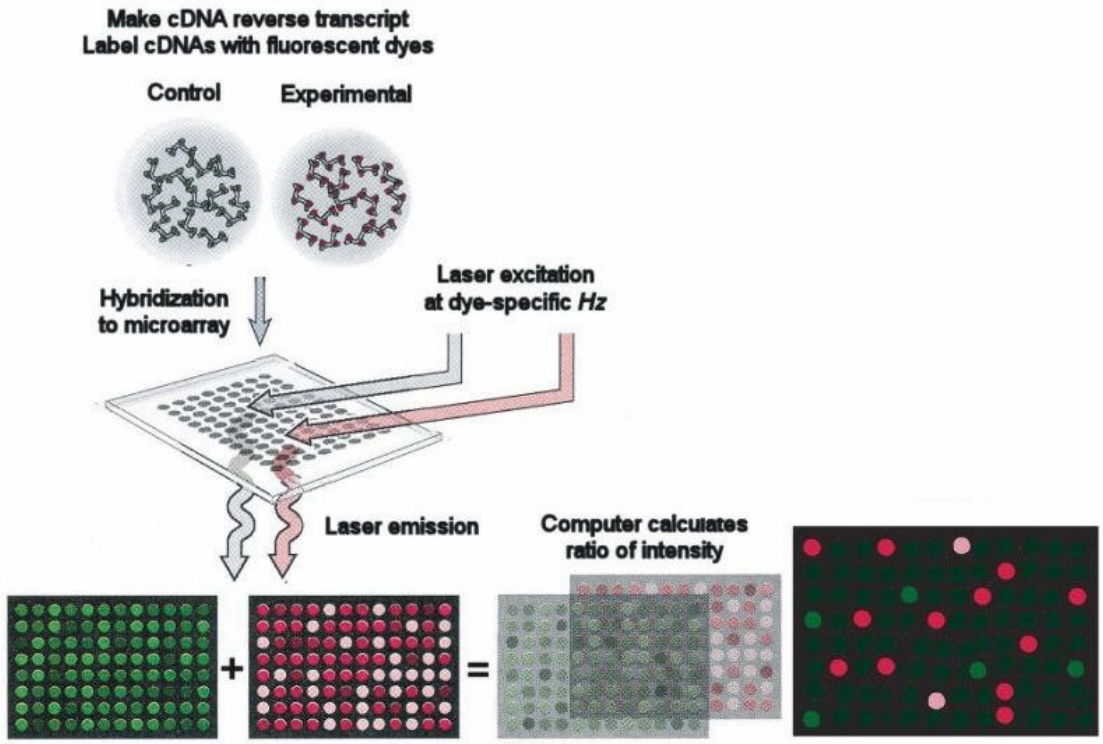
1. Στις κουκίδες του πίνακα προσκολλούμε με τη βοήθεια τεχνολογίας έκχυσης μελάνης ή με την τεχνική φωτολιθογραφίας τα *probes* που αντιστοιχούν στα γονίδια που θέλουμε να μελετήσουμε.
2. Τόσο από το δείγμα υγιών κυττάρων όσο και από το δείγμα μη υγιών κυττάρων απομονώνουμε τα mRNAs και με την διαδικασία της ανάστροφης αντιγραφής δημιουργούμε cDNAs.
3. Αυτά τα δύο δείγματα με cDNAs σημαδεύονται χρησιμοποιώντας διαφορετικά φθορίζοντα μόρια. Το δείγμα των cDNAs από μη υγιή κύτταρα σημαδεύεται με κόκκινο χρώμα (Cy5) και αποτελεί το δείγμα δοκιμής (*experiment sample*), ενώ το δείγμα των cDNAs από υγιή κύτταρα σημαδεύεται με πράσινο χρώμα (Cy3) και αποτελεί το δείγμα αναφοράς (*control sample*).
4. Στη συνέχεια αυτά τα δύο δείγματα αναμειγνύονται και με την βοήθεια της ρομποτικής (συνήθως, χρήση της τεχνικής της φωτολιθογραφίας ή της έκχυσης μελάνης) «ελευθερώνεται» το μείγμα χρωματισμένων cDNAs πάνω στον πίνακα της μικροσυστάδας, ώστε να αρχίσει η διαδικασία του υβριδισμού.
5. Έπειτα σαρώνουμε τον πίνακα χρησιμοποιώντας έναν εκτυπωτή λέιζερ, ώστε να διεγερθούν τα φθορίζοντα μόρια των cDNAs και να εντοπιστούν τα cDNAs που υπέστησαν υβριδισμό. Λόγω χρήσης δύο διαφορετικών φθοριζουσών αποχρώσεων ο πίνακας σαρώνεται δύο φορές και αποθηκεύονται οι δύο διαφορετικές φθορίζουσες εικόνες με εικονοστοιχεία που προκύπτουν. Οι εικόνες αυτές συνδυάζονται για να δημιουργηθεί η τελική εικόνα, στην οποία θα βασιστούμε για να μετρήσουμε τις εντάσεις υβριδισμού.

6. Μετράμε σε κάθε κουκίδα την ένταση του σήματος που ανιχνεύεται στα δύο μήκη φθορισμού χωριστά (πράσινα και κόκκινα φθορίζοντα μόρια). Στην ουσία υπολογίζουμε σε κάθε κουκίδα τον λόγο του πλήθους cDNAs από το δείγμα δοκιμής προς το πλήθος cDNAs από το δείγμα αναφοράς που υπέστησαν υβριδισμό. Ο λόγος αυτός είναι η ένταση του υβριδισμού στην αντίστοιχη κουκίδα. Η ένταση υβριδισμού κάθε κουκίδας διακρίνεται λαμβάνοντας υπόψη το χρώμα κάθε κουκίδας. Συγκεκριμένα:
    - Το κόκκινο χρώμα σημαίνει ότι σε αυτήν την κουκίδα μεταξύ των cDNAs που πραγματοποίησαν υβριδισμό με κάποια probes υπερέχουν τα cDNAs κόκκινης απόχρωσης έναντι των cDNAs πράσινης απόχρωσης. Μάλιστα, όσο πιο μεγάλη είναι αυτή η διαφορά τόσο πιο έντονη είναι η απόχρωση του κόκκινου χρώματος.
    - Το πράσινο χρώμα σημαίνει ότι σε αυτήν την κουκίδα μεταξύ των cDNAs που πραγματοποίησαν υβριδισμό με κάποια probes υπερέχουν τα cDNAs πράσινης απόχρωσης έναντι των cDNAs κόκκινης απόχρωσης. Μάλιστα, όσο πιο μεγάλη είναι αυτή η διαφορά τόσο πιο έντονη είναι η απόχρωση του πράσινου χρώματος.
    - Το κίτρινο χρώμα σημαίνει ότι σε αυτήν την κουκίδα μεταξύ των cDNAs που πραγματοποίησαν υβριδισμό με κάποια probes το πλήθος των cDNAs κόκκινης απόχρωσης είναι ισοδύναμο με το πλήθος των cDNAs πράσινης απόχρωσης.
    - Το μαύρο χρώμα σημαίνει ότι σε αυτήν την κουκίδα δεν περιέχεται cDNA, δηλαδή κανένα *probe* δεν πραγματοποίησε υβριδισμό με κάποιο cDNA.
7. Αποθηκεύουμε τα αποτελέσματα σε βάση δεδομένων.
8. Πραγματοποιείται εξόρυξη δεδομένων (*data mining*).

Παρατήρηση

Αξίζει να σημειώσουμε ότι οι μελέτες γονιδιακής έκφρασης μπορούν να διαιρεθούν σε δύο κατηγορίες: καταστάσεις στις οποίες τα δείγματα χρησιμοποιούνται για να παρέχουν πληροφορίες για τα γονίδια και καταστάσεις στις οποίες τα γονίδια χρησιμοποιούνται για να παρέχουν πληροφορίες για τα δείγματα. Η πρώτη περίπτωση εφαρμόζεται όταν μελετάται ο κανονισμός της γονιδιακής

έκφρασης (*regulation of gene expression*) χρησιμοποιώντας μία πολύπλοκη διαδικασία στην οποία η αλληλεπίδραση θετικών και αρνητικών σημάτων καθορίζει την κυτταρική μοίρα, ενώ η δεύτερη διαδικασία εφαρμόζεται στην μοριακή ιατρική για την ταξινόμηση ασθένειας, την διάγνωση και την πρόγνωση αλλά και σε πολυάριθμες φαρμακευτικές εφαρμογές. (Causton et al. 2003)



Εικόνα 1.4.4 Σχηματικό διάγραμμα ενός πειράματος cDNA μικροσυστάδων.
(Πηγή: www.mun.ca/biology/scarr/cDNA_microarray_Assay_of_Gene_Expression.html)

1.5 Κατανόηση των δεδομένων (*raw data*)

Οι κουκίδες του πλακιδίου της μικροσυστάδας συνθέτουν την εικόνα της μικροσυστάδας. Σε κάθε κουκίδα αυτής της εικόνας υπάρχει συγκεκριμένη ποσότητα cDNA από το δείγμα δοκιμής, η οποία συμβολίζεται με R (από το *Red*) και συγκεκριμένη ποσότητα cDNA από το δείγμα αναφοράς, η οποία συμβολίζεται με G (από το *Green*). Οι δύο αυτές ποσότητες συνήθως συνδυάζονται σε έναν απλό λογαριθμικό με βάση το 2 λόγο, δηλαδή

$$\log_2\left(R/G\right) \qquad (1.5.1)$$

ο οποίος μετράει τη σχετική πληθώρα ανάστροφης αντιγραφής (δηλαδή, το σχετικό πλήθος των cDNAs) στα δύο δείγματα. Ένας θετικός (αρνητικός) λογαριθμικός λόγος δείχνει υπερ-έκφραση (υπο-έκφραση) στο δείγμα δοκιμής σε σύγκριση με το δείγμα αναφοράς (Τσακανίκας, 2005).

1.5.1 Η σημασία των επαναληπτικών πειραμάτων

Είναι δυνατόν μία εικόνα μικροσυστάδας να μην είναι καλής ποιότητας με αποτέλεσμα να εξαιρείται από την ανάλυση, αφού δεν μπορούμε να υπολογίσουμε σωστά τους λογαριθμικούς λόγους. Για την βελτίωση της ποιότητας της εικόνας μπορούν να σχεδιαστούν επαναληπτικά πειράματα για κάθε δείγμα σε κάθε μικροσυστάδα και να εστιάσουμε στα γονίδια των οποίων οι μετρήσεις παρουσιάζουν μεγάλη αναπαραγωγικότητα (Τσακανίκας, 2005). Ωστόσο, με τις τεχνολογικές βελτιώσεις που έχουν συμβεί, η ποιότητα της εικόνας και κατ' επέκταση των δεδομένων που παίρνουμε είναι αρκετά καλή με αποτέλεσμα να αξίζει να δίνεται μεγαλύτερη έμφαση στην συλλογή μεγάλου αριθμού δειγμάτων με λιγότερες επαναλήψεις αυτών παρά στη συλλογή λιγότερων δειγμάτων με περισσότερες επαναλήψεις αυτών.

1.5.2 Κατασκευή πινάκων που περιέχουν εντάσεις υβριδισμού

Γνωρίζουμε ότι το τελευταίο στάδιο του πειράματος με cDNA μικροσυστάδες είναι η εξόρυξη των ακατέργαστων δεδομένων γνωστών με τον όρο *raw data* όπως έχει επικρατήσει στη βιβλιογραφία. Αυτά τα δεδομένα έχουν την μορφή πολύ μεγάλων πινάκων και περιέχουν διάφορες πληροφορίες για κάθε κουκίδα ενός πλακιδίου.

Με την βοήθεια των χρωμάτων των κουκίδων μπορούμε να ποσοτικοποιήσουμε την έκφραση των γονιδίων, η οποία αναπαριστάνει την δραστηριότητα των γονιδίων που εκφράζονται υπό ορισμένες συνθήκες. Το αποτέλεσμα είναι η κατασκευή ενός πίνακα με τις εκφράσεις όλων των γονιδίων της εικόνας μικροσυστάδας. Οι πληροφορίες που περιέχει ο πίνακας αυτός για κάθε κουκίδα αφορούν κυρίως (Τσακανίκας, 2005):

1. το όνομα του γονιδίου
2. τις εντάσεις φθορισμού στα δύο δείγματα (δοκιμής και αναφοράς)

3. τις θέσεις των κουκίδων (στήλη, γραμμή, πλέγμα)
4. ετικέτες (*labels*) που δείχνουν εάν το γονίδιο είναι αποδεκτό ή όχι ποιοτικά. Συνήθως, χρησιμοποιούνται οι εξής ετικέτες: $0 =$ αποδεκτό, $1 =$ μη αποδεκτό.

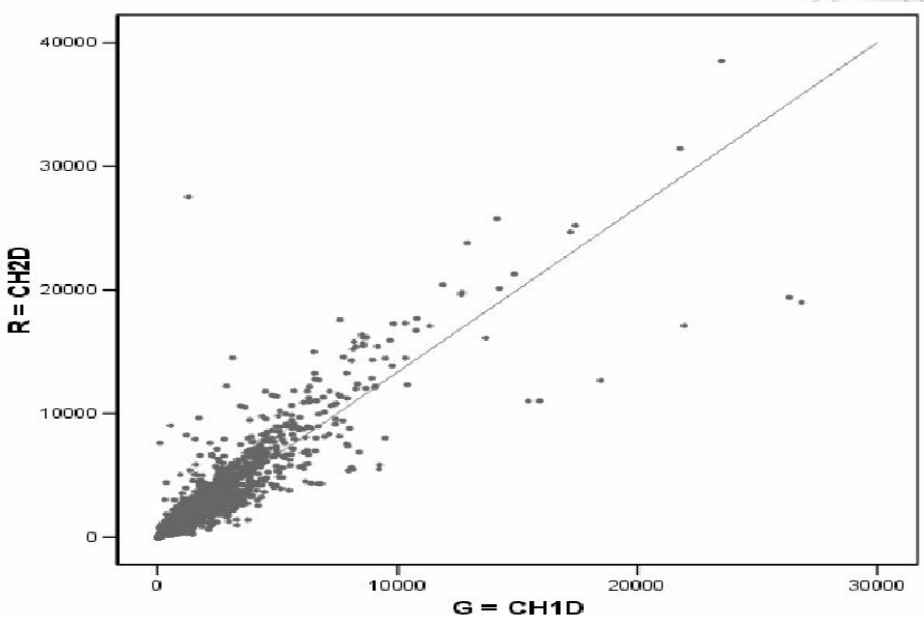
NAME	TYPE	CH1I	CH1B	CH1D	CH2I	CH2B	CH2D	CH2DN	...	GRID	ROW	COL	FLAG
EMPTY	CONTROL	13	12	1	40	36	4	4	...	1	1	20	0
EMPTY	CONTROL	16	13	3	44	30	14	14	...	1	1	21	0
YAL001C	ORF	20	20	0	118	91	27	26	...	1	2	1	1
YAL003W	ORF	5541	21	5520	4284	84	4200	4107	...	1	2	2	0
YAL005C	ORF	18	18	0	87	74	13	13	...	1	2	3	1
YAL008W	ORF	264	14	250	256	67	189	185	...	1	2	4	1
YAL010C	ORF	138	13	125	216	60	156	153	...	1	2	5	0
YAL012W	ORF	714	14	700	711	55	656	642	...	1	2	6	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Εικόνα 1.5.1 Παράδειγμα ενός πίνακα που περιέχει ακατέργαστα δεδομένα
(Πηγή:Τσακανίκας, 2005. Ανάλυση γονιδιακών εκφράσεων νεοπλασιών σε Microarrays)

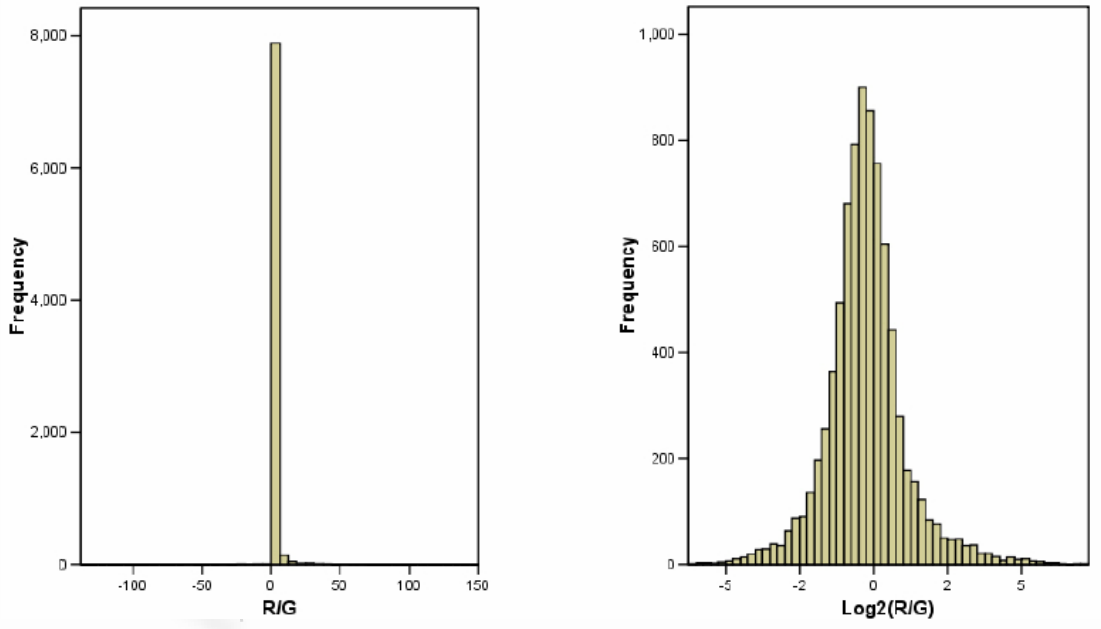
Σύμφωνα με την εικόνα 1.5.1 η πρώτη στήλη περιέχει τα ονόματα γονιδίων κάθε κουκίδας. Η στήλη με τον συμβολισμό $CH1$ περιέχει το πλήθος cDNAs στο δείγμα αναφοράς για κάθε κουκίδα, ενώ η στήλη με τον συμβολισμό $CH2$ περιέχει το πλήθος cDNAs στο δείγμα δοκιμής για κάθε κουκίδα, επίσης. Τα γράμματα I και B που συνοδεύουν τους συμβολισμούς $CH1$ και $CH2$ αναφέρονται στην κουκίδα και τον φόντο αντίστοιχα. Η στήλη με τον συμβολισμό $CH1D$ προκύπτει από τη σχέση $CH1I - CH1B$, η οποία εκφράζει την ένταση κάθε κουκίδας από τα cDNAs του δείγματος αναφοράς που περιέχει (δηλαδή, $CH1D = G$) έχοντας αφαιρέσει την ένταση του φόντου, ενώ η στήλη με τον συμβολισμό $CH2D$ προκύπτει από τη σχέση $CH2I - CH2B$, η οποία εκφράζει την ένταση κάθε κουκίδας από τα cDNAs του δείγματος δοκιμής που περιέχει (δηλαδή $CH2D = R$) έχοντας αφαιρέσει την ένταση του φόντου. Η στήλη με τον συμβολισμό $CH2DN$ περιέχει τις κανονικοποιημένες τιμές της στήλης $CH2D$, ενώ η στήλη με τον συμβολισμό $CH1DN$ περιέχει τις κανονικοποιημένες τιμές της στήλης $CH1D$. Τέλος, οι τέσσερις τελευταίες στήλες του πίνακα αφορούν αντίστοιχα το πλέγμα, την γραμμή και την στήλη του πίνακα της μικροσυστάδας στην οποία βρίσκεται το γονίδιο και αν το κάθε γονίδιο είναι αποδεκτό ή όχι ποιοτικά.

Οι ποσότητες R και G είναι τυχαίες μεταβλητές και παίρνουν ακέραιες τιμές στο διάστημα $\left(0, 2^p - 1\right)$ σύμφωνα με σάρωση στα p bit. Επιπλέον, αυτές οι τυχαίες μεταβλητές είναι συσχετισμένες σύμφωνα με το γράφημα στην εικόνα 1.5.2. Επειδή,

ο λογάριθμος με βάση το 2 του λόγου των εντάσεων $(1.5.1)$ είναι κατά προσέγγιση κανονικά κατανεμημένος, τον προτιμάμε έναντι του λόγου R/G (Τσακανίκας, 2005). Η εικόνα 1.5.3 δείχνει την διαφορετικότητα αυτών των δύο λόγων.



Εικόνα 1.5.2 Γραφική παράσταση της πράσινης έντασης G έναντι της κόκκινης έντασης R.
(Πηγή:Τσακανίκας, 2005. Ανάλυση γονιδιακών εκφράσεων νεοπλασιών σε Microarrays)



Εικόνα 1.5.3 Στο αριστερό ιστόγραμμα παρουσιάζεται η συχνότητα εμφάνισης του λόγου R/G, ενώ στο δεξί ιστόγραμμα παρουσιάζεται η συχνότητα εμφάνισης του λογαρίθμου με βάση το 2 του λόγου R/G.

(Πηγή:Τσακανίκας, 2005. Ανάλυση γονιδιακών εκφράσεων νεοπλασιών σε Microarrays)

Ο λόγος $T = R/G$ και συνεπώς ο λογάριθμος με βάση το 2 αυτού του λόγου ερμηνεύονται ως εξής (Τσακανίκας, 2005):

- Όταν $R_k/G_k > 1$, οπότε $\log_2(R_k/G_k) > 0$, τότε το k γονίδιο είναι υπερ-εκφρασμένο (*up-regulated*), διότι σε αυτό το γονίδιο το πλήθος των cDNAs από το δείγμα δοκιμής είναι μεγαλύτερο έναντι του πλήθους των cDNAs από το δείγμα αναφοράς.
- Όταν $R_k/G_k = 1$, οπότε $\log_2(R_k/G_k) = 0$, τότε το k γονίδιο είναι κανονικό (*normal*), διότι σε αυτό το γονίδιο το πλήθος των cDNAs από το δείγμα δοκιμής είναι ισοδύναμο με το πλήθος των cDNAs από το δείγμα αναφοράς.
- Όταν $R_k/G_k < 1$, οπότε $\log_2(R_k/G_k) < 0$, τότε το k γονίδιο είναι υπο-εκφρασμένο (*down-regulated*), διότι σε αυτό το γονίδιο το πλήθος των cDNAs από το δείγμα δοκιμής είναι μικρότερο έναντι του πλήθους των cDNAs από το δείγμα αναφοράς.

Λαμβάνοντας υπόψη την εικόνα 1.5.2 καταλαβαίνουμε ότι τα γονίδια που υπερ-εκφράζονται αποτελούν το «σύννεφο» σημείων που βρίσκεται πάνω από την ευθεία γραμμή, ενώ τα γονίδια που υπο-εκφράζονται αποτελούν το «σύννεφο» σημείων που βρίσκεται κάτω από την ευθεία γραμμή. Προφανώς, πάνω στην ευθεία γραμμή βρίσκονται τα κανονικά γονίδια.

1.6 Κανονικοποίηση των δεδομένων των μικροσυστάδων

Μετά την επεξεργασία και ανάλυση της εικόνας της μικροσυστάδας ακολουθεί η κανονικοποίηση των δεδομένων της μικροσυστάδας. Με τον όρο κανονικοποίηση (*normalization*) εννοούμε τον μετασχηματισμό των δεδομένων καταλλήλως με σκοπό την ελαχιστοποίηση των σφαλμάτων στις μετρήσεις εκφράσεων των γονιδίων, έτσι ώστε να είναι αξιόπιστη η σύγκριση των επιπέδων έκφρασής τους μεταξύ ενός δείγματος δοκιμής και ενός δείγματος αναφοράς σε διαφορετικές πειραματικές συνθήκες ή διαφορετικά πλακίδια (Τσακανίκας, 2005, Causton et al. 2003). Τα σφάλματα αυτά διακρίνονται σε τυχαία και συστηματικά. Τα τυχαία σφάλματα μπορεί να οφείλονται στην διαφορετική ποσότητα του RNA στα δύο δείγματα (αναφοράς και δοκιμής), ενώ τα συστηματικά σφάλματα μπορεί να οφείλονται

συνήθως στον εκτυπωτή που σαρώνει τα πλακίδια, στην φωτοευαισθησία και θερμοευαισθησία των φθοριζόντων μορίων που χρησιμοποιούνται για το μαρκάρισμα των cDNAs των δειγμάτων ή στην διαδικασία υβριδισμού (Τσακανίκας, 2005).

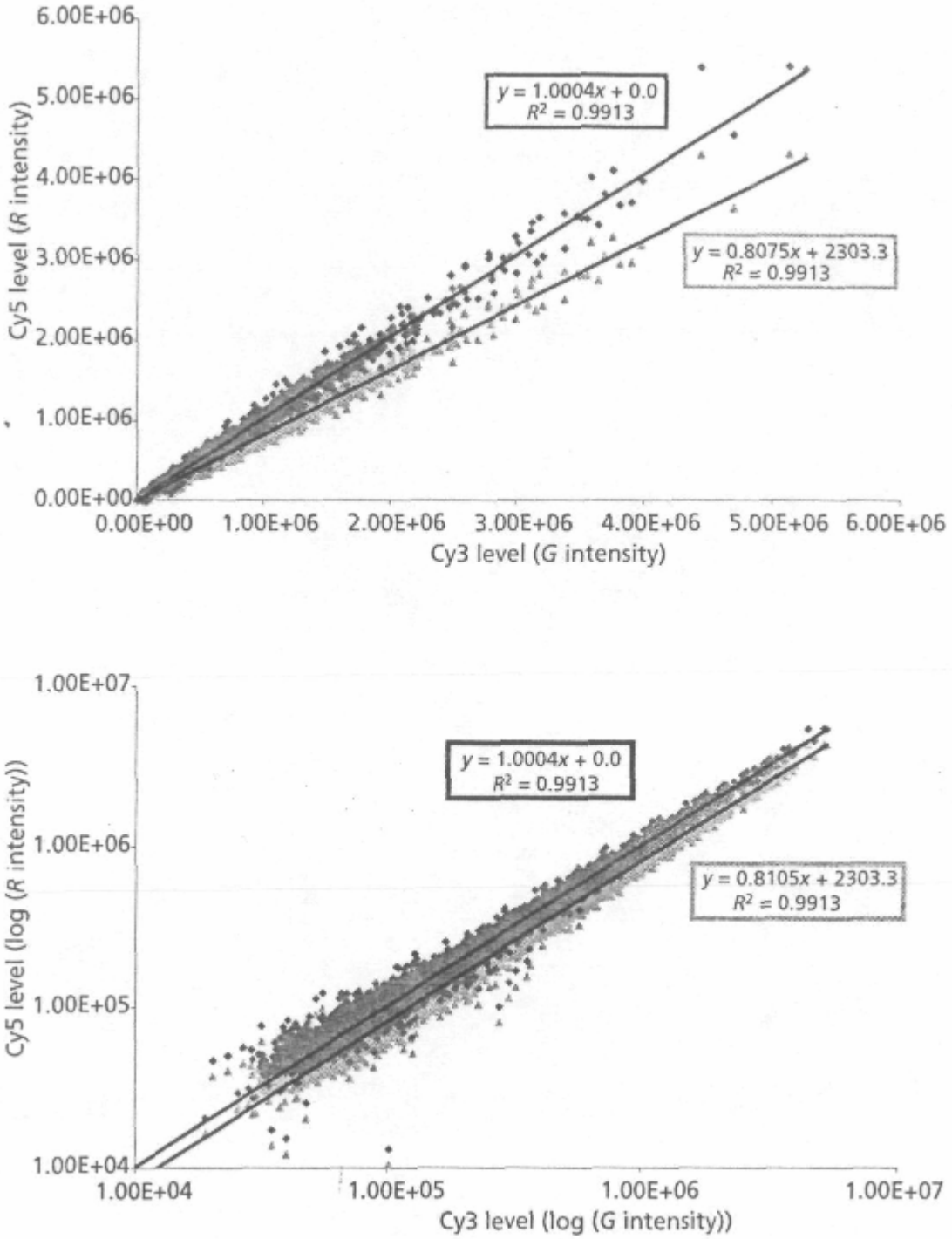
Για να κατανοήσουμε καλύτερα την σημασία των σφαλμάτων παρουσιάζουμε το παρακάτω παράδειγμα. Έστω ότι έχουμε δύο πανομοιότυπα mRNA δείγματα, όπου το ένα δείγμα μαρκάρεται με φθορίζοντα μόρια κόκκινου χρώματος, ενώ το άλλο δείγμα μαρκάρεται με φθορίζοντα μόρια πράσινου χρώματος και στην συνέχεια υφίστανται υβριδισμό στο ίδιο (πειραματικό) πλακίδιο. Είναι ιδιαίτερα σπάνιο αυτά τα δύο δείγματα να έχουν ίσες εντάσεις για τις αντίστοιχες κουκίδες του πλακιδίου (Τσακανίκας, 2005). Ωστόσο, μία τέτοια περίπτωση ισοδυναμίας εντάσεων στα δύο δείγματα μπορεί να επηρεάσει την αξιοπιστία των αποτελεσμάτων όσον αφορά τα επίπεδα έκφρασης των γονιδίων που μας ενδιαφέρουν, με αποτέλεσμα να μην εντοπιστούν μικρές βιολογικές διαφορές στα γονίδια αυτά.

1.6.1 Μέθοδοι κανονικοποίησης

Η κανονικοποίηση βασίζεται στην τυποποίηση (*scaling*) των εντάσεων ενός ή και των δύο δειγμάτων για κάθε γονίδιο έτσι ώστε να γίνουν ισοδύναμες (εικόνα 1.6.1) με αποτέλεσμα και οι λόγοι εντάσεων να είναι ισοδύναμοι μεταξύ τους. Αυτός ο μετασχηματισμός των εκφράσεων γίνεται με τέτοιον τρόπο, ώστε το μέσο επίπεδο έκφρασης όλων των υπό εξέταση γονιδίων να είναι ισοδύναμο στα δύο αυτά δείγματα. Συνεπώς, η κανονικοποίηση οδηγεί σε μέσο λόγο εκφράσεων όλων των γονιδίων ίσο με την μονάδα, δηλαδή

$$\overline{T} = \frac{\sum_{k=1}^{p} T_k}{p} = \frac{\sum_{k=1}^{p} \frac{R_i}{G_i}}{p} = \frac{\frac{R_1}{G_1} + \frac{R_2}{G_2} + \ldots + \frac{R_p}{G_p}}{p} = 1 \Leftrightarrow \frac{R_1}{G_1} + \frac{R_2}{G_2} + \ldots + \frac{R_p}{G_p} = p$$

Υπάρχουν διάφορες μέθοδοι τυποποίησης των δεδομένων (δηλαδή, των εκφράσεων των γονιδίων), οι οποίες βασίζονται σε κάποιες υποθέσεις που αφορούν τα δεδομένα και τον πειραματικό σχεδιασμό έτσι ώστε να επιλέξουμε την πιο κατάλληλη για το πείραμά μας. Παρακάτω παρουσιάζονται κάποιες από τις μεθόδους κανονικοποίησης που συνηθίζονται στην ανάλυση δεδομένων μικροσυστάδων (Causton et al., 2003).



Εικόνα 1.6.1 Τα παραπάνω διαγράμματα διασποράς παρουσιάζουν τις εκφράσεις των γονιδίων πριν την κανονικοποίηση με συντεταγμένες (G_k, R_k) και μετά την κανονικοποίηση με συντεταγμένες $(\log G_k, \log R_k)$. Μετά την κανονικοποίηση οι δύο βέλτιστες ευθείς, που προσαρμόζονται στα δεδομένα εκφράσεων, γίνονται παράλληλες.

(Πηγή: Causton et al. 2003. Microarray Gene Expression Data Analysis: A Beginner's Guide)

1.6.1.1 Ολική κανονικοποίηση εντάσεων (*Total intensity normalization*)

Θεωρείται η πιο εύκολη από τις μεθόδους κανονικοποίησης που θα παρουσιάσουμε και βασίζεται σε δύο υποθέσεις:

1. Η ποσότητα του mRNA είναι ίδια στο δείγμα αναφοράς και το δείγμα δοκιμής.
2. Οι μεταβολές που παρουσιάζονται στο επίπεδο έκφρασης των γονιδίων του πίνακα (βλέπε εικόνα 1.5.1) στα δύο δείγματα (αναφοράς και δοκιμής) θα πρέπει να εξισορροπηθούν, έτσι ώστε η συνολική ποσότητα RNA που υφίσταται υβριδισμό να είναι ίδια στα δύο δείγματα. Αυτό σημαίνει ότι η ένταση υβριδισμού θα είναι ισοδύναμη στα δύο αυτά δείγματα.

Για να ικανοποιούνται οι παραπάνω υποθέσεις θα πρέπει να υπολογιστεί ένας παράγοντας κανονικοποίησης (*normalization factor*), ο οποίος θα τυποποιήσει την ένταση κάθε γονιδίου του πίνακα. Συγκεκριμένα, ο παράγοντας κανονικοποίησης ισούται με τον λόγο του αθροίσματος των εντάσεων του δείγματος δοκιμής προς το άθροισμα των εντάσεων του δείγματος αναφοράς, δηλαδή

$$N_{Total} = \frac{\sum_{k=1}^{p} R_k}{\sum_{k=1}^{p} G_k}$$

όπου p είναι το πλήθος γονιδίων του πίνακα που μελετώνται. Στην συνέχεια σε κάθε γονίδιο η τυποποίηση της έντασής του στο δείγμα αναφοράς προκύπτει από το γινόμενο του παράγοντα κανονικοποίησης με την ένταση του γονιδίου αυτού στο δείγμα αναφοράς, ενώ η τυποποίηση της έντασής του στο δείγμα δοκιμής εξισώνεται με την έντασή του στο δείγμα δοκιμής πριν την τυποποίηση αντίστοιχα, δηλαδή

$$R'_k = R_k \text{ και } G'_k = N_{Total} \cdot G_k \text{ για } k = 1, 2, ..., p \qquad (1.6.1)$$

Τελικά, λαμβάνοντας υπόψη τις σχέσεις στο $(1.6.1)$ βρίσκουμε τον τυποποιημένο λόγο έκφρασης Τ κάθε γονιδίου, δηλαδή

$$T'_k = \frac{R'_k}{G'_k} = \frac{R_k}{N_{Total} \cdot G_k} = \frac{1}{N_{Total}} \cdot T_k \text{ για } k = 1, 2, ..., p \qquad (1.6.2)$$

και συνεπώς ο λογάριθμος με βάση το 2 της σχέσης $(1.6.2)$ είναι

$$\log_2 T'_k = \log_2 \left(\frac{1}{N_{Total}} \cdot T_k \right) = \log_2 T_k - \log_2 N_{Total} \text{ για } k = 1, 2, ..., p$$

Με την παραπάνω διαδικασία πετυχαίνουμε να έχουμε μέσο λόγο εντάσεων των δύο δειγμάτων $E(T')$ ίσο με μονάδα. Η συγκεκριμένη μέθοδος κανονικοποίησης ονομάζεται επίσης μονοπαραμετρική γραμμική κανονικοποίηση, διότι βασίζεται σε μία μόνο παράμετρο, τον παράγοντα κανονικοποίησης. Να σημειώσουμε ότι κάποιος μπορεί να ορίσει μία διαφορετική μεθοδολογία τυποποίησης, ώστε οι μέσες ή οι διάμεσες εντάσεις να είναι ισοδύναμες στα δύο δείγματα, προσέχοντας να ικανοποιούνται οι παραπάνω υποθέσεις.

1.6.1.2 Μέση λογαριθμο-κεντραρισμένη κανονικοποίηση (*Mean log-centering normalization*)

Η συγκεκριμένη μέθοδος κανονικοποίησης σχετίζεται με την προηγούμενη μέθοδο κανονικοποίησης, διότι βασίζεται στις υποθέσεις της τελευταίας και επιπλέον υποθέτουμε ότι ο μέσος των λογαρίθμων βάση μ3ε το 2 των λόγων T_k ισούται με μηδέν. Πάλι υπολογίζουμε τον παράγοντα κανονικοποίησης ως εξής

$$N_{mlc} = \frac{1}{p}\sum_{k=1}^{p} \log_2 T_k = \frac{1}{p}\sum_{k=1}^{p} \log_2 \left(\frac{R_k}{G_k}\right) \qquad (1.6.3)$$

Χρησιμοποιώντας την σχέση $(1.6.3)$, ο λογάριθμος με βάση το 2 του τυποποιημένου λόγου εντάσεων για κάθε γονίδιο προκύπτει ως εξής

$$\log_2 T_k' = \log_2 T_k - N_{mlc} \text{ για } k = 1,2,...,p \qquad (1.6.4)$$

και έτσι εγγυόμαστε ότι ο μέσος λογάριθμος του λόγου εντάσεων των δύο δειγμάτων $E(\log T)$ ισούται με μηδέν.

Αξίζει να σημειώσουμε ότι από την σχέση $(1.6.4)$ μπορούμε να καταλήξουμε στις παρακάτω σχέσεις

$$R_k' = R_k,\ G_k' = 2^{N_{mlc}} \cdot G_k \text{ και } T_k' = \frac{R_k'}{G_k'} = \frac{R_k}{2^{N_{mlc}} \cdot G_k} = \frac{1}{2^{N_{mlc}}} \cdot T_k \text{ για } k = 1,2,...,p$$

Εξάλλου, ισχύει ότι $\log_2 T_k' = \log_2 T_k - N_{mlc} \Leftrightarrow \frac{\ln T_k'}{\ln 2} = \frac{\ln T_k}{\ln 2} - \frac{N_{mlc} \cdot \ln 2}{\ln 2}$

$$\Leftrightarrow \ln T_k' = \ln T_k - N_{mlc} \cdot \ln 2$$

$$\Leftrightarrow \ln \frac{R_k}{G_k} - \ln \frac{R_k'}{G_k'} = \ln 2^{N_{mlc}}$$

$$\overset{R'_k = R_k}{\Leftrightarrow} \ln\frac{R_k}{G_k} - \ln\frac{R_k}{G_k'} = \ln 2^{N_{mlc}}$$

$$\Leftrightarrow \ln\frac{G'_k}{G_k} = \ln 2^{N_{mlc}}$$

$$\Leftrightarrow \exp\left(\ln\frac{G'_k}{G_k}\right) = \exp\left(\ln 2^{N_{mlc}}\right)$$

$$\Leftrightarrow G'_k = G_k \cdot 2^{N_{mlc}}$$

Παρατήρηση

Η συγκεκριμένη μέθοδος παρουσιάζει ευαισθησία όταν υπάρχουν ακραίες τιμές, διότι στην περίπτωση αυτή υπερεκτιμάει τον παράγοντα κανονικοποίησης N_{mlc}. Η αδυναμία αυτή αντιμετωπίζεται με επαναληπτική εφαρμογή αυτής της μεθόδου (*iterative mean log-centering*). Συγκεκριμένα, αρχικά προσαρμόζουμε τους λόγους T_k έτσι ώστε ο μέσος των λογαρίθμων με βάση το 2 των λόγων αυτών για όλα τα υπό εξέταση γονίδια να ισούται με μηδέν. Στην συνέχεια εντοπίζουμε τις ακραίες τιμές και τις εξαιρούμε και υπολογίζουμε το $E(T')$ για τα εναπομείναντα δεδομένα. Η διαδικασία αυτή επαναλαμβάνεται έως ότου δεν υπάρχουν ακραίες τιμές στα δεδομένα μας, όπου η διαδικασία θα συγκλίνει.

1.6.1.3 Γραμμική παλινδρόμηση (*Linear Regression*)

Συνηθίζεται να εφαρμόζεται όταν τα δύο δείγματα (αναφοράς και δοκιμής) σχετίζονται αρκετά μεταξύ τους. Στην περίπτωση αυτή περιμένουμε οι εκφράσεις των γονιδίων να μην μεταβάλλονται σημαντικά. Αν παραστήσουμε γραφικά τις εντάσεις του δείγματος δοκιμής έναντι των εντάσεων του δείγματος αναφοράς, τότε τα σημεία (ουσιαστικά είναι τα γονίδια) θα συγκεντρωθούν γύρω από μία ευθεία γραμμή. Η κλίση αυτής της ευθείας θα ισούται με την μονάδα αν και μόνο αν τα δύο δείγματα έχουν ίσες τις αντίστοιχες εντάσεις. Σκοπός της κανονικοποίησης με εφαρμογή γραμμικής παλινδρόμησης είναι να υπολογίσουμε την καλύτερη ευθεία που προσαρμόζεται στα δεδομένα μας (ουσιαστικά είναι οι λόγοι ή οι λογάριθμοι με βάση το 2 των λόγων των εντάσεων των δύο δειγμάτων) χρησιμοποιώντας τεχνικές

παλινδρόμησης και στη συνέχεια να προσαρμόσουμε τα δεδομένα έτσι ώστε η κλίση της ευθείας να ισούται με τη μονάδα.

Στην μέθοδο κανονικοποίησης με γραμμική παλινδρόμηση αυτή υποθέτουμε ότι:

- για κάθε γονίδιο η έντασή του στο δείγμα δοκιμής είναι γραμμική συνάρτηση της έντασής του στο δείγμα αναφοράς, οπότε προκύπτει το παρακάτω μοντέλο παλινδρόμησης

$$R_k = b_0 + b_1 \cdot G_k + e_k \text{ για } k = 1, 2, ..., p \qquad (1.6.5)$$

όπου οι άγνωστες παράμετροι b_0 και b_1 είναι η τεταγμένη και η κλίση του μοντέλου $(1.6.5)$,

- οι ποσότητες $e_1, e_2, ..., e_p$ είναι τα σφάλματα, τα οποία αποτελούν τυχαίες ασυσχέτιστες μεταβλητές
- τα σφάλματα κατανέμονται σύμφωνα με την κανονική κατανομή $N(0, \sigma^2)$, οπότε τα e_k δεν είναι μόνο ασυσχέτιστα αλλά και ανεξάρτητα.

Η κλίση και η τεταγμένη της ευθείας παλινδρόμησης εκτιμώνται αντίστοιχα από τις σχέσεις

$$\hat{b}_1 = \frac{\sum_{k=1}^{p} (R_k - \bar{R})(G_k - \bar{G})}{\sum_{k=1}^{p} (G_k - \bar{G})^2} \text{ και } \hat{b}_0 = \bar{R} - \hat{b}_1 \cdot \bar{G} \qquad (1.6.6)$$

Οι εντάσεις του δείγματος αναφοράς και του δείγματος δοκιμής τυποποιούνται αντίστοιχα ως εξής

$$G'_k = \frac{G_k - \hat{b}_0}{\hat{b}_1} \text{ και } R'_k = R_k \text{ για } k = 1, 2, ..., p \qquad (1.6.7)$$

έτσι ώστε η μέση τιμή των λόγων των τυποποιημένων εντάσεων των δύο δειγμάτων, T'_k, να ισούται με μονάδα.

Λαμβάνοντας υπόψη τις σχέσεις $(1.6.5)$ και $(1.6.6)$ ο τυποποιημένος λόγος εντάσεων κάθε γονιδίου υπολογίζεται από την παρακάτω σχέση

$$T'_k = \frac{R'_k}{G'_k} = \frac{R_k}{\frac{G_k - \hat{b}_0}{\hat{b}_1}} = \frac{\hat{b}_1}{G_k - \hat{b}_0} \cdot (\hat{b}_0 + \hat{b}_1 \cdot G_k) \text{ για } k = 1, 2, ..., p$$

1.6.1.4 Στατιστική αναλογία του Chen (*Chen's ratio statistics*)

Γνωρίζουμε ότι κάθε γονίδιο μπορεί να υπέρ- ή να υπό-εκφράζεται. Ωστόσο, όταν υπάρχουν κύτταρα που είναι συσχετισμένα μεταξύ τους, τότε είναι δυνατόν η συνολική ποσότητα mRNA να είναι ίδια στα δύο δείγματα για ένα υποσύνολο γονιδίων γνωστών ως *housekeeping* γονίδια, τα οποία θεωρούνται ότι έχουν μικρή μεταβλητότητα στις εκφράσεις τους για κάποιες συνθήκες. Στην περίπτωση αυτή σύμφωνα με τον Chen (Chen et al. 1997) υποθέτουμε ότι οι εκφράσεις αυτών των γονιδίων ακολουθούν μια κατανομή με σταθερή μέση τιμή μ και σταθερή τυπική απόκλιση σ ανεξαρτήτως δείγματος (δηλαδή, $\mu_R = \mu_G$ και $\sigma_R = \sigma_G$, όπου $\sigma_R = c \cdot \mu_R$ και $c > 0$ μια σταθερά) και έχουν πυκνότητα πιθανότητας $f_T(t)$. Χρησιμοποιείται μία επαναληπτική διαδικασία όπου κανονικοποιεί τις εντάσεις έτσι ώστε ο μέσος των λόγων των εντάσεων των δύο δειγμάτων να ισούται με μονάδα. Επιπλέον, υπολογίζει διαστήματα εμπιστοσύνης για τους λόγους T_k', έτσι ώστε να εντοπίσει τα σημαντικά γονίδια (Chen et al. 1997).

1.6.1.5 Η Lowess κανονικοποίηση (*Lowess normalization*)

Η μέθοδος της Lowess (*locally weighted scatterplot smoothing*) κανονικοποίησης εφαρμόζεται όταν παρατηρηθεί ότι ο λογάριθμος με βάση το 2 του λόγου των εντάσεων, $\log_2(T)$, παρουσιάζει σημαντική μεταβλητότητα στις χαμηλές εντάσεις. Στην περίπτωση αυτή συμπεραίνουμε ότι ο $\log_2(T)$ επηρεάζεται από τον βαθμό των εντάσεων. Για να εξετάσουμε αν υφίσταται αυτή η επίδραση χρησιμοποιούμε το $R-I$ γράφημα (*Ratio-Intensity plot*), στο οποίο αναπαριστάνουμε τις ποσότητες $\log_2(R/G)$ ως συναρτήσεις των αντίστοιχων ποσοτήτων $\log_{10}(R \times G)$. Σύμφωνα με την εικόνα 1.6.2 παρατηρούμε ότι τα σημεία στην αρχή του γραφήματος φαίνονται αρκετά διασκορπισμένα μεταξύ τους, ενώ καθώς αυξάνεται το $\log_{10}(R \times G)$ τα σημεία τείνουν σαν να συγκεντρώνονται γύρω από μία ευθεία με κλίση μηδέν. Αυτό σημαίνει ότι στις χαμηλές εντάσεις ο $\log_2(T)$ παρουσιάζει αυξημένη μεταβλητότητα η οποία μειώνεται και τείνει να σταθεροποιηθεί καθώς αυξάνονται οι εντάσεις.

Συνεπώς, το $R-I$ γράφημα μπορεί να θεωρηθεί ως ένα διαγνωστικό εργαλείο της εξάρτησης του $\log_2(T)$ από τον βαθμό των εντάσεων.

Με την εφαρμογή της Lowess κανονικοποίησης αποσκοπούμε στην εξάλειψη της επίδρασης του βαθμού της έντασης στον $\log_2(T)$. Συγκεκριμένα, η διαδικασία που ακολουθείται περιγράφεται παρακάτω:

1. Για κάθε σημείο του $R-I$ γραφήματος ορίζουμε μία στάθμιση έτσι ώστε να δίνεται λιγότερο βάρος στα σημεία που είναι απομακρυσμένα από τα γειτονικά τους σημεία. Με αυτό τον τρόπο τα απομακρυσμένα σημεία συνεισφέρουν λιγότερο στην προσαρμογή της βέλτιστης καμπύλης παλινδρόμησης. Η πιο συνηθισμένη σταθμισμένη συνάρτηση είναι η τρι-κυβική (*tri-cube weight function*), η οποία ορίζεται ως εξής:

$$w(u)=\begin{cases}\left(1-\left|u^3\right|\right)^3 & |u|<1\\ 0 & |u|>1\end{cases}$$

όπου η ποσότητα u είναι η απόσταση μεταξύ του υπό εξέταση σημείου και ενός γειτονικού του.

2. Εάν ορίσουμε $x_k=\log_{10}(R_k\cdot G_k)$ και $y_k=\log_{10}(R_k/G_k)$ για $k=1,2,...,p$, τότε εφαρμόζουμε σταθμισμένη γραμμική παλινδρόμηση σε κάθε σημείο του $R-I$ γραφήματος για να εκτιμήσουμε την βέλτιστη καμπύλη παλινδρόμησης, δηλαδή

$$y(x_k)\ \text{ή}\ y_k=\hat{b}_0+\hat{b}_1\cdot x_k \Leftrightarrow \log_2(R_k/G_k)=\hat{b}_0+\hat{b}_1\cdot\log_{10}(R_k\cdot G_k) \qquad (1.6.8)$$

$$\text{όπου}\ \hat{b}_1=\frac{\sum_{k=1}^{p}w_k(y_k-\bar{y})(x_k-\bar{x})}{\sum_{k=1}^{p}w_k(x_k-\bar{x})^2}\ \text{και}\ \hat{b}_0=\bar{y}-\hat{b}_1\cdot\bar{x} \qquad (1.6.9)$$

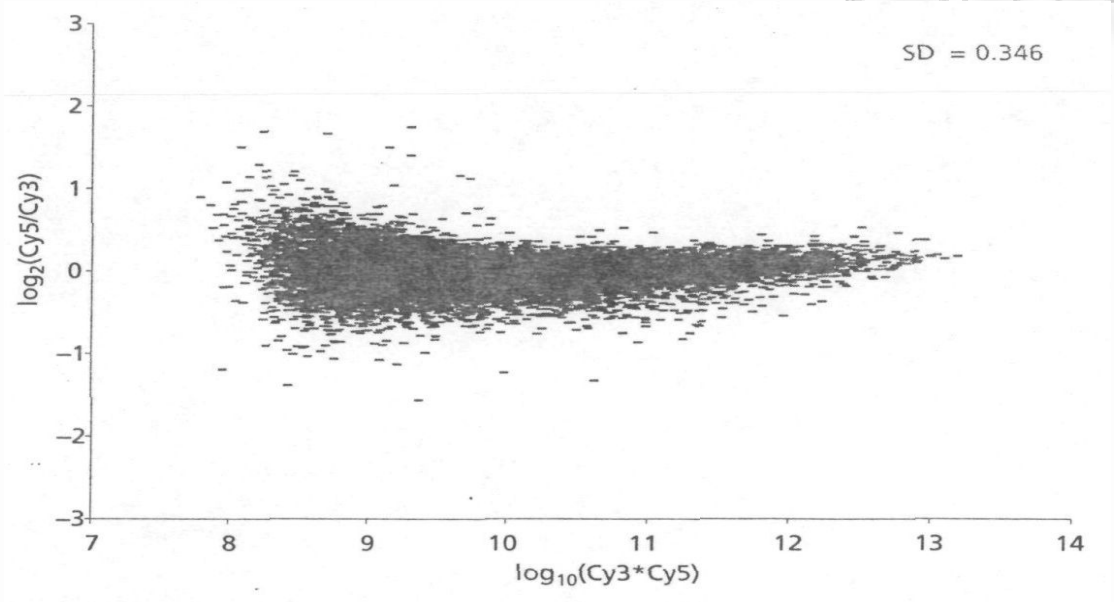
3. Στην συνέχεια λαμβάνοντας υπόψη τις σχέσεις $(1.6.8)$ και $(1.6.9)$ υπολογίζουμε τους τυποποιημένους $\log_2(T)$ ως εξής:

$$\log_2 T_k' = \log_2 T_k - y(x_k) \Leftrightarrow \log_2 T_k' = \log_2 T_k - \log_2\left(2^{y(x_k)}\right)$$

$$\Leftrightarrow \log_2 T_k' = \log_2\left(\frac{T_k}{2^{y(x_k)}}\right) = \log_2\left(\frac{R_k}{G_k}\cdot\frac{1}{2^{y(x_k)}}\right)$$

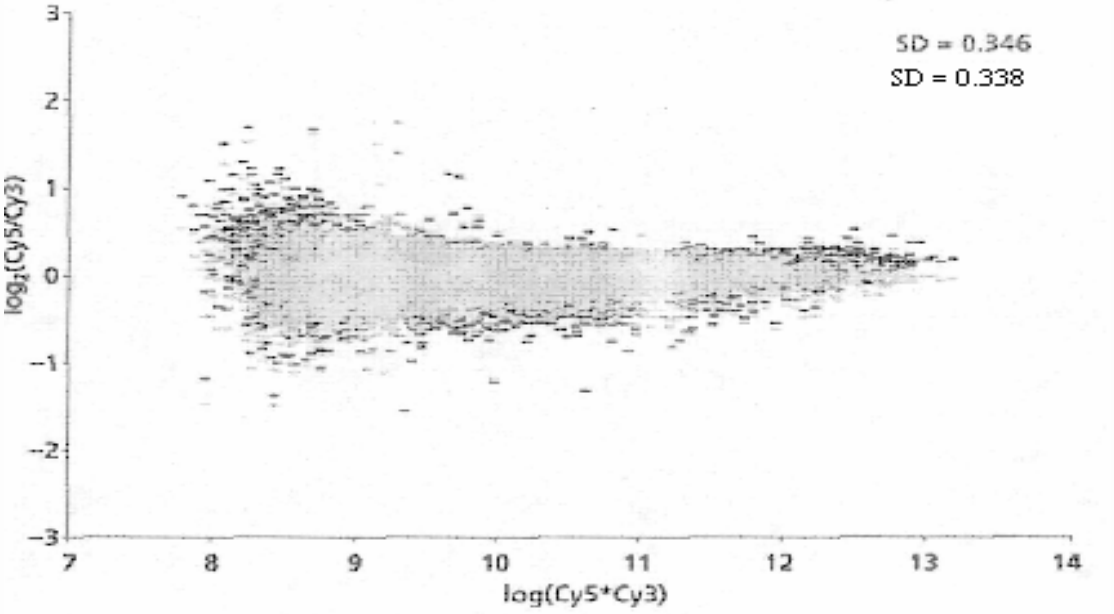
$$\Leftrightarrow \log_2 T_k' = \log_2\left(\frac{R_k'}{G_k'}\cdot\right)$$

Οι εντάσεις στο δείγμα αναφοράς και στο δείγμα δοκιμής έχουν τυποποιηθεί αντίστοιχα ως $G_k' = G_k \cdot 2^{y(x_k)}$ και $R_k' = R_k$ έτσι ώστε η μέση τιμή των $\log_2\left(T_k'\right)$ να ισούται με μηδέν, όπως γνωρίζουμε ήδη.



Εικόνα 1.6.2 Το $R-I$ γράφημα πριν εφαρμοστεί η μέθοδος της Lowess κανονικοποίησης. Η ποσότητα SD είναι η τυπική απόκλιση των σημείων του γραφήματος.

(Πηγή: Causton et al. 2003. Microarray Gene Expression Data Analysis: A Beginner's Guide)



Εικόνα 1.6.3 Το $R-I$ γράφημα μετά την εφαρμογή της μεθόδου της Lowess κανονικοποίησης.

(Πηγή: Causton et al. 2003. Microarray Gene Expression Data Analysis: A Beginner's Guide)

Μετά την εφαρμογή της Lowess κανονικοποίησης, αλλάζει η εικόνα του $R-I$ γραφήματος σύμφωνα με την εικόνα 1.6.3, αφού τα κανονικοποιημένα δεδομένα ξεχωρίζουν από τα μη κανονικοποιημένα, διότι τα πρώτα έχουν ανοιχτό γκρι χρώμα, ενώ τα δεύτερα έχουν σκούρο γκρι. Τα κανονικοποιημένα δεδομένα είναι συμμετρικά κατανεμημένα γύρω από τον $E\left(\log_2\left(T'\right)\right)=0$ και έχουν τυπική απόκλιση 0.338.

Παρατήρηση

Όλες οι παραπάνω μέθοδοι κανονικοποίησης περιγράφηκαν για όλο το σύνολο δεδομένων που μελετάμε. Στην περίπτωση αυτή λέμε ότι εφαρμόσαμε ολική κανονικοποίηση (*global normalization*). Ωστόσο, είναι δυνατόν οι παραπάνω μέθοδοι να εφαρμοστούν σε κάποιο υποσύνολο δεδομένων, οπότε εφαρμόζουμε τοπική κανονικοποίηση (*local normalization*). Στην τοπική κανονικοποίηση ισχύουν όλες οι υποθέσεις που παρουσιάσαμε στην ολική κανονικοποίηση. Η τοπική κανονικοποίηση προτιμάται όταν θέλουμε να εξαλείψουμε κάποια τοπικά συστηματικά σφάλματα, όπως τοπικές διαφορές στις συνθήκες υβριδισμού ή διαφορές ως προς το μήκος μεταξύ των κουκίδων ενός πλακιδίου (Causton et al. 2003).

1.6.2 Γενικές παρατηρήσεις στην κανονικοποίηση

Σίγουρα η κανονικοποίηση θεωρείται σημαντικό στατιστικό εργαλείο, διότι «διορθώνει» τα σφάλματα που παρουσιάζονται κατά τη διεξαγωγή του πειράματος. Ωστόσο δεν πρέπει να γίνεται αλόγιστη χρήση της ή να εφαρμόζουμε ένα πρόχειρα σχεδιασμένο πείραμα με το σκεπτικό ότι οποιοδήποτε λάθος θα εξαλειφθεί με την κανονικοποίηση, διότι υπάρχει ο κίνδυνος τα μετασχηματισμένα δεδομένα μας να υστερούν σε ακρίβεια ή να δημιουργούν μεγαλύτερη μεταβλητότητα από αυτήν που επιχειρήσαμε να εξαλείψουμε. Είναι πιο συνετό να σχεδιάζουμε προσεχτικά και σωστά το πείραμα που θα εκτελέσουμε τόσο ως προς τη συλλογή των δεδομένων όσο και ως προς τη διάθεση και το χειρισμό του απαραίτητου τεχνολογικού εξοπλισμού για το πείραμα. Με αυτόν τον τρόπο μπορούμε να μειώσουμε αρκετά τις μετά-

επεμβάσεις στα δεδομένα που αποκτούμε από τις εικόνες των μικροσυστάδων (Τσακανίκας, 2005).

Η μέθοδος κανονικοποίησης που θα επιλεγεί εξαρτάται άμεσα από την ιδιαιτερότητα και τη φύση του συνόλου των δεδομένων που μελετάμε. Είναι πιο ασφαλές να χρησιμοποιούμε την πιο ομαλή και πιο «φιλική» στα δεδομένα μέθοδο κανονικοποίησης, ώστε να διασφαλίσουμε ότι έχουμε μετασχηματίσει τα δεδομένα όσο τον δυνατόν πιο σωστά. Οι παραπάνω μέθοδοι κανονικοποίησης μπορούν να συγκεντρωθούν σε ένα πίνακα (πίνακας 1.6.1), ο οποίος παρουσιάζει τη μέθοδο κανονικοποίησης και την περίπτωση στην οποία προτείνεται να εφαρμοστεί.

Πίνακας 1.6.1

Μέθοδος Κανονικοποίησης	**Περίπτωση**
Ολική κανονικοποίηση	κλασική
Μέση λογαριθμο-κεντραρισμένη κανονικοποίηση (επαναληπτικά)	ύπαρξη ακραίων εντάσεων
Στατιστική αναλογία του Chen	ύπαρξη *housekeeping* γονιδίων
Lowess κανονικοποίηση	υψηλή μεταβλητότητα των λόγων των χαμηλών εντάσεων των δύο δειγμάτων (δοκιμής και αναφοράς)
Γραμμική παλινδρόμηση	χαμηλή μεταβλητότητα των εκφράσεων των γονιδίων λόγω συσχέτισης των δύο δειγμάτων (δοκιμής και αναφοράς)

1.7 Μέθοδοι εντοπισμού των σημαντικών γονιδίων: Φιλτράρισμα δεδομένων (*Data filtering*)

Τόσο η κανονικοποίηση όσο και το φιλτράρισμα των δεδομένων αποτελούν μεθόδους μετασχηματισμού των δεδομένων. Με το φιλτράρισμα αποσκοπούμε στο να μειώσουμε την πολυπλοκότητα και να αυξήσουμε την ποιότητας ενός συνόλου δεδομένων απομακρύνοντας δεδομένα αμφιλεγόμενα ή χαμηλής ποιότητας. Επιπλέον, επιδιώκουμε να εντοπίσουμε τα γονίδια που θεωρούνται σημαντικά ως

προς το επίπεδο έκφρασής τους, δηλαδή τα γονίδια των οποίων το επίπεδο έκφρασης αλλάζει κατά τη διάρκεια ενός πειράματος (Causton et al. 2003).

Απώτερος σκοπός της πραγματοποίησης πειραμάτων σε μικροσυστάδες είναι ο εντοπισμός των πιο σημαντικών γονιδίων του συνόλου που μελετάται. Σημαντικό θεωρείται το γονίδιο του οποίου το επίπεδο έκφρασης διαφοροποιείται (*differentially expressed gene*) όταν συγκρίνεται σε δύο διαφορετικές συνθήκες, όπως υγιή και μη υγιή κύτταρα. Έχουν προταθεί διάφορες στατιστικές μέθοδοι για τον εντοπισμό των σημαντικών γονιδίων, οι οποίες είναι δυνατόν να επηρεάσουν το σύνολο γονιδίων που θα επιλεγούν (Witten et al. 2007). Λόγω της φύσης των δεδομένων που μελετάμε θα πρέπει να είμαστε προσεχτικοί στην επιλογή της μεθόδου.

1.7.1 Η μέθοδος *fold-change*

Μία απλή μέθοδος εντοπισμού των πιο σημαντικών γονιδίων είναι η *fold-change*, η οποία εκφράζει πόσο διαφέρουν δύο μεταβλητές. Ο κλασικός ορισμός της *fold-change* μεθόδου για το k γονίδιο είναι (Witten et al. 2007):

$$FC_k = \frac{\bar{G}_{k\cdot}}{\bar{R}_{k\cdot}} \text{ για } \bar{G}_{k\cdot} = \sum_j G_{kj} \text{ και } \bar{R}_{k\cdot} = \sum_j R_{kj}$$

όπου G_{kj} και R_{kj} είναι το επίπεδο έκφρασης του k γονιδίου στο δείγμα αναφοράς και δείγμα δοκιμής αντίστοιχα στην j επανάληψη του πειράματος. Ένας άλλος ορισμός της *fold-change* μεθόδου για το k γονίδιο δίνεται παρακάτω σύμφωνα με τους Guo και Choe (Witten et al. 2007):

$$FC_k = \bar{G}_{k\cdot} - \bar{R}_{k\cdot}$$

Οι δύο παραπάνω ορισμοί της *fold-change* μεθόδου είναι γνωστοί με το όνομα FC_{ratio} και $FC_{difference}$ αντίστοιχα. Η *fold-change* διαλέγει ως σημαντικά τα γονίδια που έχουν μεγάλη διαφορά στο επίπεδο έκφρασής τους μεταξύ των δύο δειγμάτων. Για παράδειγμα, αν ορίσουμε ως κατώφλι την $FC_{difference}$, τότε ένα γονίδιο θεωρείται σημαντικό αν η έκφρασή του στο ένα δείγμα είναι πάνω από $k -$ $FC_{difference}$ μεγαλύτερη ή μικρότερη σε σχέση με την έκφρασή του στο άλλο δείγμα (McLachlan et al. 2004). Για την τιμή του k συνήθως επιλέγουμε $k = 2$ (Τσακανίκας, 2005).

Ωστόσο, αυτή η μέθοδος δεν αποτελεί κάποιο στατιστικό τεστ και επιπλέον αγνοεί την μεταβλητότητα που υπάρχει μεταξύ των επαναλήψεων του πειράματος σε κάθε

δείγμα (δοκιμής ή αναφοράς) με αποτέλεσμα να θεωρείται αναξιόπιστη και ανεπαρκής για τον εντοπισμό των σημαντικών γονιδίων (McLachlan et al. 2004).

1.7.2 Έλεγχος πολλαπλών υποθέσεων (*multiple hypothesis test*)

Για να ελέγξουμε πόσο διαφέρει κάθε γονίδιο στο δείγμα αναφοράς σε σχέση με το δείγμα δοκιμής, μπορούμε να εφαρμόσουμε έλεγχο πολλαπλών υποθέσεων (McLachlan et al. 2004).

Έστω για το k γονίδιο ορίζουμε την μηδενική υπόθεση H_{0k} : η έκφρασή του δεν διαφέρει σημαντικά στα δύο δείγματα και την εναλλακτική υπόθεση H_{1k} : η έκφρασή του διαφέρει σημαντικά στα δύο δείγματα. Αυτό ισοδυναμεί με τον έλεγχο της μηδενικής υπόθεσης $H_{0k} : T_k = 1$ έναντι της εναλλακτικής υπόθεσης $H_{1k} : T_k \neq 1$, $k = 1, 2, ..., p$. Στην συνέχεια ορίζουμε για τον παραπάνω έλεγχο την στατιστική συνάρτηση

$$T_k = \frac{R_k}{G_k}$$

Να σημειώσουμε ότι δεν πρέπει να συγχέουμε αυτό το τεστ με το $t-test$ για έναν πληθυσμό.

Η μηδενική υπόθεση θα απορρίπτεται σε επίπεδο σημαντικότητας a εάν η τιμή της συνάρτησης T_k ανήκει στην κρίσιμη περιοχή $\Gamma_k = \{T_k : |T_k| \geq c_a\}$, όπου $P\{|T_k| \geq c_a | H_k\} = a$ είναι το σφάλμα τύπου I και c_a το άνω $a-$σημείο της κατανομής που ακολουθεί η T_k. Εναλλακτικά, η μηδενική υπόθεση θα απορρίπτεται σε επίπεδο σημαντικότητας a εάν το $p-value$ είναι μεγαλύτερο από το επίπεδο σημαντικότητας, δηλαδή $P_k = P\{|T_k| \geq |t_k| | H_k\} > a$, όπου t_k είναι η παρατηρούμενη τιμή του στατιστικού T_k.

Στην πράξη ερχόμαστε αντιμέτωποι με μεγάλο πλήθος γονιδίων με αποτέλεσμα η εφαρμογή αυτής της μεθόδου να αποδειχθεί εξαιρετικά χρονοβόρα και επιπλέον να αυξήσει σημαντικά το σφάλμα τύπου I (*false positive rate*), το οποίο ορίζεται ως η πιθανότητα να απορρίψουμε λανθασμένα τουλάχιστον μία μηδενική υπόθεση, δηλαδή $1-(1-a)^p$, όπου p είναι το πλήθος των (ανεξάρτητων) γονιδίων που εξετάζουμε (McLachlan et al. 2004).

1.7.3 Επιλογή γονιδίων σύμφωνα με την εξάρτησή τους από την ένταση

Όπως έχει ήδη αναφερθεί στην Lowess κανονικοποίηση, υπάρχει η τάση οι χαμηλές τιμές του $\log_2(T_k)$ να παρουσιάζουν αρκετή μεταβλητότητα, η οποία σταθεροποιείται καθώς αυξάνεται η τιμή αυτού του λογάριθμου με αποτέλεσμα να δημιουργείται η υποψία ότι υπάρχει κάποια συσχέτιση μεταξύ των μεταβλητών R_k και G_k. Αυτή η τάση ονομάζεται εξάρτηση από την ένταση (*intensity-dependent*).

Σύμφωνα με την παράγραφο 1.5.2 ο λογάριθμος με βάση το 2 του λόγου των εντάσεων των δύο δειγμάτων ακολουθεί κατά προσέγγιση την κανονική κατανομή. Μπορούμε να υπολογίσουμε για κάθε γονίδιο ένα $Z-score$ (Causton et al. 2003), το οποίο δείχνει πόσες τυπικές αποκλίσεις είναι μεγαλύτερη ή μικρότερη η έκφραση κάθε γονιδίου από την μέση έκφραση. Αρχικά θα εκτιμήσουμε την μέση τιμή και την τυπική απόκλιση των λογαρίθμων με βάση το 2 των λόγων των εντάσεων των δύο δειγμάτων για όλο το σύνολο των γονιδίων που μελετάμε, δηλαδή

$$\hat{\mu} = \frac{\sum_{k=1}^{p} \log_2(T_k)}{p} \text{ και } \hat{\sigma} = \sqrt{\frac{\sum_{k=1}^{p} (\log_2 T_k - \mu)^2}{p-1}} \text{ αντίστοιχα}$$

Στην συνέχεια για το k γονίδιο υπολογίζουμε το $Z-score$ ως εξής:

$$z_k = \frac{\log_2 T_k - \hat{\mu}}{\hat{\sigma}} \quad k = 1, 2, ..., p$$

Σε επίπεδο σημαντικότητας a το k γονίδιο θα θεωρείται σημαντικό αν και μόνο αν για το $Z-score$ του ισχύει ότι $|z_k| > z_{a/2}$. Συνεπώς, γι' αυτό το γονίδιο καταλαβαίνουμε ότι κατ' απόλυτη τιμή η έκφρασή του είναι μεγαλύτερη από $z_{a/2}$ τυπικές αποκλίσεις από το ολικό μέσο επίπεδο έκφρασης.

1.7.4 Ανάλυση της διακύμανσης (*Analysis of variation, ANOVA*)

Η ANOVA είναι ένα χρήσιμο στατιστικό εργαλείο στην μελέτη μεγάλου συνόλου δεδομένων. Μπορούμε να εκτιμήσουμε ποιοι παράγοντες επηρεάζουν και πόσο κατά μέσο όρο την μεταβλητή απόκρισης, να συγκρίνουμε αυτούς τους παράγοντες ως προς την μέση επίδρασή τους στην μεταβλητή απόκρισης, να μελετήσουμε την διακύμανση εντός των επιπέδων των παραγόντων (*within-groups variability*), καθώς

επίσης και την διακύμανση μεταξύ των παραγόντων (*between-groups variability*) (Causton et al. 2003).

Με την χρήση της μεθόδου ANOVA επιδιώκουμε να ελέγξουμε εάν διαφέρουν σημαντικά οι μέσες εκφράσεις των γονιδίων. Να σημειώσουμε ότι στην περίπτωση που συγκρίνουμε μόνο δύο γονίδια, τότε η ANOVA συμπίπτει με το $t-test$. Βασικό χαρακτηριστικό της ANOVA είναι ότι μπορούμε να διαμερίσουμε την ολική μεταβλητότητα (*total sum of squares, SSTO*) των δεδομένων στην μεταβλητότητα που οφείλεται στο τυχαίο σφάλμα (*total within-groups variability* ή *error sum of squares, SSE*) και σε επιμέρους μεταβλητότητες που οφείλονται στους παράγοντες που μελετάμε. Μεγάλη απόκλιση της μεταβλητότητας που οφείλεται στο παράγοντα Α από την μεταβλητότητα που οφείλεται στο σφάλμα σημαίνει ότι ο παράγοντας Α συνεισφέρει σημαντικά στην ολική μεταβλητότητα των δεδομένων, οπότε επηρεάζει το (λογαριθμικό) επίπεδο έκφρασης των γονιδίων (Causton et al. 2003).

Στην ανάλυση μικροσυστάδων η μεταβλητή απόκρισης είναι το επίπεδο έκφρασης του mRNA που αντιστοιχεί σε ένα συγκεκριμένο γονίδιο στον πίνακα και επηρεάζεται από ένα σύνολο παραγόντων που συνεισφέρουν στην τυχαία και συστηματική μεταβλητότητα των μετρήσεων (Causton et al. 2003). Οι παράγοντες που πιθανώς να επηρεάζουν την μεταβλητή απόκρισης είναι ο πίνακας (*array*), η απόχρωση (*dye*), το δείγμα mRNAs (*variety*) και το γονίδιο (*gene*). Συνεπώς, το μοντέλο ANOVA που θα μελετήσουμε δίνεται παρακάτω (Causton et al. 2003):

$$y_{ijkr} = y_0 + A_i + D_j + V_k + G_r + (AD)_{ij} + (AV)_{ik} + (AG)_{ir} + (DV)_{jk} + (DG)_{jr} + (VG)_{kr} + e_{ijkr}$$

όπου y_0 είναι μία σταθερά που εκφράζει την ολική μέση επίδραση των παραγόντων

A_i είναι η επίδραση του i πίνακα, $i = 1, 2, ..., a$

D_j είναι η επίδραση της j απόχρωσης, $j = 1, 2, ..., d$

V_k είναι η επίδραση του k δείγματος, $k = 1, 2, ..., v$

G_r είναι η επίδραση του r γονιδίου, $r = 1, 2, ..., p$

$(AD)_{ij}$ είναι η αλληλεπίδραση του i πίνακα με την j απόχρωση και εκφράζει την επίδραση της j απόχρωσης στην απόκριση, όταν μελετάμε τον i πίνακα.

$(AV)_{ik}$ είναι η αλληλεπίδραση του i πίνακα με το k δείγμα και εκφράζει την επίδραση του k δείγματος στην απόκριση, όταν μελετάμε τον i πίνακα

$(AG)_{ir}$ είναι η αλληλεπίδραση του i πίνακα με το r γονίδιο και εκφράζει την επίδραση του r γονιδίου στην απόκριση, όταν μελετάμε τον i πίνακα

$(DV)_{jk}$ είναι η αλληλεπίδραση της j απόχρωσης με το k δείγμα και εκφράζει την επίδραση του k δείγματος στην απόκριση, όταν χρησιμοποιούμε την j απόχρωση

$(DG)_{jr}$ είναι η αλληλεπίδραση της j απόχρωσης με το r γονίδιο και εκφράζει την επίδραση του r γονιδίου στην απόκριση, όταν χρησιμοποιούμε την j απόχρωση

$(VG)_{kr}$ είναι η αλληλεπίδραση του k δείγματος με το r γονίδιο και εκφράζει την επίδραση του r γονιδίου στην απόκριση, όταν μελετάμε το k δείγμα

e_{ijkr} είναι το τυχαίο σφάλμα, για το οποίο υποθέτουμε ότι ακολουθεί την κανονική κατανομή με μέση τιμή 0 και σταθερή διακύμανση.

Ο πίνακας 1.7.1 (Παράρτημα Α) δείχνει τον σχεδιασμό αυτού του πειράματος, ενώ ο πίνακας 1.7.2 (Παράρτημα Α) παρουσιάζει την γενική μορφή του πίνακα ANOVA. Από όλες τις παραμέτρους του μοντέλου η αλληλεπίδραση (VG) μας ενδιαφέρει περισσότερο, γιατί αφορά την αλληλεπίδραση των πιο σημαντικών μεταβλητών. Για να ελέγξουμε αν υπάρχουν σημαντικά γονίδια, αρκεί να εφαρμόσουμε το $F-test$ (Causton et al. 2003) με το οποίο εξετάζουμε αν όλα τα γονίδια εκφράζονται το ίδιο κατά μέσο όρο για κάθε δείγμα (συνήθως, αναφοράς ή δοκιμής) ή υπάρχει τουλάχιστον ένα γονίδιο που διαφοροποιείται. Ουσιαστικά εξετάζουμε την μηδενική υπόθεση $H_0:(VG)_{kr}=0$ για όλα τα ζεύγη k,r έναντι της εναλλακτικής υπόθεσης $H_1:(VG)_{kr}\neq 0$ για κάποιες τιμές των k,r χρησιμοποιώντας την στατιστική συνάρτηση

$$F=\frac{SSVG}{SSE}$$

η οποία ακολουθεί κατανομή F με $(v-1)(p-1)$ και $adv(p-1)$ βαθμούς ελευθερίας υπό την μηδενική υπόθεση.

Η απόρριψη της μηδενικής υπόθεσης δίνει το έναυσμα για να εντοπίσουμε αυτά τα γονίδια που διαφοροποιούνται κατά μέσο όρο από τα υπόλοιπα. Στην περίπτωση αυτή το κριτήριο για τον εντοπισμό των σημαντικών γονιδίων είναι το $t-test$ με το οποίο μπορούμε να συγκρίνουμε ανά δύο τα γονίδια για κάθε απόχρωση για το αν

διαφοροποιούνται ή όχι κατά μέσο όρο. Συγκεκριμένα εξετάζουμε την μηδενική υπόθεση

$$H_0 : (VG)_{kr} - (VG)_{kr'} = 0 \Leftrightarrow H_0 : (VG)_{kr} = (VG)_{kr'} \quad k = 1,2,...,v \text{ και } r \neq r' = 1,2,...,p$$

δηλαδή τα γονίδια r και r' δεν διαφοροποιούνται όταν χρησιμοποιούμε την απόχρωση k, έναντι της εναλλακτικής υπόθεσης

$$H_0 : (VG)_{kr} - (VG)_{kr'} \neq 0 \Leftrightarrow H_0 : (VG)_{kr} \neq (VG)_{kr'} \quad k = 1,2,...,v \text{ και } r \neq r' = 1,2,...,p$$

δηλαδή τα γονίδια r και r' διαφοροποιούνται όταν χρησιμοποιούμε την απόχρωση k.

Γι' αυτόν τον έλεγχο χρησιμοποιούμε την παρακάτω στατιστική συνάρτηση

$$t = \frac{\left[\left(\widehat{VG}\right)_{kr} - \left(\widehat{VG}\right)_{kr'}\right] - 0}{\sqrt{\hat{V}\left[\left(\widehat{VG}\right)_{kr} - \left(\widehat{VG}\right)_{kr'}\right]}}$$

η οποία ακολουθεί την κατανομή Student με $(v-1)(p-1)$ βαθμούς ελευθερίας υπό την μηδενική υπόθεση. Αν απορριφθεί η μηδενική υπόθεση, τότε το επόμενο βήμα είναι να εξετάσουμε το πρόσημο και το μέγεθος της εκτιμώμενης διαφοράς $\left(\widehat{VG}\right)_{kr} - \left(\widehat{VG}\right)_{kr'}$. Συγκεκριμένα,

- μία μεγάλη θετική τιμή σημαίνει ότι το γονίδιο r είναι το πιο σημαντικό, διότι η έκφρασή του διαφοροποιείται περισσότερο από την έκφραση του γονιδίου r', όταν χρησιμοποιούμε την απόχρωση k.
- μία μεγάλη αρνητική τιμή σημαίνει ότι το γονίδιο r' είναι το πιο σημαντικό, διότι η έκφρασή του διαφοροποιείται περισσότερο από την έκφραση του γονιδίου r, όταν χρησιμοποιούμε την απόχρωση k.

Κεφάλαιο 2

Ομαδοποίηση εκφράσεων γονιδίων

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο αρχικά θα παρουσιάσουμε τα μέτρα απόστασης που συνήθως χρησιμοποιούνται στην ομαδοποίηση εκφράσεων γονιδίων. Στη συνέχεια θα αναπτύξουμε τις ιεραρχικές και μη ιεραρχικές μεθόδους ομαδοποίησης που προτιμώνται από την ιατρική κοινότητα, όταν μελετάμε εκφράσεις γονιδίων. Το κεφάλαιο ολοκληρώνεται με την ανάλυση των μέτρων αξιολόγησης των μεθόδων ομαδοποίησης, τα οποία αποτελούν σημαντικό κριτήριο στην απόφαση για την μέθοδο ομαδοποίησης που θα επιλεγεί ως η πιο κατάλληλη για το σετ δεδομένων που μελετάμε.

2.2 Πίνακας δεδομένων γονιδιακών εκφράσεων

Μέσω των γονιδιακών εκφράσεων μπορούμε να διαγιγνώσκουμε τις συνθήκες της ασθένειας, να εντοπίσουμε τις επιδράσεις των θεραπειών και στο προσεχές μέλλον να ανακαλύψουμε και άλλες σημαντικές εφαρμογές τους. Τα δεδομένα γονιδιακών εκφράσεων συνήθως αναπαρίστανται σε έναν πίνακα όπου οι στήλες αντιπροσωπεύουν τα γονίδια και οι γραμμές τους ασθενείς, τους ιστούς ή τα χρονικά σημεία. Οι τιμές των επιπέδων έκφρασης των γονιδίων μπορεί να είναι τυποποιημένες (*normalized*), απόλυτοι αριθμοί ή ποσοστά. Σε κάθε ασθενή (ιστό ή χρονικό σημείο) αντιστοιχεί μία τιμή από κάθε γονίδιο και όλες αυτές οι τιμές συμπληρώνουν το προφίλ (*profile)* του ασθενή. Όλες οι τιμές κάθε γονιδίου αποτελούν ένα διάνυσμα που ονομάζεται σχέδιο έκφρασης *(expression pattern)* του γονιδίου (Sharan et al. 2002).

Πίνακας 2.2.1

Επίπεδα εκφράσεων p γονιδίων σε n ασθενείς

gene / patient	x_1	x_2	...	x_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{np}

Μέσω της ομαδοποίησης εντοπίζουμε τις ομάδες γονιδίων με παρόμοιο σχέδιο έκφρασης (*similar expression pattern).* Το μέτρο απόστασης ή ομοιότητας ορίζεται μεταξύ δύο διανυσμάτων σχεδίων έκφρασης και συνήθως χρησιμοποιείται ο συντελεστής συσχέτισης του Pearson. Στόχος της ομαδοποίησης είναι να διαιρέσει τα γονίδια σε ομάδες (*subsets)* που ονομάζονται συστάδες (*clusters*), έτσι ώστε να ικανοποιούνται δύο κριτήρια (Sharan et al. 2002):

1. ομοιογένεια (*homogeneity*) κάθε συστάδας, όπου τα γονίδια στην ίδια συστάδα είναι ιδιαίτερα όμοια (*highly similar)* μεταξύ τους
2. ετερογένεια (*heterogeneity)* μεταξύ συστάδων, όπου τα γονίδια από διαφορετικές συστάδες έχουν χαμηλή ομοιότητα (*low similarity*) μεταξύ τους.

Έστω $N = \{x_1, x_2, ..., x_p\}$ ένα σύνολο από p γονίδια και έστω $C = \{C_1, C_2, ..., C_k\}$ ένα σύνολο από k συστάδες που προέκυψαν από την ομαδοποίηση των p γονιδίων. Το σύνολο C ονομάζεται ομαδοποίηση (*clustering ή clustering solution*). Δύο στοιχεία x_i και x_j αποτελούν ζεύγος σύμφωνα με το C αν και μόνο αν ανήκουν στην ίδια συστάδα. Τα δεδομένα εισαγωγής (*input data*) για την ομαδοποίηση είναι συνήθως σε δύο μορφές (Sharan et al. 2002):

1. πρωτογενή δεδομένα (*fingerprint data*), όπου κάθε γονίδιο αποτελεί ένα διάνυσμα πραγματικών τιμών το οποίο ονομάζεται *fingerprint* ή *pattern* του γονιδίου και περιέχει n μετρήσεις, δηλαδή όσοι είναι οι ασθενείς. Ο πίνακας που περιέχει αυτά τα δεδομένα είναι διάστασης $n \times p$.

2. δεδομένα ομοιότητας (*similarity data*), δηλαδή δεδομένα που αντιπροσωπεύουν τον βαθμό ομοιότητας των γονιδίων. Τα δεδομένα αυτά προκύπτουν από τα πρωτογενή δεδομένα εφαρμόζοντας κάποιο μέτρο απόστασης ή ομοιότητας και διαμορφώνουν έναν συμμετρικό πίνακα διάστασης $p \times p$.

Αν και τα πρωτογενή δεδομένα περιέχουν όλη την πληροφορία, ωστόσο για την ομαδοποίηση χρησιμοποιούμε τα δεδομένα ομοιότητας, τα οποία περιέχουν περιληπτική πληροφορία κατάλληλη για ομαδοποίηση (Sharan et al. 2002).

2.3 Μέτρα απόστασης

Μέτρο απόστασης ορίζεται ουσιαστικά μία συνάρτηση $d_{ij} = d(x_i, x_j)$ δύο διανυσμάτων $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})'$ και $\mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{jn})'$ που υπολογίζει την απόσταση μεταξύ των συνιστωσών τους μέσα σε ένα χώρο. Το μέτρο απόστασης συνήθως ικανοποιεί τρεις ιδιότητες (Erfaneh and Yonhghong, 2009):

I1. $d_{ij} \geq 0$ για κάθε i, j και $d_{ij} = 0$ όταν $i = j$

I2. $d_{ij} \leq d_{is} + d_{sj}$ ονομάζεται τριγωνική ανισότητα

I3. $d_{ij} = d_{ji}$ αποτελεί την συμμετρική ιδιότητα

Στην πράξη πολλά μέτρα δεν ικανοποιούν την τριγωνική ανισότητα, ενώ σε κάποιες περιπτώσεις μπορεί να μην είναι απαραίτητη η συμμετρική ιδιότητα (Κούτρας, 2005).

Συνήθως στους αλγόριθμους της ομαδοποίησης γονιδίων χρησιμοποιείται είτε η ευκλείδεια απόσταση

$$d_{ij} = \sqrt{\sum_{r=1}^{n} \left(x_{ir} - x_{jr} \right)^2}$$

είτε ο συντελεστής συσχέτισης του Pearson

$$r_{ij} = r\left(x_i, x_j \right) = \frac{\sum_{r=1}^{n} \left(x_{ir} - \overline{x}_i \right)\left(x_{jr} - \overline{x}_j \right)}{\sqrt{\sum_{r=1}^{n} \left(x_{ir} - \overline{x}_i \right)^2} \sqrt{\sum_{r=1}^{n} \left(x_{jr} - \overline{x}_j \right)^2}}$$

με $r_{ij} \in [-1,1]$, όπου ως μέτρο απόστασης ορίζεται η ποσότητα $d_{ij} = 1 - r_{ij}$, οπότε παίρνει τιμές στο διάστημα $d_{ij} \in [0,2]$. Ο συντελεστής συσχέτισης του Pearson αποτελεί μέτρο ομοιότητας και εκφράζει τον βαθμό ομοιότητας μεταξύ δύο γονιδίων, αφού δεν δίνει έμφαση στο μέγεθος των μετρήσεων των γονιδίων, αλλά στο πόσο και πως συσχετίζονται τα γονίδια αυτά ανά δύο (Daxin et al. 2004).

Όταν τυποποιούμε τα δεδομένα η αποτελεσματικότητα του αλγόριθμου ομαδοποίησης είναι ισοδύναμη είτε εφαρμόσουμε την απόσταση του Pearson είτε την ευκλείδεια απόσταση. Ωστόσο, και τα δύο αυτά μέτρα έχουν μερικά μειονεκτήματα. Συγκεκριμένα όσον αφορά την ευκλείδεια απόσταση (Daxin et al. 2004):

- όταν έχουμε διαφορετική κλίμακα μέτρησης έχει αποδειχθεί ότι δεν δίνει συνεπή αποτελέσματα.
- επίσης, μεταβλητές με μεγάλες απόλυτες τιμές ή πολλές ακραίες τιμές συνεισφέρουν περισσότερο σε σχέση με μεταβλητές με μικρότερες απόλυτες τιμές με συνέπεια το αποτέλεσμα να είναι παραπλανητικό.

Και στις δύο περιπτώσεις το πρόβλημα αντιμετωπίζεται με την τυποποίηση των μεταβλητών ώστε να γίνουν συγκρίσιμες και η απόστασή τους αξιόπιστη.

Όσον αφορά την απόσταση του Pearson μερικά μειονεκτήματα είναι τα εξής (Daxin et al. 2004):

- το αποτέλεσμα μπορεί να είναι παραπλανητικό όταν υπάρχουν πολλές ακραίες τιμές, γιατί στην περίπτωση αυτή τα άτομα στα οποία αντιστοιχίζονται οι ακραίες τιμές συνεισφέρουν περισσότερο στην απόσταση των αντίστοιχων γονιδίων σε σχέση με τα άτομα στα οποία αντιστοιχίζονται μικρότερες τιμές.
- για τον υπολογισμό του συντελεστή συσχέτισης του Pearson υποθέτουμε ότι τα δεδομένα ακολουθούν προσεγγιστικά την κατανομή Gauss, με αποτέλεσμα να μην είναι αξιόπιστος σε αντίθετη περίπτωση.

Στην πρώτη περίπτωση το πρόβλημα αντιμετωπίζεται με την εφαρμογή του συντελεστή συσχέτισης του Jacknife,

$$Jacknife(x_i, x_j) = \min\left\{r_{ij}^{(1)}, r_{ij}^{(2)}, ..., r_{ij}^{(k)}, ..., r_{ij}^{(n)}\right\}$$

όπου $r_{ij}^{(k)}$ είναι η συσχέτιση Pearson των γονιδίων x_i και x_j όταν διαγραφεί το $k-$οστό άτομο. Ωστόσο, ο συντελεστής αυτός χρησιμοποιείται σπάνια, γιατί είναι χρονοβόρος στον υπολογισμό του. Στην δεύτερη περίπτωση χρησιμοποιείται ο συντελεστής συσχέτισης του Spearman, όπου χρησιμοποιείται η τάξη της

παρατήρησης (*rank*) αντί η ίδια η παρατήρηση. Όμως, έχει αποδειχθεί ότι γενικά η συσχέτιση του Pearson υπερέχει του Spearman (Daxin et al.2004).

Εμπειρικά έχει αποδειχθεί ότι και τα δύο μέτρα απόστασης είναι αποτελεσματικά. Συνήθως, η ευκλείδεια απόσταση προτείνεται όταν έχουμε log ratio δεδομένα, δηλαδή λογαριθμικές εκφράσεις γονιδίων σύμφωνα με την τύπο $(1.5.1)$, ενώ η απόσταση του Pearson προτείνεται όταν έχουμε απόλυτες τιμές, όπως στην περίπτωση δεδομένων Affymetrix (Patrik D'haeseleer, 2005).

Έστω ότι έχουμε p γονίδια $x_1, x_2, ..., x_p$, τότε οι αποστάσεις τους $d_{ij} = d(x_i, x_j)$, i, j=1,2,...,p τοποθετούνται στον συμμετρικό πίνακα $D_p = \left[d_{ij} \right]$ που ονομάζεται πίνακας εγγύτητας (*proximity matrix*).

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1p} \\ d_{21} & d_{22} & \cdots & d_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ d_{p1} & d_{p2} & \cdots & d_{pp} \end{pmatrix}$$

Αν εφαρμόσουμε την ευκλείδεια απόσταση, τότε η κύρια διαγώνιος θα είναι $diag(0,0,...,0)$ σύμφωνα με την ιδιότητα I1, ενώ θα είναι $diag(1,1,...,1)$ αν εφαρμόσουμε τον συντελεστή συσχέτισης του Pearson. Και με τα δύο μέτρα ο πίνακας D είναι συμμετρικός σύμφωνα με την ιδιότητα I3, οπότε συνήθως παραλείπουμε τα στοιχεία που βρίσκονται πάνω από την κύρια διαγώνιο (Κούτρας, 2005).

2.4 Ιεραρχικές Μέθοδοι Ομαδοποίησης

Οι ιεραρχικές μέθοδοι ομαδοποίησης (*Hierarchical Clustering Methods*) βασίζονται στην απόσταση μεταξύ δύο γονιδίων ή δύο συστάδων ή μίας συστάδας με ένα γονίδιο χωρίς να γνωρίζουμε εκ των προτέρων το πλήθος των συστάδων που θα σχηματιστούν, γι' αυτό θεωρείται ως μία προσέγγιση χωρίς επιτήρηση (*unsupervised*) στην αρχική ανάλυση των δεδομένων. Η ιεραρχική ομαδοποίηση αναπαριστάνεται γραφικά σ' ένα δενδρόγραμμα, όπου μπορούμε να εντοπίσουμε τις αποστάσεις μεταξύ των συστάδων, να βρούμε το βέλτιστο αριθμό συστάδων και να ερμηνεύσουμε τη δομή της ομαδοποίησης. Σε κάθε βήμα της μεθόδου οι ομάδες που

σχηματίζονται συνδέονται με μία γραμμή η οποία αντιστοιχεί σε κάποια τιμή του κάθετου άξονα που περιέχει τις αποστάσεις (Erfaneh and Yonghong, 2009).

Οι ιεραρχικές μέθοδοι ομαδοποίησης διακρίνονται σε δύο βασικές κατηγορίες: στις συσσωρευτικές μεθόδους (*agglomerative methods ή bottom-up approach*) και στις διαιρετικές μεθόδους (*divisive methods ή top-down approach*). Στη συνέχεια παρουσιάζουμε τις κυριότερες συσσωρευτικές μεθόδους.

2.4.1 Συσσωρευτικές μέθοδοι

Σύμφωνα με τους R. A. Johnson και D. W. Wichern (1998) στο βιβλίο τους Applied Multivariate Statistical Analysis, στις συσσωρευτικές μεθόδους ξεκινάμε από p ομάδες, δηλαδή όσα είναι τα γονίδια, όπου κάθε ομάδα περιέχει ένα γονίδιο και με συσσωρευτικό αλγόριθμο καταλήγουμε βήμα-βήμα σε μία ομάδα που περιέχει p γονίδια. Ο συσσωρευτικός αλγόριθμος με την μορφή βημάτων έχει ως εξής:

ΒΗΜΑ 1^{Ο}: αρχίζουμε με p ομάδες, όπου κάθε ομάδα περιέχει ένα στοιχείο και υπολογίζουμε τον πίνακα αποστάσεων D_p χρησιμοποιώντας κάποιο μέτρο απόστασης (έστω την ευκλείδεια απόσταση),

ΒΗΜΑ 2^{Ο}: εντοπίζουμε τις δύο ομάδες με την μικρότερη απόσταση d_{ij}. Έστω οι ομάδες U και V, οι οποίες έχουν απόσταση d_{UV},

ΒΗΜΑ 3^{Ο}: συγχωνεύουμε αυτές τις δύο ομάδες σε μία ομάδα (UV). Ανανεώνουμε τον πίνακα D_p διαγράφοντας τις γραμμές και τις στήλες που αφορούν τις ομάδες U και V και προσθέτοντας μία γραμμή και μία στήλη για τις αποστάσεις μεταξύ της ομάδας (UV) και τις υπόλοιπες ομάδες-στοιχεία,

ΒΗΜΑ 4^{Ο}: επαναλαμβάνουμε τα βήματα 2 και 3 συνολικά $p-1$ φορές, έτσι ώστε οι αρχικές p ομάδες να συγχωνευτούν σε μία ομάδα, οπότε τερματίζεται ο αλγόριθμος.

Γνωρίζουμε ήδη πώς να υπολογίσουμε τις αποστάσεις μεταξύ στοιχείων, αλλά δεν γνωρίζουμε πώς μπορούμε να υπολογίσουμε τις αποστάσεις μεταξύ ομάδων. Σύμφωνα με τους Lance και Williams (McLachlan et al. 2004) ο γενικός τύπος που

εκφράζει την απόσταση μεταξύ μίας νέας ομάδας u και μιας ήδη υπάρχουσας ομάδας t είναι

$$d(u,t) = a_r d(r,t) + a_s d(s,t) + \beta d(r,s) + \gamma \left| d(r,t) - d(s,t) \right| \qquad (2.4.1)$$

όπου οι παράμετροι a_r, a_s, β και γ ορίζονται διαφορετικά σε κάθε συσσωρευτική μέθοδο.

2.4.1.1 Μέθοδος της απλής συνένωσης (Single Linkage Method)

Στην μέθοδο αυτή η απόσταση μεταξύ δύο ομάδων ορίζεται ως η μικρότερη απόσταση μεταξύ ενός στοιχείου της μιας ομάδας και ενός στοιχείου της άλλης ομάδας. Γι' αυτό η μέθοδος αυτή ονομάζεται και «μέθοδος του κοντινότερου γείτονα». Οι παράμετροι στον γενικό τύπο της απόστασης ορίζονται ως $a_r = a_s = \frac{1}{2}, \beta = 0$ και $\gamma = -\frac{1}{2}$ (McLachlan et al. 2004), οπότε ο τύπος έχει τη μορφή

$$d(u,t) = \frac{1}{2}\left(d(r,t) + d(s,t)\right) - \frac{1}{2}\left|d(r,t) - d(s,t)\right| = \min\left\{d(r,t), d(s,t)\right\}$$

Ένα μειονέκτημα της μεθόδου αυτής είναι ότι υπάρχει ο κίνδυνος να συμβεί το «φαινόμενο της αλυσίδας» (Κούτρας, 2005). Συγκεκριμένα, δύο ομάδες που είναι εμφανώς διαφορετικές θα συγχωνευτούν αν υπάρχει κάποιο σημείο ή ένα σύνολο σημείων που τις συνδέει. Έτσι, μπορεί να δημιουργηθούν μερικές πολύ μεγάλες ομάδες και κάποιες άλλες πολύ μικρές. Πλεονέκτημα της μεθόδου είναι ότι δεν επηρεάζεται από την ύπαρξη ακραίων τιμών (Causton et al. 2003).

2.4.1.2 Μέθοδος της πλήρους συνένωσης (Complete Linkage Method)

Σε αντίθεση με την απλή συνένωση, στην πλήρη συνένωση λαμβάνουμε υπόψη την μέγιστη απόσταση μεταξύ ενός στοιχείου της μίας ομάδας και ενός στοιχείου της άλλης ομάδας. Εναλλακτική ορολογία της μεθόδου αυτής είναι «μέθοδος του μακρινότερου γείτονα» λόγω του ορισμού της απόστασης μεταξύ δύο ομάδων. Ο γενικός τύπος της απόστασης είναι

$$d(u,t) = \frac{1}{2}\left(d(r,t) + d(s,t)\right) + \frac{1}{2}\left|d(r,t) - d(s,t)\right| = \max\left\{d(r,t), d(s,t)\right\}$$

όπου $a_r = a_s = \frac{1}{2}, \beta = 0$ και $\gamma = \frac{1}{2}$ στη σχέση $(2.4.1)$ (McLachlan et al. 2004).

Στην μέθοδο αυτή υπάρχει ο κίνδυνος εμφανώς όμοιες ομάδες να μην μπορούν να συγχωνευτούν όταν υπάρχει κάποιο ζεύγος σημείων που απέχουν αρκετά μεταξύ τους (Κούτρας, 2005). Σε αντίθεση με την απλή συνένωση, η πλήρης συνένωση επηρεάζεται από την ύπαρξη ακραίων τιμών (Causton et al. 2003). Η μέθοδος αυτή μπορεί να ξεχωρίσει μεγάλες ομάδες, αλλά αποτυγχάνει να εντοπίσει κάποιες μικρές (Κούτρας, 2005).

2.4.1.3 Μέθοδος UPGMAA (Unweighted Pair Group Method Arithmetic Average)

Η απόσταση μεταξύ δύο ομάδων ορίζεται ως ο σταθμισμένος αριθμητικός μέσος όρος των αποστάσεων μεταξύ μιας υπάρχουσας ομάδας και των στοιχείων της νέας ομάδας (Holmes, 2009). Συγκεκριμένα ο γενικός τύπος της απόστασης μεταξύ των ομάδων είναι

$$d(u,t)=\frac{n_r}{n_r+n_s}d(r,t)+\frac{n_s}{n_r+n_s}d(s,t)$$

όπου έχουμε θέσει $a_r=\frac{n_r}{n_r+n_s}, a_s=\frac{n_s}{n_r+n_s}, \beta=0$ και $\gamma=0$ στη σχέση $(2.4.1)$ και n_r και n_s δηλώνουν το πλήθος στοιχείων που περιέχουν οι ομάδες r και s αντίστοιχα (Teknomo, 2009).

Η μέθοδος UPGMAA προτιμάται ιδιαίτερα από τους βιολόγους και χρησιμοποιείται ευρέως στην ανάλυση έκφρασης γονιδίων (Daxin et al. 2004). Δημιουργός της μεθόδου είναι ο Eisen.

2.4.1.4 Μέθοδος WPGMA (Weighted Pair Group Method Average)

Σε αυτήν την περίπτωση η απόσταση μεταξύ δύο ομάδων ορίζεται ο μέσος όρος των αποστάσεων μεταξύ κάθε στοιχείου της μιας ομάδας και κάθε στοιχείου της άλλης ομάδας (Κούτρας, 2005). Εναλλακτικός ορισμός της μεθόδου είναι «μέθοδος της μέσης συνένωσης» (*Average Linkage Method*). Οι παράμετροι ορίζονται ως $a_r=a_s=\frac{1}{2}$ και $\beta=\gamma=0$ στη σχέση $(2.4.1.1)$ (Teknomo, 2009),

$$d(u,t)=\frac{1}{2}d(r,t)+\frac{1}{2}d(s,t)$$

Αξίζει να σημειωθεί ότι τόσο η Μέθοδος της πλήρους συνένωσης όσο και η Μέθοδος της μέσης συνένωσης χρησιμοποιούνται ευρύτατα στην ανάλυση δεδομένων σε μικροσυστάδες (*microarray data analysis*) με ικανοποιητικά αποτελέσματα (Naghieh and Yonghong, 2009). Επιπλέον, οι ομάδες που δημιουργούνται έχουν συνήθως την ίδια εντός-συστάδας διακύμανση (*within-cluster variance*) (Holmes, 2009). Η συγκεκριμένη μέθοδος προτιμάται αντί της UPGMAA όταν υπάρχει έντονη ανισορροπία στο μέγεθος των συστάδων (Holmes, 2009).

2.4.1.5 Μέθοδος της Ward συνένωσης (Ward Linkage Method)

Η συγκεκριμένη μέθοδος διαφέρει από τις προηγούμενες, αφού επιδιώκει να τοποθετήσει τα στοιχεία σε ομάδες στις οποίες η διακύμανση θα είναι όσο το δυνατόν πιο μικρή, διότι έτσι χάνεται όσο το δυνατόν λιγότερη πληροφορία από την συγχώνευση των ομάδων σε μία ομάδα (Johnson et al. 1998). Η μέθοδος βασίζεται στο άθροισμα τετραγώνων των αποκλίσεων ESS_k, που υπολογίζεται για κάθε ομάδα που προκύπτει ως το άθροισμα τετραγώνων της απόκλισης κάθε στοιχείου της ομάδας από το κέντρο βάρους της. Ουσιαστικά με την μέθοδο αυτή αντιμετωπίζουμε την ομαδοποίηση των στοιχείων ως ανάλυση διακύμανσης χωρίς να χρησιμοποιούμε κάποιο μέτρο απόστασης, οπότε ούτε πίνακα αποστάσεων.

Το άθροισμα των ESS_k δίνει το συνολικό άθροισμα τετραγωνικών αποκλίσεων ESS, που αποτελεί το «κλειδί» για μια σωστή ομαδοποίηση. Σε κάθε βήμα της μεθόδου συγχωνεύουμε όλες τις ομάδες που προκύπτουν, αλλά τελικά καταλήγουμε σε εκείνη τη συγχώνευση που αυξάνει ελάχιστα το ESS, οπότε αυξάνει ελάχιστα την απώλεια πληροφορίας (*loss of information*) λόγω συγχώνευσης. Να σημειώσουμε ότι αρχικά, όπου κάθε στοιχείο αποτελεί μία ανεξάρτητη ομάδα, το ESS ισούται με μηδέν, καθώς το κέντρο βάρους κάθε ομάδας είναι η ίδια η παρατήρηση με αποτέλεσμα $ESS_k=0$ για $k=1,2,...,p$. Στο τελευταίο βήμα της μεθόδου, όπου όλα πλέον τα στοιχεία αποτελούν μία ομάδα, ισχύει ότι $ESS=\sum_{i=1}^{p}(x_i-\bar{x})'(x_i-\bar{x})$, όπου x_i είναι η πολυδιάστατη μέτρηση του i-οστού στοιχείου και $\bar{x}$ είναι ο μέσος όρος των μετρήσεων όλων των στοιχείων.

Οι παράμετροι στη σχέση $(2.4.1)$ ορίζονται ως $a_r = \frac{n_r + n_t}{n_r + n_s + n_t}$, $\alpha_s = \frac{n_s + n_t}{n_r + n_s + n_t}$, $\beta = -\frac{n_t}{n_r + n_s + n_t}$ και $\gamma = 0$ (McLachlan et al. 2004), όπου n_r, n_s και n_t δηλώνουν το πλήθος στοιχείων που περιέχουν οι ομάδες r, s και t αντίστοιχα, οπότε ο γενικός τύπος απόστασης για αυτή τη μέθοδο είναι

$$d(u,t) = \frac{n_r + n_t}{n_r + n_s + n_t} d(r,t) + \frac{n_s + n_t}{n_r + n_s + n_t} d(s,t) - \frac{n_t}{n_r + n_s + n_t} d(r,s)$$

Η μέθοδος της Ward συνένωσης είναι πολύ καλός οδηγός στην επιλογή του βέλτιστου πλήθους συστάδων (Johnson et al. 1998). Επιπλέον, συνήθως δημιουργεί συστάδες παρόμοιου μεγέθους (Holmes, 2009). Σύμφωνα με μελέτες που έκαναν οι Jain και Dubes (McLachlan et al. 2004) έχει αποδειχθεί ότι αυτή η μέθοδος υπερισχύει έναντι των άλλων συσσωρευτικών μεθόδων όσον αφορά την ομαδοποίηση των γονιδίων.

2.4.1.6 Μέθοδος UPGMC (Unweighted Pair Group Method Centroid)

Σε αυτή τη μέθοδο ως απόσταση μεταξύ δύο ομάδων ορίζουμε την ευκλείδεια απόσταση μεταξύ των κέντρων των ομάδων (Holmes, 2009). Εναλλακτικός ορισμός της μεθόδου είναι «μέθοδος των κέντρων βάρους» (*Centroid Method*). Δηλαδή, ο γενικός τύπος της απόστασης μεταξύ των ομάδων είναι

$$d(u,t) = \frac{n_r}{n_r + n_s} d(r,t) + \frac{n_s}{n_r + n_s} d(s,t) - \frac{n_r n_s}{(n_r + n_s)^2} d(r,s)$$

όπου $a_r = \frac{n_r}{n_r + n_s}, a_s = \frac{n_s}{n_r + n_s}, \beta = -\frac{n_r n_s}{(n_r + n_s)^2}$ και $\gamma = 0$ (Teknomo, 2009).

Να σημειώσουμε ότι $d(u,t) = d\left(\bar{x}(u), \bar{x}(t)\right) = \sqrt{\sum_{i=1}^{n} \left(\bar{x}_i(u) - \bar{x}_i(t)\right)^2}$ όπου το κέντρο βάρους της ομάδας u είναι το σημείο $\bar{x}(u) = \left(\bar{x}(u)_1, \bar{x}(u)_2, ..., \bar{x}(u)_n\right)$ το οποίο έχει ως $i-$οστή συντεταγμένη την $\bar{x}_i(u) = \frac{1}{n_u} \sum_{j \in u} x_{ij}$ (Κούτρας, 2005).

2.4.1.7 Μέθοδος WPGMC (Weighted Pair Group Method Centroid)

Σε αντίθεση με την μέθοδο των κέντρων βάρους η απόσταση μεταξύ δύο ομάδων ορίζεται ως η διάμεσος της ευκλείδειας απόστασης μεταξύ των κέντρων των ομάδων. Την συγκεκριμένη μέθοδο συχνά την συναντάμε με τον όρο «μέθοδος της διάμεσης συνένωσης» (Median Linkage Method). Ο γενικός τύπος της απόστασης μεταξύ δύο ομάδων είναι

$$d(u,t)=\frac{1}{2}d(r,t)+\frac{1}{2}d(s,t)-\frac{1}{4}d(r,s)$$

όπου έχουμε θέσει $a_r=\frac{1}{2}, a_s=\frac{1}{2}, \beta=-\frac{1}{2^2}$ και $\gamma=0$ στη σχέση $(2.4.1)$. Η συγκεκριμένη μέθοδος υπερέχει της προηγούμενης μεθόδου για συστάδες με έντονη ανισορροπία στο μέγεθός τους (Holmes, 2009).

Τόσο η μέθοδος της διάμεσης συνένωσης όσο και η μέθοδος των κέντρων βάρους μπορούν να δημιουργήσουν δενδρόγραμμα που περιέχει αναστροφές (*inversions*), δηλαδή ένωση ενός στοιχείου με μία ήδη υπάρχουσα ομάδα σε απόσταση μικρότερη από εκείνη της προηγούμενης συνένωσης (Johnson et al. 1998). Αυτό συμβαίνει όταν η απόσταση από την ένωση των ομάδων r και s με την ομάδα t είναι μικρότερη από την απόσταση είτε μεταξύ των ομάδων r και t είτε μεταξύ των ομάδων s και t.

2.4.2 Διαιρετικές μέθοδοι

Στις διαιρετικές μεθόδους ξεκινάμε από μία συστάδα που περιέχει p στοιχεία και με τη βοήθεια του διαιρετικού αλγορίθμου καταλήγουμε σε p συστάδες, όπου καθεμία περιέχει ένα στοιχείο. Δηλαδή, πρόκειται για μία διαδικασία αντίστροφη της συσσωρευτικής μεθόδου (Κούτρας, 2005). Σε κάθε βήμα του αλγορίθμου σχηματίζονται υποομάδες, που προκύπτουν από τον διαχωρισμό των ομάδων που ήδη υπάρχουν και είναι περισσότερο απομακρυσμένες.

Ο διαιρετικός αλγόριθμος απαιτεί περισσότερους υπολογισμούς σε σχέση με τον συσσωρευτικό αλγόριθμο με αποτέλεσμα οι διαιρετικές μέθοδοι να μην χρησιμοποιούνται συχνά στην πράξη (Κούτρας, 2005). Αν λάβουμε υπόψη ότι στο πρώτο στάδιο ενός συσσωρευτικού αλγορίθμου οι δυνατές συνενώσεις δύο στοιχείων

από τα p διαθέσιμα είναι $\frac{p(p-1)}{2}$, ενώ στο πρώτο στάδιο του διαιρετικού αλγορίθμου οι δυνατές διαιρέσεις των p στοιχείων σε δύο ομάδες είναι $2^{p-1}-1=\frac{(2^p-2)}{2}$, τότε καταλαβαίνουμε πόσο χρονοβόρος υπολογιστικά είναι ο διαιρετικός αλγόριθμος (Κούτρας, 2005).

Είναι γεγονός ότι οι ιεραρχικές μέθοδοι ομαδοποίησης χρησιμοποιούνται ευρύτατα λόγω της εύκολης εφαρμογής τους και ερμηνείας των αποτελεσμάτων τους (McLachlan et al. 2004). Ωστόσο, σημαντικό μειονέκτημα είναι τόσο το υπολογιστικό κόστος, που περιλαμβάνει το πλήθος αποστάσεων που πρέπει να υπολογιστούν και το πλήθος πινάκων αποστάσεων που πρέπει να κατασκευαστούν, όσο και το ακαθόριστο κριτήριο τερματισμού του αλγόριθμου σε περίπτωση μεγάλου πλήθους γονιδίων (Erfaneh and Yonghong, 2009).

Επιπλέον, υπάρχει ο κίνδυνος σχηματισμού μικρού πλήθους ομάδων με πολλά γονίδια, αλλά και ομάδων με ένα μόνο γονίδιο λόγο της ιδιαιτερότητας κυρίως του συσσωρευτικού αλγόριθμου να μην χωρίζει σε επόμενα βήματα γονίδια που αποτελούν ήδη μία ομάδα από τα πρώτα βήματα (Κούτρας, 2005). Γενικά οι ομάδες που θα σχηματίσει ο αλγόριθμος εξαρτώνται από την μέθοδο υπολογισμού των αποστάσεων μεταξύ των ομάδων, οπότε εφαρμόζοντας διάφορες μεθόδους αποστάσεων μεταξύ των ομάδων μπορούμε να δούμε ποια αποτελέσματα συμφωνούν, ώστε να έχουμε πετύχει μία «φυσική» (*natural*) ομαδοποίηση (Johnson et al. 1998).

Τέλος, ένα ακόμα μειονέκτημα στην ιεραρχική ομαδοποίηση είναι πώς θα αποφασίσουμε σε ποιο ύψος του άξονα αποστάσεων θα «κόψουμε» το δενδρόγραμμα για να ορίσουμε το πλήθος των συστάδων (McLachlan et al. 2004).

2.4.3 Δενδρόγραμμα και θερμικός χάρτης (*Heat map*)

Γνωρίζουμε ότι το δενδρόγραμμα αναπαριστάνει τις σχέσεις μεταξύ των γονιδίων και με την βοήθεια της κλίμακας αποστάσεων αναδεικνύει πόσο μοιάζουν οι συστάδες που προκύπτουν. Στην πραγματικότητα το δενδρόγραμμα δεν υφίσταται για τις γονιδιακές εκφράσεις, ωστόσο είναι ένα πολύ χρήσιμο εργαλείο που μας βοηθάει

να εντοπίσουμε τις βέλτιστες συστάδες γονιδίων, να τις συγκρίνουμε μεταξύ τους, να εντοπίσουμε ποια γονίδια αντιπροσωπεύουν περισσότερο και ποια λιγότερο την κάθε συστάδα και τέλος να κάνουμε υποθέσεις για την ποιότητα των δεδομένων (Eisen et al. 1998).

Για μία ολοκληρωμένη μελέτη των ομαδοποιημένων γονιδίων συνηθίζεται να συνδυάζεται το δενδρόγραμμα με έναν θερμικό χάρτη, ο οποίος περιέχει τα δεδομένα γονιδίων ως σημεία με χρώμα και έτσι με ποσοτικό και ποιοτικό τρόπο μπορούμε να αναλύσουμε και να κατανοήσουμε τα δεδομένα (Eisen et al. 1998).

Συγκεκριμένα, εφαρμόζεται ιεραρχική ομαδοποίηση τόσο στα γονίδια όσο και στα δείγματα (δηλαδή στους ασθενείς ή τους ιστούς) και έπειτα γίνεται αναδιάταξη ξεχωριστά των γραμμών και των στηλών του πρωτογενούς πίνακα δεδομένων σύμφωνα με τη μέση τιμή τους και το αντίστοιχο δενδρόγραμμα, έτσι ώστε να βρίσκονται όσο πιο κοντά γίνεται οι παρόμοιες τιμές αποκαλύπτοντας τις λειτουργικές σχέσεις μεταξύ γονιδίων και δειγμάτων (Wilkinson et al. 2009). Οι γραμμές του θερμικού χάρτη μικροσυστάδων (*microarray heat map*) αντιπροσωπεύουν τα δείγματα, ενώ οι στήλες αντιπροσωπεύουν τα γονίδια. Το δενδρόγραμμα των δειγμάτων τοποθετείται κατά μήκος του άξονα y του θερμικού χάρτη, ενώ το δενδρόγραμμα των γονιδίων τοποθετείται κατά μήκος του άξονα x. Το χρώμα των ψηφίδων (*pixels*) του θερμικού χάρτη αντιπροσωπεύει το επίπεδο έκφρασης των γονιδίων σε κάθε δείγμα. Υψηλό επίπεδο έκφρασης γονιδίων (δηλαδή, το επίπεδο έκφρασής τους έχει μεγαλύτερη τιμή από το μέσο επίπεδο έκφρασής τους) αντιπροσωπεύεται με κόκκινες ψηφίδες για κάποια δείγματα, ενώ χαμηλό επίπεδο έκφρασης (δηλαδή, το επίπεδο έκφρασής τους έχει μικρότερη τιμή από το μέσο επίπεδο έκφρασής τους) αντιπροσωπεύεται με πράσινες ή μπλε ψηφίδες. Τέλος, οι σκοτεινοί τόνοι του κόκκινου ή του πράσινου χρώματος σε κάποια δείγματα δηλώνουν ότι το επίπεδο έκφρασης των γονιδίων αυτών σχεδόν συμπίπτει με το μέσο επίπεδο έκφρασής τους (Wilkinson et al. 2009).

Παρατήρηση

Τα p γονίδια θα «δώσουν» 2^{p-1} γραμμικές διατάξεις (*linear orderings*) που συμφωνούν με τη δομή του δενδρογράμματος. Ως βέλτιστη γραμμική διάταξη ορίζουμε τη διάταξη που μεγιστοποιεί την ομοιότητα μεταξύ γειτονικών γονιδίων, δηλαδή γονίδια με παρόμοιο επίπεδο έκφρασης θα είναι γειτονικά (*adjacent*). Για να

πετύχουμε αυτή τη μεγιστοποίηση χρησιμοποιούμε στα γονίδια σταθμίσεις (*weights*), όπως το κατά μέσο όρο επίπεδο έκφρασής τους και τα διατάσσουμε κατά αύξουσα διάταξη (Eisen et al. 1998).

2.5 Μη Ιεραρχικές Μέθοδοι Ομαδοποίησης

Στις μεθόδους αυτές γνωρίζουμε εκ των προτέρων το πλήθος k των αρχικών ομάδων ή το προσδιορίζουμε μέσω ιεραρχικής ομαδοποίησης. Στη συνέχεια μετακινούμε τα στοιχεία μεταξύ των ομάδων έως ότου πετύχουμε τις επιθυμητές ομάδες. Εναλλακτικά, αντί για k ομάδες μπορούμε να έχουμε k μητρικά σημεία (*seed points*), όπου στη συνέχεια γύρω από αυτά ταξινομούμε τα υπόλοιπα $p-k$ στοιχεία που τα προσεγγίζουν περισσότερο σχηματίζοντας έτσι τις τελικές k ομάδες (Causton et al. 2003). Οι πιο γνωστές μέθοδοι μη ιεραρχικής ομαδοποίησης που χρησιμοποιούνται στην ανάλυση γονιδιακών εκφράσεων είναι η k-means και η PAM (Erfaneh and Yonghong, 2009). Ωστόσο, συχνά χρησιμοποιείται και η μέθοδος SOM (Self-Organized Maps), η οποία μπορεί να θεωρηθεί ως η γραφική αναπαράσταση των ομάδων που προκύπτουν από μη ιεραρχική ομαδοποίηση (Erfaneh and Yonghong, 2009).

2.5.1 Η k-means ομαδοποίηση

Οι πιο γνωστοί αλγόριθμοι που έχουν αναπτυχθεί για την k-means είναι του Forgy και του McQueen. Έχει αποδειχθεί ότι ο αλγόριθμος του McQueen είναι πιο γρήγορος, διότι από τις πρώτες επαναλήψεις πλησιάζει πολύ κοντά στην τελική λύση, ενώ στις υπόλοιπες επαναλήψεις οι διαφοροποιήσεις που προκύπτουν στην σύσταση των ομάδων οφείλονται στη μετακίνηση μικρού αριθμού παρατηρήσεων που «συνορεύουν» με κάποιες ομάδες. Συγκεκριμένα, σύμφωνα με τον παρακάτω πίνακα (Κούτρας, 2005) έχει βρεθεί ότι

Πίνακας 2.5.1

Διαδικασία	**Forgy**	**McQueen**
υπολογισμοί αποστάσεων	nk	$(2n-k)k$
συγκρίσεις αποστάσεων	$n(k-1)$	$(2n-k)(k-1)$
ανανεώσεις κέντρων ομάδων	0	$n-k$

Παρακάτω παρουσιάζουμε τον αλγόριθμο του McQueen μέσω βημάτων (Κούτρας, 2005)

ΒΗΜΑ 1^{Ο}: επιλέγουμε k μητρικά σημεία από το πλήθος p στοιχείων,

ΒΗΜΑ 2^{Ο}: κάθε στοιχείο από τα $p-k$ που έμειναν το κατατάσσουμε στο μητρικό σημείο-ομάδα από το οποίο έχει την μικρότερη απόσταση. Μόλις τοποθετήσουμε σε κάθε ομάδα από ένα στοιχείο υπολογίζουμε το κέντρο βάρους της ομάδας,

ΒΗΜΑ 3^{Ο}: μόλις κατατάξουμε όλα τα $p-k$ στοιχεία στις κατάλληλες ομάδες και υπολογίσουμε το τελικό κέντρο βάρους για κάθε ομάδα εξετάζουμε για τελευταία φορά αν πρέπει να κινηθούν τα στοιχεία μεταξύ των ομάδων λαμβάνοντας υπόψη την απόστασή τους από τα κέντρα βάρους των ομάδων. Με αυτόν τον τρόπο καταλήγουμε σε ισοπληθείς περίπου ομάδες.

Στην περίπτωση που χρησιμοποιήσουμε k μητρικές ομάδες, υπολογίζουμε το κέντρο βάρους τους και εφαρμόζουμε τα βήματα 2 και 3 όπως περιγράφηκαν παραπάνω (Κούτρας, 2005).

Να σημειώσουμε ότι ως κέντρο βάρους (*centroid*) μίας ομάδας ορίζουμε το διάνυσμα των μέσων ανά γονίδιο για όλα τα στοιχεία της ομάδας. Κάθε στοιχείο κατατάσσεται στην ομάδα από της οποίας το κέντρο απέχει λιγότερο. Για να βρούμε αυτή την απόσταση συνήθως χρησιμοποιούμε την ευκλείδεια απόσταση (Κούτρας, 2005).

Η επιλογή των αρχικών k μητρικών σημείων μπορεί να επηρεάσει τις ομάδες που θα σχηματιστούν με κίνδυνο να καταλήξουμε σε «αφύσικη» ομαδοποίηση των

στοιχείων (Causton et al. 2003). Για να αποφύγουμε αυτό το πρόβλημα επαναλαμβάνουμε τον αλγόριθμο για διαφορετικά μητρικά σημεία κάθε φορά, έτσι ώστε να καταλήξουμε σε εκείνα τα μητρικά σημεία γύρω από τα οποία θα συγκεντρωθούν τα υπόλοιπα $p-k$ στοιχεία σχηματίζοντας ομάδες με την ελάχιστη εντός ομάδας διακύμανση (*within-clusters variance*). Όμως, με αυτόν τον τρόπο η $k-$means μέθοδος γίνεται υπολογιστικά χρονοβόρα ιδιαίτερα αν το πλήθος των στοιχείων είναι πάρα πολύ μεγάλο (Κούτρας, 2005). Να σημειώσουμε ότι στην ομαδοποίηση των γονιδίων λόγω της ιδιαίτερης «φύσης» τους είναι δύσκολο εκ των προτέρων να γνωρίζουμε το πλήθος ομάδων των γονιδίων που φαίνεται λογικό να υφίσταται με αποτέλεσμα να επαναλαμβάνουμε τον αλγόριθμο έως ότου καταλήξουμε στη βέλτιστη ομαδοποίηση (Daxin et al. 2004).

Επιπλέον, η ύπαρξη ακραίων τιμών (*outliers*) μπορεί να επηρεάσει τον αλγόριθμο με αποτέλεσμα να σχηματιστούν ομάδες με πολύ διεσπαρμένα άτομα (Causton et al. 2003). Αυτό σημαίνει ότι η απόσταση των στοιχείων κάθε ομάδας από το κέντρο βάρους της θα είναι μεγάλη και αυτό είναι ένδειξη ότι η ομαδοποίηση δεν είναι ιδανική.

Ιδιαίτερο χαρακτηριστικό της $k-$means μεθόδου ομαδοποίησης είναι ότι εντός κάθε ομάδας τα στοιχεία έχουν όσο το δυνατόν πιο μικρή απόσταση από το κέντρο βάρους της ομάδας, ενώ μεταξύ των ομάδων τα στοιχεία της μιας ομάδας απέχουν όσο το δυνατόν περισσότερο από το κέντρο βάρους της άλλης ομάδας. Ωστόσο, η συγκεκριμένη μέθοδος δεν μας πληροφορεί για τη σχέση μεταξύ των ομάδων που σχηματίζονται, όπως συμβαίνει με το δενδρόγραμμα που προκύπτει από κάποια μέθοδο ιεραρχικής ομαδοποίησης (Erfaneh and Yonghong, 2009).

Τέλος, λόγω της ιδιαίτερης «φύσης» των γονιδιακών εκφράσεων να παρουσιάζουν «θόρυβο» (*noisy data*) υπάρχει κίνδυνος με την εφαρμογή του $k-$means αλγόριθμου να καταλήξουμε σε ομάδες με «θόρυβο», όπου η εντός ομάδας διακύμανση θα είναι ιδιαίτερα υψηλή, και αυτό αποτελεί ένδειξη ότι η ομαδοποίηση δεν είναι σωστή (Daxin et al. 2004).

2.5.2 Η PAM (*Partitioning Around Medoids*) ομαδοποίηση

Σε αντίθεση με τον $k-$means αλγόριθμο αρχικά χρησιμοποιούμε αλγόριθμο για την επιλογή των k medoids (αντί k κέντρων βάρους στην $k-$means μέθοδο). Ως

medoid ορίζουμε εκείνο το στοιχείο της συστάδας που έχει τη μικρότερη μέση απόσταση σε σχέση με τα άλλα στοιχεία της συστάδας αυτής. Σκοπός της μεθόδου είναι να μειωθεί το τετραγωνικό σφάλμα σε κάθε συστάδα που σχηματίζεται, δηλαδή το άθροισμα των αποστάσεων των στοιχείων της συστάδας από το κέντρο της (Theodoridis and Koutroubas, 2006).

Ο αλγόριθμος της PAM μεθόδου για την επιλογή των k αντιπροσωπευτικών medoids και την δημιουργία των συστάδων διαιρείται σε δύο στάδια και περιγράφεται παρακάτω βήμα-βήμα (Theodoridis and Koutroubas, 2006):

➢ Built-step

Είναι το πρώτο στάδιο του αλγορίθμου, όπου επιλέγουμε τα αρχικά k medoids ως εξής

ΒΗΜΑ 1^{Ο}: επιλέγουμε τυχαία k στοιχεία από το σετ δεδομένων ως αρχικά medoids,

ΒΗΜΑ 2^{Ο}: υπολογίζουμε την απόσταση των υπολοίπων $p-k$ στοιχείων από τα k επιλεγμένα στοιχεία και γύρω από αυτά τοποθετούμε τα πλησιέστερα στοιχεία σχηματίζοντας έτσι k συστάδες,

ΒΗΜΑ 3^{Ο}: βρίσκουμε την αντικειμενική συνάρτηση (*objective function, OF*) ή αλλιώς το «συνολικό κόστος» (*total cost*) της επιλογής αυτών των σημείων ως medoids. Αυτή η συνάρτηση ορίζεται ως το άθροισμα των αποστάσεων όλων των σημείων από το medoid της συστάδας στην οποία ανήκουν, δηλαδή

$$OF = \sum_{i=1}^{p} d\left(i, m_W\right)$$

όπου m_W είναι το medoid της συστάδας $W = 1, 2, ..., k$.

➢ Swap-step

Είναι το δεύτερο στάδιο του αλγορίθμου, όπου εξετάζουμε αν μπορούν να συμβούν αντικαταστάσεις των παραπάνω medoids, έτσι ώστε να καταλήξουμε στα τελικά medoids που θα χρησιμοποιήσουμε στην ανάλυσή μας. Παρακάτω περιγράφεται αυτό το στάδιο

ΒΗΜΑ 1^{Ο}: έστω ότι διαλέγουμε τυχαία κάποιο από τα υπόλοιπα $p-k$ στοιχεία ως non-medoids και αντικαθιστούμε προσωρινά κάποια από τα k medoids,

ΒΗΜΑ 2^{Ο}: επαναλαμβάνουμε τα βήματα 2 και 3 του Built-step,

ΒΗΜΑ 3^{Ο}: βρίσκουμε τη διαφορά μεταξύ της OF του Built-step (OF_BS) και της OF του Swap-step (OF_SS). Αν η διαφορά είναι θετική, δηλαδή $(\text{OF\_BS}) - (\text{OF\_SS}) > 0$, τότε σωστά έγινε η αντικατάσταση στο βήμα 1 του Swap-step, οπότε τώρα το non-medoid γίνεται medoid. Διαφορετικά αναιρούμε την αντικατάσταση. Αν η διαφορά είναι μηδενική, τότε είτε κρατήσουμε την αντικατάσταση είτε όχι, η OF θα είναι ίδια και στις δύο περιπτώσεις,

ΒΗΜΑ 4^{Ο}: επαναλαμβάνουμε τα βήματα του Swap-step όσες φορές χρειαστεί έως ότου να μην συμβαίνουν πλέον αντικαταστάσεις των medoids.

➢ Final-step

Αφού καταλήξουμε στα k αντιπροσωπευτικά medoids, κατατάσσουμε τα υπόλοιπα $p-k$ στοιχεία στα πλησιέστερα medoids. Δηλαδή, το στοιχείο i τοποθετείται στην συστάδα V_i, όταν το medoid m_{V_i} αυτής της συστάδας βρίσκεται πιο κοντά στο στοιχείο i από οποιοδήποτε άλλο medoid m_W. Συγκεκριμένα θα πρέπει να ισχύει ότι $d\left(i, m_{V_i}\right) \le d\left(i, m_W\right)$ για $W = 1, 2, ..., k$.

Στη ομαδοποίηση στοιχείων με χρήση PAM λαμβάνουμε υπόψη τις παρακάτω ποσότητες έτσι ώστε να συγκρίνουμε μεταξύ τους τις συστάδες και να μπορούμε να τις αξιολογήσουμε ως προς την αξιοπιστία τους γενικότερα (Nagpaul, 1999):

➢ το μέγεθος κάθε συστάδας,

➢ η εντός-συστάδας μέγιστη απόσταση από το medoid (*Maximum Distance to medoid*), MD, που ορίζεται ως η μέγιστη από τις αποστάσεις των στοιχείων μιας συστάδας C από το medoid της, δηλαδή $\text{MD} = \max d_{ij}$ όπου $i, j \in C$ και j το medoid της συστάδας,

➢ η εντός-συστάδας μέση απόσταση από το medoid (*Average Distance to medoid*), AD, που ορίζεται ως η μέση απόσταση όλων των στοιχείων μιας συστάδας C από το medoid της, δηλαδή $\text{AD} = \dfrac{1}{N_j} \sum_{i \in C} d_{ij}$ όπου $i, j \in C$, j το medoid της συστάδας και N_j το πλήθος αντικειμένων της συστάδας εκτός του medoid,

➢ η διάμετρος (*Diameter*) της συστάδας, D, που ορίζεται ως η μέγιστη από τις αποστάσεις μεταξύ δύο στοιχείων που ανήκουν σε μία συστάδα C, δηλαδή $\mathrm{D} = \max d_{ik}$, όπου $i,k \in C$ και

➢ ο διαχωρισμός (*Separation*) μιας συστάδας, S, που ορίζεται ως η μικρότερη από τις αποστάσεις μεταξύ δύο στοιχείων που δεν ανήκουν στην ίδια συστάδα, δηλαδή $\mathrm{S} = \min d_{il}$, όπου $i \in C$, $l \notin C$.

Τέλος, οι συστάδες που δημιουργούνται με χρήση της μεθόδου PAM μπορούν να χαρακτηρισθούν ως απομονωμένες συστάδες (*isolated clusters*) και να διακριθούν σε $L-$συστάδες ή L^*-συστάδες, αν ικανοποιούν τους παρακάτω ορισμούς (Nagpaul, 1999):

➢ μία συστάδα C θεωρείται $L-$συστάδα αν και μόνο αν για κάθε i στοιχείο, όπου $i \in C$, ικανοποιείται η σχέση $\max d_{ij} \le \min d_{ih}$ όπου $i, j \in C$ και $h \notin C$, δηλαδή $D^i \le S^i$ για κάθε στοιχείο $i \in C$. Συγκεκριμένα, στην συστάδα C εντοπίζουμε ποια στοιχεία απέχουν περισσότερο μεταξύ τους (έστω τα στοιχεία i και j) και έπειτα εντοπίζουμε εκείνη τη συστάδα, όπου το h στοιχείο απέχει λιγότερο από το i στοιχείο της συστάδας C και ταυτόχρονα αυτή η απόσταση είναι μεγαλύτερη από την απόσταση μεταξύ των στοιχείων i και j της συστάδας C.

➢ μία συστάδα C θεωρείται L^*-συστάδα αν και μόνο για κάθε i στοιχείο, όπου $i \in C$, ικανοποιείται η σχέση $\max d_{ij} \le \min d_{lh}$ όπου $i, j, l \in C$ και $h \notin C$, δηλαδή $D^i \le S^l$ όπου $i, l \in C$. Είναι προφανές ότι αν $l = i$, τότε η L^*-συστάδα συμπίπτει με την $L-$συστάδα.

Επιπλέον, για τις απομονωμένες συστάδες πρέπει να σημειώσουμε ότι αυτή η ιδιότητα εξαρτάται από:

(α) την εσωτερική δομή των συστάδων, δηλαδή τα στοιχεία που περιέχει κάθε συστάδα, αλλά και τις αποστάσεις μεταξύ των στοιχείων αυτών.

(β) την απόσταση κάθε συστάδας σε σχέση με τις άλλες συστάδες.

Ο PAM αλγόριθμος έχει αποδειχθεί (Theodoridis and Koutroubas, 2006) ότι υπερέχει από τον $k-$means αλγόριθμο στα εξής σημεία:

(α) είναι πιο ανθεκτικός (*robust*) όταν έχουμε θορυβώδη δεδομένα (*noisy data*) και ακραίες τιμές (*outliers*),

(β) χρησιμοποιείται πίνακας αποστάσεων D, οπότε μπορούμε να μελετήσουμε τις αποστάσεις μεταξύ των σημείων. Η PAM είναι η μόνη μη ιεραρχική μέθοδος ομαδοποίησης που χρησιμοποιεί πίνακα αποστάσεων.
(γ) προσπαθεί να ελαχιστοποιήσει ένα άθροισμα αποστάσεων αντί για ένα άθροισμα από τετράγωνα ευκλείδειων (συνήθως) αποστάσεων.

2.5.3 Η SOM (*Self-Organizing Maps*) ομαδοποίηση

Πρόκειται για την γραφική αναπαράσταση μεγάλου συνόλου δεδομένων σε έναν χάρτη μικρών διαστάσεων, όπου τα δεδομένα είναι εμφανώς ομαδοποιημένα. Η μέθοδος θυμίζει την $k-$means ομαδοποίηση (Hautaniemi et al. 2003). Ουσιαστικά είναι μία μέθοδος μείωσης της διάστασης των δεδομένων κατασκευάζοντας έναν χάρτη που είναι συνήθως δύο διαστάσεων, όπου απεικονίζουμε τις ομοιότητες δεδομένων μέσω της ομαδοποίησής τους. Συνεπώς, με την μέθοδο SOM πετυχαίνουμε όχι μόνο να μειώσουμε τη διάσταση των δεδομένων, αλλά και να αναδείξουμε τις ομοιότητές τους. Γενικά, η SOM είναι εύκολη και γρήγορη στην εφαρμογή της χωρίς πολύπλοκο αλγόριθμο και παρέχει αποτελέσματα με ακρίβεια (*accuracy*) και ανθεκτικότητα (*robustness*) (Erfaneh and Yonghong, 2009).

Ο αλγόριθμος της μεθόδου SOM περιγράφεται παρακάτω βήμα-βήμα (Sayad, 2010):

ΒΗΜΑ 1^{Ο}: Κατασκευάζουμε έναν εξάγωνο ή ορθογώνιο χάρτη που περιέχει κόμβους (*nodes*), οι οποίοι είναι συνδεδεμένοι μεταξύ τους σαν σε πλέγμα (*grid*).

ΒΗΜΑ 2^{Ο}: Σε κάθε κόμβο καθορίζουμε ένα αρχικό διάνυσμα αναφοράς (*reference vector* ή *codebook vector*) από n βάρη (*weights*), όσα είναι τα δείγματα (*samples*).

ΒΗΜΑ 3^{Ο}: Από τον $n \times p$ πίνακα πρωτογενών δεδομένων (*data raw matrix*) επιλέγουμε τυχαία ένα διάνυσμα-στήλη (*input vector*). Συνηθίζουμε να τυποποιούμε ως προς τα γονίδια αυτόν τον πίνακα, όταν υπάρχουν πολλές ακραίες τιμές.

ΒΗΜΑ 4^{Ο}: Μέσω ευκλείδειας απόστασης υπολογίζουμε τις αποστάσεις του διανύσματος-στήλη από όλα τα διανύσματα αναφοράς και βρίσκουμε

τον κόμβο που είναι πιο κοντά σε αυτό το διάνυσμα-στήλη. Αυτός ο κόμβος ονομάζεται BMU (Best Matching unit). Αν V είναι το διάνυσμα-στήλη και W_{xy} το διάνυσμα αναφοράς στη θέση (x, y) του χάρτη, τότε η ευκλείδεια απόσταση ορίζεται ως

$$D = \sqrt{\sum_{i=1}^{n} \left(V_i - W_{xy,i}\right)^2}$$

Συνεπώς ο BMU εντοπίζεται από τη σχέση $d\left(V, W_{ts}\right) = \min_{i,j} d\left(V, W_{ij}\right)$ $i = 1, 2, ..., X$ και $j = 1, 2, ..., Y$ (Hautaniemi et al. 2003).

ΒΗΜΑ 5^{Ο}: Υπολογίζουμε την ακτίνα της γειτονιάς που σχηματίζεται γύρω από τον BMU-κόμβο και καθορίζει το μέγεθος της γειτονιάς. Στην αρχή ($t = 0$) του αλγόριθμου η ακτίνα αυτή συμπίπτει με την ακτίνα του πλέγματος. Μετά από κάθε επανάληψη των βημάτων 3 ως 4 η ακτίνα της γειτονιάς μειώνεται. Η ακτίνα υπολογίζεται από την συνάρτηση (www.ai-junkie.com/ann/som/som1.html)

$$\sigma(t) = \sigma_0 \exp\left(-\frac{t}{\lambda}\right)$$

όπου $\sigma(t)$ είναι η ακτίνα της γειτονιάς στην t επανάληψη

σ_0 είναι η ακτίνα-μήκος του πλέγματος στην αρχή και ορίζεται από την παρακάτω σχέση

$$\sigma_0 = \frac{\max\{\text{πλάτος άξονα x, πλάτος άξονα y}\}}{2}$$

λ είναι σταθερά του χρόνου επανάληψης, η οποία εξαρτάται από την ποσότητα σ_0 και το πλήθος επαναλήψεων N του αλγόριθμου. Υπολογίζεται από την παρακάτω σχέση (www.ai-junkie.com/ann/som/som1.html)

$$\lambda = \frac{N}{\log(\sigma_0)}$$

ΒΗΜΑ 6^{Ο}: Τώρα προσαρμόζουμε το διάνυσμα αναφοράς κάθε κόμβου στη γειτονιά του BMU, αλλά και του BMU, έτσι ώστε να απέχουν όλο και λιγότερο από αυτό το διάνυσμα-στήλη. Συνεπώς, οι κόμβοι που βρίσκονται πιο κοντά στο BMU θα δέχονται μεγαλύτερη προσαρμογή στο διάνυσμα

αναφοράς τους σε σχέση με τους κόμβους που βρίσκονται πιο μακριά. Το καινούριο διάνυσμα αναφοράς υπολογίζεται από την συνάρτηση

$$W(t+1) = W(t) + \Theta(t)L(t)\big(V(t) - W(t)\big)$$

όπου $W(t)$ είναι το διάνυσμα αναφοράς στην επανάληψη t

$\Theta(t)$ είναι ο δείκτης επίδρασης (*influence rate*) των κόμβων

$L(t)$ είναι ο δείκτης εκπαίδευσης (*learning rate*)

$V(t) - W(t)$ είναι η απόκλιση του διανύσματος-στήλη από το διάνυσμα αναφοράς του κόμβου στην επανάληψη t του αλγορίθμου

ΒΗΜΑ 7^{Ο}: Η διαδικασία επαναλαμβάνεται από το βήμα 2 ως και το βήμα 6 έως ότου φτάσουμε σε σύγκλιση (*convergence*) και αυτό σημαίνει ότι η ομαδοποίηση των γονιδίων έχει ολοκληρωθεί, οπότε παρόμοια γονίδια έχουν ομαδοποιηθεί σε γειτονικούς κόμβους και δεν υπάρχουν μετακινήσεις των γονιδίων μεταξύ των κόμβων (θυμίζει $k-$means αλγόριθμο) (Hautaniemi et al. 2003).

Παρατηρήσεις

1. Η μεταβλητή $\sigma(t)$ είναι μία συνάρτηση εκθετικής μείωσης (*exponential decay function*), η οποία μειώνεται σε κάθε επανάληψη του αλγόριθμου με αποτέλεσμα η γειτονιά να μικραίνει (Sayad, 2010). Επιπλέον, σε κάθε επανάληψη του αλγόριθμου ορίζεται ένα «νέο» BMU, το οποίο εξαρτάται από το διάνυσμα-στήλη που επιλέγουμε. Ο αλγόριθμος τερματίζει μόλις η γειτονιά οριστεί από ένα μόνο κόμβο. Γενικά, κάθε φορά που εντοπίζουμε έναν BMU θα πρέπει να καθορίσουμε ποιοι κόμβοι θα ορίζουν τη γειτονιά του BMU. Μόλις υπολογίσουμε την ακτίνα της γειτονιάς, εφαρμόζουμε Πυθαγόρειο Θεώρημα για να εντοπίσουμε τους κόμβους που είναι εντός της ακτίνας της γειτονιάς (www.ai-junkie.com/ann/som/som1.html).

2. Ο δείκτης επίδρασης $\Theta(t)$ είναι συνάρτηση του βήματος επανάληψης του αλγόριθμου και εκφράζει πόσο επιδρά η απόσταση ενός κόμβου από το BMU της γειτονιάς του στην προσαρμογή του διανύσματος αναφοράς του και ορίζεται ως $\Theta(t) = \exp\left(-\frac{D^2}{2\sigma^2(t)}\right)$, όπου D είναι η απόσταση του κόμβου από το BMU της

γειτονιάς του και η ποσότητα $\sigma(t)$ έχει οριστεί ήδη στο **ΒΗΜΑ 5^{Ο}**. Μάλιστα, το πόσο αυξάνεται ή μειώνεται αυτή η επίδραση όταν μειώνεται ή αυξάνεται αντίστοιχα η απόσταση του κόμβου από το BMU της γειτονιάς του θυμίζει Gaussian καμπύλη. Να σημειώσουμε ότι μετά από κάθε επανάληψη του αλγόριθμου ο δείκτης επίδρασης μειώνεται, γι' αυτό θεωρείται και συνάρτηση εκθετικής εξασθένισης. Φυσικά, κόμβοι που βρίσκονται κοντά στο BMU κόμβο επηρεάζονται περισσότερο (άρα έχουν μεγαλύτερο $\Theta(t)$) σε σχέση με κόμβους που βρίσκονται πιο μακριά (www.ai-junkie.com/ann/som/som1.html).

3. Ο δείκτης $L(t)$ υπολογίζεται από τη σχέση $L(t) = L_0 \exp\left(-\frac{t}{\lambda}\right)$, $t = 1, 2, \ldots$, όπου συνήθως ορίζουμε το $L_0 = 0.05$ (Hautaniemi et al. 2003). Ο δείκτης αυτός μειώνεται σε κάθε επανάληψη του αλγόριθμου, έτσι ώστε στις τελευταίες επαναλήψεις να πλησιάζει το μηδέν (ιστοσελίδες, 1). Σε κάθε επανάληψη του αλγόριθμου καθορίζει πόσο γρήγορα θα συγκλίνει ο αλγόριθμος. Προφανώς και αυτός ο δείκτης είναι συνάρτηση εκθετικής εξασθένισης.

4. Δεν υπάρχει κάποιος αυστηρός κανόνας για την επιλογή του πλήθους των κόμβων. Σύμφωνα με τους Vesanto et al (2000) συνήθως ορίζεται ως $5 \cdot \sqrt{p}$, όπου p είναι το πλήθος των γονιδίων (Hautaniemi et al. 2003).

5. Για τον προσδιορισμό των αρχικών διανυσμάτων αναφοράς η PCA θεωρείται ότι υπερέχει σε σχέση με την τυχαία επιλογή τους από τον πίνακα δεδομένων (*input data matrix*). Συγκεκριμένα επιλέγουμε τα δύο πρώτα ιδιοδιανύσματα που προκύπτουν και εφαρμόζουμε τον παρακάτω τύπο (Abe et al. 1999):

$$W_{xy}(0) = x_{aver} + 5\sigma_1 \left[b_1 \frac{\left(x - \frac{X}{2}\right)}{X} + b_2 \frac{\left(y - \frac{Y}{2}\right)}{Y} \right]$$

όπου $W_{xy}(0)$ είναι το αρχικό διάνυσμα αναφοράς του κόμβου που βρίσκεται στη θέση (x, y) του χάρτη

x_{aver} είναι ένα διάνυσμα $n \times 1$ που εκφράζει το μέσο επίπεδο έκφρασης των γονιδίων για κάθε δείγμα

X είναι η πρώτη διάσταση του χάρτη, οπότε θα ισχύει ότι $x = \{1, 2, \ldots, X\}$

Y είναι η δεύτερη διάσταση του χάρτη και ορίζεται από τη σχέση $Y = X\frac{\sigma_1}{\sigma_2}$, οπότε θα ισχύει ότι $y = \{1, 2, ..., Y\}$

σ_1, σ_2 είναι η πρώτη και η δεύτερη ιδιοτιμή των ιδιοδιανυσμάτων b_1 και b_2 αντίστοιχα.

Σημαντικό μειονέκτημα της SOM ομαδοποίησης είναι ότι δεν γνωρίζουμε τις αποστάσεις μεταξύ των κόμβων έτσι ώστε να ομαδοποιήσουμε σε κοινή συστάδα παρόμοιους κόμβους, δηλαδή εκείνους τους κόμβους που έχουν την ελάχιστη απόσταση μεταξύ τους. Η συγκεκριμένη αδυναμία αντιμετωπίζεται με τη χρήση του U-matrix (*Unified Distance Matrix*), ο οποίος είναι ένας χάρτης που παρουσιάζει γραφικά τις αποστάσεις μεταξύ των γειτονικών κόμβων, κάτι το οποίο δεν μπορούμε να διακρίνουμε κατευθείαν από έναν SOM χάρτη. Ουσιαστικά ο U-matrix διαφέρει από τον SOM χάρτη μόνο ως προς τα χρώματα που χρησιμοποιούμε και τις χρωματισμένες περιοχές που έχουμε ορίσει. Συγκεκριμένα, η απόσταση μεταξύ γειτονικών κόμβων υπολογίζεται και παρουσιάζεται στον U-matrix με διαφορετικά χρώματα. Σκούρες αποχρώσεις του γκρι αντιστοιχούν σε μεγάλη απόσταση μεταξύ των γειτονικών κόμβων, οπότε μεγάλη απόσταση μεταξύ των αντίστοιχων διανυσμάτων αναφοράς, ενώ ανοιχτές αποχρώσεις του γκρι αντιστοιχούν σε μικρή απόσταση μεταξύ των γειτονικών κόμβων. Οι περιοχές που ορίζονται από κόμβους ανοιχτών αποχρώσεων του γκρι αποτελούν τις συστάδες, ενώ περιοχές που ορίζονται από κόμβους σκούρων αποχρώσεων του γκρι αποτελούν τα όρια των συστάδων (*cluster borders*). Συνεπώς, ο U-matrix είναι ένα χρήσιμο εργαλείο για κάποιον που θέλει να εντοπίσει συστάδες γονιδίων χωρίς να έχει εκ των προτέρων κάποια γνώση για τις συστάδες αυτές (Hollmen, 1996).

Σε αντίθεση με τον $k-$means αλγόριθμο, ο αλγόριθμος της SOM ομαδοποίησης μπορεί να αντιμετωπίσει δεδομένα με υψηλό «θόρυβο» (*noisy data*) και να κατασκευάσει ένα αξιόπιστο χάρτη ομαδοποίησης των δεδομένων. Ωστόσο, αν δεν καθορίσουμε σωστά από την αρχή το πλήθος κόμβων που ορίζουν τον γεωμετρικό χώρο του χάρτη, τότε κινδυνεύουμε να καταλήξουμε σε ομαδοποίηση γονιδίων που απέχει από την «φυσική» τους (Daxin et al. 2004).

Τέλος, αν κάποια γονίδια παρουσιάζουν μεγάλη διακύμανση στις τιμές τους, τότε ο αλγόριθμος της SOM δεν θα είναι αποτελεσματικός, γιατί υπάρχει ο κίνδυνος να τοποθετήσει τα περισσότερα γονίδια σε μία ή δύο ομάδες μόνο (Daxin et al. 2004).

2.6 Αξιολόγηση της ομαδοποίησης γονιδίων

Πριν την ερμηνεία των αποτελεσμάτων της ομαδοποίησης των γονιδίων επιβάλλεται να αξιολογήσουμε τις μεθόδους ομαδοποίησης που χρησιμοποιήσαμε, ώστε να καταλήξουμε στην πιο κατάλληλη για τα δεδομένα που αναλύουμε και τη φύση του προβλήματος που μελετάμε. Η προσεχτική αξιολόγηση της ομαδοποίησης είναι το κλειδί για την σωστή ερμηνεία των αποτελεσμάτων.

Από τις πιο συνηθισμένες και δύσκολες ερωτήσεις που καλείται να απαντήσει κάθε αναλυτής ως προς την ομαδοποίηση είναι οι παρακάτω:

(α) πώς θα αποφασίσουμε ποια μέθοδος ομαδοποίησης είναι πιο αξιόπιστη

(β) πώς θα καταλήξουμε στο κατάλληλο πλήθος συστάδων

(γ) πώς θα διακρίνουμε μία κατάλληλη συστάδα από μία ακατάλληλη ως προς τα στοιχεία που περιέχει

Οι παραπάνω ερωτήσεις μπορούν να απαντηθούν ικανοποιητικά και αξιόπιστα χρησιμοποιώντας τρεις μετρήσεις-κριτήρια, που χρησιμοποιούνται ευρέως στην αξιολόγηση της ομαδοποίησης εκφράσεων γονιδίων (Brock et al. 2008), όπου αξιολογούμε όλες τις μεθόδους ομαδοποίησης που εφαρμόσαμε.

2.6.1 Εσωτερικά μέτρα (*Internal measures*)

Περιλαμβάνουν κάποιους δείκτες που εξετάζουν το κατάλληλο πλήθος συστάδων και την καταλληλότητα κάθε στοιχείου στην συστάδα που τοποθετήθηκε. Για τον υπολογισμό των εσωτερικών μέτρων λαμβάνουμε υπόψη τα εξής χαρακτηριστικά (Brock et al. 2008):

- <u>πυκνότητα (*compactness*)</u>

  Αναφέρεται στο πόσο ικανοποιητική είναι η σύσταση κάθε συστάδας. Βασίζεται στην ομοιογένεια, δηλαδή στην όσο το δυνατόν μικρότερη εντός-συστάδας απόσταση (*intra-cluster distance*) των στοιχείων κάθε συστάδας. Δεν δίνει καλά αποτελέσματα όταν η ομαδοποίηση είναι περίπλοκη, οπότε δεν φαίνονται ξεκάθαρα οι συστάδες που σχηματίζονται (Handl et al. 2005).

- διαχωρισμός (*separation*)

  Μελετάει πόσο διαφέρουν μεταξύ τους οι συστάδες. Βασίζεται στην ανομοιογένεια, δηλαδή στην όσο το δυνατόν μεγαλύτερη εκτός-συστάδας απόσταση (*inter-cluster distance*) μεταξύ των στοιχείων διαφορετικών συστάδων. Στον υπολογισμό της εκτός συστάδας απόστασης λαμβάνει υπόψη τα κέντρα βάρους (*centroids*) των συστάδων.

- συνδεσιμότητα (*connectivity*)

  Εξετάζει σε τι βαθμό οι συστάδες περιέχουν στοιχεία που είναι γειτονικά μεταξύ τους. Συστάδες που περιέχουν όσο το δυνατόν περισσότερα γειτονικά στοιχεία θεωρούνται κατάλληλες.

Οι πιο συνηθισμένοι δείκτες εσωτερικών μέτρων είναι οι εξής :

2.6.1.1 Δείκτης Silhouette (*Silhouette index*)

Αυτός ο δείκτης ορίζει πόσο σωστά έχουν κατανεμηθεί τα στοιχεία στις συστάδες που ανήκουν σύμφωνα με κάποιον αλγόριθμο. Συγκεκριμένα (Κούτρας, 2005), έστω ένα στοιχείο i, το οποίο ανήκει στην συστάδα A και $a(i)$ η μέση απόσταση του στοιχείου αυτού από τα υπόλοιπα στοιχεία της συστάδας αυτής. Έστω η συστάδα C, όπου $i \notin C$, τότε $d(i,C)$ είναι η μέση απόσταση του στοιχείου i από τα στοιχεία της συστάδας C. Γενικά υπολογίζουμε την απόσταση $d(i,C)$ για όλες τις ομάδες $C \neq A$ στις οποίες δεν ανήκει το στοιχείο i. Επιλέγουμε την μικρότερη απόσταση $d(i,C)$, δηλαδή την $b(i) = \min_{C \neq A} d(i,C)$ που αντιστοιχεί στη συστάδα B και υπολογίζουμε τον δείκτη Silhouette $s(i)$ (ή silhouette width) γι' αυτό το στοιχείο ως εξής (Brock et al. 2008):

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \in [-1,1]$$

Σύμφωνα με το αποτέλεσμα του δείκτη δίνεται η ακόλουθη ερμηνεία (Κούτρας, 2005):

Πίνακας 2.6.1

Αποτέλεσμα	**Ερμηνεία**
$1-\frac{a(i)}{b(i)}$, $b(i)>a(i)$	σωστά τοποθετήσαμε το στοιχείο i στην συστάδα Α, διότι η συστάδα Β, που θεωρείται η δεύτερη καλύτερη συστάδα μετά την Α, δεν βρίσκεται πολύ κοντά στο στοιχείο αυτό.
0 , $b(i)=a(i)$	δεν είναι ξεκάθαρο αν το στοιχείο i πρέπει να τοποθετηθεί στην συστάδα Α ή Β. Πρόκειται για μία ενδιάμεση περίπτωση (*intermediate case*)
$\frac{b(i)}{a(i)}-1$, $b(i)<a(i)$	λάθος τοποθετήσαμε το στοιχείο i στην συστάδα Α, διότι η συστάδα Β βρίσκεται πιο κοντά στο στοιχείο αυτό σε σχέση με την Α, οπότε θα έπρεπε να τοποθετηθεί στην συστάδα Β.

Η χρήση του γραφήματος Silhouette ανακεφαλαιώνει οπτικά την πληροφορία που δίνουν τα silhouettes για την καταλληλότητα κάθε στοιχείου στην συστάδα που βρίσκεται. Συνεπώς, εισάγοντας όλα τα silhouettes σε ένα κοινό γράφημα μπορούμε να αξιολογήσουμε ολόκληρη την ομαδοποίηση των στοιχείων. Σε αυτό το γράφημα αναπαριστάνουμε στον άξονα y τις συστάδες με τα στοιχεία που περιέχουν τα οποία εμφανίζονται κατά φθίνουσα διάταξη σύμφωνα με το $s(i)$ που τους αντιστοιχεί. Στον άξονα x ορίζουμε τις ποσότητες $s(i)$. Όσο πιο μεγάλος είναι ο δείκτης $s(i)$ τόσο πιο ικανοποιητικά θεωρείται ότι έχει τοποθετηθεί το στοιχείο i στην συστάδα που ανήκει (Nagpaul, 1999).

Μέσω του Silhouette plot μπορούμε να αποφασίσουμε και για το πλήθος των συστάδων αν το συνδυάσουμε με το συνολικό μέσο silhouette μήκος (*overall average silhouette width*) $\overline{s}(k)$ του silhouette plot, όπου k το πλήθος συστάδων που βρήκαμε με κάποια μέθοδο ομαδοποίησης (Nagpaul, 1999). Η ποσότητα $\overline{s}(k)$ ορίζεται ως ο μέσος όρος των silhouette widths $s(i)$ όλων των στοιχείων, δηλαδή

$$\overline{s}(k)=\frac{\sum_{i=1}^{p} s(i)}{p}$$

για κάποιο k. Συγκεκριμένα η επιλογή εκείνου του k που μεγιστοποιεί το $\overline{s}(k)$ θεωρείται ως το ιδανικό πλήθος συστάδων και καθορίζεται από την ποσότητα $SC = \max \overline{s}(k)$, όπου $k = 2,3,...,p$ για το οποίο μπορεί να κατασκευαστεί silhouette plot (Nagpaul, 1999). Η ερμηνεία των αποτελεσμάτων της ποσότητας SC, αλλά και της ποσότητας $\overline{s}_j(k) = \frac{1}{n_j}\sum_{i=1}^{n_j} s_j(i)$, που εκφράζει το μέσο όρο των silhouette widths των στοιχείων της j συστάδας, δίνεται παρακάτω (Nagpaul, 1999):

Πίνακας 2.6.2

Αποτέλεσμα	**Ερμηνεία**
$0.71-1$	δυνατή δομή των συστάδων ή της j συστάδας για αυτό το k
$0.51-0.7$	λογική δομή των συστάδων ή της j συστάδας για αυτό το k
$0.26-0.5$	αδύναμη δομή των συστάδων ή της j συστάδας για αυτό το k
-1 έως 0.25	καμία ουσιώδη δομή των συστάδων ή της j συστάδας για αυτό το k

2.6.1.2 <u>Δείκτης Συνδεσιμότητας</u> (*Connectivity index*)

Η ιδέα στην οποία βασίζεται ο δείκτης συνδεσιμότητας (*Connectivity index*) είναι ότι τα γειτονικά στοιχεία θα πρέπει να βρίσκονται στην ίδια συστάδα. Οπότε αυτός ο δείκτης υπολογίζει σε τι βαθμό ο διαχωρισμός (*partition*) των στοιχείων στις συστάδες που ορίστηκαν δείχνει ότι τα γειτονικά στοιχεία είναι στην ίδια συστάδα.

Έστω $nn_{i(j)}$ είναι το στοιχείο j που είναι γειτονικό του στοιχείου i και έστω ότι $x_{i,nn_{i(j)}}$ μία ποσότητα που ορίζεται ως

$$x_{i,nn_{i(j)}} = \begin{cases} 0\text{ , αν τα στοιχεία i και j είναι στην ίδια συστάδα} \\ \frac{1}{j}\text{, διαφορετικά} \end{cases}$$

Έστω ότι p στοιχεία έχουν ομαδοποιηθεί σε k συστάδες που ορίζει το διάνυσμα $C=\{C_1,C_2,...,C_l\}$. Τότε ο δείκτης συνδεσιμότητας αυτών των συστάδων ορίζεται ως

$$conn(C)=\sum_{i=1}^{p}\sum_{j=1}^{L}x_{i,nn_{i(j)}}\in[0,\infty)$$

όπου L είναι μία παράμετρος που καθορίζει το πλήθος των γειτονικών στοιχείων που συνεισφέρουν στην μέτρηση του δείκτη συνδεσιμότητας (Brock et al. 2008).

Τα αποτελέσματα του δείκτη συνδεσιμότητας ερμηνεύονται ως εξής :

Πίνακας 2.6.3

Αποτέλεσμα	**Ερμηνεία**
$conn(C)\approx 0$	οι γειτονικές παρατηρήσεις είναι στην ίδια συστάδα. Η ομαδοποίηση είναι αξιόλογη
$conn(C)\gg 0$	οι γειτονικές παρατηρήσεις δεν είναι στην ίδια συστάδα. Όχι αξιόλογη ομαδοποίηση

2.6.1.3 Δείκτης Dunn (*Dunn index*)

Ο συγκεκριμένος δείκτης υπολογίζει την αναλογία (*ratio*) της μικρότερης απόστασης μεταξύ των στοιχείων που δεν είναι στην ίδια συστάδα (*inter-cluster distance*) και της μεγαλύτερης απόστασης μεταξύ των στοιχείων που είναι στην ίδια συστάδα (*intra-cluster distance*) και ορίζεται ως εξής:

$$D(C)=\frac{\min\limits_{C_k,C_l\in C,C_k\neq C_l}\left\{\min\limits_{i\in C_k,j\in C_l}d(i,j)\right\}}{\max\limits_{C_m\in C}diam(C_m)}\in[0,\infty)$$

όπου $diam(C_m)$ είναι η μέγιστη απόσταση μεταξύ των παρατηρήσεων της συστάδας C_m και $C=\{C_1,C_2,...,C_n\}$ είναι ένα σύνολο που περιέχει τις n συστάδες.

Στόχος του δείκτη Dunn είναι η μεγιστοποίηση της εκτός-συστάδας απόστασης και η ελαχιστοποίηση της εντός-συστάδας απόστασης. Η ομαδοποίηση θεωρείται ικανοποιητική ως προς το πλήθος και τη σύσταση των συστάδων όταν η τιμή του δείκτη Dunn είναι όσο το δυνατόν πιο μακριά από το μηδέν. Συγκεκριμένα, τα αποτελέσματα του δείκτη ερμηνεύονται παρακάτω (Brock et al. 2008):

Πίνακας 2.6.4

Αποτέλεσμα	Ερμηνεία
$0 < D(C) < 1$	δεν έχουμε ορίσει σωστά τις συστάδες, γιατί ικανοποιούν μικρή εκτός-συστάδας απόσταση και μεγάλη εντός-συστάδας απόσταση
$D(C) = 1$	κάποια στοιχεία ανήκουν σε δύο συστάδες, γιατί η εκτός-συστάδας απόσταση ισούται με την εντός-συστάδας απόσταση
$D(C) > 1$	έχουμε ορίσει σωστά τις συστάδες, γιατί ικανοποιούν μεγάλη εκτός-συστάδας απόσταση και μικρή εντός-συστάδας απόσταση

Παρατήρηση

Οι δείκτες Dunn και Silhouette συνδυάζουν την πυκνότητα και τον διαχωρισμό, όχι όμως γραμμικά. Αυτά τα δύο χαρακτηριστικά δηλώνουν αντίθετες τάσεις, καθώς η πυκνότητα έχει την τάση να αυξάνει το πλήθος των συστάδων, ενώ ο διαχωρισμός έχει την τάση να το μειώνει (Brock et al. 2008).

2.6.2 Μέτρα σταθερότητας (*Stability measures*)

Εκτιμούν την προβλεπτική δύναμη ενός αλγορίθμου ομαδοποίησης που εφαρμόσαμε, στο να δημιουργήσει βέλτιστες συστάδες ως προς το πλήθος και τη σύστασή τους. Πρόκειται για μία ειδική εκδοχή των εσωτερικών μέτρων, όπου μετράει τη σταθερότητα των συστάδων συγκρίνοντας κάθε συστάδα με τις υπόλοιπες, όταν αφαιρούμε κάθε φορά ένα δείγμα (Handl et al. 2005). Μικρή τιμή των μέτρων σταθερότητας συνεπάγεται ότι η μέθοδος ομαδοποίησης που εφαρμόσαμε έχει μεγάλη προβλεπτική δύναμη (Brock et al. 2008). Μάλιστα, όταν οι μεταβλητές που μελετάμε είναι υψηλά συσχετισμένες, τότε τα αποτελέσματα αυτών των μέτρων θεωρούνται ικανοποιητικά (Brock et al. 2008).

Οι μετρήσεις σταθερότητας που χρησιμοποιούνται κυρίως είναι οι εξής:

2.6.2.1 APN (Average Proportion of Non-overlap)

Μετράει το μέσο ποσοστό στοιχείων που δεν βρίσκονται στην ίδια συστάδα έχοντας πραγματοποιήσει ομαδοποίηση που βασίζεται αρχικά σε όλο το σύνολο δεδομένων και έπειτα στο σύνολο δεδομένων που προκύπτει όταν αποκλείσουμε ένα τυχαίο δείγμα.

Έστω $C^{i,0}$ η συστάδα που περιέχει το στοιχείο i όταν ομαδοποιήσουμε τα στοιχεία όλου του συνόλου δεδομένων και $C^{i,k}$ η συστάδα που περιέχει το στοιχείο i όταν ομαδοποιήσουμε το σύνολο δεδομένων που προκύπτει αν εξαιρέσουμε το k-οστό δείγμα. Τότε το μέτρο APN υπολογίζεται ως εξής (Brock et al. 2008):

$$APN(C) = \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}\left\{1 - \frac{n\left(C^{i,k} \cap C^{i,0}\right)}{n\left(C^{i,0}\right)}\right\} \in [0,1]$$

όπου:

n είναι το πλήθος των δειγμάτων

$n\left(C^{i,0}\right)$ είναι το πλήθος γονιδίων που περιέχονται στην συστάδα $C^{i,0}$

$n\left(C^{i,k} \cap C^{i,0}\right)$ είναι το πλήθος κοινών γονιδίων που περιέχονται στις συστάδες $C^{i,0}$ και $C^{i,k}$.

2.6.2.2 AD (Average Distance)

Υπολογίζει τη μέση απόσταση μεταξύ στοιχείων που ανήκουν στην ίδια συστάδα, όταν η ομαδοποίηση βασίζεται αρχικά σε όλο το σύνολο δεδομένων και έπειτα στο σύνολο δεδομένων που προκύπτει όταν εξαιρέσουμε τυχαία ένα δείγμα. Ο τύπος του AD ορίζεται ως εξής (Brock et al. 2008):

$$AD(C) = \frac{1}{pn}\sum_{i=1}^{p}\sum_{k=1}^{n}\frac{1}{n\left(C^{i,0}\right)n\left(C^{i,k}\right)}\left\{\sum_{\substack{i\in C^{i,0}\\ j\in C^{i,k}}} d(i,j)\right\} \in [0,\infty)$$

2.6.2.3 ADM (Average Distance between Means)

Με το συγκεκριμένο μέτρο υπολογίζεται η μέση απόσταση μεταξύ των κέντρων βάρους (*centroids*) των συστάδων, όταν η ομαδοποίηση βασίζεται αρχικά σε όλο το

σετ δεδομένων και έπειτα στο σύνολο δεδομένων που προκύπτει όταν εξαιρέσουμε τυχαία ένα δείγμα. Ο τύπος του ADM ορίζεται ως εξής (Brock et al. 2008):

$$ADM\left(C\right)=\frac{1}{pn}\sum_{i=1}^{p}\sum_{k=1}^{n}d\left(\overline{x}_{C^{i,k}},\overline{x}_{C^{i,0}}\right)\in\left[0,\infty\right)$$

όπου $\overline{x}_{C^{i,k}}$ είναι το κέντρο βάρους της συστάδας $C^{i,k}$

$\overline{x}_{C^{i,0}}$ είναι το κέντρο βάρους της συστάδας $C^{i,0}$

2.6.2.4 FOM (Figure Of Merit)

Για να εκτιμήσουμε με το συγκεκριμένο μέτρο την προβλεπτική ικανότητα ενός αλγόριθμου ομαδοποίησης αφαιρούμε από το σύνολο των δεδομένων ένα τυχαίο δείγμα, έστω Ε, ομαδοποιούμε τα γονίδια βασισμένοι στα υπόλοιπα δεδομένα και υπολογίζουμε την εντός-συστάδας διακύμανση (*within-cluster variance*). Συγκεκριμένα ο ορισμός του μέτρου FOM δίνεται παρακάτω (Yeung et al. 2001):

Έστω ότι μελετάμε τα επίπεδα έκφρασης p γονιδίων σε n ασθενείς. Εφαρμόζουμε κάποιο αλγόριθμο ομαδοποίησης των γονιδίων για τους ασθενείς $1,2,...,r-1,r+1,...,n$ εξαιρώντας τον ασθενή r, στον οποίο θα βασιστούμε για να εκτιμήσουμε την προβλεπτική ισχύ (*predictive power*) του αλγόριθμου ομαδοποίησης που εφαρμόσαμε. Επίσης, υποθέτουμε ότι από την ομαδοποίηση των p γονιδίων στον ασθενή r προκύπτουν k συστάδες, $C_1,C_2,...,C_k$. Έστω $R\left(x,r\right)$ είναι το επίπεδο έκφρασης του γονιδίου x στον ασθενή r και $\mu_{C_i}\left(r\right)$ είναι το μέσο επίπεδο έκφρασης (*average expression level*) των γονιδίων στην συστάδα C_i. Τότε το μέτρο FOM για k συστάδες και υπό τον ασθενή r ορίζεται ως η εκτίμηση της ρίζας της μέσης τετραγωνικής απόκλισης του επιπέδου έκφρασης κάθε γονιδίου από το μέσο επίπεδο έκφρασης των γονιδίων στην συστάδα που ανήκουν, δηλαδή

$$FOM\left(r,k\right)=\sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i}\left[R\left(x,r\right)-\mu_{C_i}\left(r\right)\right]^2}$$

Η συνολική προβλεπτική ισχύς του αλγόριθμου ομαδοποίησης που εφαρμόσαμε εκτιμάται από το μέτρο AFOM (Aggregate Figure Of Merit), όπου λαμβάνει υπόψη όλους τους ασθενείς που εξαιρούμε κάθε φορά από το σύνολο δεδομένων, δηλαδή

$$\text{AFOM ή } FOM\left(k\right)=\sum_{r=1}^{n}FOM\left(r,k\right)$$

Παρατηρήσεις

Κάποιες παρατηρήσεις σχετικά με τα παραπάνω που δίνονται στο άρθρο των Yeung et al (2001) είναι οι εξής:

1. Για τον υπολογισμό του $FOM(k)$ αν δεν λάβουμε υπόψη κάποιον από τους n ασθενείς, τότε δεν θα υπάρξει σημαντική επίδραση στο αποτέλεσμα για την ποιότητα της ομαδοποίησης.

2. Μπορούμε να βασιστούμε σε οποιονδήποτε από τους n ασθενείς για να υπολογίσουμε το $FOM(r,k)$.

3. Υπάρχει η γενική πεποίθηση ότι η average linkage και complete linkage δίνουν καλύτερη FOM σε σχέση με την single linkage, οπότε είναι και οι προτεινόμενες μέθοδοι για την ομαδοποίηση των γονιδίων.

4. Η FOM δεν μπορεί να εφαρμοστεί για τη σύγκριση αλγορίθμων ομαδοποίησης που δίνουν διαφορετικό πλήθος συστάδων ή έχουν υπολογιστεί με διαφορετικό μέτρο απόστασης.

2.6.3 Βιολογικά μέτρα (*Biological measures*)

Εξετάζουν πόσο ικανός είναι ο αλγόριθμος ομαδοποίησης που εφαρμόσαμε στο να δημιουργεί συστάδες βιολογικής αξίας. Εφαρμόζονται συνήθως σε ανάλυση δεδομένων με μικροσυστάδες όπου οι παρατηρήσεις μπορεί να είναι γονίδια, τμήμα (*portion*) των γονιδίων που περιέχει μια ακολουθία βάσεων που μπορούν να κωδικοποιήσουν την πρωτεΐνη (*open reading frames, ORF*), σύντομη ακολουθία της μεταγραμμένης ακολουθίας cDNA (*expressed sequence tag, EST*) ή ανάλυση της έκφρασης γονιδίων σε σειρά (*serial analysis of gene expression, SAGE*).

Τα κύρια βιολογικά μέτρα αναλύονται παρακάτω (Brock et al. 2008)

2.6.3.1 *Δείκτης βιολογικής ομοιογένειας (Biological Homogeneity Index, BHI)*

Ο συγκεκριμένος δείκτης υπολογίζει πόσο ομογενείς βιολογικά είναι οι συστάδες. Έστω $B=\{B_1,B_2,...,B_F\}$ ένα σύνολο από F συναρτησιακές ομάδες (*functional classes*) και έστω $B(i)$ η συναρτησιακή ομάδα που περιέχει το γονίδιο i. Ομοίως, έστω $B(j)$ η συναρτησιακή ομάδα που περιέχει το γονίδιο j. Να σημειώσουμε ότι

είναι δυνατόν περισσότερες από μία συναρτησιακές ομάδες να περιέχουν το γονίδιο i ή/ και j. Τότε η δείκτρια συνάρτηση $I\left[B(i)=B(j)\right]$ ορίζεται ως εξής:

$$I\left[B(i)=B(j)\right]=\begin{cases}1\text{ , αν }B(i)\text{ και }B(j)\text{ ταιριάζουν}\\0\text{ , διαφορετικά}\end{cases}$$

Ελπίζουμε ότι τα γονίδια που ανήκουν στην ίδια συστάδα θα ανήκουν και στις ίδιες συναρτησιακές ομάδες. Λαμβάνοντας υπόψη την ομαδοποίηση $C=\{C_1,C_2,...,C_k\}$ του αλγόριθμου που εφαρμόσαμε, και το σύνολο Β των συναρτησιακών ομάδων, τότε το BHI ορίζεται ως

$$BHI(C,B)=\frac{1}{k}\sum_{t=1}^{k}\frac{1}{n_t(n_t-1)}\sum_{i\neq j\in C_t}I\left[B(i)=B(j)\right]\in[0,1]$$

όπου $n_t=n(C_t\cap B)$ είναι το πλήθος κοινών γονιδίων που έχει η συστάδα C_t με το σύνολο βιολογικών ομάδων B.

Όσο πιο μεγάλη είναι η τιμή του δείκτη BHI τόσο πιο μεγάλη είναι βιολογικά η ομοιογένεια των συστάδων.

2.6.3.2 *Δείκτης βιολογικής σταθερότητας (Biological Stability Index, BSI)*

Ορίζεται παρόμοια με τα μέτρα σταθερότητας (*stability measures*) και εξετάζει την αξιοπιστία του αλγόριθμου ομαδοποίησης που εφαρμόσαμε για γονίδια με παρόμοια βιολογική λειτουργικότητα.

Ένα δείγμα αφαιρείται κάθε φορά από το σύνολο δεδομένων και η συστάδα με τα γονίδια παρόμοια συναρτησιακής έκφρασης (*functional annotation*) συγκρίνεται με την συστάδα με γονίδια παρόμοιας, επίσης, συναρτησιακής έκφρασης λαμβάνοντας υπόψη όλο το σύνολο δεδομένων. Ο δείκτης BSI ορίζεται ως

$$BSI(C,B)=\frac{1}{F}\sum_{k=1}^{F}\frac{1}{n(B_k)\left[n(B_k)-1\right]}\frac{1}{n}\sum_{l=1}^{n}\sum_{i\neq j\in B_k}\frac{n\left(C^{i,0}\cap C^{j,l}\right)}{n\left(C^{i,0}\right)}\in[0,1]$$

όπου F είναι το πλήθος των συναρτησιακών ομάδων

n είναι το πλήθος των δειγμάτων

$n(B_k)$ είναι το πλήθος γονιδίων που περιέχει η συναρτησιακή ομάδα k

Όσο πιο μεγάλη είναι η τιμή του δείκτη BSI τόσο πιο σταθερές είναι οι συστάδες των γονιδίων που είναι συναρτησιακά εκφρασμένα (functionally *annotated*)

Παρατηρήσεις

1. Για να αποφασίσουμε ποια από τις ιεραρχικές μεθόδους θα ξεχωρίσουμε, ώστε να την συγκρίνουμε με τις υπόλοιπες μεθόδους ομαδοποίησης, που έχουμε περιγράψει παραπάνω, θα βασιστούμε τόσο στα τρία κριτήρια που περιγράψαμε παραπάνω όσο και στον συντελεστή συσσώρευσης (*agglomerative coefficient*), AC, ο οποίος χρησιμοποιείται στην αξιολόγηση της ιεραρχικής ομαδοποίησης. Ο συντελεστής συσσώρευσης ορίζεται ως

$$AC = \sqrt{\frac{1}{p}\sum_{i=1}^{p} l(i)} \in [0,1]$$

όπου $l(i)$ είναι η απόσταση του i στοιχείου από την πιο πρόσφατη συστάδα και p είναι το πλήθος των στοιχείων που ομαδοποιούμε. Όταν ο συντελεστής αυτός έχει τιμή κοντά στο μηδέν, τότε συμπεραίνουμε ότι στην συγκεκριμένη μέθοδο ομαδοποίησης η συνοχή των στοιχείων είναι ασθενής. Ενώ, όταν ο συντελεστής έχει τιμή κοντά στη μονάδα, τότε συμπεραίνουμε ότι στην συγκεκριμένη μέθοδο έχουμε ισχυρή ομαδοποίηση. Πρέπει να σημειώσουμε ότι ο συντελεστής συσσώρευσης επηρεάζεται από το πλήθος δεδομένων σε κάθε συστάδα και αυξάνεται καθώς αυξάνεται το μέγεθος των συστάδων με αποτέλεσμα να δίνει παραπλανητικό αποτέλεσμα όταν οι συστάδες είναι άνισες ως προς το πλήθος στοιχείων που περιέχουν (Κούτρας, 2005).

2. Παράλληλα με τον συντελεστή συσσώρευσης επιβάλλεται να μελετήσουμε την εντός-συστάδας μεταβλητότητα (*intra-cluster variability*) και την εκτός-συστάδας μεταβλητότητα (*inter-cluster variability*) των συστάδων που προκύπτουν. Σύμφωνα με τους Shamir και Sharan (Chen et al. 2002) η ομοιογένεια ορίζεται ως η μέση απόσταση κάθε στοιχείου από το κέντρο βάρους της συστάδας που ανήκει, δηλαδή

$$H_{aver} = \frac{1}{p}\sum_{i=1}^{p} S(i,R) \qquad (2.6.1)$$

όπου $S(i,R)$ είναι η απόσταση του i στοιχείου από το κέντρο βάρους της συστάδας R στη οποία ανήκει και p είναι το πλήθος των στοιχείων που ομαδοποιούμε. Όσο πιο μικρή είναι η τιμή που δίνει η σχέση $(2.6.1)$, τότε έχουμε μεγάλη ομοιογένεια. Η ετερογένεια (Chen et al. 2002) ορίζεται ως η σταθμισμένη μέση απόσταση μεταξύ

των κέντρων βάρους δύο συστάδων με σταθμίσεις το γινόμενο του πλήθους στοιχείων αυτών των συστάδων, δηλαδή

$$S_{aver} = \frac{1}{\sum_{i \neq j} N_i N_j} \sum_{i \neq j} N_i N_j S\left(\overline{x}_i, \overline{x}_j\right) \qquad (2.6.2)$$

όπου N_i είναι το πλήθος στοιχείων της i συστάδας

$\overline{x}_i$ είναι το κέντρο βάρους της i συστάδας και

$S\left(\overline{x}_i, \overline{x}_j\right)$ είναι η απόσταση μεταξύ των κέντρων βάρους των συστάδων i και j.

Όσο πιο μεγάλη είναι η τιμή που δίνει η σχέση $(2.6.2)$, τότε έχουμε μεγάλη ετερογένεια.

Ωστόσο, τόσο η ομοιογένεια όσο και η ετερογένεια περιέχονται στα εσωτερικά μέτρα, που έχουμε ήδη περιγράψει (Brock et al. 2008).

2.7 Γενικές παρατηρήσεις

Η ομαδοποίηση είναι το πρώτο βήμα της ανάλυσης των δεδομένων γονιδιακών εκφράσεων, καθώς μπορούμε να εντοπίσουμε ποια γονίδια συνεκφράζονται (*co-expressed genes*) με παρόμοιο τρόπο, πόσο διαφέρουν οι ομάδες μεταξύ τους, πόσο πολύπλοκη είναι η ομαδοποίηση των γονιδίων κτλ. Συνεπώς, ο αλγόριθμος που θα επιλέξουμε να χρησιμοποιήσουμε θα πρέπει να παρέχει όσο το δυνατόν πιο «φυσικό» (*natural*) πλήθος των ομάδων. Στη συνέχεια δίνουμε μερικές παρατηρήσεις από το βιβλίο των Daxin et al. (2004).

Στην ανάλυση γονιδιακών εκφράσεων ενδιαφέρον έχει τόσο η ομαδοποίηση γονιδίων (*genes-based clustering*) όσο και η ομαδοποίηση δειγμάτων (*samples-based clustering*), αλλά και γονιδίων και δειγμάτων ταυτόχρονα (*subspace clustering*). Η πρώτη περίπτωση είναι και η πιο συνηθισμένη, γιατί έχει ενδιαφέρον να μελετήσουμε την σχέση μεταξύ των συστάδων όσον αφορά την ομοιότητά τους, αλλά και τη σχέση μεταξύ των γονιδίων εντός της συστάδας ώστε να εντοπίσουμε ποιο γονίδιο είναι το περισσότερο και ποιο το λιγότερο αντιπροσωπευτικό της συστάδας. Γι' αυτό επιβάλλεται η χρήση ενός αλγόριθμου ομαδοποίησης που θα μας διευκολύνει να δούμε γραφικά όλες τις παραπάνω σχέσεις. Συνήθως, αυτό είναι δυνατό με την ιεραρχική ομαδοποίηση όπου δημιουργείται το δενδρόγραμμα.

Λόγω της πολυπλοκότητας των πειραμάτων με μικροσυστάδες υπάρχει η πιθανότητα να καταλήξουμε με ένα σύνολο δεδομένων γονιδιακών εκφράσεων που περιέχει αρκετό «θόρυβο» (*noise*) με αποτέλεσμα να δυσκολευτούμε στην ομαδοποίησή τους. Η δυσκολία αυτή μπορεί να ξεπεραστεί με την εφαρμογή ενός αλγόριθμου που αντιμετωπίζει αυτόν τον «θόρυβο» πραγματοποιώντας αξιόπιστη ομαδοποίηση έτσι ώστε η ερμηνεία των αποτελεσμάτων να είναι ξεκάθαρη και κατανοητή.

Ένας άλλος κίνδυνος που προέρχεται από την πολυπλοκότητα των πειραμάτων με μικροσυστάδες είναι να παρουσιαστούν κάποιες συστάδες «σφηνωμένες» (*embedded*) σε άλλες συστάδες με αποτέλεσμα να μην μπορούμε να ερμηνεύσουμε αυτή την ομαδοποίηση. Γι' αυτό χρειαζόμαστε έναν αλγόριθμο που θα μπορεί να αντιμετωπίσει αυτή τη δυσκολία.

Είναι γνωστό όταν μελετάμε εκφράσεις γονιδίων δεν έχουμε εκ των προτέρων επαρκή γνώση για την ιδανική ομαδοποίησή τους, γι' αυτό συνηθίζεται να ξεκινάμε με κάποια μέθοδο ιεραρχικής ομαδοποίησης, ώστε να εντοπίσουμε το βέλτιστο πλήθος συστάδων και έπειτα εφαρμόζουμε k – means ομαδοποίηση χρησιμοποιώντας τις ομάδες που έχουμε ήδη βρει, καθώς η k – means θεωρείται ως η πιο αξιόπιστη μέθοδος ομαδοποίησης των γονιδίων.

2.8 Αντιμετώπιση των ελλειπουσών τιμών (*missing values*)

Στην ανάλυση εκφράσεων γονιδίων με μικροσυστάδες οι ελλείπουσες τιμές προκύπτουν για διάφορους λόγους, όπως ανεπαρκή ανάλυση (*resolution*), φθορά της εικόνας (*image corruption*) ή απλά λόγω σκόνης ή εκδορών στην φωτογραφική διαφάνεια. Επιπλέον, οι ελλείπουσες τιμές μπορεί να προκύπτουν συστηματικά λόγω ρομποτικών (*robotic*) μεθόδων που χρησιμοποιούνται στην δημιουργία συστάδων με μικροεκφράσεις (McLachlan et al. 2004).

Όταν ένας πίνακας δεδομένων περιέχει ελλείπουσες τιμές, τότε πρέπει να λάβουμε υπόψη ότι δεν μπορούν να οριστούν όλες οι μέθοδοι ανάλυσης των δεδομένων. Για παράδειγμα, τα μέτρα απόστασης (όπως, η ευκλείδεια απόσταση, η συσχέτιση του Pearson, η ελαχιστοποίηση της διακύμανσης) δεν μπορούν να υπολογιστούν όταν υπάρχουν διανύσματα-σχέδια εκφράσεων γονιδίων με ελλείπουσες τιμές. Όμως, είναι δυνατόν να υπολογιστούν στην περίπτωση που απορρίψουμε τα δείγματα με ελλείπουσες τιμές και βασιστούμε στα υπόλοιπα δείγματα που έχουν τιμές.

Δυστυχώς, σε κάποιες άλλες μεθόδους μία τέτοια γενίκευση μπορεί να είναι δύσκολο να εφαρμοστεί με αποτέλεσμα είτε να είμαστε αναγκασμένοι να αντικαταστήσουμε τις ελλείπουσες τιμές με κάποια αυθαίρετη τιμή (για παράδειγμα, με μηδέν) ή να τα εκτιμήσουμε μέσω των υπόλοιπων δεδομένων (Causton et al. 2003).

Συνήθως, οι ελλείπουσες τιμές αντικαθίστανται με μηδέν, όταν πρόκειται για log-μετασχηματισμένα δεδομένα, ή με το μέσο επίπεδο έκφρασης των γονιδίων κατά μήκος των δειγμάτων. Ωστόσο, τέτοιες προσεγγίσεις δεν θεωρούνται ιδανικές, διότι δεν λαμβάνουν υπόψη τη συσχέτιση των δεδομένων. Μια εναλλακτική, αλλά όχι οικονομική λύση είναι η επανάληψη του πειράματος. Υπάρχουν πιο αξιόπιστες μέθοδοι αντικατάστασης (*imputation*) των ελλειπουσών τιμών, όπου λαμβάνουν υπόψη τη συσχέτιση των δεδομένων (McLachlan et al. 2004).

Έστω έχουμε έναν πίνακα δεδομένων A με διάσταση $N \times M$, όπου N είναι το πλήθος των γονιδίων και M είναι το πλήθος των δειγμάτων (McLachlan et al. 2004). Παρακάτω παρουσιάζονται οι μέθοδοι αντικατάστασης των ελλειπουσών τιμών που πρότεινε ο Troyanskaya (McLachlan et al. 2004):

2.8.1 Η μέθοδος του μέσου όρου των γραμμών (row average method)

Σύμφωνα με αυτή τη μέθοδο σε κάθε γονίδιο που περιέχει ελλείπουσες τιμές αντικαθιστούμε αυτές τις ελλείπουσες τιμές με το μέσο όρο των επιπέδων έκφρασής του για όλα τα υπόλοιπα δείγματα για τα οποία έχει τιμή (Causton et al. 2003).

2.8.2 Η μέθοδος των k κοντινότερων γειτόνων (k-nearest neighbors method)

Χρησιμοποιώντας ένα μέτρο ομοιότητας βρίσκουμε τα k γονίδια που έχουν παρόμοιο σχέδιο έκφρασης με το i γονίδιο, χωρίς να λάβουμε υπόψη το επίπεδο έκφρασής τους στο j δείγμα, στο οποίο λείπει το επίπεδο έκφρασης του i γονιδίου (McLachlan et al. 2004). Αυτά τα k γονίδια πρέπει να έχουν επίπεδο έκφρασης στο j δείγμα απαραιτήτως. Ως άριστο θεωρείται το k που κυμαίνεται από το 10 ως το 20 (Hastie et al. 1999, Causton et al. 2003). Βρίσκουμε το σταθμισμένο μέσο επίπεδο έκφρασης αυτών των k γονιδίων για το j δείγμα μόνο. Η στάθμιση μπορεί να είναι ανάλογη του μέτρου ομοιότητας (όπως, ευκλείδεια απόσταση, συσχέτιση Pearson, ελαχιστοποίηση διακύμανσης) που χρησιμοποιήθηκε (McLachlan et al. 2004).

Δηλαδή, αν χρησιμοποιήσουμε τον συντελεστή συσχέτισης του Pearson για τον εντοπισμό των k γονιδίων που «μοιάζουν» ως προς το σχέδιο έκφρασής τους με το i υπό εξέταση γονίδιο, τότε η ελλείπουσα τιμή y_{ij} του γονιδίου αυτού υπολογίζεται από την σχέση

$$\overline{y}_{\cdot j} = \frac{w_1 \cdot y_{1j} + w_2 \cdot y_2 + ... + w_k \cdot y_{kj}}{w_1 + w_2 + ... + w_k} = \frac{\sum_{m=1}^{k} w_m \cdot y_{mj}}{\sum_{m=1}^{k} w_m}$$

όπου $w_m = r_{mi} \frac{1}{\sum_{m=1}^{k} r_{mi}}$ είναι η στάθμιση της τιμής y_{mj} για το γονίδιο m

(Hourani and Emary, 2009).

2.8.3 Singular Value Decomposition (SVD)

Αρχικά αντικαθιστούμε όλες τις ελλείπουσες τιμές χρησιμοποιώντας την row-average μέθοδο, διότι ο SVD αλγόριθμος δεν μπορεί να εφαρμοστεί σε μη ολοκληρωμένους πίνακες. Έπειτα ο SVD αλγόριθμος υπολογίζει ορθογώνιες εκφράσεις γονιδίων (*orthogonal expression patterns*), οι οποίες αν συνδυαστούν γραμμικά μπορούν να προσεγγίσουν τις εκφράσεις των γονιδίων του πίνακα A. Η SVD του πίνακα A είναι

$$A = U_1 \Lambda U_2^T$$

όπου U_1 και U_2 είναι ορθογώνιοι πίνακες.

Ο πίνακας U_1 είναι διάστασης $N \times M$ και οι στήλες του αποτελούν τα ιδιοδείγματα (*eigen-assays*), τα οποία δημιουργούν μία ορθογώνια βάση με τα προφίλ των ασθενών. Συνεπώς, τα ιδιοδείγματα είναι ανά δύο ορθογώνια, δηλαδή $\left(u_i^1\right)' \cdot u_j^1 = \sum_{k=1}^{N} u_{ki}^1 \cdot u_{kj}^1 = 1$ για $i \neq j = 1, 2, ..., M$ και έχουν μήκος ίσο με μηδέν, δηλαδή $\left(u_i^1\right)' \cdot u_i^1 = \sum_{k=1}^{N} \left(u_{ki}^1\right)^2 = 0$ για $i = j = 1, 2, ..., M$. Οι γραμμές του U_2 είναι τα ιδιοδιανύσματα (*eigenvectors*) ή ιδιογονίδια (*eigen-genes*) του πίνακα $A^T A$, τα οποία αντιστοιχούν στις ιδιοτιμές (*eigenvalues*) του διαγώνιου πίνακα Λ και δημιουργούν

την ορθογώνια βάση με τα σχέδια έκφρασης των γονιδίων. Κάθε ιδιοδιάνυσμα έχει τη δική του ιδιοτιμή. Τα μεγαλύτερα k ιδιοδιανύσματα (που θεωρούνται και τα πιο σημαντικά) επιλέγονται εμπειρικά από τον πίνακα U_2 και συνεπώς η διάσταση του πίνακα Α μειώνεται σε $K \times M$. Το i γονίδιο, που έχει ελλείπουσα τιμή στο j δείγμα, εξισώνεται με τα k ιδιοδιανύσματα με μορφή παλινδρόμησης, δηλαδή $y_{ij} = a + b_1 e_1 + b_2 e_2 + ... + b_k e_k$, έχοντας αγνοήσει όλες τις τιμές εκφράσεων για το j δείγμα. Ουσιαστικά, εκτιμάμε την ελλείπουσα τιμή y_{ij} ως γραμμικό συνδυασμό των k ιδιοδιανυσμάτων σταθμισμένα με τους συντελεστές παλινδρόμησης. Ο SVD αλγόριθμος επαναλαμβάνεται έως ότου η συνολική διακύμανση του μειωμένου πίνακα A να συγκλίνει (*converge*) σε μία επαρκώς μικρή αυθαίρετη τιμή (McLachlan et al. 2004 και Causton et al. 2003).

2.8.4 Principal Component Analysis (PCA)

Στη συγκεκριμένη μέθοδο αρχικά επιλέγουμε τα k πιο σημαντικά ιδιοδιανύσματα και έπειτα προβάλουμε (*project*) αυτά τα ιδιοδιανύσματα στο i γονίδιο, το οποίο έχει ελλείπουσα τιμή στο j δείγμα, έτσι ώστε να δούμε πόσο αποδίδει το καθένα στο i γονίδιο. Τέλος, ανακατασκευάζουμε (*reconstruct*) την ελλείπουσα τιμή από το j στοιχείο των k ιδιοδιανυσμάτων, το οποίο θεωρούμε ότι είναι ανάλογο του ποσού που αποδίδει το αντίστοιχο ιδιοδιάνυσμα στο i γονίδιο (Causton et al. 2003).

Ο Troyanskaya σύγκρινε αυτές τις μεθόδους μεταξύ τους ως προς την πολυπλοκότητα και την ακρίβεια των αποτελεσμάτων χρησιμοποιώντας την κανονικοποιημένη ρίζα του μέσου τετραγωνικού σφάλματος (*normalized Root Mean Square error-RMS*). Κατέληξε στο συμπέρασμα ότι η row-average μέθοδος είναι η γρηγορότερη, αλλά όχι η ακριβέστερη (McLachlan et al. 2004). Τελικά, προτείνει την k -nearest neighbors μέθοδο (k – NN μέθοδος) ως την ακριβέστερη, αλλά και την πιο ανθεκτική (*robust*) απέναντι στο αυξημένο ποσοστό των ελλειπουσών τιμών και συγκεκριμένα για k που κυμαίνεται από 10 ως 20 (Causton et al. 2003). Ωστόσο, σημειώνει ότι οι ερευνητές πρέπει να είναι προσεχτικοί όταν αντικαθιστούν τις ελλείπουσες τιμές με τιμές που προκύπτουν από μεθόδους, όπως αυτές που περιγράφηκαν παραπάνω, γιατί υπάρχει ο κίνδυνος οι τιμές αυτές να μην συμφωνούν

με τη βιολογική «φύση» των γονιδίων που εξετάζουμε, με αποτέλεσμα η ανάλυση που θα ακολουθήσει να είναι παραπλανητική (McLachlan et al. 2004).

Κεφάλαιο 3

Εφαρμογή στην ομαδοποίηση εκφράσεων γονιδίων

3.1 Εισαγωγή

Στο κεφάλαιο αυτό επιδιώκουμε να ομαδοποιήσουμε 37 γονίδια εφαρμόζοντας και εξετάζοντας τις μεθόδους που περιγράψαμε στο κεφάλαιο 2. Το σύνολο δεδομένων που μελετάμε αποτελείται από 367 γυναίκες που υποβλήθηκαν σε εγχείριση για αφαίρεση του καρκίνου στο μαστό και 37 γονίδια, που εντοπίστηκαν στον καρκινικό ιστό που αφαιρέθηκε από κάθε γυναίκα. Για κάθε ασθενή έχει καταγραφεί το επίπεδο έκφρασης των περισσοτέρων γονιδίων με αποτέλεσμα να υπάρχουν κάποιες ασθενείς με ελλείπουσες τιμές σε ορισμένα γονίδια. Ωστόσο, υπάρχουν 51 ασθενείς για τις οποίες δεν καταγράφηκε το επίπεδο έκφρασης για κανένα από τα γονίδια. Αυτό το σύνολο ασθενών επιβάλλεται να αφαιρεθεί από τα δεδομένα μας για να μπορούμε να ομαδοποιήσουμε τα γονίδια όσο το δυνατόν πιο σωστά. Συνεπώς, το τελικό σύνολο δεδομένων που αναλύουμε παρακάτω αποτελείται από 316 γυναίκες ασθενείς και 37 γονίδια.

3.2 Ιεραρχική ομαδοποίηση με χρήση θερμικού χάρτη

Η $k-\text{NN}$ μέθοδος θεωρείται η προτιμότερη για αντικατάσταση των ελλειπουσών τιμών σε εκφράσεις γονιδίων (Troyanskaya, 2001), οπότε θα την εφαρμόσουμε. Εξάλλου, γνωρίζουμε ότι στη $k-\text{NN}$ μέθοδο επιβάλλεται τα k γονίδια που θα επιλέξουμε όχι μόνο να μοιάζουν με το υπό εξέταση i γονίδιο, αλλά επίσης να μην έχουν ελλείπουσες τιμές στο j δείγμα, όπου στην περίπτωσή μας υπάρχουν αυτά τα γονίδια. Δεν θα εφαρμόσουμε τη row-average μέθοδο, διότι δεν θεωρείται ακριβής.

Η μέθοδος PCA θα ήταν η εναλλακτική επιλογή στην περίπτωση που δεν υπήρχε κανένα γονίδιο με τιμή στο j δείγμα, έτσι ώστε να θεωρηθεί υποψήφιο να περιληφθεί στα k γονίδια που αναζητάμε για να αντικαταστήσουμε την ελλείπουσα τιμή y_{ij} του γονιδίου i για το j δείγμα, οπότε δεν μπορούσε να εφαρμοστεί η $k-\text{NN}$ μέθοδος. Να σημειώσουμε ότι η PCA και SVD είναι παρόμοιες τεχνικές ανάλυσης που συνηθίζουμε να εφαρμόζουμε στην ανάλυση γονιδιακών εκφράσεων (Wall et al., 2003, Chapter 5), αλλά η SVD εφαρμόζεται κυρίως στην περίπτωση δεδομένων χρονοσειρών.

Έχοντας εκτιμήσει τις ελλείπουσες τιμές μπορούμε να εφαρμόσουμε τις μεθόδους ομαδοποίησης που έχουμε περιγράψει, να επιλέξουμε τις βέλτιστες, να καθορίσουμε το βέλτιστο πλήθος συστάδων και να τις αξιολογήσουμε.

Στο γράφημα 3.1 παρουσιάζουμε έναν θερμικό χάρτη που περιέχει τις συσχετίσεις του Pearson μεταξύ των γονιδίων ανά δύο και δύο δενδρογράμματα, τα οποία είναι ταυτόσημα, διότι και οι δύο άξονες του χάρτη περιέχουν τα ίδια γονίδια. Λαμβάνοντας υπόψη τα χρώματα του θερμικού χάρτη και το εύρος των τιμών στις οποίες αντιστοιχούν σύμφωνα με το «κλειδί» χρωμάτων παρατηρούμε ότι αρκετά γονίδια φαίνονται να είναι αρνητικά συσχετισμένα μεταξύ τους (πράσινο χρώμα), ενώ υπάρχουν περίπου πέντε ομάδες γονιδίων που είναι θετικά συσχετισμένα (αποχρώσεις του κόκκινου) και βρίσκονται κοντά στην κύρια διαγώνιο του χάρτη (έντονο κόκκινο). Το μαύρο χρώμα εκφράζει ασυσχέτιστα γονίδια, ενώ οι πολύ σκούροι τόνοι του κόκκινου και του πράσινου εκφράζουν ασθενώς συσχετισμένα γονίδια, τα οποία φαίνονται να είναι πολύ περισσότερα από τα αρνητικά συσχετισμένα γονίδια.

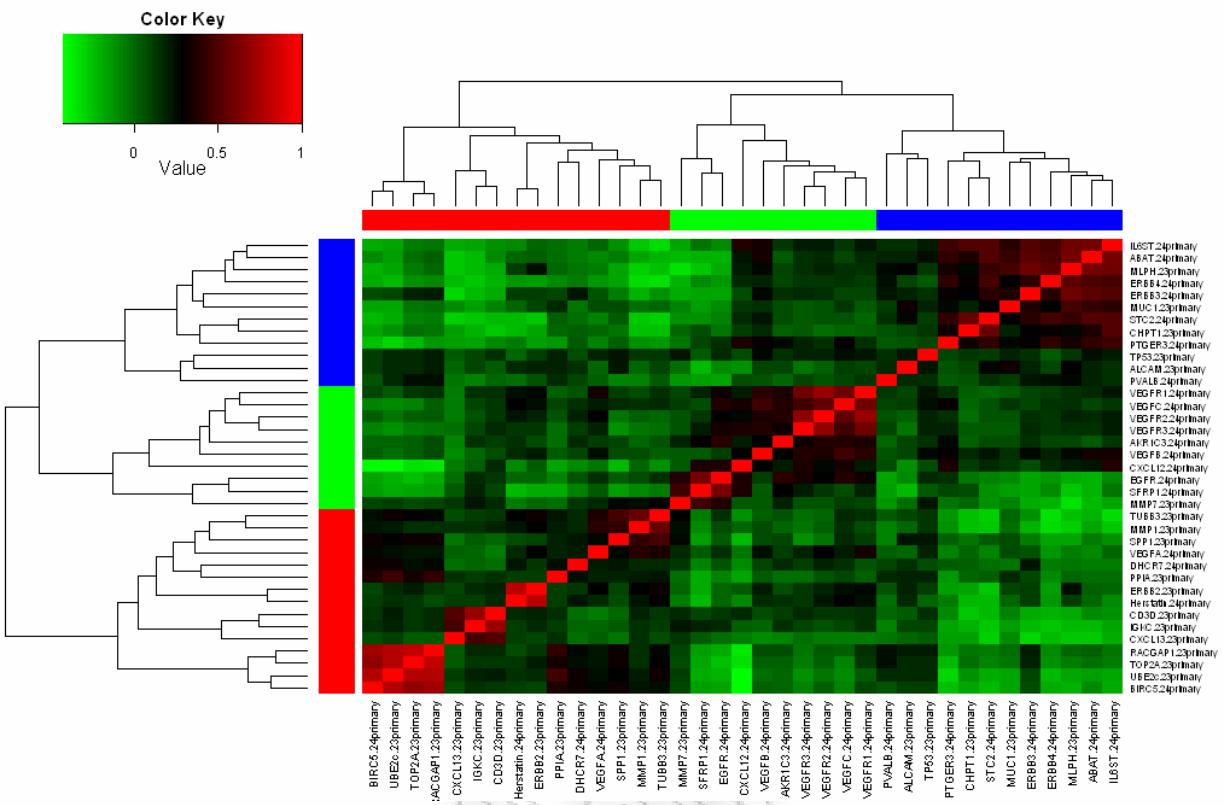
Εύκολα μπορούμε να διακρίνουμε τις 3 συστάδες γονιδίων που προκύπτουν με μικρή ανισορροπία ως προς το πλήθος, στις οποίες αντιστοιχίσαμε χρωματιστή μπάρα για να τις ξεχωρίζουμε. Παρατηρούμε ότι σε κάθε συστάδα υπάρχει η τάση τα γονίδια να είναι από ασθενώς συσχετισμένα ως ασυσχέτιστα μεταξύ τους. Ωστόσο, τα συμπεράσματα αλλάζουν όταν συσχετίζουμε μεταξύ τους γονίδια διαφορετικών συστάδων. Συγκεκριμένα, τα γονίδια της συστάδας με το μπλε χρώμα τείνουν ως επί το πλείστον να έχουν από μέτρια ως ισχυρή αρνητική συσχέτιση με τα γονίδια της συστάδας με το κόκκινο χρώμα, ενώ μόλις τα γονίδια DHCR7, PPIA, ERBB2, Herstatin, RACGAP1, TOP2A, UBERC και BIRC5 φαίνονται να είναι από ασθενώς συσχετισμένα ως ασυσχέτιστα με το γονίδιο ERBB3, όπως επίσης και το γονίδιο

MLPH με τα γονίδια ERBB2, Herstatin και το γονίδιο MUC1 με τα γονίδια PPIA, ERBB2. Τα γονίδια της συστάδας με το πράσινο χρώμα είναι κατά πλειοψηφία ασθενώς συσχετισμένα με τα γονίδια της κόκκινης ομάδας. Μόνο τα γονίδια CXCL12, EGFR και MMP7 φαίνονται να έχουν ισχυρή αρνητική συσχέτιση με τα γονίδια BIRC5, UBERC, TOP2A και RACGAP1, όπως επίσης και τα γονίδια Herstatin, ERBB2, PPIA με το γονίδιο MMP7 και το γονίδιο SPP1 με το CXCL12. Τέλος, τα γονίδια της πράσινης ομάδας έχουν την τάση να είναι από ασθενώς συσχετισμένα ως ασυσχέτιστα με τα γονίδια της μπλε ομάδας. Ωστόσο, διακρίνουμε ένα σύνολο από μέτρια ως ισχυρά αρνητικά συσχετισμένα γονίδια, όπου πρόκειται για τα γονίδια PTGER3, STC2, MUC1, ERBB3, ERBB4, MLPH, ABAT και IL6ST με τα γονίδια EGFR, SFRP1 και MMP7.

Χρησιμοποιούμε την απόσταση του Pearson για να υπολογίσουμε τις αποστάσεις μεταξύ των γονιδίων για κάθε ασθενή και εφαρμόζουμε ιεραρχική ομαδοποίηση, ενώ παράλληλα κατασκευάζουμε θερμικό χάρτη για κάθε μέθοδο (ή μέτρο απόστασης μεταξύ των συστάδων). Τα αποτελέσματα δείχνουν ότι τόσο η centroid όσο και η median linkage παρουσιάζουν αναστροφές (*inversions*) στη σύνδεση των συστάδων, οπότε θεωρούνται ακατάλληλες για την ομαδοποίηση των γονιδίων. Η UPGMA όχι μόνο περιέχει πάρα πολλές αναστροφές, αλλά επιπλέον ο τρόπος που δημιουργεί γραφικά τα γονίδια δεν θυμίζει δενδρόγραμμα σύμφωνα με τον ορισμό που γνωρίζουμε. Επίσης, η single linkage θεωρείται ακατάλληλη διότι δεν μπορούμε να διακρίνουμε τις συστάδες που προκύπτουν, επειδή οι συγχωνεύσεις των συστάδων στα τελευταία στάδια κατασκευής του δενδρογράμματος τείνουν να γίνονται σχεδόν στο ίδιο ύψος απόστασης! Τελικά, η complete linkage, ward linkage και average linkage σχηματίζουν ευδιάκριτες συστάδες (βλέπε Παράρτημα Β1), οπότε μπορούμε να διακρίνουμε το βέλτιστο πλήθος συστάδων, να μελετήσουμε ως προς τα γονίδια που περιέχουν οι βέλτιστες συστάδες τους και να καταλήξουμε στην πιο αντιπροσωπευτική μέθοδο.

Γράφημα 3.1

Θερμικός χάρτης με συσχετίσεις 37 γονιδίων



Ως προς το μέγεθος και το περιεχόμενο των συστάδων καταλήγουμε στις τρεις συστάδες σύμφωνα με εκείνο το σημείο του δενδρογράμματος που παρατηρείται η μεγαλύτερη μεταβολή στην απόσταση όταν μετακινούμαστε στο επόμενο επίπεδο συνένωσης και τον περιορισμό σχηματισμού συστάδων χωρίς έντονη ανισορροπία στο μέγεθός τους και για τις τρεις αποδεκτές μεθόδους ιεραρχικής ομαδοποίησης των γονιδίων. Η complete linkage και η average linkage είναι οι μόνες που περιέχουν συστάδες ίδιες ως προς το περιεχόμενο, άρα και ως προς το μέγεθος. Η ward linkage συμφωνεί ως προς το περιεχόμενο με τις άλλες δύο μεθόδους μόνο στην πρώτη ομάδα. Όσον αφορά τις άλλες δύο συστάδες η ward linkage περιέχει στην δεύτερη συστάδα το γονίδιο PTGER3, το οποίο περιέχεται στην τρίτη συστάδα της complete και average linkage. Αξίζει να σημειώσουμε ότι για βέλτιστο πλήθος τις τρεις συστάδες και οι τρεις συσσωρευτικές αυτές μέθοδοι έχουν συστάδες παρόμοιου μεγέθους και περιεχομένου και αυτό μας ενθαρρύνει να τις συγκρίνουμε για το συγκεκριμένο πλήθος συστάδων.

Χρησιμοποιώντας τα εσωτερικά μέτρα, τα μέτρα σταθερότητας, τα βιολογικά μέτρα και τον συντελεστή συσσώρευσης θα αποφασίσουμε ποια από τις τρεις αυτές μεθόδους θα επιλέξουμε ως την πιο κατάλληλη.

Πίνακας 3.1

	Internal Measures			Stability Measure				
	Conn.	Dunn	Silh.	APN	AD	ADM	FOM	AC
Average	15.488	0.593	0.219	0	0.656	0	2.570	0.544
Complete	15.488	0.593	0.219	0	0.656	0.017	2.570	0.665
Ward	17.921	0.511	0.211	0.067	0.679	3.258	2.563	0.899

Biological Measures		
	BHI	BSI
Average	0.220	0.299
Complete	0.220	0.299
Ward	0.209	0.308

Αρχικά παρατηρούμε ότι τόσο στα εσωτερικά και βιολογικά μέτρα όσο και στα μέτρα σταθερότητας η average έχει ίδιες τιμές με την complete linkage, διότι έχουν τις ίδιες συστάδες ως προς το περιεχόμενο. Ωστόσο, διαφέρουν μόνο ως προς τον δείκτη ADM, ο οποίος βασίζεται στο κέντρο βάρους των συστάδων, διότι η απόσταση μεταξύ των συστάδων ορίζεται διαφορετικά σε αυτές τις μεθόδους (η average linkage λαμβάνει υπόψη την μέση απόσταση μεταξύ ενός στοιχείου της μιας συστάδας και ενός στοιχείου της άλλης συστάδας, ενώ η complete linkage λαμβάνει υπόψη την μεγαλύτερη απόσταση).

Λαμβάνοντας υπόψη τις εσωτερικές μετρήσεις παρατηρούμε ότι σύμφωνα με τον δείκτη συνδεσιμότητας καμία από τις υπό εξέταση συσσωρευτικές μεθόδους δεν οδήγησε σε αξιόλογη ομαδοποίηση, διότι ο συγκεκριμένος δείκτης έχει τιμή πολύ μεγαλύτερη από το μηδέν και αυτό είναι ένδειξη ότι τα γειτονικά γονίδια δεν βρίσκονται στην ίδια συστάδα, οπότε ο διαχωρισμός των γονιδίων στις συστάδες που ορίστηκαν δεν είναι ικανοποιητικός. Εξετάζοντας τον δείκτη Dunn παρατηρούμε ότι η τιμή του είναι μικρότερη από τη μονάδα και στις τρεις συσσωρευτικές μεθόδους, δηλαδή η μικρότερη απόσταση μεταξύ των γονιδίων που δεν βρίσκονται στην ίδια

συστάδα είναι αρκετά μικρότερη σε σχέση με τη μεγαλύτερη απόσταση μεταξύ των γονιδίων που βρίσκονται στην ίδια συστάδα. Οπότε συμπεραίνουμε ότι σε καμία συσσωρευτική μέθοδο δεν ορίστηκαν σωστά οι συστάδες ως προς το πλήθος και τη σύσταση τους λόγω μικρής εκτός-συστάδας απόστασης και μεγάλης εντός-συστάδας απόστασης κάτι που θεωρείται ανεπιθύμητο στην ομαδοποίηση. Τέλος, σύμφωνα με τον δείκτη Silhouette και οι τρεις συσσωρευτικές μέθοδοι έχουν όχι μόνο τιμή αρκετά μεκρότερη από το 0.51, που θεωρείται η ελάχιστη ικανοποιητική τιμή για τον δείκτη αυτό σύμφωνα με τον πίνακα 2.6.2 της ενότητας 2.6.1, αλλά επίσης τιμή μικρότερη από το 0.25 με αποτέλεσμα να θεωρήσουμε ότι δεν υπάρχει καμία ουσιώδης δομή στις συστάδες αυτές λόγω λάθος κατανομής των περισσοτέρων γονιδίων στις συστάδες που ανήκουν. Αυτό σημαίνει ότι τα περισσότερα γονίδια τοποθετήθηκαν λανθασμένα στην συστάδα που ανήκουν, διότι η γειτονική τους συστάδα βρίσκεται πιο κοντά σε αυτά σε σχέση με την συστάδα που ανήκουν, οπότε έπρεπε να είχαν τοποθετηθεί στη γειτονική τους συστάδα. Να σημειώσουμε ότι οι εσωτερικοί δείκτες είναι ευαίσθητοι σε περίπλοκες ομαδοποιήσεις. Στην περίπτωσή μας τα αποτελέσματα θεωρούνται αξιόπιστα, διότι καμία ομαδοποίηση δεν είναι περίπλοκη.

Εξετάζουμε τα μέτρα σταθερότητας των συστάδων. Σύμφωνα με τον δείκτη APN παρατηρούμε ότι και στις τρεις συσσωρευτικές μεθόδους που μελετάμε κανένα γονίδιο δεν θα μετακινηθεί από τη συστάδα που βρίσκεται σύμφωνα με την ομαδοποίηση όλου του σετ δεδομένων, ώστε να τοποθετηθεί σε κάποια άλλη συστάδα σύμφωνα με την ομαδοποίηση του σετ δεδομένου όταν αφαιρέσουμε τυχαία κάποια ασθενή. Δηλαδή, το πλήθος των κοινών γονιδίων που περιέχονται στις συστάδες $C^{i,0}$ και $C^{i,l}$ με $i=1,2,...,37$ και $l=1,2,...,316$ είναι ίδιο με το πλήθος των γονιδίων που περιέχονται στη συστάδα $C^{i,0}$ με αποτέλεσμα ο λόγος

$$\frac{n\left(C^{i,0}\cap C^{i,l}\right)}{n\left(C^{i,0}\right)}$$

να ισούται με μονάδα και τελικά ο δείκτης APN να ισούται με μηδέν. Αυτό συνεπάγεται υψηλή σταθερότητα των συστάδων. Ανάλογο συμπέρασμα περί μεγάλης σταθερότητας των συστάδων προκύπτει και από τους δείκτες AD και ADM. Μεταξύ των τριών συσσωρευτικών μεθόδων η average linkage και complete linkage φαίνονται να έχουν την υψηλότερη σταθερότητα στις συστάδες με αμελητέα διαφορά από τις άλλες δύο μεθόδους σύμφωνα με τον δείκτη AD, ενώ η μόνη μέθοδος που έχει δείκτη ADM με μηδενική τιμή όταν λαμβάνουμε υπόψη το κέντρο βάρους κάθε

συστάδας είναι η average linkage. Τέλος, σύμφωνα με τον δείκτη FOM η ward linkage έχει την μικρότερη εντός-συστάδας απόσταση, οπότε έχει και την μεγαλύτερη προβλεπτική δύναμη στο να δημιουργεί ικανοποιητικές συστάδες ως προς το πλήθος και τη σύστασή τους σε σχέση με τις άλλες δύο συσσωρευτικές μεθόδους, οι οποίες δεν διαφέρουν σημαντικά μεταξύ τους. Ωστόσο, γνωρίζουμε ήδη από την ενότητα 2.6.2.Δ (βλέπε παρατήρηση 4) ότι για να είναι αξιόπιστο το αποτέλεσμα του δείκτη FOM θα πρέπει οι αλγόριθμοι ομαδοποίησης που συγκρίνουμε να μην δίνουν συστάδες με ιδιαίτερα διαφορετικό πλήθος στοιχείων ούτε να έχουν υπολογιστεί με διαφορετικό μέτρο απόστασης. Στην συγκεκριμένη περίπτωση ικανοποιούνται και οι δύο περιορισμοί για την ward linkage, καθώς οι συστάδες έχουν μικρή ανισορροπία και το μέτρο απόστασης που εφαρμόστηκε είναι η απόσταση Pearson. Επιπλέον, μην ξεχνάμε ότι οι δείκτες σταθερότητας είναι πιο αξιόπιστοι όταν οι μεταβλητές που ομαδοποιούμε είναι υψηλά συσχετισμένες. Σύμφωνα με τον θερμικό χάρτη των συσχετίσεων Pearson τα ασθενώς συσχετισμένα γονίδια υπερέχουν σημαντικά από τα υψηλά συσχετισμένα (έντονο κόκκινο ή πράσινο) με αποτέλεσμα να μην μπορούμε να είμαστε απόλυτα σίγουροι για την αξιοπιστία των αποτελεσμάτων των δεικτών σταθερότητας.

Σύμφωνα με τον συντελεστή συσσώρευσης, AC, τη μεγαλύτερη ισχύ στην ομαδοποίηση φαίνεται να έχει η ward linkage με διαφορά από τις άλλες μεθόδους. Λόγω της μικρής ανισορροπίας στο μέγεθος των συστάδων το αποτέλεσμα αυτό είναι αξιόπιστο, διότι όπως έχουμε ήδη αναφέρει στην ενότητα 2.6, ο συγκεκριμένος συντελεστής παρουσιάζει ευαισθησία όταν εφαρμόζεται σε συστάδες διαφορετικού μεγέθους.

Τέλος, λαμβάνοντας υπόψη τους βιολογικούς δείκτες παρατηρούμε ότι η average και complete linkage έχουν την ίδια τιμή τόσο στον δείκτη βιολογικής ομοιογένειας (*BHI*) όσο και στον δείκτη βιολογικής σταθερότητας (*BSI*), όπως ήταν αναμενόμενο. Και στους δύο δείκτες η τιμή είναι ιδιαίτερα χαμηλή. Σύμφωνα με τον δείκτη *BHI* συμπεραίνουμε ότι το ποσοστό γονιδίων που βρίσκονται στην ίδια συστάδα και έχουν παρόμοια βιολογική λειτουργία είναι ιδιαίτερα χαμηλό (περίπου το 22% των γονιδίων), ενώ σύμφωνα με τον δείκτη *BSI* συμπεραίνουμε ότι η αξιοπιστία και των δύο αλγορίθμων ομαδοποίησης που εφαρμόσαμε για γονίδια με παρόμοια βιολογική λειτουργικότητα είναι χαμηλή, διότι οι συστάδες περιέχουν πολύ μικρό ποσοστό γονιδίων παρόμοιας βιολογικής λειτουργικότητας (γύρω στο 30%). Η ward linkage έχει αμελητέα μικρότερη τιμή από τους παραπάνω αλγορίθμους ως προς τον δείκτη

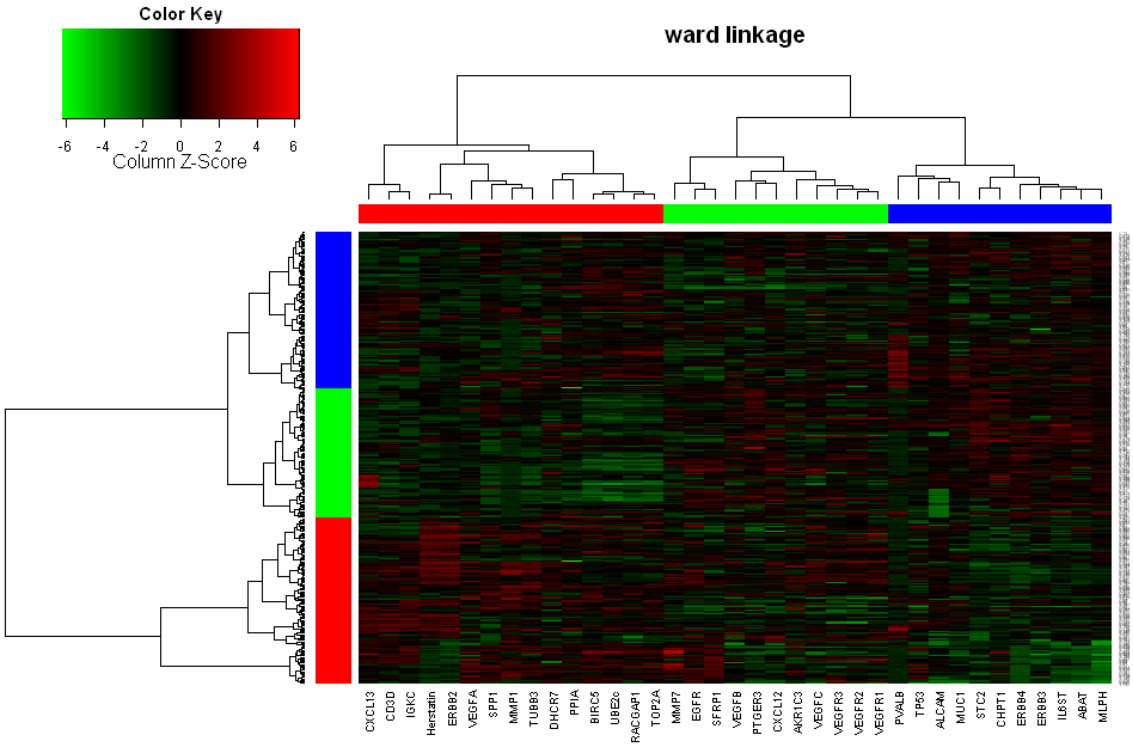
BHI, οπότε καταλαβαίνουμε ότι ούτε η ward linkage δημιούργησε συστάδες με σημαντική ομοιογένεια βιολογικά. Επίσης, η τιμή της ward linkage είναι αμελητέα μεγαλύτερη στον δείκτη *BSI* με αποτέλεσμα να εξάγουμε τα ίδια συμπεράσματα με τους παραπάνω αλγορίθμους ομαδοποίησης ως προς την βιολογική σταθερότητα των συστάδων.

Ανακεφαλαιώνοντας τα αποτελέσματα των παραπάνω δεικτών συμπεραίνουμε ότι και οι τρεις συσσωρευτικές μέθοδοι δημιουργούν συστάδες υψηλής σταθερότητας, αλλά χαμηλής αξιοπιστίας ως προς την κατανομή των γονιδίων που περιέχουν λόγω μικρής εκτός-συστάδας απόστασης και μεγάλης εντός-συστάδας απόστασης. Δηλαδή, ενώ η συντριπτική πλειοψηφία των γονιδίων έπρεπε να βρίσκεται σε γειτονική συστάδα λόγω μεγάλης εντός-συστάδας απόστασης της συστάδας που ανήκουν και μικρής εκτός-συστάδας απόστασης μεταξύ της συστάδας που ανήκουν και της γειτονικής τους συστάδας, τελικά δεν μετακινούνται στη γειτονική ή σε άλλη συστάδα όταν αφαιρούμε τυχαία από το σετ των δεδομένων μία ασθενή κάθε φορά. Ως προς τους δείκτες σταθερότητας η average και complete linkage δημιουργούν συστάδες με την υψηλότερη σταθερότητα και με αμελητέα διαφορά μεταξύ τους, ενώ ως προς τους εσωτερικούς δείκτες και οι τρεις συσσωρευτικές μέθοδοι δημιουργούν συστάδες χαμηλής αξιοπιστίας ως προς την κατανομή των γονιδίων που περιέχουν. Όμως, και στους βιολογικούς δείκτες τα συμπεράσματα είναι απογοητευτικά λόγω των σημαντικά χαμηλών τιμών τους και στις τρεις συσσωρευτικές μεθόδους με αποτέλεσμα οι συστάδες που δημιούργησαν οι αλγόριθμοι ομαδοποίησης να χαρακτηρίζονται από χαμηλή βιολογική ομοιογένεια και σταθερότητα. Η ward linkage υπερέχει αμελητέα στον δείκτη *BSI*, ενώ μειονεκτεί επίσης αμελητέα στον δείκτη *BHI*.

Στην απόφαση για την συσσωρευτική μέθοδο ομαδοποίησης που θα επιλέξουμε ως την πιο αντιπροσωπευτική για το συγκεκριμένο σετ δεδομένων που μελετάμε θα βοηθήσουν οι θερμικοί χάρτες των συσσωρευτικών μεθόδων.

Παρακάτω δίνεται η ιεραρχική ομαδοποίηση 37 γονιδίων και 316 γυναικών με ward linkage σε θερμικό χάρτη

Γράφημα 3.2



Η ελάχιστη υπεροχή της ward linkage έναντι της complete linkage και average linkage επιβεβαιώνεται όχι μόνο από τα μέτρα αξιολόγησης που αναλύσαμε παραπάνω σύμφωνα με τον πίνακα 3.1, αλλά και από τα δενδρογράμματα του θερμικού χάρτη. Συγκεκριμένα, μπορούμε να διακρίνουμε εύκολα τις τρεις συστάδες ασθενών και τις τρεις συστάδες γονιδίων μικρής ανισορροπίας ως προς το μέγεθός τους, οπότε μπορούμε να ερμηνεύσουμε σωστά αυτόν τον θερμικό χάρτη. Μεγάλο ενδιαφέρον παρουσιάζουν οι μεγάλες αποστάσεις μεταξύ των συστάδων των γονιδίων και των συστάδων των ασθενών, όπως προκύπτουν στα δύο τελευταία στάδια της συγχώνευσης των συστάδων με αποτέλεσμα να υποθέσουμε ότι η ανομοιογένεια μεταξύ των συστάδων είναι μεγάλη και αυτό θεωρείται επιθυμητό για να μπορούμε να χαρακτηρίσουμε μία ομαδοποίηση κατάλληλη προς μελέτη. Αξίζει να υπενθυμίσουμε ότι η ward linkage υπερισχύει των άλλων συσσωρευτικών μεθόδων, διότι οδηγεί στη δημιουργία βέλτιστου πλήθους συστάδων με τάση ισορροπίας ως προς το μέγεθός τους (Jain and Dubes, 1988).

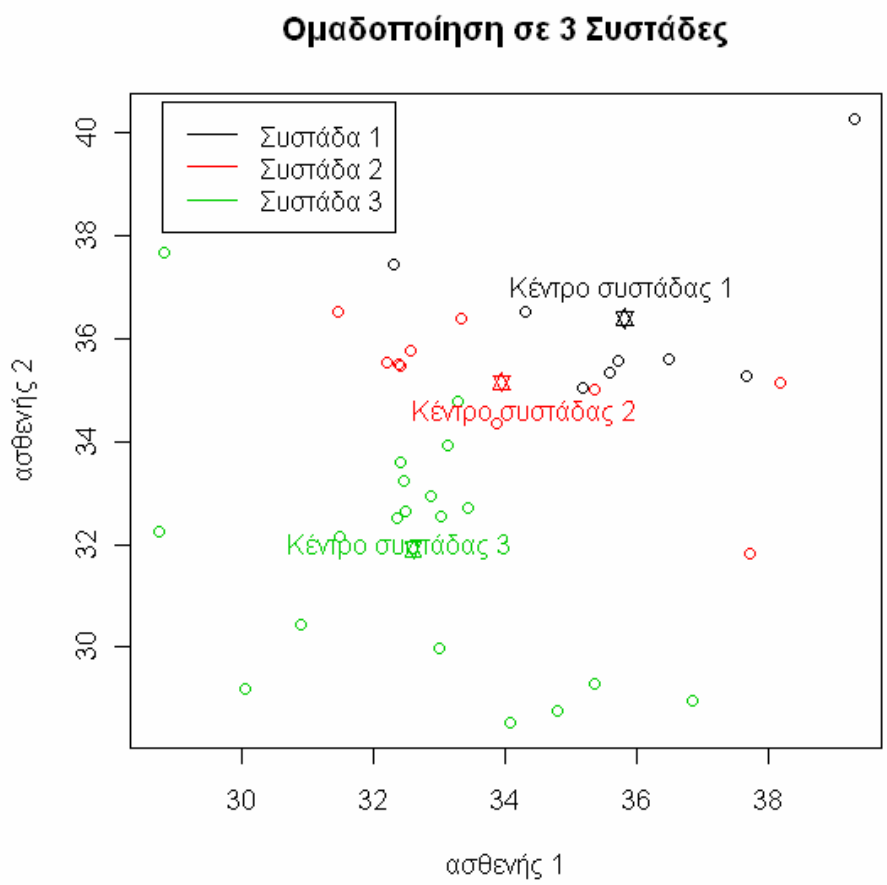
Σύμφωνα με το γράφημα 3.2 κατά μήκος κάθε ασθενούς βλέπουμε πως εκφράζεται κάθε γονίδιο σύμφωνα με το μέσο επίπεδο έκφρασής του. Στις ασθενείς της συστάδας με το κόκκινο χρώμα η πλειοψηφία των γονιδίων της συστάδας με το κόκκινο χρώμα τείνει να υπερεκφράζεται και της συστάδας με το μπλε χρώμα τείνει να υποκφράζεται αντίστοιχα, ενώ συμβαίνει το ακριβώς αντίθετο με τα γονίδια αυτών των συστάδων για τις ασθενείς της συστάδας με το πράσινο χρώμα. Επιπλέον, στις ασθενείς της συστάδας με το πράσινο χρώμα τείνουν να υπερεκφράζονται και τα περισσότερα από τα γονίδια της πράσινης συστάδας. Αντίθετα, στις ασθενείς της συστάδας με το μπλε χρώμα δεν υπάρχει η τάση τα γονίδια να υπερεκφράζονται ή να υποεκφράζονται σε κάποια από τις συστάδες, αλλά φαίνεται σαν να υπάρχει ομοιομορφία υποέκφρασης και υπερέκφρασης σε κάθε συστάδα γονιδίων.

3.3 Μη ιεραρχικές μέθοδοι ομαδοποίησης

3.3.1 k -means ομαδοποίηση

Στη συνέχεια χρησιμοποιούμε το πλήθος συστάδων που βρήκαμε στην ιεραρχική ομαδοποίηση των γονιδίων με average linkage για να πραγματοποιήσουμε k -means ομαδοποίηση των γονιδίων αυτών. Από τη στιγμή που δεν έχουμε εκ των προτέρων γνώση για το πλήθος και το περιεχόμενο των συστάδων με γονίδια επιβάλλεται να ξεκινήσουμε την ομαδοποίηση με κάποια ιεραρχική μέθοδο, ώστε να βρούμε το πλήθος των συστάδων στις οποίες διαιρούνται τα γονίδια και έπειτα να εφαρμόσουμε k -means ομαδοποίηση βασισμένοι στο πλήθος συστάδων που βρήκαμε. Παρακάτω παρουσιάζονται γραφικά τα ομαδοποιημένα γονίδια σε τρεις συστάδες μαζί με το κέντρο βάρους τους. Τα χρώματα μάς διευκολύνουν να ξεχωρίζουμε τις συστάδες και το σύμβολο «αστεράκι» αντιπροσωπεύει το κέντρο βάρους, δηλαδή το μέσο επίπεδο έκφρασης των γονιδίων κάθε συστάδας.

Γράφημα 3.3



Παρατηρούμε ότι οι τρεις συστάδες γονιδίων ξεχωρίζουν μεταξύ τους. Επιπλέον, οι γειτονικές συστάδες βρίσκονται πολύ κοντά, οπότε έχουν μικρή εκτός-συστάδας απόσταση. Να σημειώσουμε ότι και στις τρεις συστάδες λίγα γονίδια βρίσκονται γύρω από το κέντρο βάρους της συστάδας τους, καθώς τα περισσότερα είναι απομακρυσμένα από αυτό σαν ακραίες τιμές και αυτό συνεπάγεται μεγάλη εντός-συστάδας απόσταση. Γονίδια που βρίσκονται πολύ κοντά στο κέντρο βάρους της συστάδας τους έχουν παρόμοιο επίπεδο έκφρασης (ανενεργά γονίδια) με το μέσο επίπεδο έκφρασης σε αυτή τη συστάδα , ενώ γονίδια απομακρυσμένα από το κέντρο βάρους της συστάδας τους έχουν μεγαλύτερο (overexpressed) ή μικρότερο (underexpressed) επίπεδο έκφρασης σε σχέση με το μέσο επίπεδο έκφρασης σε αυτή τη συστάδα αναλόγως αν βρίσκονται δεξιά ή αριστερά του κέντρου βάρους τους αντίστοιχα.

Μπορούμε να παρουσιάσουμε σε πίνακα τα γονίδια που περιέχει κάθε συστάδα, σύμφωνα με το γράφημα 3.3, όπως παρακάτω:

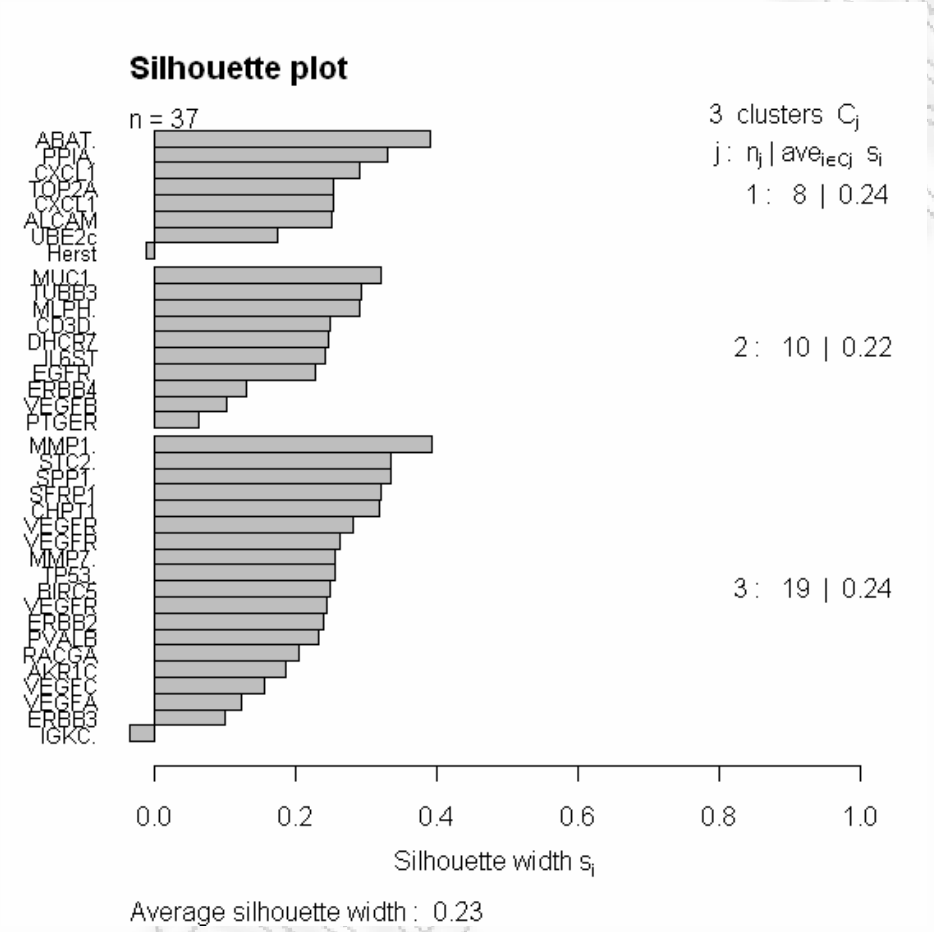
Πίνακας 3.2

Ομαδοποίηση 37 γονιδίων χρησιμοποιώντας K-means		
Συστάδα 1	**Συστάδα 2**	**Συστάδα 3**
PPIA MLPH ERBB2 MUC1 CXCL12 IL6ST VEGFB CHPT1	UBE2c TP53 TOP2A RACGAP1 ALCAM DHCR7 ERBB3 SFRP1 STC2 VEGFA	MMP1 IGKC CXCL13 TUBB3 SPP1 CD3D MMP7 ABAT AKR1C3 BIRC5 EGFR ERBB4 PTGER3 PVALB Herstatin VEGFC VEGFR1 VEGFR2 VEGFR3

Η πρώτη και η δεύτερη συστάδα απέχουν ελάχιστα ως προς το πλήθος των γονιδίων που περιέχουν. Ωστόσο και οι δύο συστάδες απέχουν σημαντικά από την τρίτη συστάδα, η οποία περιέχει 19 γονίδια. Λαμβάνοντας υπόψη την εντός-συστάδας απόσταση (*intra-cluster distance*) παρατηρούμε ότι η πρώτη και η δεύτερη συστάδα δεν διαφέρουν σημαντικά, διότι έχουν εντός-συστάδας απόσταση 7898,57 και 7955,79 αντίστοιχα, ενώ διαφέρουν σημαντικά από την τρίτη που έχει εντός-συστάδας απόσταση 18838,73, δηλαδή σχεδόν τη διπλάσια. Τόσο η μεγάλη εντός-συστάδας απόσταση των δύο «παρόμοιων» συστάδων όσο και η εξαιρετικά μεγάλη εντός-συστάδας απόσταση της τρίτης συστάδας μας δημιουργούν υποψίες ότι τα γονίδια δεν πρέπει να έχουν κατανεμηθεί σωστά. Για να εξετάσουμε αν πράγματι τα

γονίδια δεν έχουν κατανεμηθεί σωστά στις συστάδες που ανήκουν, θα χρησιμοποιήσουμε το γράφημα των silhouettes.

Γράφημα 3.4



Πίνακας 3.3

Συστάδα	1	2	3
Μέγεθος	8	10	19
Average silhouette width	0.2418807	0.2170527	0.2350883

Πίνακας 3.4

Περιγραφικά Στατιστικά για τα Silhouette widths					
Minimum	**Maximum**	**Mean**	**Median**	**1st Quartile**	**3rd Quartile**
-0.0336	0.39290	0.23170	0.24960	0.18670	0.29130

Αρχικά λαμβάνουμε υπόψη τα περιγραφικά στατιστικά των silhouette widths. Συγκεκριμένα, ο μέσος όρος των silhouette widths είναι 0.23170, δηλαδή κατά μέσο όρο δεν υπάρχει καμία ουσιώδης δομή των συστάδων αυτών. Δηλαδή, κατά μέσο όρο τοποθετήσαμε λάθος τα γονίδια στην συστάδα που ανήκουν, ενώ θα έπρεπε να τοποθετηθούν στην γειτονική τους συστάδα. Το μικρότερο silhouette width είναι -0.0336, που είναι μεν αρνητική τιμή, αλλά πολύ κοντά στο μηδέν, ενώ το μεγαλύτερο silhouette width είναι 0.39290, δηλαδή πιθανώς να μην τοποθετήσαμε σωστά το συγκεκριμένο γονίδιο στη συστάδα που ανήκει. Το silhouette plot παριστάνει γραφικά το silhouette width κάθε γονιδίου στη συστάδα που ανήκει. Παρατηρούμε ότι το γονίδιο IGKC έχει το μικρότερο silhouette width, ενώ το γονίδιο ABAT έχει το μεγαλύτερο. Λόγω του αρνητικού silhouette width που αντιστοιχεί στο IGKC συμπεραίνουμε ότι δεν έπρεπε να βρίσκεται στη τρίτη συστάδα. Ο μέσος όρος των silhouette widths κάθε συστάδας μας πληροφορεί, επίσης, για την καταλληλότητα της ομαδοποίησης των γονιδίων σε τρεις συστάδες με την k-means μέθοδο. Στην πρώτη και τρίτη συστάδα ο μέσος όρος των silhouette widths είναι $\overline{s}_1(3) = \overline{s}_3(3) = 0.24$, δηλαδή η δομή αυτών των συστάδων είναι οριακά αδύναμη, ενώ στη δεύτερη συστάδα ο μέσος όρος των silhouette widths είναι $\overline{s}_1(3) = 0.22$, δηλαδή δεν υπάρχει καμία ουσιώδης δομή σε αυτή τη συστάδα. Τέλος, ο συνολικός μέσος όρος των silhouette widths είναι ένας δείκτης που αξιολογεί γενικά την καταλληλότητα αυτής της ομαδοποίησης. Επειδή, ολικός μέσος όρος των silhouette widths είναι

$$\overline{s}(3) = \frac{1}{3}\sum_{j=1}^{3}\overline{s}_j(3) = \frac{1}{37}\sum_{i=1}^{37}s_i = 0.25$$

συμπεραίνουμε ότι η δομή των συστάδων είναι αδύναμη για $k=3$, οπότε η συγκεκριμένη ομαδοποίηση δεν θεωρείται κατάλληλη. Να σημειώσουμε ότι η ομαδοποίηση θα ήταν ικανοποιητική αν ο ολικός μέσος όρος των silhouette widths ήταν τουλάχιστον 0.51. Ωστόσο, μπορούμε να βρούμε εκείνο το πλήθος συστάδων που θεωρείται ιδανικό χρησιμοποιώντας τον δείκτη $SC = \max \overline{s}(k)$. Συγκεκριμένα,

πραγματοποιούμε k-means ομαδοποίηση για διάφορα $k \geq 2$ και μελετάμε τον αντίστοιχο ολικό μέσο όρο των silhouette widths.

Πίνακας 3.5

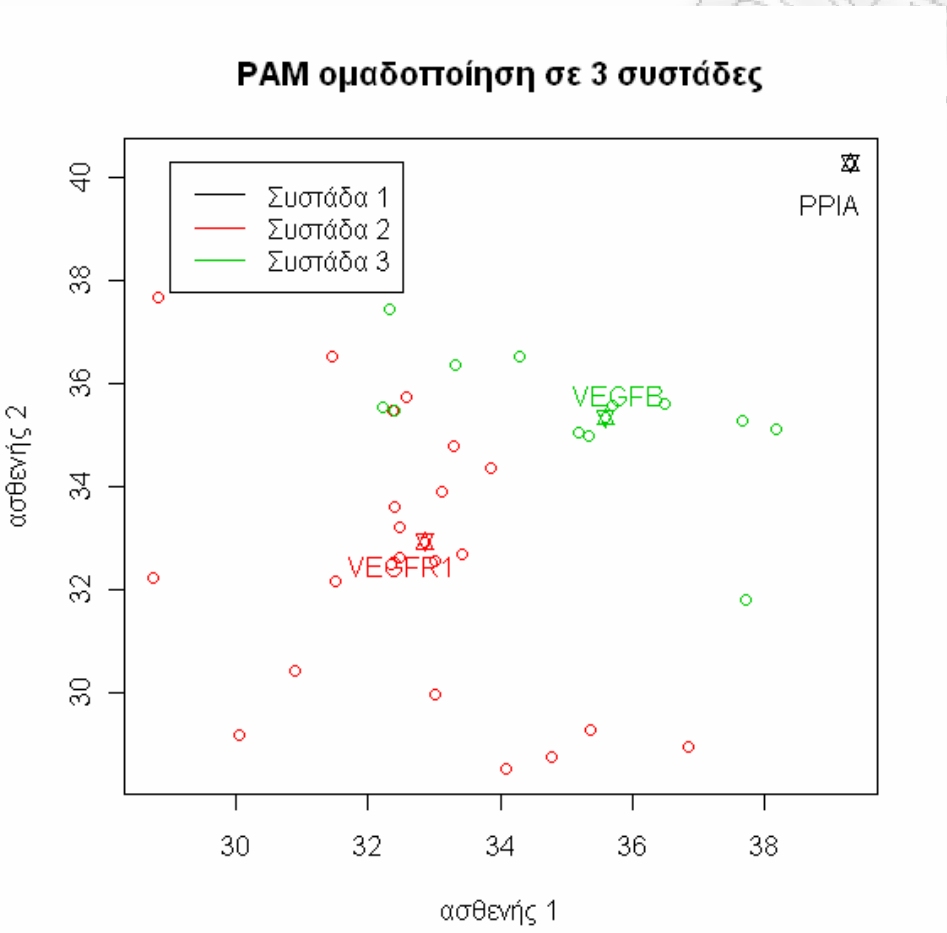
k – **means ομαδοποίηση**	
k	$\overline{s}(k)$
2	0.39
3	0.23
4	0.22
5	0.18
6	0.19

Σύμφωνα με τον παραπάνω πίνακα παρατηρούμε ότι $SC = \max \overline{s}(2) = 0.39$, όπου για $k = 2$ η δομή των συστάδων θεωρείται αδύναμη. Λαμβάνοντας υπόψη τη μεγάλη εντός-συστάδας απόσταση, το πολύ μικρό μέσο silhouette width κάθε συστάδας, το γράφημα των silhouette widths για κάθε k και τέλος τον δείκτη SC συμπεραίνουμε ότι η ομαδοποίηση αυτών των γονιδίων με την k – means μέθοδο δεν θεωρείται αξιόλογη.

3.3.2 PAM ομαδοποίηση

Παρακάτω ομαδοποιούμε τα γονίδια με την μέθοδο PAM για $k=3$ συστάδες παρουσιάζουμε γραφικά τις ομάδας μαζί με το αντίστοιχο medoid.

Γράφημα 3.5



Σύμφωνα με το γράφημα και οι τρεις συστάδες γονιδίων ξεχωρίζουν μεταξύ τους. Ωστόσο η συστάδα 1 περιέχει μόνο ένα γονίδιο, που είναι το ίδιο το medoid της συστάδας αυτής και βρίσκεται αρκετά απομακρυσμένο από τη γειτονική του συστάδα. Για τις υπόλοιπες συστάδες να σημειώσουμε ότι υπάρχουν λίγα γονίδια που βρίσκονται κοντά στο medoid της συστάδας, καθώς τα περισσότερα γονίδια βρίσκονται αρκετά απομακρυσμένα από αυτό και θυμίζουν ακραίες τιμές. Τα τελευταία γονίδια ευθύνονται για την μεγάλη τιμή στην εντός-συστάδας απόσταση.

Τα γονίδια που περιέχονται σε κάθε συστάδα δίνονται στον επόμενο πίνακα, όπου με πλάγια και έντονη γραφή έχουμε δηλώσει το medoid κάθε συστάδας:

Πίνακας 3.6

Ομαδοποίηση 37 γονιδίων χρησιμοποιώντας PAM			
Συστάδα 1	**Συστάδα 2**		**Συστάδα 3**
PPIA	IGKC TOP2A CXCL13 CD3D AKR1C3 DHCR7 ERBB4 PVALB VEGFC VEGFR3 UBE2c RACGAP1	MMP7 BIRC5 EGFR Herstatin ***VEGFR1*** MMP1 TUBB3 SPP1 ABAT PTGER3 VEGFR2	MUC1 VEGFA TP53 ERBB2 ALCAM IL6ST SFRP1 MLPH CHPT1 CXCL12 ERBB3 STC2 ***VEGFB***

Στον παρακάτω πίνακα παρουσιάζονται χρήσιμες πληροφορίες για κάθε συστάδα:

Πίνακας 3.7

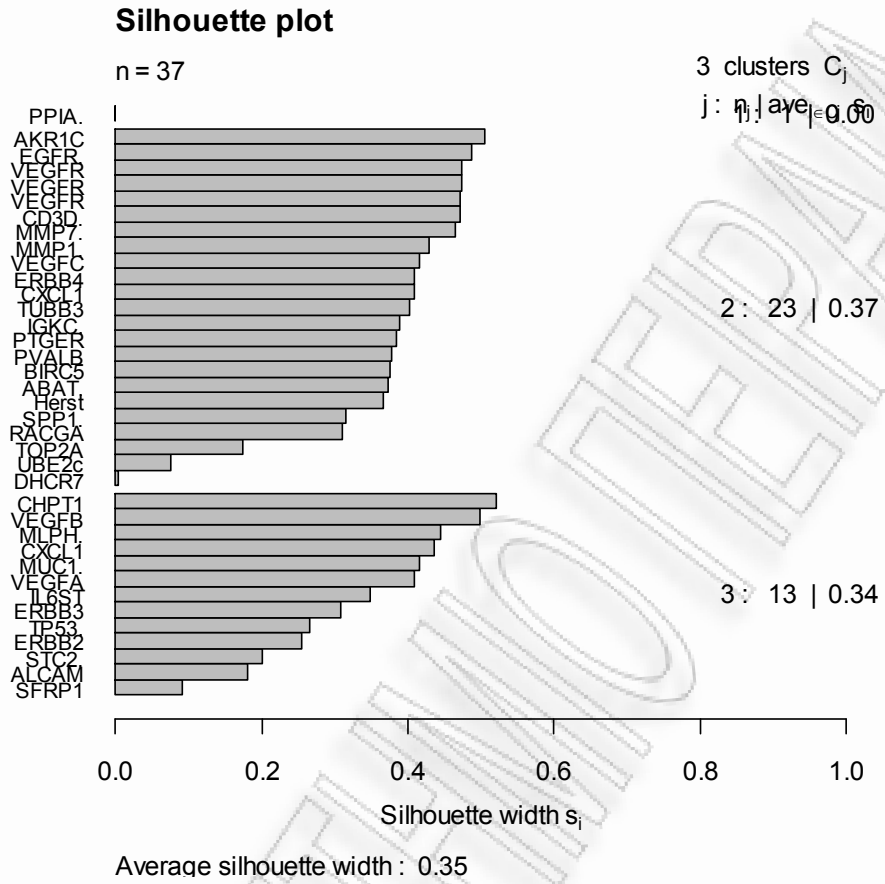
Ομαδοποίηση με PAM					
Συστάδα	**Μέγεθος**	**Μέγιστη απόσταση**	**Μέση απόσταση**	**Διάμετρος**	**Διαχωρισμός**
1	1	0	0	0	30.231
2	23	26.473	15.325	35.828	13.479
3	13	23.717	14.914	37.182	13.479

Ως προς το μέγεθος των συστάδων η κατανομή των γονιδίων δεν είναι ικανοποιητική, καθώς υπάρχει έντονη ανισορροπία. Η πρώτη συστάδα περιέχει μόνο ένα γονίδιο, το οποίο είναι και το medoid αυτής της συστάδας με αποτέλεσμα, τόσο η εντός-συστάδας μέγιστη απόσταση από το medoid, όσο και η εκτός-συστάδας μέση απόσταση από το medoid και η διάμετρος της συστάδας να ισούται με μηδέν. Μεταξύ των δύο τελευταίων συστάδων, η τρίτη έχει τη μικρότερη εντός-συστάδας μέση και μέγιστη απόσταση από το medoid και τη μικρότερη διάμετρο. Αυτό σημαίνει ότι στην συγκεκριμένη συστάδα τα γονίδια βρίσκονται πιο κοντά στο medoid, αλλά και μεταξύ τους σε σχέση με την δεύτερη συστάδα. Τέλος, ο διαχωρισμός έχει την ίδια τιμή στις δύο τελευταίες συστάδες, διότι βρίσκονται πιο κοντά μεταξύ τους σε σχέση με την πρώτη συστάδα σύμφωνα με το παραπάνω γράφημα.

Επιπλέον πληροφορίες που μπορούμε να έχουμε από την ομαδοποίηση με PAM είναι η αντικειμενική συνάρτηση (*Objective Function, OF*) στο build-step και swap-step, αλλά και αν οι συστάδες που προκύπτουν θεωρούνται απομονωμένες (isolated clusters). Συγκεκριμένα, η αντικειμενική συνάρτηση στο build-step είναι $OF\_BS = 14.776$ και στο swap-step είναι $OF\_SS = 14.767$. Επειδή η διαφορά μεταξύ αυτών των δύο ποσοτήτων είναι θετική, καταλαβαίνουμε ότι τα medoids της μεθόδου δεν είναι τα αρχικά medoids που ορίσαμε τυχαία στο build-step, αλλά είναι τα non-medoids που ορίσαμε τυχαία στο swap-step στο τελευταίο βήμα του αλγορίθμου και με αυτά αντικαταστήσαμε τα πρώτα. Επιπλέον, καμία συστάδα δεν θεωρείται απομονωμένη, δηλαδή και για τις τρεις συστάδες δεν ισχύει η σχέση $D^i \le S^i$ ή $D^i \le S^l$, δηλαδή πιο κοντά βρίσκονται μεταξύ τους οι συστάδες παρά τα γονίδια κάθε συστάδας. Αυτό συνεπάγεται μικρή ανομοιογένεια και ομοιογένεια συστάδων, γεγονός που καθιστά τη συγκεκριμένη ομαδοποίηση ανεπιθύμητη.

Τελικά, οι άνισες συστάδες, τα μεγάλα αποτελέσματα σε κάθε παράμετρο σύμφωνα με τον παραπάνω πίνακα και η απουσία απομονωμένων συστάδων δημιουργούν υποψίες ότι η ομαδοποίηση των γονιδίων με τη μέθοδο PAM δεν πρέπει να είναι ικανοποιητική. Για να το εξετάσουμε αυτό θα εφαρμόσουμε πάλι silhouette plot.

Γράφημα 3.6



Πίνακας 3.8

Συστάδα	1	2	3
Μέγεθος	1	23	13
Average silhouette width	0	0.3723	0.3362

Πίνακας 3.9

Περιγραφικά Στατιστικά για τα Silhouette widths					
Minimum	**Maximum**	**Mean**	**Median**	**1st Quartile**	**3rd Quartile**
0.0000	0.5198	0.3495	0.3888	0.3086	0.4447

Σύμφωνα με τα περιγραφικά στατιστικά των silhouette widths ο μέσος όρος των silhouette widths είναι 0.3495, δηλαδή κατά μέσο όρο τα γονίδια δεν πρέπει να τοποθετήθηκαν σωστά στη συστάδα που ανήκουν και ίσως θα έπρεπε να τοποθετηθούν στη γειτονική συστάδα τους. Το μικρότερο silhouette width είναι 0, οπότε δεν είναι ξεκάθαρο αν το γονίδιο αυτό πρέπει να τοποθετηθεί στην συστάδα που ανήκει ή στη γειτονική του που είναι η συστάδα 3 σύμφωνα με το γράφημα 3.5, ενώ το μεγαλύτερο silhouette width είναι 0.5198, δηλαδή πιθανώς τοποθετήσαμε σωστά το συγκεκριμένο γονίδιο στη συστάδα που ανήκει. Το silhouette plot παριστάνει γραφικά το silhouette width κάθε γονιδίου στη συστάδα που ανήκει. Παρατηρούμε ότι το γονίδιο PPIA έχει το μικρότερο silhouette width, ενώ το γονίδιο CHPT1 έχει το μεγαλύτερο. Στην πρώτη συστάδα ο μέσος όρος των silhouette widths είναι $\bar{s}_1(3) = 0$ και συμπεραίνουμε ότι η συγκεκριμένη συστάδα δεν έχει φυσικά καμία ουσιώδη δομή, αφού περιέχει μόνο ένα γονίδιο, ενώ στις δύο τελευταίες συστάδες ο μέσος όρος των silhouette widths είναι $\bar{s}_2(3) = 0.3723$ και $\bar{s}_3(3) = 0.3362$ αντίστοιχα, δηλαδή η δομή αυτών των συστάδων είναι αδύναμη για $k = 3$. Τέλος, ο συνολικός μέσος όρος των silhouette widths είναι

$$\bar{s}(3) = \frac{1}{3}\sum_{j=1}^{3}\bar{s}_j(3) = \frac{1}{37}\sum_{i=1}^{37}s_i = 0.35$$

και συμπεραίνουμε ότι η δομή των συστάδων είναι αδύναμη για $k = 3$, οπότε η συγκεκριμένη ομαδοποίηση δεν θεωρείται κατάλληλη. Ωστόσο, πάλι μπορούμε να βρούμε εκείνο το πλήθος συστάδων που θεωρείται ιδανικό χρησιμοποιώντας τον δείκτη $SC = \max \bar{s}(k)$, οπότε πραγματοποιούμε PAM ομαδοποίηση για διάφορα $k \geq 2$ και μελετάμε τον αντίστοιχο συνολικό μέσο όρο των silhouette widths.

Πίνακας 3.10

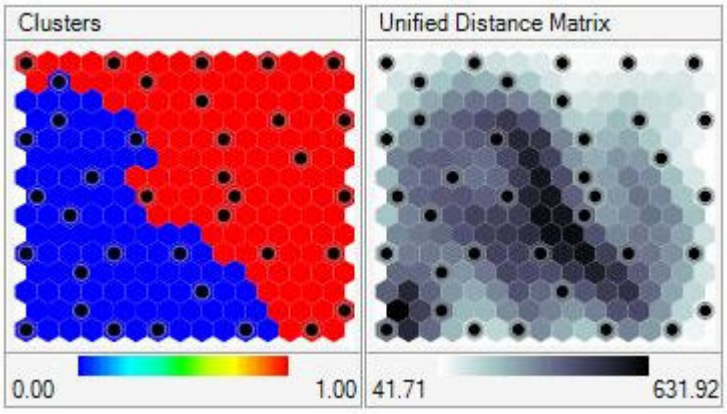
k – **means ομαδοποίηση**	
k	$\bar{s}(k)$
2	0.38
3	0.35
4	0.17
5	0.17
6	0.17

Παρατηρούμε ότι μετά το $k=3$ ο συνολικός μέσος όρος των silhouette widths μειώνεται σημαντικά και καθιστά μη ουσιώδη τη δομή των συστάδων. Μεταξύ των k το μεγαλύτερο δείκτη SC δίνει το $k=2$ (με διαφορά κατά 0.03 από το $k=3$) και συγκεκριμένα $SC=\max \bar{s}(2)=0.38$, όπου η δομή των συστάδων θεωρείται αδύναμη. Οπότε συνολικά η ομαδοποίηση των γονιδίων με τη μέθοδο PAM δεν οδηγεί σε αξιόλογες συστάδες.

3.3.3 SOM ομαδοποίηση

Εφαρμόζουμε SOM ομαδοποίηση χρησιμοποιώντας ένα ορθογώνιο πλέγμα από 225 κόμβους. Να σημειώσουμε ότι το μέγεθος του πλέγματος δεν επηρεάζει το αποτέλεσμα της ομαδοποίησης, αρκεί να είναι τουλάχιστον $5\cdot\sqrt{p}$ σύμφωνα με Vesanto (2000), όπου p είναι το πλήθος των γονιδίων που ομαδοποιούμε. Παρακάτω παρουσιάζονται οι συστάδες που προκύπτουν από τη μέθοδο SOM, όπως επίσης και ο U-matrix σύμφωνα με το πρόγραμμα *Peltarion Synapse 1.3.6*.

Γράφημα 3.7



Οι δύο συστάδες αντιπροσωπεύονται με διαφορετικό χρώμα. Μεγάλο ενδιαφέρον στην ανάλυση της συγκεκριμένης ομαδοποίησης παρουσιάζει ο U-matrix, όπου μπορούμε να διακρίνουμε τα όρια κόμβων που διαιρεί τον πλέγμα σε δύο μέρη, άρα σε δύο συστάδες μικρής ανισορροπίας ως προς το μέγεθός τους. Η μπάρα με τις αποχρώσεις του γκρι κάτω από τον U-matrix ορίζει το μέγεθος των αποστάσεων

μεταξύ των γειτονικών κόμβων, όπου η ελάχιστη απόσταση ισούται με 41.71, ενώ η μέγιστη απόσταση ισούται με 631.92 . Είναι προφανές ότι οι 37 μαύροι κύκλοι μέσα στους αντίστοιχους κόμβους αντιπροσωπεύουν τα γονίδια. Οι «κενοί» κόμβοι περιέχουν μόνο το διάνυσμα αναφοράς που ορίστηκε στο τελικό στάδιο του αλγόριθμου. Σύμφωνα με τις αποχρώσεις του γκρι στον U-matrix παρατηρούμε ότι η συστάδα με το μπλε χρώμα φαίνεται να έχει τη μεγαλύτερη εντός-συστάδας απόσταση, διότι έχει τους περισσότερους κόμβους με γκρι και σκούρες γκρι αποχρώσεις, οπότε στη συγκεκριμένη συστάδα τα γειτονικά γονίδια βρίσκονται απομακρυσμένα μεταξύ τους. Ωστόσο, εκτός από το χρώμα που αντιστοιχεί σε κάθε απόσταση, αν λάβουμε υπόψη και το μέγεθος της απόστασης, τότε παρατηρούμε ότι η τιμή 41.71 που αντιστοιχεί στην μικρότερη απόσταση γειτονικών γονιδίων είναι ουσιαστικά μεγάλη, ενώ η τιμή 631.92 που αντιστοιχεί στη μεγαλύτερη απόσταση γειτονικών γονιδίων είναι σημαντικά μεγάλη. Και οι δύο περιπτώσεις συνεπάγονται την ύπαρξη μεγάλης και σημαντικά μεγάλης εντός-συστάδας απόστασης στην κόκκινη και μπλε συστάδα αντίστοιχα. Λαμβάνοντας υπόψη τις σκούρες γκρι αποχρώσεις των κόμβων, που αποτελούν το όριο των συστάδων αυτών, συμπεραίνουμε ότι οι δύο συστάδες απέχουν σημαντικά μεταξύ τους, δηλαδή η εκτός-συστάδας απόσταση φαίνεται να είναι μεγάλη και αυτό είναι επιθυμητό στην ομαδοποίηση.

Για να θεωρήσουμε ως κατάλληλη την παρούσα ομαδοποίηση θα πρέπει η εντός-συστάδας απόσταση να είναι μικρή και αυτό δεν υφίσταται. Η μεγάλη εκτός-συστάδας απόσταση είναι ένδειξη ότι οι συστάδες διαφέρουν σημαντικά μεταξύ τους, ενώ η μεγάλη εντός-συστάδας απόσταση είναι ένδειξη ότι τα γειτονικά γονίδια δεν βρίσκονται στην ίδια συστάδα, οπότε υπάρχει η υποψία ότι η πλειοψηφία των γονιδίων θα έπρεπε να βρίσκεται στη γειτονική συστάδα.

Στη συνέχεια παρουσιάζονται σε πίνακα τα γονίδια που περιέχει κάθε συστάδα

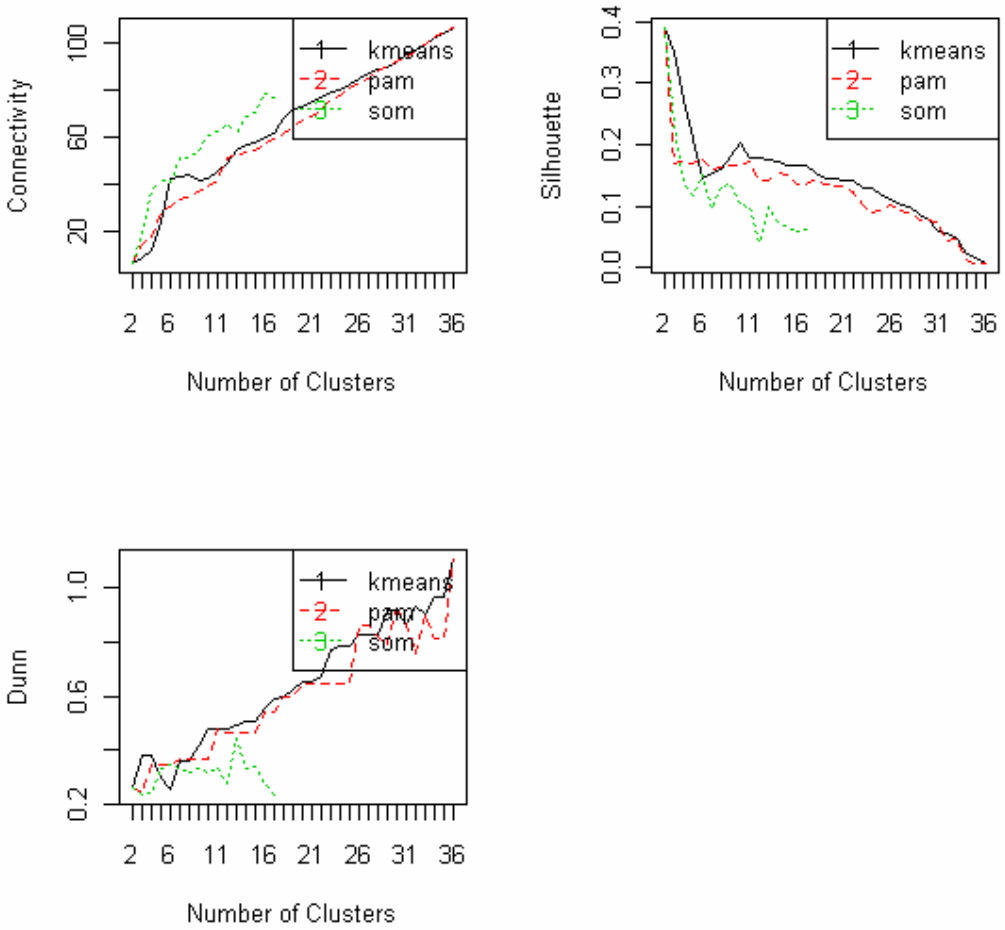
Πίνακας 3.11

Ομαδοποίηση 37 γονιδίων χρησιμοποιώντας SOM				
Συστάδα 1		**Συστάδα 2**		
PPIA UBE2C TP53 MLPH TOP2A RACGAP1 CHPT1 ERBB2 MUC1	ALCAM CXCL12 DHCR7 ERBB3 IL6ST STC2 VEGFA VEGFB	MMP1 IGKC CXCL13 TUBB3 SPP1 CD3D MMP7	ABAT AKR1C3 BIRC5 EGFR ERBB4 PTGER3 PVALB	SFRP1 Herstatin VEGFC VEGFR1 VEGFR2 VEGFR3

3.4 Αξιολόγηση μη ιεραρχικών μεθόδων ομαδοποίησης

Αντί να περιοριστούμε στην αξιολόγηση των μεθόδων ομαδοποίησης $k-$means, PAM και SOM για τρεις συστάδες αυστηρά, μπορούμε να μελετήσουμε τη συμπεριφορά των μέτρων αξιολόγησης για όλα τα δυνατά πλήθη συστάδων που μπορούν να δώσουν οι παραπάνω μέθοδοι. Η απόφαση στην επιλογή της κατάλληλης μη ιεραρχικής μεθόδου θα βασιστεί στα αποτελέσματα των μέτρων αξιολόγησης για τρεις συστάδες, ωστόσο θα έχουμε ήδη μία γενικότερη εικόνα για την ποιότητα της ομαδοποίησης που δίνει κάθε μέθοδος για διάφορα πλήθη συστάδων. Αρχικά θα εξετάσουμε τα εσωτερικά μέτρα σύμφωνα με τα παρακάτω γραφήματα

Γράφημα 3.8



➢ <u>Connectivity index</u>

Ο συγκεκριμένος δείκτης και στις τρεις μεθόδους ομαδοποίησης ξεκινάει από την τιμή 5.862 που αντιστοιχεί στις δύο συστάδες και τείνει να αυξάνεται σχεδόν γραμμικά φτάνοντας στην τιμή 106.372 που αντιστοιχεί στις 36 συστάδες για την k – means και PAM, ενώ φτάνει στην τιμή 85.185 για την SOM που αντιστοιχεί στις 22 συστάδες. Μετά τις δύο συστάδες ο δείκτης αυξάνεται απότομα στην PAM και SOM και παίρνει αρκετά μεγάλη τιμή με αποτέλεσμα η ομαδοποίηση σε τουλάχιστον τρεις συστάδες να μην κρίνεται αξιόλογη γι' αυτές τις μεθόδους, διότι τα γειτονικά γονίδια τείνουν να μην βρίσκονται στην ίδια συστάδα. Τελικά, σύμφωνα με τον δείκτη συνδεσιμότητας ως άριστη επιλογή θεωρείται η ομαδοποίηση των γονιδίων σε δύο συστάδες με την μέθοδο k – means.

Παρόλο που η ομαδοποίηση κρίνεται ιδιαίτερα αξιόλογη όταν αυτός ο δείκτης έχει τιμή κοντά στο μηδέν για κάποιο πλήθος συστάδων (πίνακας 2.6.3 της ενότητας 2.6.1), η τιμή 5.861 είναι αρκετά μικρή και έτσι μπορούμε να θεωρήσουμε επιθυμητή την ομαδοποίηση σε δύο συστάδες με την μέθοδο $k-$means.

➢ Dunn index

Και στις τρεις μεθόδους ο δείκτης Dunn ξεκινάει από την τιμή 0.2643 για δύο συστάδες και αυξάνεται σχεδόν γραμμικά στις μεθόδους $k-$means και PAM φτάνοντας την τιμή 1.108, ενώ στην SOM αυξομειώνεται φτάνοντας την τιμή 0.298 που αντιστοιχεί στις 22 συστάδες και την μέγιστη τιμή 0.442 για 13 συστάδες. Σύμφωνα με το γράφημα η μέθοδος SOM για κανένα πλήθος συστάδων δεν έχει ικανοποιητικό δείκτη Dunn, διότι σύμφωνα με τον πίνακα 2.6.4 της ενότητας 2.6.1 οι συστάδες χαρακτηρίζονται από χαμηλή εκτός-συστάδας απόσταση και υψηλή εντός-συστάδας απόσταση με αποτέλεσμα οι συστάδες να μην ορίζονται σωστά. Σε ανάλογο συμπέρασμα καταλήγουμε και για τις άλλες δύο μεθόδους. Ωστόσο, μετά τις 26 συστάδες, όπου ο δείκτης προσεγγίζει τη μονάδα, για κάποια γονίδια δεν είναι προφανές αν πρέπει να τοποθετηθούν ή όχι στη γειτονική τους συστάδα λόγω του γεγονότος ότι η εντός-συστάδας απόσταση ισούται με την εκτός-συστάδας απόσταση. Όμως, χωρίς αμφιβολία ένα τόσο μεγάλο πλήθος συστάδων είναι ανεπιθύμητο σε μία ομαδοποίηση.

Τελικά, σύμφωνα με τον δείκτη Dunn και οι τρεις μέθοδοι ομαδοποίησης σχηματίζουν συστάδες με μικρή εκτός-συστάδας απόσταση και μεγάλη εντός-συστάδας απόσταση, οπότε οδηγούν σε ακατάλληλη ομαδοποίηση ως προς το πλήθος και τη σύσταση των συστάδων.

➢ Silhouette index

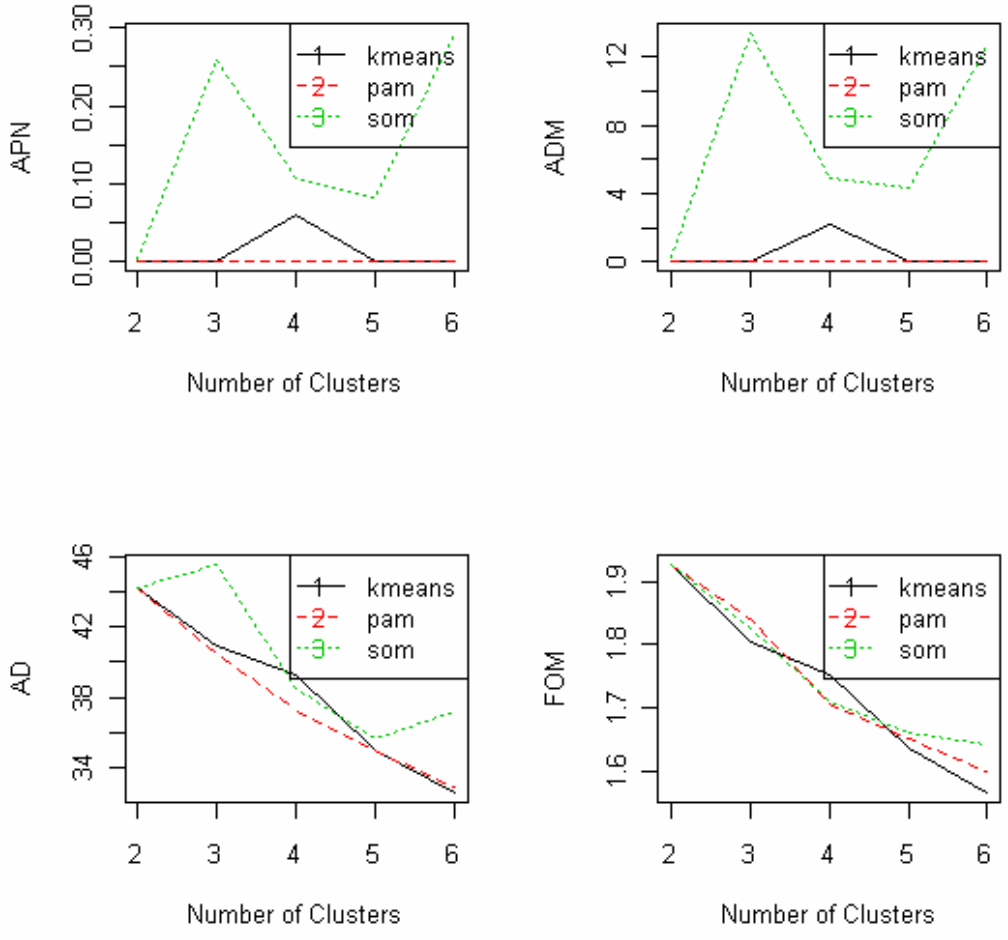
Ο δείκτης αυτός και στις τρεις μεθόδους ομαδοποίησης ξεκινάει από την τιμή 0.390 για δύο συστάδες και μειώνεται φτάνοντας στην τιμή 0.007 που αντιστοιχεί στις 36 συστάδες για την $k-$means και PAM, ενώ φτάνει στην τιμή 0.014 για την SOM που αντιστοιχεί στις 22 συστάδες. Σύμφωνα με το γράφημα παρατηρούμε ότι και στις τρεις μεθόδους ο συγκεκριμένος δείκτης παίρνει τιμές μικρότερες από 0.4 για κάθε πλήθος συστάδων με αποτέλεσμα να θεωρήσουμε ότι η δομή των συστάδων είναι αδύναμη για οποιοδήποτε πλήθος συστάδων (πίνακας 2.6.2 της ενότητας 2.6.1),

διότι τα περισσότερα γονίδια έπρεπε να τοποθετηθούν στη γειτονική τους συστάδα αντί στη συστάδα που ανήκουν.

Τελικά, και ο δείκτης Silhouette δεν κατάφερε να αναδείξει κάποια από τις μεθόδους ομαδοποίησης ως κατάλληλη.

Στη συνέχεια θα μελετήσουμε τα γραφήματα που αντιστοιχούν στους δείκτες σταθερότητας.

Γράφημα 3.9



➢ APN index

Ο συγκεκριμένος δείκτης ξεκινάει από το μηδέν και για τις μεθόδους $k-$means και PAM. Στην $k-$means μέθοδο η τιμή του δείκτη παραμένει σταθερή μέχρι και τις τρεις συστάδες, για να αυξηθεί περίπου στην τιμή 0.05 στις 4 συστάδες και έπειτα να μειωθεί στο μηδέν έως και τις 6 συστάδες, δηλαδή συνολικά για οποιοδήποτε πλήθος

συστάδων το μέσο ποσοστό γονιδίων που δεν βρίσκονται στην ίδια συστάδα είναι σχεδόν μηδέν, όταν ομαδοποιήσουμε τα γονίδια έχοντας αποκλείσει τυχαία κάθε φορά μία ασθενή από το σετ δεδομένων, οπότε συμπεραίνουμε ότι η μέθοδος αυτή έχει μεγάλη προβλεπτική δύναμη στο να δημιουργεί ικανοποιητικές συστάδες ως προς το πλήθος και τη σύστασή τους. Στην μέθοδο PAM η τιμή του δείκτη παραμένει σταθερή στο μηδέν κατά μήκος του πλήθους συστάδων που μελετάμε, οπότε και σε αυτή τη μέθοδο τα συμπεράσματα είναι ανάλογα με αυτά της $k-$means μεθόδου. Ωστόσο, ο δείκτης διαφέρει σημαντικά στην μέθοδο SOM, όπου ξεκινάει μεν από το μηδέν για δύο συστάδες, αλλά αυξάνεται απότομα στην τιμή 0.258 για τρεις συστάδες και μειώνεται σταδιακά για να αυξηθεί πάλι απότομα στην τιμή 0.294 για έξι συστάδες.

Συγκριτικά με τις μεθόδους $k-$means και PAM η μέθοδος SOM δεν θεωρείται ότι δεν έχει ικανοποιητική προβλεπτική δύναμη, διότι ο δείκτης παρουσιάζει ευαισθησία όταν αυξάνεται το πλήθος των συστάδων.

- <u>ADM index</u>

Το γράφημα του δείκτη ADM δεν διαφέρει από το γράφημα του δείκτη APN παρά μόνο ως προς την κλίμακα μέτρησης του δείκτη και για τις τρεις μεθόδους ομαδοποίησης. Συγκεκριμένα, στην $k-$means μέθοδο ο δείκτης ξεκινάει από την τιμή μηδέν για δύο συστάδες και παραμένει σταθερή, για να αυξηθεί στην τιμή 2.224 για τέσσερις συστάδες και να μειωθεί στην τιμή 0.08 για έξι συστάδες, ενώ στην PAM μέθοδο ο δείκτης παραμένει σταθερός στο μηδέν. Είναι προφανές ότι με την μέθοδο $k-$means ως άριστη θεωρείται η ομαδοποίηση σε δύο ή τρεις συστάδες, ενώ με την PAM μέθοδο οποιοδήποτε πλήθος συστάδων και αν επιλέξουμε θα δημιουργήσει συστάδες υψηλής σταθερότητας, αφού η μέση απόσταση μεταξύ των γονιδίων που ανήκουν στην ίδια συστάδα είναι μηδέν, όταν η ομαδοποίηση βασίζεται αρχικά σε όλο το σετ δεδομένων και έπειτα στο σετ δεδομένων που προκύπτει όταν εξαιρούμε τυχαία μία ασθενή κάθε φορά. Ωστόσο, με την μέθοδο SOM ο δείκτης έχει παρόμοια συμπεριφορά με τον δείκτη APN με αποτέλεσμα να μην θεωρείται ικανή να σχηματίσει συστάδες υψηλής σταθερότητας συγκριτικά με τις μεθόδους $k-$means και PAM.

➢ AD index

Σε αντίθεση με τον δείκτη ADM, όταν στον υπολογισμό της μέσης απόστασης μεταξύ των γονιδίων που ανήκουν στην ίδια συστάδα δεν λαμβάνουμε υπόψη το κέντρο βάρους των συστάδων, παρατηρείται όχι μόνο έντονη ευαισθησία καθώς αυξάνεται το πλήθος των συστάδων. Και στις τρεις μεθόδους ο δείκτης ξεκινάει από σημαντικά υψηλή τιμή και μειώνεται σχεδόν γραμμικά για τις μεθόδους $k-$means και PAM φτάνοντας την επίσης υψηλή τιμή 32.564 και 32.891 αντίστοιχα, ενώ στην μέθοδο SOM αυξάνεται στην τιμή 45.568 για τρεις συστάδες για να μειωθεί στη συνέχεια στην τιμή 35.684 για πέντε συστάδες και να αυξηθεί στην τιμή 37.162 για έξι συστάδες.

Συνολικά, σύμφωνα με τον δείκτη AD η απουσία των κέντρων βάρους των συστάδων στον υπολογισμό της μέσης απόστασης καθιστά και τις τρεις μεθόδους ακατάλληλες να δημιουργήσουν σταθερές συστάδες με ικανοποιητικό μέγεθος και σύσταση.

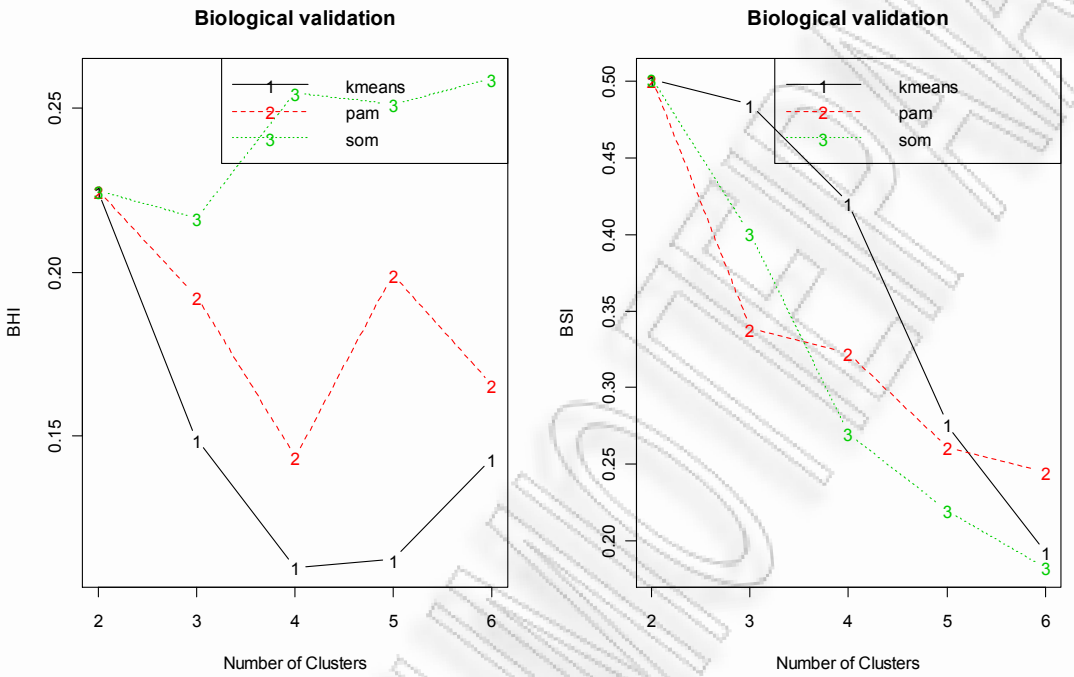
➢ FOM index

Το γράφημα του FOM δείκτη δεν διαφέρει σημαντικά με το γράφημα του δείκτη AD τουλάχιστον στις μεθόδους $k-$means και PAM. Και στις τρεις μεθόδους ο δείκτης αυτός ξεκινάει περίπου από την τιμή 1.925 και μειώνεται σχεδόν γραμμικά στην τιμή 1.564, 1.599 και 1.641 αντίστοιχα για την $k-$means, PAM και SOM μέθοδο. Ουσιαστικά ο δείκτης FOM είναι η εντός-συστάδας τυπική απόκλιση των γονιδίων σύμφωνα με τον αλγόριθμο ομαδοποίησης που εφαρμόσαμε, οπότε παρατηρούμε ότι και στις τρεις μεθόδους ομαδοποίησης καθώς αυξάνεται το πλήθος των συστάδων τα γειτονικά γονίδια τείνουν να παραμένουν στην ίδια συστάδα και επιπλέον η απόστασή τους τείνει να μειώνεται σχεδόν γραμμικά.

Σύμφωνα με τον δείκτη FOM δεν είναι προφανές ποια από τις τρεις μεθόδους ομαδοποίησης θα προτιμήσουμε διότι ο δείκτης δεν διαφέρει, όμως είναι προφανές ότι η επιλογή πέντε ή έξι συστάδων θεωρείται κατάλληλη για την ομαδοποίηση των γονιδίων διότι τότε ο δείκτης έχει τη μικρότερη τιμή.

Τέλος, θα μελετήσουμε τα γραφήματα που αντιστοιχούν στους βιολογικούς δείκτες.

Γράφημα 3.10



➢ <u>BHI index</u>

Και στις τρεις μεθόδους ο δείκτης βρίσκεται στην τιμή 0.225 όταν μελετάμε δύο συστάδες. Στις τρεις συστάδες ο δείκτης υπερέχει στην SOM έναντι των άλλων, αλλά μειώνεται ελάχιστα στην τιμή 0.216. Έπειτα ακολουθεί χωρίς σημαντική διαφορά η PAM όπου ο δείκτης μειώνεται στην τιμή 0.192, ενώ συγκριτικά τη μεγαλύτερη μείωση φαίνεται να παρουσιάζει ο δείκτης για την $k-$means όπου φτάνει στην τιμή 0.148. Μετά τις τρεις συστάδες ο δείκτης συνεχίζει να υπερέχει χαρακτηριστικά στην SOM και φτάνει στην τιμή 0.259 στις έξι συστάδες όπου είναι η μέγιστη τιμή. Συνεπώς, η SOM διακρίνεται για την δημιουργία συστάδων υψηλότερης βιολογικής ομοιογένειας σε σχέση με τις άλλες δύο μεθόδους.

➢ <u>BSI index</u>

Όταν μελετάμε δύο συστάδες και στις τρεις μεθόδους ο δείκτης βρίσκεται αρκετά ψηλά περίπου στην τιμή 0.50 και αυτό αποτελεί ένδειξη ότι στις δύο συστάδες περίπου το 50% των γονιδίων παρουσιάζουν βιολογική σταθερότητα και στις τρεις μεθόδους ομαδοποίησης. Στην συνέχεια ο δείκτης μειώνεται και στις τρεις μεθόδους

φτάνοντας σε ιδιαίτερα χαμηλές τιμές στις έξι συστάδες, Μόνο στην $k-$means ο δείκτης κυμαίνεται συγκριτικά ικανοποιητικά μεταξύ 0.4 και 0.5 μέχρι και τις τέσσερις συστάδες. Τέλος, στις έξι συστάδες ο δείκτης σχεδόν συμπίπτει στην ίδια τιμή για την SOM και $k-$means , ενώ υπερέχει στην PAM. Συνεπώς, η $k-$means φαίνεται να υπερέχει έναντι των άλλων μεθόδων στην δημιουργία συστάδων βιολογικής σταθερότητας όταν το πλήθος τους κυμαίνεται από τρεις ως πέντε συστάδες.

Παρατήρηση

Τόσο στους δείκτες σταθερότητας όσο και στους βιολογικούς δείκτες περιοριζόμαστε αναγκαστικά στις έξι ομάδες, διότι ο αλγόριθμος που χρησιμοποιείται για τον υπολογισμό αυτών των δεικτών καθυστερεί πάρα πολύ να δώσει αποτελέσματα για περισσότερες συστάδες.

Μπορούμε να συγκεντρώσουμε σε έναν πίνακα τις άριστες μεθόδους ομαδοποίησης και τα πλήθη συστάδων που κρίνονται κατάλληλα σύμφωνα με τα αποτελέσματα των δεικτών αξιολόγησης των μη ιεραρχικών μεθόδων ομαδοποίησης

Πίνακας 3.12

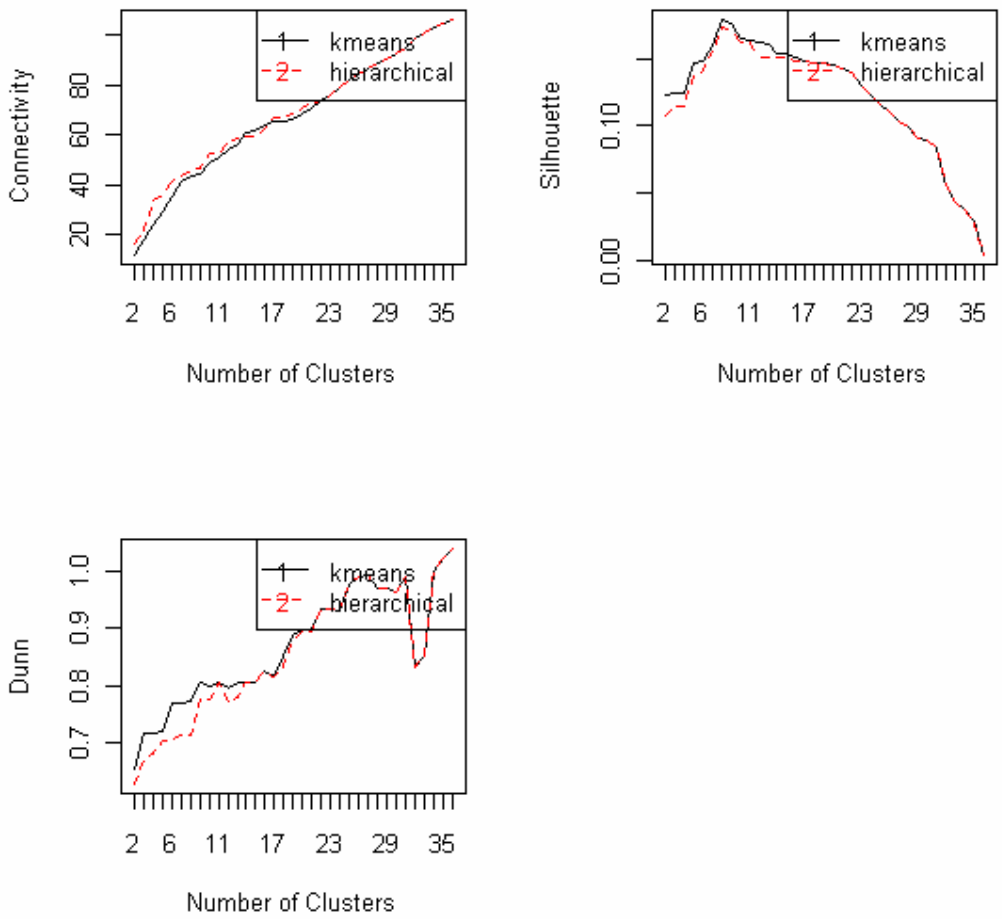
Internal Measures		**Stability Measures**		**Biological Measures**	
Connectivity	$k-$means $k=2$ ή 3	**APN**	PAM ή $k-$means $k=2$ ως 6	**BHI**	SOM $k=6$
Dunn	καμία	**ADM**	PAM ή $k-$means $k=2$ ως 6	**BSI**	$k-$means $k=2$ ή 3
Silhouette	καμία	**AD**	καμία		
		FOM	PAM ή $k-$means ή SOM $k=5$ ή 6		

Μεταξύ των μεθόδων $k-$means, PAM και SOM κατάλληλη για τη δημιουργία τριών συστάδων γονιδίων θεωρείται η $k-$means μέθοδος, διότι οι συστάδες που δημιουργεί διακρίνονται για την δυνατή συνδεσιμότητά τους, δηλαδή περιέχουν όσο το δυνατόν περισσότερα γειτονικά γονίδια, και τη μέτρια ανισορροπία ως προς το μέγεθός τους. Επιπλέον, έχει υψηλή προβλεπτική δύναμη, δηλαδή δημιουργεί σταθερές και ποιοτικές συστάδες και τέλος έχει υψηλότερη βιολογική σταθερότητα και ελάχιστα καλύτερη βιολογική ομοιογένεια συγκριτικά. Αξίζει να σημειώσουμε ότι κοινό χαρακτηριστικό και των τριών μεθόδων ομαδοποίησης είναι η υψηλή εντός-συστάδας απόσταση και η χαμηλή εκτός-συστάδας απόσταση, δηλαδή οι συστάδες έχουν μικρή ομοιογένεια και ανομοιογένεια κάτι που δεν θεωρείται επιθυμητό στην ομαδοποίηση με αποτέλεσμα η βιολογική αξία της ομαδοποίησης και το γράφημα ομαδοποίησης να αποτελούν αναγκαστικά τα τελικά κριτήρια στην αξιολόγηση της.

3.5 Σύγκριση Ward linkage με *k* - means μέθοδο μέσω μέτρων αξιολόγησης

Για να έχει νόημα η συγκεκριμένη σύγκριση θα πρέπει και στους δύο αλγορίθμους ομαδοποίησης να έχει εφαρμοστεί το ίδιο μέτρο απόστασης. Στην ward linkage χρησιμοποιήσαμε την απόσταση του Pearson, όπου συνηθίζεται στις συσσωρευτικές μεθόδους ομαδοποίησης γονιδίων, ενώ στην $k-$means μέθοδο χρησιμοποιήσαμε ευκλείδεια απόσταση, που συνηθίζεται στην μέθοδο αυτή. Θα τυποποιήσουμε τα δεδομένα και θα εφαρμόσουμε ευκλείδεια απόσταση, όπου η ευκλείδεια απόσταση των τυποποιημένων δεδομένων συμπίπτει με την απόσταση Pearson τυποποιημένων ή μη δεδομένων.

Γράφημα 3.11



➢ Connectivity index

Ο δείκτης ξεκινάει από την τιμή 12.143 για την $k-$means μέθοδο και από την τιμή 16.561 για την ward linkage και αυξάνεται σχεδόν γραμμικά φτάνοντας την τιμή 106.372 και για τις δύο μεθόδους. Για τις 16 πρώτες συστάδες η τιμή του δείκτη συνδεσιμότητας για την ward linkage είναι ελάχιστα μεγαλύτερη σε σχέση με την τιμή του δείκτη για την $k-$means μέθοδο, δηλαδή η $k-$means μέθοδος έχει ελάχιστα καλύτερο διαχωρισμό των γονιδίων στις συστάδες που ανήκουν σε σχέση με την ward linkage. Ωστόσο, ο διαχωρισμός αυτός δεν μπορεί να θεωρηθεί αξιόλογος λόγω των μεγάλων τιμών του δείκτη σε αυτές τις συστάδες. Μετά τις 21 συστάδες ο δείκτης αυτός έχει την ίδια τιμή και για τις δύο μεθόδους για κάθε πλήθος συστάδων, δηλαδή ο διαχωρισμός των γονιδίων σε αυτές τις συστάδες είναι κοινός

και για τις δύο μεθόδους. Όμως, αυτός ο διαχωρισμός πάλι δεν μπορεί να θεωρηθεί αξιόλογος λόγω των σημαντικά μεγάλων τιμών του δείκτη.

Τελικά, παρόλο που η $k-$means μέθοδος έχει μικρότερο δείκτη συνδεσιμότητας από την ward linkage για τις πρώτες 16 συστάδες, δεν χαρακτηρίζεται από ικανοποιητική συνδεσιμότητα των γειτονικών γονιδίων.

➢ Dunn index

Σε αντίθεση με τον δείκτη συνδεσιμότητας, ο δείκτης Dunn έχει μεγαλύτερες τιμές στην $k-$means μέθοδο για τις πρώτες 13 συστάδες αλλά και για 18 ως 21 συστάδες. Δηλαδή σε αυτά τα πλήθη συστάδων η $k-$means μέθοδος ορίζει πιο σωστά τις συστάδες σε σχέση με την ward linkage. Παράλληλα, λαμβάνοντας υπόψη το μέγεθος των τιμών και την συμπεριφορά του δείκτη γραφικά και στις δύο μεθόδους παρατηρούμε ότι ο δείκτης ξεκινάει από την τιμή 0.654 στην $k-$means μέθοδο και από την τιμή 0.627 για την ward linkage και αυξάνεται σταδιακά για να μειωθεί στις 32 συστάδες και στη συνέχεια να αυξηθεί πάλι. Μετά τις 25 συστάδες και οι δύο μέθοδοι τείνουν να ορίζουν συστάδες που ικανοποιούν μεγάλη εκτός-συστάδας απόσταση και μικρή εντός-συστάδας απόσταση, όπου η $k-$means μέθοδος υπερέχει ελάχιστα έναντι της ward linkage. Ωστόσο, ένα πλήθος συστάδων 25 και άνω θεωρείται υπερβολικά μεγάλο για να το επιλέξουμε για την ομαδοποίηση των γονιδίων. Μετά τις 27 συστάδες ο δείκτης Dunn είναι κοινός και για τις δύο μεθόδους και προσεγγίζει τη μονάδα, δηλαδή και οι δύο μέθοδοι τείνουν να ορίζουν το ίδιο ικανοποιητικά τις συστάδες σύμφωνα με την εντός- και εκτός-συστάδας απόσταση.

Συνολικά, παρόλο που η $k-$means μέθοδος έχει μεγαλύτερο δείκτη Dunn, δεν καταφέρνει να ορίσει ιδιαίτερα σωστά τις συστάδες, διότι ικανοποιούν σχετικά μικρή εκτός-συστάδας απόσταση και σχετικά μεγάλη εντός-συστάδας απόσταση.

➢ Silhouette index

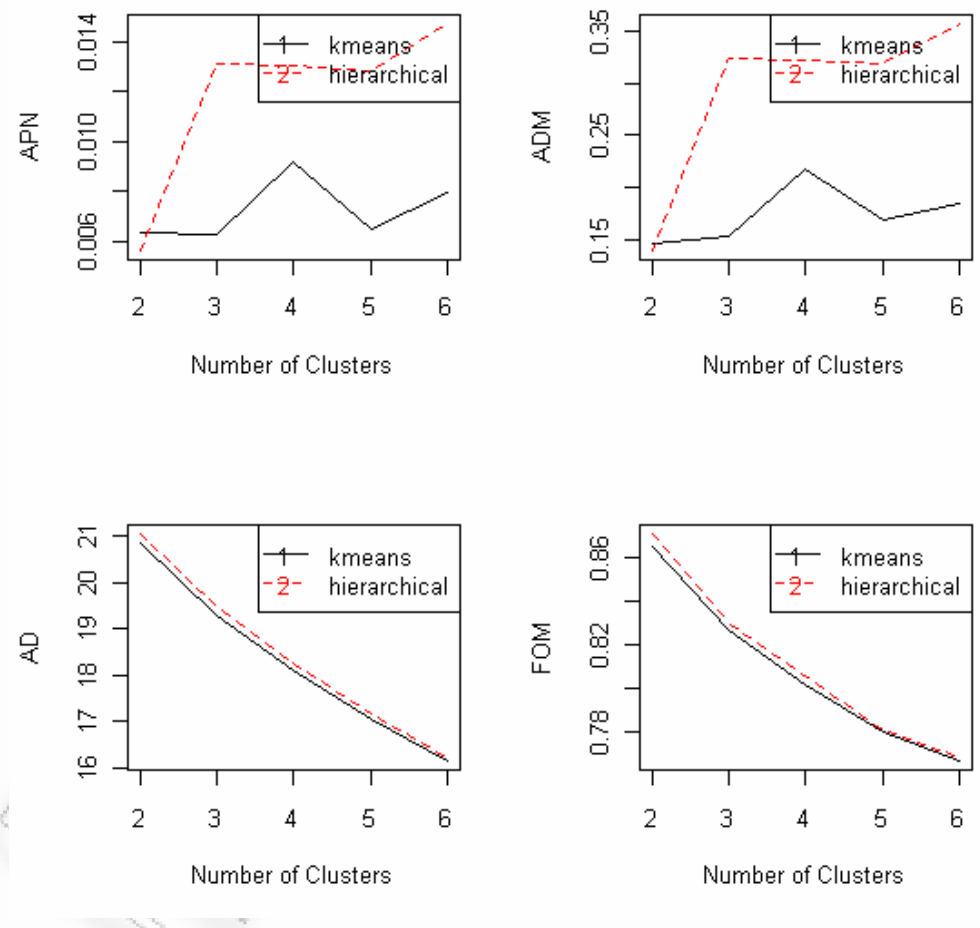
Στην $k-$means μέθοδο ο δείκτης Silhouette ξεκινάει από την τιμή 0.124, ενώ στην ward linkage ο δείκτης αυτός ξεκινάει από την τιμή 0.107, αυξάνεται μέχρι τις 8 συστάδες και στη συνέχεια μειώνεται φτάνοντας την τιμή 0.004 και στις δύο μεθόδους. Η $k-$means μέθοδος υπερέχει ελάχιστα έναντι της ward linkage και σε αυτόν τον δείκτη, δηλαδή στις συστάδες που ορίζει τα γονίδια έχουν κατανεμηθεί καλύτερα. Μετά τις 22 συστάδες ο δείκτης είναι κοινός και στις δύο μεθόδους, οπότε

σε αυτές τις συστάδες, που ορίζουν οι υπό εξέταση μέθοδοι, τα γονίδια κατανέμονται το ίδιο.

Ο δείκτης Silhouette δεν ξεπερνάει την τιμή 0.20 και στις δύο μεθόδους, οπότε καμία από αυτές τις μεθόδους δεν καταφέρνει να ορίσει συστάδες με ουσιώδη δομή για οποιοδήποτε πλήθος συστάδων.

Στην συνέχεια παρουσιάζουμε και αναλύουμε τα γραφήματα που αντιστοιχούν στους δείκτες σταθερότητας

Γράφημα 3.12



➢ APN

Στην $k-$means μέθοδο ο δείκτης ξεκινάει από την τιμή 0.0064 για 2 συστάδες, παραμένει σταθερός για 3 συστάδες, αυξάνεται στην τιμή 0.0092 για 4 συστάδες για να μειωθεί στις 5 συστάδες και τέλος ν' αυξηθεί πάλι στις 6 συστάδες φτάνοντας την τιμή 0.0079, ενώ στην ward linkage ο δείκτης ακολουθεί παρόμοια συμπεριφορά αλλά σε σχετικά μεγαλύτερες τιμές για κάθε πλήθος συστάδων. Παρατηρούμε ότι στην $k-$means μέθοδο ο δείκτης δεν ξεπερνάει την τιμή 0.010, ενώ στην ward linkage ο δείκτης δεν ξεπερνάει την τιμή 0.015.

Και στις δύο μεθόδους οι τιμές του δείκτη APN είναι ιδιαίτερα χαμηλές και προσεγγίζουν σημαντικά το μηδέν με αποτέλεσμα να μην είναι προφανές αν τελικά η $k-$means μέθοδος υπερέχει στην προβλεπτική δύναμη να δημιουργεί συστάδες που διακρίνονται για υψηλή σταθερότητα.

➢ ADM

Το γράφημα του δείκτη ADM διαφέρει από το γράφημα του δείκτη APN μόνο ως προς την κλίμακα. Στην $k-$means μέθοδο ο δείκτης ADM κινείται στο διάστημα τιμών από 0.147 ως 0.218, ενώ στην ward linkage ο δείκτης αυτός κινείται στο διάστημα τιμών από 0.140 ως 0.358. Παρατηρούμε ότι και ο δείκτης ADM έχει μεγαλύτερες τιμές για κάθε πλήθος συστάδων στην ward linkage έναντι της $k-$means μεθόδου. Όμως, σε αντίθεση με τον δείκτη APN που παίρνει τιμές αυστηρά στο διάστημα $[0,1]$, στον συγκεκριμένο δείκτη, όπου επιτρέπεται να έχει τιμές μεγαλύτερες ή ίσες του μηδενός, μπορούμε να διακρίνουμε την υπεροχή της $k-$means μεθόδου να δημιουργεί συστάδες σταθερές και βέλτιστες ως προς το πλήθος και τη σύστασή τους σε σύγκριση με την ward linkage κυρίως για τουλάχιστον τρεις συστάδες.

Συνεπώς, όταν μελετάμε την μέση απόσταση μεταξύ των κέντρων βάρους των συστάδων μπορούμε να ξεχωρίσουμε την $k-$means μέθοδο ως την πιο κατάλληλη.

➢ AD

Ο δείκτης AD ξεκινάει από την τιμή 20.845 στην $k-$means μέθοδο και από την τιμή 21.044 στην ward linkage και μειώνεται σχεδόν γραμμικά και στις δύο μεθόδους φτάνοντας την τιμή 16.141 και 16.244 αντίστοιχα. Παρόλο που ο συγκεκριμένος δείκτης έχει ελάχιστα μεγαλύτερες τιμές στην ward linkage για κάθε πλήθος συστάδων, οι ευθείες που αντιπροσωπεύουν αυτές τις δύο υπό εξέταση μεθόδους ομαδοποίησης τείνουν να συμπέσουν η μία πάνω στην άλλη και αυτό αποτελεί ένδειξη ότι δεν μπορούμε να ξεχωρίσουμε με βεβαιότητα κάποια από τις δύο μεθόδους ως την πιο κατάλληλη, διότι και οι δύο μέθοδοι έχουν σχεδόν την ίδια μέση απόσταση μεταξύ των κέντρων βάρους των συστάδων όταν αφαιρούμε μία ασθενή τυχαία κάθε φορά από το σετ δεδομένων. Ωστόσο, λαμβάνοντας υπόψη το μέγεθος των τιμών του δείκτη και στις δύο μεθόδους παρατηρούμε ότι είναι ιδιαίτερα υψηλές με αποτέλεσμα καμία από τις μεθόδους αυτές να μην κρίνεται αξιόλογη με κριτήριο την σταθερότητα των συστάδων.

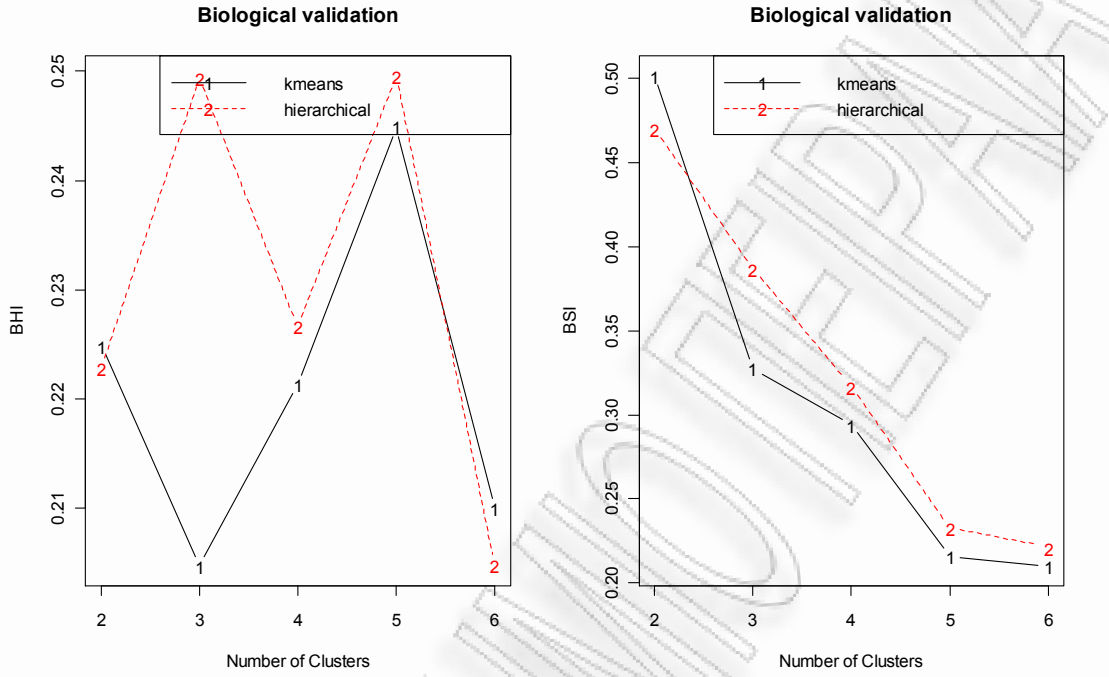
Τελικά, όταν στον υπολογισμό της μέσης απόστασης μεταξύ των συστάδων δεν λαμβάνουμε υπόψη τα κέντρα βάρους των συστάδων παρά μόνο τα γονίδια που περιέχουν οι συστάδες, τότε και οι δύο μέθοδοι κρίνονται ανίκανες να δημιουργήσουν σταθερές συστάδες.

➢ FOM

Το γράφημα αυτού του δείκτη διαφέρει από το γράφημα του δείκτη AD μόνο ως προς την κλίμακα τιμών για τις 6 συστάδες που μελετάμε. Στην $k-$means μέθοδο ο δείκτης ξεκινάει από την τιμή 0.865 και μειώνεται σχεδόν γραμμικά φτάνοντας στην τιμή 0.766, ενώ στην ward linkage ο δείκτης ξεκινάει από την τιμή 0.871 και μειώνεται επίσης σχεδόν γραμμικά φτάνοντας στην τιμή 0.768. Και σε αυτήν την περίπτωση ο δείκτης διαφέρει αμελητέα μεταξύ των μεθόδων, διότι για κάθε πλήθος συστάδων και οι δύο αυτές μέθοδοι έχουν σχεδόν την ίδια εντός-συστάδας διακύμανση κάθε φορά που αφαιρούμε τυχαία μία ασθενή από το σύνολο δεδομένων, με αποτέλεσμα να μην μπορούμε να αποφασίσουμε ποια από αυτές τις δύο μεθόδους είναι ικανή να δημιουργήσει σταθερές συστάδες.

Παρακάτω μελετάμε τα γραφήματα που αντιστοιχούν στους βιολογικούς δείκτες

Γράφημα 3.13



➢ <u>BHI</u>

Σύμφωνα με το γράφημα ο δείκτης φαίνεται να διαφέρει σημαντικά μόνο στις τρεις συστάδες, καθώς στην ward linkage αντιστοιχεί στην τιμή 0.49, ενώ στην k – means αντιστοιχεί στην τιμή 0.205. Δηλαδή, στην ward linkage οι τρεις συστάδες περιέχουν το 25% των γονιδίων που βρίσκονται στην ίδια συστάδα και στην ίδια βιολογική ομάδα, ενώ το αντίστοιχο ποσοστό στην k – means για τις τρεις συστάδες είναι περίπου 20%. Αν και η ward linkage υπερέχει ελάχιστα από την k – means στις τρεις συστάδες καμία μέθοδος δεν καταφέρνει να ορίσει σημαντική βιολογική ομοιογένεια, καθώς ένα σημαντικό ποσοστό γονιδίων που βρίσκονται στην ίδια συστάδα δεν φαίνονται να βρίσκονται και στην ίδια βιολογική ομάδα. Μελετώντας τη συμπεριφορά του δείκτη γα τις υπόλοιπες συστάδες παρατηρούμε ότι και στις δύο μεθόδους μεταβάλλεται απότομα, αλλά σε κανένα πλήθος συστάδων δεν καταφέρνει να αναδείξει ικανοποιητική βιολογική ομοιογένεια..

➢ BSI

Ο συγκεκριμένος δείκτης εκφράζει ικανοποιητική βιολογική σταθερότητα των γονιδίων στις δύο συστάδες και για τις δύο μεθόδους, αφού στην ward linkage αντιστοιχεί στην τιμή 0.4698 και στην $k-$means αντιστοιχεί στην τιμή 0.5016. Με ελάχιστη διαφορά συμπεραίνουμε ότι οι δύο συστάδες που σχηματίζει η $k-$means χαρακτηρίζονται από αρκετά ικανοποιητική σταθερότητα βιολογικά ως προς τα γονίδια που περιέχουν. Στις τρεις συστάδες ο δείκτης μειώνεται απότομα και στις δύο μεθόδους και συνεχίζει να μειώνεται μέχρι και τις έξι συστάδες. Αυτό είναι ένδειξη ότι οι μέθοδοι δεν καταφέρνουν να δημιουργήσουν βιολογικά σταθερές συστάδες όταν το πλήθος των συστάδων αυξάνεται.

Στη συνέχεια συγκεντρώνουμε σε πίνακα τα αποτελέσματα από όλους τους παραπάνω δείκτες και για τις δύο μεθόδους για να διακρίνουμε ποια μέθοδος υπερέχει για κάποιο πλήθος συστάδων**.**

Πίνακας 3.13

Internal Measures		**Stability Measures**		**Biological Measures**	
Connectivity	καμία	**APN**	ward ή $k-$means $k=2$ ως 6	**BHI**	ward $k=$ 3 ή 5
Dunn	καμία	**ADM**	$k-$means $k=2$ ή 3	**BSI**	ward για $k=3$ $k-$means για $k=2$
Silhouette	καμία	**AD**	καμία		
		FOM	ward ή $k-$means $k=2$ ως 6		

Ωστόσο, το ενδιαφέρον μας επικεντρώνεται κυρίως στο να αποφασίσουμε ποια από τις δύο αυτές μεθόδους θα επιλέξουμε ως η πιο κατάλληλη για τη δημιουργία τριών συστάδων. Για τον σκοπό αυτό δίνουμε τα αποτελέσματα των παραπάνω δεικτών και για τις δύο μεθόδους, αλλά μόνο για τρεις συστάδες.

Πίνακας 3.14

	Internal Measures			Stability Measures				Biological Measures	
	Conn.	**Dunn**	**Silh.**	**APN**	**ADM**	**AD**	**FOM**	**BSI**	**BHI**
k – **means**	17.686	0.717	0.124	0.0063	0.152	19.290	0.827	0.327	0.205
ward linkage	20.693	0.670	0.114	0.0131	0.324	19.486	0.831	0.386	0.249

Internal Measures

Η k – means φαίνεται να υπερέχει ελάχιστα. Ωστόσο, και στις δύο μεθόδους οι τρεις συστάδες χαρακτηρίζονται από ιδιαίτερα υψηλή εντός-συστάδας απόσταση και χαμηλή εκτός-συστάδας απόσταση, δηλαδή οι συστάδες αυτές δεν έχουν ικανοποιητική σύσταση (άρα έχουν χαμηλή πυκνότητα) και δεν διαχωρίζονται σημαντικά μεταξύ τους. Επιπλέον, η συνδεσιμότητα των συστάδων είναι σχεδόν ασήμαντη και αυτό συνεπάγεται ότι οι συστάδες δεν έχουν ουσιώδη δομή, αφού περιέχουν πολύ ελάχιστα γειτονικά γονίδια.

Stability Measures

Και σε αυτούς τους δείκτες η k – means φαίνεται να υπερέχει, αλλά όχι σημαντικά. Αν εξαιρέσουμε τον δείκτη APN, τότε αντανακλάται μικρή προβλεπτική δύναμη και από τις δύο μεθόδους ομαδοποίησης σύμφωνα με τους υπόλοιπου δείκτες, οι οποίοι είναι πιο αξιόπιστοι διότι βασίζονται είτε στην απόσταση μεταξύ των γονιδίων (AD) είτε στην απόσταση μεταξύ των κέντρων βάρους των συστάδων (ADN) είτε στην τετραγωνική απόκλιση του επιπέδου έκφρασης κάθε γονιδίου από το μέσο επίπεδο έκφρασης των γονιδίων της συστάδας που ανήκει (FOM). Αυτό σημαίνει ότι οι τρεις συστάδες που προκύπτουν από αυτούς τους αλγόριθμους ομαδοποίησης δεν είναι ικανοποιητικές ως προς τη σύστασή τους με αποτέλεσμα να παρουσιάζεται μικρή σταθερότητα όταν συγκρίνουμε κάθε συστάδα με τις υπόλοιπες έχοντας αφαιρέσει ένα δείγμα ασθενών κάθε φορά.

Biological Measures

Στους βιολογικούς δείκτες η ward linkage φαίνεται να υπερέχει πολύ ελάχιστα από την k – means. Όμως, και σε αυτούς τους δείκτες τα αποτελέσματα δεν φαίνονται να είναι γενικά ικανοποιητικά και για τις δύο μεθόδους. Συγκεκριμένα η βιολογική ομοιογένεια των τριών συστάδων είναι χαμηλή, αφού ένα σημαντικό ποσοστό (περίπου τουλάχιστον το 70%) γονιδίων που βρίσκονται στην ίδια συστάδα δεν φαίνονται να βρίσκονται και στην ίδια βιολογική ομάδα. Δηλαδή, η σύσταση των συστάδων αυτών δεν είναι ικανοποιητική, αφού η πλειοψηφία των γονιδίων που περιέχουν δεν έχουν παρόμοια βιολογική λειτουργικότητα. Τέλος, οι συστάδες παρουσιάζουν μικρή βιολογική σταθερότητα, διότι κάθε φορά που αφαιρούμε ένα δείγμα ασθενών και συγκρίνουμε τις συστάδες μεταξύ τους παρατηρούμε ότι η «βιολογική» τους σύνθεση αλλάζει. Συνεπώς, καμία μέθοδος δεν καταφέρνει να δημιουργήσει συστάδες ιδιαίτερης βιολογικής αξίας.

Συνυπολογίζουμε τα συμπεράσματα και των τριών μέτρων αξιολόγησης και συμπεραίνουμε ότι για τον σχηματισμό τριών συστάδων από γονίδια είναι προτιμότερο να εφαρμόσουμε την k – means μέθοδο ομαδοποίησης.

Κεφάλαιο 4

Μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης

4.1 Εισαγωγή

Το παρόν κεφάλαιο καλύπτει μία επισκόπηση της θεωρίας της Ανάλυσης Επιβίωσης. Αρχικά επεξηγούμε τη χρησιμότητα της Ανάλυσης Επιβίωσης στη μελέτη των γονιδιακών εκφράσεων και αναλύουμε δύο σημαντικές έννοιές της: την λογοκρισία και τα τερματικά σημεία μίας έρευνας. Έπειτα παρουσιάζουμε ορισμένες εισαγωγικές έννοιες που αποτελούν την βάση για την μαθηματική στατιστική (εκτιμητική και έλεγχος υποθέσεων) της Ανάλυσης Επιβίωσης. Στην συνέχεια μελετάμε την Kaplan-Meier και Nelson-Aalen εκτίμηση της συνάρτησης επιβίωσης και ολοκληρώνουμε το κεφάλαιο παρουσιάζοντας και αναλύοντας κάποιες μεθόδους για τη σύγκριση των συναρτήσεων επιβίωσης, όπως ο Log-rank και Gehan-Wilcoxon έλεγχος, αλλά και οι στρωματοποιημένοι έλεγχοι (τοπικοί και ολικός).

4.2 Ανάλυση επιβίωσης στην μελέτη εκφράσεων γονιδίων

Χρησιμοποιώντας τις γνώσεις μας για την αποκωδικοποίηση (*transcription*) του ανθρώπινου γονιδιώματος μπορούμε να κατανοήσουμε τον καρκίνο. Συγκεκριμένα, οι εκφράσεις γονιδίων αποτελούν έναν σημαντικό προγνωστικό παράγοντα της επιβίωσης των ασθενών με καρκίνο. Τα δεδομένα από μικροσυστάδες (*microarray data*) συνήθως χαρακτηρίζονται ως υψηλής διάστασης (*high-dimensionality*), καθώς το πλήθος των μεταβλητών $\left(p \sim 1000\right)$ ξεπερνάει σημαντικά το πλήθος των ασθενών $\left(n \sim 100\right)$ και η μελέτη τέτοιων δεδομένων θεωρείται πρόκληση για την πρόβλεψη της επιβίωσης των ασθενών λόγω της δυσκολίας ανάλυσης τέτοιου σετ δεδομένων.

Έχουν προταθεί κάποιες μέθοδοι που μοντελοποιούν την επιβίωση λαμβάνοντας υπόψη την υψηλή διάσταση των δεδομένων. Φυσικά, για την επιλογή της κατάλληλης μεθόδου για το σύνολο δεδομένων που μελετάμε χρησιμοποιούμε κάποια μέτρα αξιολόγησης (*evaluation measures*) (Wessel et al., 2008).

Ο χρόνος επιβίωσης είναι μια σημαντική μεταβλητή που τα τελευταία χρόνια χρησιμοποιείται ιδιαίτερα στην ανάλυση δεδομένων. Στην μελέτη γονιδιακών εκφράσεων λαμβάνουμε υπόψη και άλλες μεταβλητές, όπως η ηλικία των ασθενών, το μέγεθος και το στάδιο του όγκου, τον χρόνο επιβίωσης των ασθενών**,** έτσι ώστε να έχουμε όσο το δυνατόν περισσότερες πληροφορίες για τα δεδομένα που μελετάμε. Προφανώς, η μεταβλητή «χρόνος επιβίωσης» χρησιμοποιείται σ' έναν τομέα ανάλυσης ιδιαίτερου ενδιαφέροντος γνωστός ως Ανάλυση Επιβίωσης.

Αντικείμενο της Ανάλυσης Επιβίωσης είναι η μελέτη του χρόνου επιβίωσης, που ορίζεται ως το χρονικό μήκος από την αρχή της παρατήρησης ενός ατόμου έως ότου συμβεί το υπό εξέταση ενδεχόμενο ή έως το τέλος της μελέτης (τερματικό σημείο). Συνήθως, το ενδεχόμενο που εξετάζεται είναι ο θάνατος του ασθενή ή η επανεμφάνιση των συμπτωμάτων μιας ασθένειας, όταν μελετάμε δεδομένα από τον χώρο της ιατρικής. Κύριος στόχος είναι η πρόβλεψη του χρόνου πραγματοποίησης του ενδεχομένου, γνωστός ως χρόνος επιβίωσης, χρησιμοποιώντας ως επεξηγηματικές μεταβλητές (*explanatory variables*) τις εκφράσεις γονιδίων.

4.3 Λογοκριμένα δεδομένα και τερματικά σημεία

Είναι δυνατόν το υπό μελέτη ενδεχόμενο να μην συμβεί για κάποια άτομα, οπότε προκύπτουν λογοκριμένα δεδομένα και δεν μπορούμε να εφαρμόσουμε κλασικές μεθόδους παλινδρόμησης. Συνεπώς, επιβάλλεται η εφαρμογή μεθόδων που λαμβάνουν υπόψη λογοκριμένα δεδομένα. Στη μελέτη δεδομένων επιβίωσης η παρουσία λογοκριμένων δεδομένων (*censored data*) δυσκολεύει αρκετά την ανάλυση, γι' αυτό πρέπει να λαμβάνονται σοβαρά υπόψη και να μελετώνται επαρκώς. Είναι γεγονός ότι η λογοκρισία συμβαίνει αναπόφευκτα και συνηθίζεται το φαινόμενο ένας μεγάλος αριθμός ασθενών να έχει λογοκριμένους χρόνους.

Τα λογοκριμένα δεδομένα μας πληροφορούν μερικώς για τον χρόνο επιβίωσης κάθε ατόμου. Συγκεκριμένα, ο χρόνος επιβίωσης ενός ατόμου μπορεί να είναι είτε λογοκριμένος από δεξιά (*right censoring*), εάν γνωρίζουμε ότι είναι μεγαλύτερος από κάποιο προκαθορισμένο χρόνο U , είτε λογοκριμένος από αριστερά (*left censoring*),

εάν γνωρίζουμε ότι είναι μικρότερος από κάποιο χρόνο U ή λογοκριμένος σε διάστημα (*interval censoring*), εάν γνωρίζουμε ότι ο χρόνος ζωής ενός ατόμου είναι εντός ενός διαστήματος $[a,b]$ με $a < b$.

Οι λογοκριμένοι χρόνοι από δεξιά μελετώνται περισσότερο, διότι θεωρούνται ως το πιο σοβαρό και συχνά εμφανιζόμενο πρόβλημα στην Ανάλυση Επιβίωσης. Επιπλέον, αυτοί οι χρόνοι διακρίνονται στους μη-ενημερωτικούς (*uniformative*), αν η λογοκρισία δεν σχετίζεται με κάποιο ενδεχόμενο που αφορά την έρευνα, όπως το να εγκαταλείψει κάποιος ασθενής την έρευνα λόγω μετακόμισης, και στους πληροφοριακούς (*informative*), αν η λογοκρισία σχετίζεται με κάποιο ενδεχόμενο που αφορά την έρευνα, όπως το να εγκαταλείψει κάποιος ασθενής την έρευνα λόγω σοβαρών παρενεργειών από τη θεραπεία που λαμβάνει. Είναι αναγκαίο να διακρίνουμε τους ενημερωτικούς από τους μη ενημερωτικούς λογοκριμένους χρόνους από δεξιά για να γίνει σωστή ανάλυση (Shoemaker and Lin, 2005).

Αξίζει να σημειώσουμε ότι σε αντίθεση με τους λογοκριμένους χρόνους, που τους συναντάμε ευρέως στην ανάλυση επιβίωσης, τα τερματικά σημεία της έρευνας δεν χρησιμοποιούνται επαρκώς. Το φαινόμενο αυτό συνηθίζεται στις έρευνες που αφορούν γονιδιακές εκφράσεις. Αντί των τερματικών σημείων συνηθίζεται η ομαδοποίηση των ασθενών σε δύο κατηγορίες: ασθενείς με βραχυπρόθεσμη (*short-term*) και μακροπρόθεσμη (*long-term*) επιβίωση. Με αυτόν τον τρόπο δημιουργείται μία διχοτόμος μεταβλητή (*dichotomous variable*), όπου για την ανάλυσή της υπάρχουν αναρίθμητες μέθοδοι. Ωστόσο, ένα σημαντικό μειονέκτημα αυτής της μεθόδου είναι ότι χάνουμε σημαντικά μεγάλη πληροφορία γύρω από τον ακριβή χρόνο επιβίωσης κάθε ασθενή και επιπλέον υπάρχει πάντα ο κίνδυνος οι ομάδες αυτές να χαρακτηρίζονται από έντονη εντός-ομάδας μεταβλητότητα. Για παράδειγμα, έστω ότι τα δύο χρόνια επιβίωσης είναι το σημείο αναφοράς (*cut off*) για την διαίρεση των ασθενών στις δύο προαναφερθείσες κατηγορίες. Κάποιος ασθενής που επιβίωσε πάνω από δύο χρόνια θα τοποθετηθεί στην ίδια κατηγορία (*long-term survival*) με κάποιον ασθενή που επιβίωσε δέκα χρόνια, ενώ κάποιος που έζησε λιγότερα από δύο χρόνια θα τοποθετηθεί στην άλλη κατηγορία.

Ένα σύνηθες ερώτημα που καλείται να απαντήσει η ιατρική κοινότητα στην περίπτωση που αποδειχθεί για μία συγκεκριμένη μελέτη ότι το προφίλ έκφρασης του ασθενή επηρεάζει τον χρόνο επιβίωσής του, είναι κατά πόσο αυτό το αποτέλεσμα συμπίπτει με την πραγματικότητα και δεν είναι προϊόν των μεθόδων ανάλυσης που

εφαρμόστηκαν. Αυτό το αποτέλεσμα μπορεί να αξιολογηθεί μέσω της αναπαραγωγικότητας (*reproducibility)* των δεδομένων.

Πρόσφατα υπάρχουν υποψίες ότι μπορεί να υπάρχουν ουσιώδεις διαφορές στα αποτελέσματα γονιδιακών εκφράσεων που παίρνουμε από συστάδες DNA (*cDNA arrays*) από εργαστήρια μικρής κλίμακας και στα αποτελέσματα που παίρνουμε από συστάδες ολιγονουκλεοτιδίων (*oligonucleotide arrays)* και κυρίως των συστάδων *Affymetrix* (Shoemaker and Lin, 2005). Παρόμοια ασυμφωνία αποτελεσμάτων παρουσιάστηκαν και από μελέτες που αφορούσαν πλατφόρμες σε συστάδες «σημαδεμένων» (*spotted*) ή «τυπωμένων» (*printed*) ολιγονουκλεοτιδίων (Shoemaker and Lin, 2005). Τέλος, ανάλογο πρόβλημα μπορεί να παρουσιαστεί όταν στην ίδια πλατφόρμα χρησιμοποιούμε ακολουθίες συστάδων διαφορετικού τύπου.

Το συμπέρασμα είναι ότι ο τύπος των συστάδων ανεξαρτήτως του τύπου της πλατφόρμας μπορεί να επηρεάσει σημαντικά τις γονιδιακές εκφράσεις που θα εξαχθούν με αποτέλεσμα να διαφέρει η συμπερασματολογία της ανάλυσης των αντίστοιχων σετ δεδομένων. Αυτό το πρόβλημα είναι ένα χαρακτηριστικό παράδειγμα έλλειψης αναπαραγωγικότητας που σκοπός της είναι ο εντοπισμός κοινών αποτελεσμάτων από την ανάλυση του αντίστοιχου σετ δεδομένων. Αξίζει να σημειωθεί ότι αυτές οι δυσκολίες εμφανίζονται συνήθως σε μελέτες που αφορούν καρκίνο στο στήθος, καρκίνο στον πνεύμονα και το μεγάλο Β-κυτταρικό λέμφωμα (*large B-cell lymphoma)* (Shoemaker and Lin, 2005).

4.4 Εισαγωγικές έννοιες

Ο χρόνος επιβίωσης περιγράφεται συνήθως από την μη αρνητική συνεχή τυχαία μεταβλητή T, δηλαδή $T \in R_T \subseteq [0,\infty)$. Η **αθροιστική συνάρτηση κατανομής** του χρόνου ζωής ορίζεται από τη σχέση

$$F(t) = P(T \le t) = \int_0^t f(u)du \quad t \ge 0 \qquad (4.4.1)$$

και ερμηνεύεται ως η πιθανότητα ο χρόνος ζωής του ατόμου να είναι το πολύ ίσος με t. Είναι προφανές ότι $F(t) = 0$ για $t < 0$, όπως επίσης και ότι $F(0) = 0$ και $\lim_{t\to\infty} F(t) = 1$. Συμπεραίνουμε ότι η συνάρτηση $F(t)$ είναι αύξουσα συνάρτηση και συνεχής από δεξιά.

Η ποσότητα $1-F(t)=1-P(T\le t)=P(T>t)=P(T\ge t)=\int_t^{\infty} f(u)du$ δίνει την **συνάρτηση επιβίωσης** $S(t)$, δηλαδή την πιθανότητα ο χρόνος ζωής του ατόμου να είναι τουλάχιστον ίσος με t. Από την παραπάνω σχέση που συνδέει την συνάρτηση επιβίωσης με την αθροιστική συνάρτηση του χρόνου ζωής συμπεραίνουμε ότι $S(t)=1$ για $t<0$, όπως επίσης και ότι $S(0)=1$ και $\lim_{t\to\infty} S(t)=0$. Η συνάρτηση $S(t)$ είναι φθίνουσα συνεχής από αριστερά.

Η **συνάρτηση πυκνότητας** του χρόνου ζωής υπολογίζεται τόσο μέσω της συνάρτησης $F(t)$ όσο και μέσω της συνάρτησης $S(t)$. Συγκεκριμένα, ισχύει ότι

$$f(t)=F'(t)=-S'(t)=\lim_{\Delta t\downarrow 0}\frac{P(t\le T<t+\Delta t)}{\Delta t}\quad t\ge 0 \qquad (4.4.2)$$

Μέσω των συναρτήσεων $S(t)$ και $f(t)$ υπολογίζεται η **συνάρτηση κινδύνου**, όπου είναι μία ιδιαίτερα σημαντική ποσότητα στην ανάλυση επιβίωσης και εκφράζει την πιθανότητα θανάτου ενός ατόμου σ΄ ένα μικρό χρονικό διάστημα, δοθέντος ότι ο χρόνος ζωής του είναι τουλάχιστον ίσος με t, δηλαδή

$$h(t)=\frac{f(t)}{S(t)}=\frac{1}{P(T\ge t)}\cdot\lim_{\Delta t\downarrow 0}\frac{P(t\le T<t+\Delta t)}{\Delta t}=\lim_{\Delta t\downarrow 0}\frac{P(t\le T<t+\Delta t/T\ge t)}{\Delta t}\quad t\ge 0$$

όπως, επίσης, μέσω της σχέσης $(4.4.2)$ αποδεικνύεται ότι

$$h(t)=\frac{f(t)}{S(t)}=-\frac{S'(t)}{S(t)}=-\frac{d\log S(t)}{dt}\quad t\ge 0 \qquad (4.4.3)$$

Ανάλογα με τον αν το άτομο γερνάει ή βελτιώνεται η υγεία του με την πάροδο του χρόνου, η συνάρτηση $h(t)$ είναι αύξουσα συνάρτηση ή φθίνουσα συνάρτηση του χρόνου ζωής αντίστοιχα. Σε κάθε περίπτωση συμπεραίνουμε ότι ο χρόνος ζωής T έχει την ιδιότητα *IFR* (*Increasing Failure Rate*) ή την ιδιότητα *DFR* (*Decreasing Failure Rate*) αντίστοιχα.

Ολοκληρώνοντας την συνάρτηση κινδύνου στο διάστημα $(0,t)$ για $t\ge 0$ βρίσκουμε την **αθροιστική συνάρτηση κινδύνου**, $H(t)$. Συγκεκριμένα, ισχύει ότι

$$H(t)=\int_0^t h(u)du\quad t\ge 0 \qquad (4.4.4)$$

Η συνάρτηση $H(t)$ συνδέεται με τις συναρτήσεις $F(t)$ και $S(t)$ μέσω της σχέσης

$$H(t)=-\log S(t)=-\log(1-F(t))\quad t\ge 0 \qquad (4.4.5)$$

Εξάλλου αποδεικνύεται μέσω της σχέση $(4.4.3)$ ότι

$$H(t)=\int_0^t h(u)du=-\int_0^t \frac{S'(u)}{S(u)}du=-\int_0^t \frac{d\log S(u)}{du}du=-\log S(t) \quad t\geq 0$$

Από τη σχέση $(4.4.5)$ συμπεραίνουμε ότι $0\leq H(t)<\infty$, όπως επίσης και ότι $H(0)=0$ και $\lim_{t\to\infty} H(t)=\infty$.

(Αντζουλάκος, 2009)

4.5 Μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης

Στην ανάλυση η στρατηγική που ακολουθείται είναι αρχικά η ομαδοποίηση των ασθενών σε συστάδες που διακρίνονται για την μικρή εντός-συστάδας μεταβλητότητα και την μεγάλη εκτός-συστάδας μεταβλητότητα και έπειτα η γραφική αναπαράσταση των χρόνων ζωής των ομαδοποιημένων ασθενών σε μορφή καμπύλων, όπου απώτερος σκοπός είναι να δείξουμε ότι η διαφορά στην πιθανότητα επιβίωσης μεταξύ των συστάδων είναι στατιστικά σημαντική.

Για τη γραφική αναπαράσταση και σύγκριση των χρόνων επιβίωσης των ομαδοποιημένων ασθενών χρησιμοποιούνται οι Kaplan-Meier καμπύλες και οι έλεγχοι λογαρίθμων τάξεων (*log-rank test*), όπου η λογική αυτών των ελέγχων δεν διαφέρει από τη λογική του Mantel-Haenszel ελέγχου. Η μέθοδος Kaplan-Meier είναι μία μη παραμετρική τεχνική εκτίμησης της πιθανότητας ότι κάποιο άτομο θα ζήσει τουλάχιστον περισσότερο από κάποιον γνωστό χρόνο. Το log-rank τεστ βασίζεται στην κατασκευή πινάκων συνάφειας με μεταβλητές τις συστάδες ασθενών και την κατάσταση επιβίωσής τους (με επίπεδα το πλήθος θανάτων και το πλήθος επιζώντων πέραν μίας χρονικής στιγμής) και συνδυάζουμε την πληροφορία από όλους αυτούς τους πίνακες χρησιμοποιώντας το Mantel-Haenszel στατιστικό. Το πλήθος αυτών των πινάκων συνάφειας συμπίπτει με το πλήθος των διαφορετικών χρονικών στιγμών που συνέβη ο θάνατος.

Ωστόσο, ένα μειονέκτημα που πηγάζει από τον συνδυασμό Kaplan-Meier καμπυλών και Log-rank ελέγχων είναι ότι ουσιαστικά αξιολογούμε την αποτελεσματικότητα της ομαδοποίησης και δεν κάνουμε κάποια πρόβλεψη για την επιβίωση των ασθενών. Συγκεκριμένα, η ομαδοποίηση θεωρείται αποτελεσματική αν οι συστάδες διαφέρουν σημαντικά ως προς την πιθανότητα επιβίωσης των αντίστοιχων ασθενών. Για να μπορούμε να κάνουμε προβλέψεις για την επιβίωση

των ομαδοποιημένων ασθενών θα πρέπει να κατασκευάσουμε ένα κατάλληλο μοντέλο στο οποίο θα περιλάβουμε όλες τις απαραίτητες μεταβλητές που ολοκληρώνουν την επάρκειά του. Ένα τέτοιο μοντέλο είναι το μοντέλο παλινδρόμησης του Cox (*Cox regression model*), το οποίο είναι γνωστό και ως μοντέλο αναλογικού κινδύνου (*proportional hazards model)* ή μοντέλο *PH*.
(Shoemaker and Lin, 2005)

4.5.1 Kaplan - Meier εκτιμητής της συνάρτησης επιβίωσης

Έστω ένα δείγμα από n χρόνους ζωής στους οποίους υπάρχουν k $(k \le n)$ διαφορετικοί πλήρεις χρόνοι ζωή διατεταγμένοι ως $t_1 < t_2 < ... < t_k$. Τότε έχουμε τις παρακάτω πληροφορίες για τα άτομα της έρευνας:

- d_j: το πλήθος θανάτων που συμβαίνουν τη χρονική στιγμή t_j (φυσικά ισχύει $d_j \ge 1$)
- r_j: το πλήθος ατόμων σε κίνδυνο τη χρονική στιγμή t_j. Κάποια από αυτά τα άτομα δύναται να πεθάνουν ή να λογοκριθούν τη χρονική t_j ή μεταγενέστερη
- c_j: το πλήθος ατόμων που λογοκρίνονται στο διάστημα $\left[t_j, t_{j+1}\right)$, όπου $j = 1, 2, ...k$

Να σημειώσουμε ότι στο διάστημα $\left[t_k, t_{k+1}\right)$ αντιστοιχούν c_k λογοκριμένες παρατηρήσεις, που αντιπροσωπεύουν τα άτομα που έχουν διαφύγει στο διάστημα αυτό, όπου αξιωματικά ορίζουμε $t_{k+1} = \infty$. Επιπλέον, τη χρονική στιγμή $t_0 = 0$, όπου αρχίζει η έρευνα, οι παραπάνω ποσότητες ορίζονται ως $d_0 = 0$, $r_0 = n$ και $c_0 \ge 0$. Τέλος, οι ποσότητες αυτές συνδέονται μέσω των παρακάτω σχέσεων:

- $r_j = r_{j-1} - d_{j-1} = r_{j-2} - d_{j-2} - d_{j-1} = ... = n - \sum_{i=1}^{j-1} d_i$, όπου $j = 1, 2, ..., k$
- $d_1 + d_2 + ... + d_k =$ συνολικό πλήθος πλήρων χρόνων ζωής
- $q_j = \frac{d_j}{r_j}$, $j = 1, 2, ..., k$, δηλαδή το ποσοστό ατόμων που πεθαίνουν τη στιγμή t_j.

Οπότε το ποσοστό των ατόμων που επιβιώνουν τη στιγμή t_j ορίζεται ως $p_j = 1 - q_j$.

Στο σημείο αυτό θα θυμήσουμε τον ορισμό της εμπειρικής συνάρτησης επιβίωσης για πλήρη δεδομένα. Θεωρούμε ένα δείγμα από n πλήρεις χρόνους στους οποίους

υπάρχουν k $(k \le n)$ διαφορετικοί πλήρεις χρόνοι ζωή διατεταγμένοι ως $t_1 < t_2 < ... < t_k$. Τότε η συνάρτηση επιβίωσης εκτιμάται από την **εμπειρική συνάρτηση επιβίωσης** (*empirical survival function, ESF*)

$$\tilde{S}(t) = \begin{cases} 1 & t \le t_1 \\ \dfrac{n - \sum_{i=1}^{j} d_i}{n} & t_j < t \le t_{j+1} \ , j = 1, 2, ..., k-1 \\ 0 & t > t_k \end{cases}$$

όπου d_j είναι το πλήθος των πλήρων χρόνων ζωής που ισούνται με t_j και ισχύει ότι

$d_1 + d_2 + ... + d_k = n$

δηλαδή

$$\tilde{S}(t) = \frac{n - \{\pi\lambda\acute{\eta}\theta o\varsigma \ \theta\alpha\nu\acute{\alpha}\tau\omega\nu \ \acute{\epsilon}\omega\varsigma \ \kappa\alpha\iota \ \tau o\nu \ \chi\rho\acute{o}\nu o < t\}}{n} = \frac{1}{n}\sum_{j=1}^{n} I\{t_j > t\} \quad t \ge 0$$

όπου αξιωματικά ορίζουμε ότι στο διάστημα $[t_0, t_1)$ δεν υπάρχει κάποιος πλήρης χρόνος ζωής, οπότε και κανένας θάνατος. Να σημειώσουμε ότι η εμπειρική συνάρτηση επιβίωσης είναι μία μη παραμετρική εκτίμηση της συνάρτησης $S(t)$, η οποία είναι φθίνουσα συνάρτηση και συνεχής από αριστερά και η τιμή της μειώνεται κατά $\frac{d_j}{n}$ μετά τη χρονική στιγμή t_j.

Η εμπειρική συνάρτηση κατανομής χρησιμοποιείται για την εκτίμηση της συνάρτησης $S(t)$, όταν μελετάμε αποκλειστικά πλήρους χρόνους ζωής. Ωστόσο, όταν στο δείγμα με χρόνους ζωής υπάρχουν και λογοκριμένοι χρόνοι, τότε η εμπειρική συνάρτηση κατανομής αδυνατεί να εκτιμήσει την συνάρτηση $S(t)$, διότι λόγω διαφυγών στο διάστημα $[0, t)$ δεν γνωρίζουμε αν τα άτομα αυτά πέθαναν ή όχι, οπότε δεν μπορούμε να γνωρίζουμε τον πραγματικό αριθμό των χρόνων ζωής που είναι μεγαλύτεροι ή ίσοι του t. Το πρόβλημα αυτό αντιμετωπίζεται με την εφαρμογή του **εκτιμητή Kaplan-Meier**, ο οποίος ορίζεται ως εξής

$$\hat{S}(t) = \prod_{j:t_j < t} \left(1 - \frac{d_j}{r_j}\right) \quad t \ge 0$$

όπου για $t \le t_1$ ισχύει $\hat{S}(t) = 1$. Λόγω της δομής του ο Kaplan-Meier εκτιμητής ονομάζεται επίσης και **εκτιμητής γινομένου ορίου** (*product limit estimator, PLE*) (Αντζουλάκος, 2005)

Ο τύπος αυτός βασίζεται στην λογική ότι για να επιβιώσει κάποιος έως τον χρόνο t θα πρέπει πρώτα να επιβιώσει έως τον χρόνο t_1. Έπειτα θα πρέπει να επιβιώσει στο διάστημα t_1 ως t_2 δοθέντος ότι έχει επιβιώσει έως τον χρόνο t_1 και ούτω καθ' εξής. Μεταξύ των χρονικών σημείων t_{j-1} και t_j δεν συμβαίνει κάποιος θάνατος, οπότε η πιθανότητα θανάτου μεταξύ αυτών των χρονικών στιγμών ισούται με μηδέν. Η υπό συνθήκη πιθανότητα να πεθάνει κάποιος τη στιγμή t_j, δοθέντος ότι ήταν ζωντανός έως και τη στιγμή t_{j-1} ισούται με $\frac{d_j}{r_j}$, οπότε η υπό συνθήκη πιθανότητα να ζήσει κάποιος πέρα από τη στιγμή t_j, δοθέντος ότι ήταν ζωντανός έως και τη στιγμή t_{j-1} ισούται με $1 - d_j / r_j$. Συνεπώς, το γινόμενο των υπό συνθήκη πιθανοτήτων για όλες τις χρονικές στιγμές έως τον χρόνο t ορίζουν την εκτίμηση της πιθανότητας επιβίωσης έως τον χρόνο t, δηλαδή τον Kaplan-Meier εκτιμητή (Rodriquez, 2005).

Ο Kaplan-Meier εκτιμητής είναι μία φθίνουσα σκαλωτή συνάρτηση συνεχής από αριστερά, όπου η τιμή της δεν αλλάζει στα σημεία που παρατηρούνται λογοκριμένοι χρόνοι, παρά μόνο στα σημεία που συμβαίνουν θάνατοι (πλήρεις χρόνοι) και μειώνεται ανάλογα του παράγοντα $\left(r_j - d_j\right)/r_j$ αμέσως μετά τον χρόνο t_j που συνέβη ο θάνατος. Συνεπώς, μία διαφορετική εκδοχή υπολογισμού του Kaplan-Meier εκτιμητή είναι ο παρακάτω τύπος

$$\hat{S}(t) = \prod_{i=1}^{j} \frac{r_i - d_i}{r_i} = \hat{S}(t_j)\frac{r_j - d_j}{r_j}$$

όπου $\hat{S}(t_j) = \prod_{i=1}^{j-1} \frac{r_i - d_i}{r_i} \quad j = 1, 2, ..., k$.

Παρατηρήσεις

1. Για τον υπολογισμό του Kaplan-Meier εκτιμητή βασιζόμαστε σε δύο θεμελιώδεις υποθέσεις (Altman, 2009):

(1) οι χρόνοι επιβίωσης είναι ανεξάρτητοι μεταξύ τους και

(2) οι λογοκριμένοι χρόνοι είναι ανεξάρτητοι από τους χρόνους επιβίωσης.

2. Στην περίπτωση που δεν υπάρχουν λογοκριμένοι χρόνοι, ο εκτιμητής Kaplan-Meier συμπίπτει με την εμπειρική συνάρτηση επιβίωσης, δηλαδή $\hat{S}(t) = \tilde{S}(t)$.

3. Όταν ο μέγιστος παρατηρούμενος χρόνος τ είναι λογοκριμένος, τότε ο Kaplan-Meier εκτιμητής δεν μπορεί να οριστεί πέρα από τον χρόνο αυτό. Επιπλέον, σε αυτή την περίπτωση στον (μέγιστο) πλήρη χρόνο $t_k < \tau$ το πλήθος θανάτων διαφέρει από το πλήθος ατόμων σε κίνδυνο, δηλαδή $d_k \neq r_k$ και φυσικά το πλήθος διαφυγών είναι μεγαλύτερο του μηδενός.

4. Αντιθέτως, όταν ο μέγιστος παρατηρούμενος χρόνος $t_k = \tau$ είναι πλήρης, τότε ο Kaplan-Meier εκτιμητής μπορεί να οριστεί και πέρα από τον χρόνο αυτό. Επίσης, σε αυτόν τον χρόνο το πλήθος θανάτων ισούται με το πλήθος ατόμων σε κίνδυνο, δηλαδή $d_k = r_k$ και φυσικά το πλήθος διαφυγών ισούται με μηδέν.

(Αντζουλάκος, 2009)

4.5.2 Nelson-Aalen εκτιμητής της συνάρτησης επιβίωσης και της αθροιστικής συνάρτησης επιβίωσης

Ο εκτιμητής Kaplan-Meier παίζει κεντρικό ρόλο στην ανάλυση των περισσοτέρων βιοϊατρικών μελετών. Ωστόσο, ένας εναλλακτικός εκτιμητής προτάθηκε από τον Nelson και μελετήθηκε από τον Aalen γνωστός ως Nelson-Aalen εκτιμητής. Σύμφωνα με Andersen και Fleming-Harrington και οι δύο εκτιμητές έχουν παρόμοιες ασυμπτωτικές ιδιότητες. Συγκεκριμένα, έχει αποδειχθεί ότι είναι ασυμπτωτικά ισοδύναμοι εκτιμητές (Colosimo et al. 2002).

Λαμβάνοντας υπόψη τη σχέση $(4.4.5)$ συμπεραίνουμε ότι ο Kaplan-Meier εκτιμητής της αθροιστικής συνάρτησης επιβίωσης είναι

$$\hat{H}(t) = -\log \hat{S}(t) \quad t \geq 0$$

Στην συνέχεια αναλύουμε την παραπάνω σχέση ως εξής

$$\hat{H}(t) = -\log \hat{S}(t) = -\log\left(\prod_{j:t_j<t}\left(1-\frac{d_j}{r_j}\right)\right) = -\sum_{j:t_j<t}\log\left(1-\frac{d_j}{r_j}\right) = \sum_{j:t_j<t}\left(\frac{d_j}{r_j}+\frac{d_j^2}{2r_j^2}+\ldots\right) \qquad (4.5.1)$$

Η προσέγγιση πρώτης τάξης για την σχέση $(4.5.1)$ είναι η ποσότητα

$$\sum_{j:t_j<t} \frac{d_j}{r_j}$$

η οποία ονομάζεται Nelson-Aalen εκτιμητής της αθροιστικής συνάρτησης κινδύνου $H(t)$ (Αντζουλάκος, 2009). Ουσιαστικά πρόκειται για το άθροισμα των συναρτήσεων κινδύνου για όλες τις χρονικές στιγμές που συμβαίνουν θάνατοι έως τον χρόνο t και ερμηνεύεται ως ο αναμενόμενος αριθμός θανάτων στο διάστημα $(0,\ t]$ ανά μονάδα κινδύνου (Rodriquez, 2005).

Αν λάβουμε πάλι υπόψη τη σχέση $(4.4.5)$, τότε παίρνουμε τον Nelson-Aalen εκτιμητής της συνάρτησης επιβίωσης (Αντζουλάκος, 2009), δηλαδή

$$\hat{H}_{NA}(t) = -\log \hat{S}_{NA}(t) \Leftrightarrow \hat{S}_{NA}(t) = \exp\left(-\hat{H}_{NA}(t)\right) \quad t \geq 0$$

Συγκρίνοντας τους δύο εκτιμητές έχει αποδειχθεί ότι ο Nelson-Aalen εκτιμητής της αθροιστικής συνάρτησης κινδύνου έχει καλύτερη συμπεριφορά και ιδιότητες από τον αντίστοιχο Kaplan-Meier εκτιμητή, όταν μελετάμε μικρό δείγμα χρόνων επιβίωσης (Colosimo et al. 2002). Δηλαδή, ισχύει ότι

$$\hat{H}_{NA}(t) \leq \hat{H}(t) \quad t \geq 0 \qquad (4.5.2)$$

Αν λάβουμε πάλι υπόψη τη σχέση $(4.5.2)$ θα συμπεράνουμε ότι ο Kaplan-Meier εκτιμητής της συνάρτησης επιβίωσης υπερέχει του αντίστοιχου Nelson-Aalen εκτιμητή (Colosimo et al. 2002). Συγκεκριμένα ισχύει ότι

$$\hat{H}_{NA}(t) \leq \hat{H}(t) \Leftrightarrow -\log \hat{S}_{NA}(t) \leq -\log \hat{S}_{NA}$$
$$\Leftrightarrow \hat{S}_{NA}(t) \geq \hat{S}(t) \quad t \geq 0$$

4.6 Σύγκριση m συναρτήσεων επιβίωσης: Mantel-Haenszel έλεγχος

Έστω ότι συγκρίνουμε δύο ή περισσότερες ομάδες ατόμων με βάση τις αντίστοιχες συναρτήσεις επιβίωσης. Ουσιαστικά συγκρίνουμε τα επίπεδα μιας μεταβλητής ως προς τους χρόνους επιβίωσης. Για παράδειγμα μπορούμε να ελέγξουμε αν ο χρόνος ζωής των ασθενών διαφοροποιείται (στατιστικά) σημαντικά ή όχι μεταξύ m διαφορετικών θεραπειών.

Θεωρούμε ένα δείγμα n χρόνων επιβίωσης, που υπάρχουν συνολικά στις ομάδες που μελετάμε, στους οποίους εντοπίζουμε k $(k \leq n)$ διαφορετικούς πλήρεις χρόνους

ζωής διατεταγμένους, δηλαδή $t_1 < t_2 < ... < t_k$. Έστω οι ποσότητες d_{ij} και r_{ij}, που ερμηνεύονται ως το πλήθος θανάτων και το πλήθος ατόμων σε κίνδυνο αντίστοιχα την χρονική στιγμή t_j στην ομάδα i, όπου $1 \le j \le k$ και $1 \le i \le m$. Επίσης, θεωρούμε τις ποσότητες

$$d_j = \sum_{i=1}^{m} d_{ij} \text{ και } r_j = \sum_{i=1}^{m} r_{ij}$$

που ορίζονται ως το συνολικό πλήθος θανάτων και συνολικό πλήθος ατόμων σε κίνδυνο αντίστοιχα την χρονική στιγμή t_j, όταν λαμβάνουμε υπόψη όλες τις υπό εξέταση ομάδες. Οι ποσότητες αυτές συνοψίζονται στον πίνακα 4.6.1. Δηλαδή, για κάθε πλήρη χρόνο t_j κατασκευάζουμε έναν αντίστοιχο πίνακα. Στην περίπτωσή μας μελετάμε k διαφορετικούς πλήρεις χρόνους ζωής, οπότε θα κατασκευάσουμε k αντίστοιχους πίνακες (Rodriquez, 2005).

Πίνακας 4.6.1

Κατάσταση / **Ομάδα**	**# θανάτων τη στιγμή t_j**	**# ατόμων που επιζούν πέρα της στιγμής t_j**	**# ατόμων σε κίνδυνο τη στιγμή t_j**
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
2	d_{2j}	$r_{2j} - d_{2j}$	r_{2j}
...	...	...	...
m	d_{mj}	$r_{mj} - d_{mj}$	r_{mj}
Σύνολο (pooled sample)	d_j	$r_j - d_j$	r_j

Τώρα υποθέτουμε ότι τα περιθώρια αθροίσματα d_j, $r_j - d_j$, r_{1j}, r_{2j},..., r_{kj} είναι σταθερά και ότι ισχύει η μηδενική υπόθεση

$$H_0 : S_1(t) = S_2(t) = ... = S_m(t) \quad \text{για όλα τα } t \le t_k$$

δηλαδή, ανεξαρτησία «Ομάδας» και «Κατάστασης». Τότε τα εσωτερικά κελιά του πίνακα μπορούν να πάρουν οποιαδήποτε τυχαία τιμή αρκεί να ικανοποιούνται τα περιθώρια αθροίσματα. Τελικά το διάνυσμα τυχαίων μεταβλητών $\mathbf{D}_j = \left(d_{1j}, d_{2j}, ..., d_{mj}\right)'$ για τη χρονική στιγμή t_j ακολουθεί την πολυδιάστατη υπεργεωμετρική κατανομή με παραμέτρους $v = d_j$, $a_1 = r_{1j}$, $a_2 = r_{2j}$,..., $a_m = r_{mj}$, δηλαδή $\mathbf{D}_j \sim MultHg\left(d_j, r_{1j}, r_{2j}, ..., r_{mj}\right)$. Συνεπώς, για $1 \le j \le k$ και $1 \le i \le m$ ορίζονται οι παρακάτω ποσότητες

$$E\left(d_{ij}\right) = E_{ij} = r_{ij}\frac{d_j}{r_j}$$

$$V\left(d_{ij}\right) = V_{ij} = \frac{r_{ij}\left(r_i - r_{ij}\right)d_j\left(r_i - d_j\right)}{r_j^2\left(r_j - 1\right)}$$

$$Cov\left(d_{ij}, d_{sj}\right) = V_{is,j} = -\frac{r_{ij}r_{sj}d_j\left(r_j - d_j\right)}{r_j^2\left(r_j - 1\right)} \quad \text{για } 1 \le i \ne s \le m$$

δηλαδή, τα πλήθη θανάτων δύο διαφορετικών ομάδων είναι αρνητικά συσχετισμένα μεταξύ τους, όταν συμβαίνουν την ίδια χρονική στιγμή και

$$Cov\left(d_{ij}, d_{sj^*}\right) = V_{ij,sj^*} = 0 \text{ για } 1 \le i \ne s \le m \text{ και } 1 \le j \ne j^* \le k \qquad (4.6.1)$$

δηλαδή, τα πλήθη θανάτων δύο διαφορετικών ομάδων είναι ασυσχέτιστα μεταξύ τους, όταν συμβαίνουν σε διαφορετικές χρονικές στιγμές (Αντζουλάκος, 2009 και Rodriquez, 2005).

Για να μπορούμε να συγκρίνουμε σε κάθε ομάδα το παρατηρούμενο πλήθος θανάτων με το αναμενόμενο ταυτόχρονα για όλες τις χρονικές στιγμές και όχι μόνο για τη χρονική στιγμή t_j, θα χρησιμοποιήσουμε την στατιστική συνάρτηση

$$\mathbf{D} = \sum_{j=1}^{k}\left(\vec{d}_j - \vec{E}_j\right) = \left(D_1, D_2, ..., D_m\right)'$$

Υποθέτοντας ότι οι τυχαίες μεταβλητές d_{ij} είναι ανεξάρτητες σύμφωνα με τη σχέση $(4.6.1)$ συμπεραίνουμε ότι

$$E\left(\mathbf{D}\right) = 0 \qquad (4.6.2)$$

$$\sigma_{ii} = Cov\left(D_i, D_i\right) = V\left(D_i\right) = \sum_{j=1}^{k} V_{ij} \quad 1 \le i \le m \qquad (4.6.3)$$

$$\sigma_{is} = Cov\left(D_i, D_s\right) = \sum_{j=1}^{k} V_{is,j} \quad 1 \le i \ne s \le m \qquad (4.6.4)$$

Να σημειώσουμε ότι τα στοιχεία του διανύσματος **D** είναι εξαρτημένα αφού $\sum_{i=1}^{m} D_i = 0$ (Αντζουλάκος, 2009 και Rodriquez, 2005).

Οι Mantel-Haenszel πρότειναν για τον έλεγχο της ισότητας των m συναρτήσεων επιβίωσης, δηλαδή

$H_0 : S_1(t) = S_2(t) = ... = S_m(t)$ για όλα τα $t \le t_k$ έναντι

H_1 : τουλάχιστον ένα από τα $S_r(t)$ είναι διαφορετικό από τα υπόλοιπα για όλα τα $t \le t_k$

την εφαρμογή της παρακάτω δευτεροβάθμιας μορφής

$$Q = \mathbf{D}'\mathbf{V}^{-1}\mathbf{D}$$

η οποία ακολουθεί προσεγγιστικά χ^2 κατανομή με $m-1$ βαθμούς ελευθερίας, όπου έχουμε παραλείψει την i ομάδα από τον υπολογισμό των **D** και **V**. Δηλαδή, ο έλεγχος μπορεί να βασιστεί σε οποιεσδήποτε $m-1$ ομάδες χωρίς να μεταβληθεί το αποτέλεσμα. Η ποσότητα $\mathbf{V} = \left(\sigma_{is}\right)_{(m-1)\times(m-1)}$ είναι ο πίνακας διακυμάνσεων των στοιχείων του τυχαίου διανύσματος $\left(D_1, D_2, ..., D_{m-1}\right)'$. Φυσικά, η μηδενική υπόθεση απορρίπτεται όταν ικανοποιείται η ανισότητα $Q > \chi^2_{m-1,a}$ σε προκαθορισμένο επίπεδο σημαντικότητας a (Αντζουλάκος, 2009 και Rodriquez, 2005).

Παρατήρηση

Στην περίπτωση που μελετάμε δύο ομάδες, δηλαδή $m = 2$, η στατιστική συνάρτηση που εφαρμόζουμε για τον έλεγχο της ισότητας των δύο συναρτήσεων επιβίωσης, δηλαδή

$$H_0 : S_1(t) = S_2(t) \text{ έναντι } H_1 : S_1(t) \ne S_2(t) \text{ για όλα τα } t \le t_k$$

ορίζεται ως

$$Q = \mathbf{D}'\mathbf{V}^{-1}\mathbf{D} = \frac{\left(\sum_{j=1}^{k}\left(d_{1j} - E_{1j}\right)\right)^2}{\sum_{j=1}^{k} V_{1j}} = \frac{\left(\sum_{j=1}^{k}\left(d_{2j} - E_{2j}\right)\right)^2}{\sum_{j=1}^{k} V_{2j}}$$

η οποία ακολουθεί προσεγγιστικά χ^2 κατανομή με 1 βαθμό ελευθερίας. Η μηδενική υπόθεση απορρίπτεται όταν ικανοποιείται η ανισότητα $Q > \chi^2_{1,a}$ σε προκαθορισμένο επίπεδο σημαντικότητας a (Αντζουλάκος, 2009).

Αξίζει να σημειώσουμε ότι για μεγάλα δείγματα χρόνων ζωής μπορούμε να εφαρμόσουμε την στατιστική συνάρτηση

$$Z = \frac{\mathbf{D} - E(\mathbf{D})}{\sqrt{V(\mathbf{D})}} = \frac{\sum_{j=1}^{k}\left(d_{1j} - E_{1j}\right)}{\sqrt{\sum_{j=1}^{k} V_{1j}}} = \sqrt{Q}$$

η οποία ακολουθεί προσεγγιστικά την τυποποιημένη κανονική κατανομή σύμφωνα με το κεντρικό οριακό θεώρημα. Εξάλλου, γνωρίζουμε ότι η $Z^2 \sim \chi_1^2$, οπότε εύκολα συμπεραίνουμε ότι $Z^2 = Q \overset{\alpha}{\sim} \chi_1^2$ και η μηδενική υπόθεση απορρίπτεται όταν ικανοποιείται η ανισότητα $Q > \chi^2_{1,a} = \left(z_{\alpha/2}\right)^2$ (Αντζουλάκος, 2009).

4.7 Log-rang και Gehan-Wilcoxon έλεγχος

Ο Mantel-Haenszel έλεγχος ανήκει στην οικογένεια των γραμμικών βαθμωτών ελέγχων (*linear rank tests*) και γι' αυτό συχνά ονομάζεται *log-rank* ή *Savage test* (Rodriguez, 2005).

Αν στα στοιχεία του τυχαίου διανύσματος **D** εισάγουμε τα βάρη ή σκορ (*weights or scores*) w_j, τα οποία είναι σταθερές ποσότητες, τότε ισχύει ότι

$$\mathbf{D} = \left(D_1, D_2, ..., D_m\right)' = \left(\sum_{j=1}^{k} w_j\left(d_{1j} - E_{1j}\right), \sum_{j=1}^{k} w_j\left(d_{2j} - E_{2j}\right), ..., \sum_{j=1}^{k} w_j\left(d_{mj} - E_{mj}\right)\right)$$

οπότε οι σχέσεις $(4.6.2) - (4.6.4)$ μετατρέπονται αντίστοιχα στις παρακάτω σχέσεις

$$E(\mathbf{D}) = 0 \qquad (4.7.1)$$

$$\sigma_{ii} = Cov\left(D_i, D_i\right) = V\left(D_i\right) = \sum_{j=1}^{k} w_j^2 V_{ij} \quad 1 \le i \le m \qquad (4.7.2)$$

$$\sigma_{is} = Cov\left(D_i, D_s\right) = \sum_{j=1}^{k} w_j^2 V_{is,j} \quad 1 \le i \ne s \le m \qquad (4.7.3)$$

και τελικά η στατιστική συνάρτηση Q δίνεται πάλι από την παρακάτω σχέση

$$Q = \mathbf{D}'\mathbf{V}^{-1}\mathbf{D} \overset{\alpha}{\sim} \chi^2_{m-1}$$

Στον *log-rank* έλεγχο θέτουμε $w_j = 1$, δηλαδή δίνεται ίδιο βάρος στους θανάτους που συμβαίνουν σε κάθε χρονική στιγμή, ενώ στον *Gehan-Wilcoxon* έλεγχο θέτουμε $w_j = r_j$ για $1 \le j \le k$, δηλαδή δίνουμε διαφορετικό βάρος στους θανάτους που συμβαίνουν σε κάθε χρονική στιγμή και εξαρτάται από το πλήθος των ατόμων σε κίνδυνο στην αντίστοιχη χρονική στιγμή (Αντζουλάκος, 2009).

Ας θεωρήσουμε ένα γενικό μοντέλο πιθανότητας, όπου για κάποια παρατηρούμενη (πλήρη) χρονική στιγμή t_j, ο κίνδυνος θανάτου κάθε ασθενή της ομάδας 1 ορίζεται ως $h_1(t_j)$, ενώ ο κίνδυνος θανάτου κάθε ασθενή της ομάδας 2 ορίζεται ως $h_2(t_j)$. Στην περίπτωση που θεωρούμε ότι ο κίνδυνος θανάτου κάθε ασθενή είναι σταθερός σε κάθε πλήρη χρόνο ζωής και ίσος με $h_1(t)$ και $h_2(t)$ για τις ομάδες 1 και 2 αντίστοιχα, τότε για τον έλεγχο της μηδενικής υπόθεσης $H_0 : h_1(t) = h_2(t)$ έναντι της εναλλακτικής $H_1 : h_1(t) \ne h_2(t)$ μπορεί να αποδειχθεί ότι ο *log-rank* έλεγχος έχει μεγαλύτερη ισχύ από τον *Gehan-Wilcoxon* έλεγχο. Ωστόσο, αν για τη χρονική στιγμή t_j ισχύει ότι $h_1(t_j) = r_j h_1(t)$ και $h_2(t_j) = r_j h_2(t)$ για τις ομάδες 1 και 2 αντίστοιχα, τότε αποδεικνύεται ότι ο *Gehan-Wilcoxon* έλεγχος έχει μεγαλύτερη ισχύ (Sawyer, 2005).

Υπάρχει κάποιο κριτήριο επιλογής μεταξύ των δύο ελέγχων; Ο *Gehan-Wilcoxon* έλεγχος υπερέχει έναντι του *log-rank* όταν ο κίνδυνος θανάτου είναι ιδιαίτερα υψηλός στους πρώτους χρόνους, οπότε δίνουμε έμφαση στους πιο πρόσφατους θανάτους, οι οποίοι αντιστοιχούν στο αριστερό άκρο των πλήρων χρόνων ζωής διότι υπάρχουν περισσότερα άτομα υπό μελέτη, οπότε περισσότερη πληροφορία για τις συναρτήσεις επιβίωσης των ομάδων. Αντιθέτως, όταν θέλουμε να δώσουμε ίδια έμφαση σε όλους τους θανάτους, τότε εφαρμόζουμε *log-rank* έλεγχο (Αντζουλάκος, 2009 και Sawyer, 2005).

Μπορούμε επίσης να εξετάσουμε μέσω γραφήματος ποιο από τα δύο κριτήρια ισχύει, οπότε ποιος από τους δύο ελέγχους πρέπει να εφαρμοστεί. Ουσιαστικά εξετάζουμε αν ισχύει η **υπόθεση του αναλογικού κινδύνου** (*proportional hazard*

assumption), δηλαδή αν ο λόγος των συναρτήσεων κινδύνου δύο ομάδων που συγκρίνουμε είναι σταθερός κατά μήκος των χρονικών σημείων που εξετάζουμε (Miller, 2005). Εξετάζουμε, λοιπόν, αν ισχύει η σχέση

$$h_1(t) = ch_2(t) \Leftrightarrow \frac{h_1(t)}{h_2(t)} = c \quad t \geq 0$$

όπου $c > 0$ είναι μία σταθερά. Εναλλακτικά μπορούμε να χρησιμοποιήσουμε τις συναρτήσεις επιβίωσης των ομάδων, οπότε θα ισχύει ότι

$$S_1(t) = \left[S_2(t)\right]^c \quad t \geq 0$$

Συγκεκριμένα, αν $c > 1$ τότε $h_1(t) > h_2(t) \Leftrightarrow S_1(t) < S_2(t) \quad t \geq 0$

ενώ αν $c < 1$ τότε $h_1(t) < h_2(t) \Leftrightarrow S_1(t) > S_2(t) \quad t \geq 0$

οπότε και στις δύο περιπτώσεις οι καμπύλες δεν τέμνονται.

Όταν ικανοποιείται η υπόθεση του αναλογικού κινδύνου, τότε θα παρατηρήσουμε γραφικά ότι οι καμπύλες των συναρτήσεων κινδύνου των δύο ομάδων είναι σχεδόν παράλληλες μεταξύ τους σε κάθε χρονική στιγμή, οπότε εφαρμόζουμε τον *log-rank* έλεγχο. Ωστόσο, όταν οι καμπύλες τέμνονται, τότε η υπόθεση του αναλογικού κινδύνου παραβιάζεται, οπότε εφαρμόζουμε τον *Gehan-Wilcoxon* έλεγχο (Αντζουλάκος, 2009).

4.8 Στρωματοποιημένοι έλεγχοι

Όταν θέλουμε να συγκρίνουμε m συναρτήσεις επιβίωσης λαμβάνοντας υπόψη και κάποιο άλλον παράγοντα, όπως οι ηλικιακές ομάδες των ασθενών, το φύλο ή το νοσοκομείο, τότε εφαρμόζουμε στρωματοποιημένους ελέγχους (*stratified tests*), όπου συγκρίνουμε τις συναρτήσεις επιβίωσης σε κάθε στρώμα του παράγοντα και είναι επιθυμητό οι συναρτήσεις επιβίωσης να διαφέρουν στα στρώματα του παράγοντα.

Οι στρωματοποιημένοι έλεγχοι αναδεικνύουν το πραγματικό αποτέλεσμα της σύγκρισης των m συναρτήσεων επιβίωσης, αφού λαμβάνουν υπόψη και άλλη συμμεταβλητή (για παράδειγμα το φύλο των ασθενών) που μπορεί να επηρεάζει (*compound*) τους χρόνους ζωής των ασθενών στις ομάδες (για παράδειγμα οι m θεραπείες), οπότε οι υπό εξέταση ασθενείς κατανέμονται σε στρώματα ομογενή-εντός και ετερογενή-εκτός, ώστε το αποτέλεσμα της σύγκρισης των συναρτήσεων επιβίωσης να αντανακλά την πραγματική διαφορετικότητα τους μεταξύ των θεραπειών που χορηγούνται και όχι μεταξύ των στρωμάτων.

Ο ορισμός των στρωματοποιημένων ελέγχων είναι μία επέκταση του Mantel-Haenszel ελέγχου που αναπτύξαμε στο κεφάλαιο 4.6, όταν λαμβάνουμε υπόψη και άλλη μεταβλητή. Θεωρούμε, λοιπόν, ένα δείγμα n χρόνων επιβίωσης, που υπάρχουν συνολικά στις ομάδες που μελετάμε, στους οποίους εντοπίζουμε k $(k \le n)$ διαφορετικούς πλήρεις χρόνους ζωής διατεταγμένους, δηλαδή $t_1 < t_2 < ... < t_k$. Έστω ότι στο g στρώμα έχουν παρατηρηθεί οι διατεταγμένοι διαφορετικοί πλήρεις χρόνοι ζωής $t_{1,g} < t_{2,g} < ... < t_{k_g,g}$, όπου $1 \le j \le k_g$ και $1 \le g \le G$.

Συγκεκριμένα, για την ταυτόχρονη σύγκριση των m συναρτήσεων επιβίωσης σε κάθε στρώμα εφαρμόζουμε τον ολικό έλεγχο (*global or overall test*) της μηδενική υπόθεσης

$$H_0 : S_{1g}(t) = S_{2g}(t) = ... = S_{mg}(t), \text{ για όλα τα } t \le t_k \text{ και } 1 \le g \le G$$

όπου g είναι κάποιο συγκεκριμένο στρώμα που εξετάζουμε. Για κάθε χρονική στιγμή $t_{j,g}$ κατασκευάζουμε τον παρακάτω πίνακα

Πίνακας 4.8.1

Κατάσταση / **Ομάδα**	**# θανάτων τη στιγμή $t_{j,g}$**	**# ατόμων που επιζούν πέρα της στιγμής $t_{j,g}$**	**# ατόμων σε κίνδυνο τη στιγμή $t_{j,g}$**
1	$d_{1j,g}$	$r_{1j,g} - d_{1j,g}$	$r_{1j,g}$
2	$d_{2j,g}$	$r_{2j,g} - d_{2j,g}$	$r_{2j,g}$
...	...	...	...
m	$d_{mj,g}$	$r_{mj,g} - d_{mj,g}$	$r_{mj,g}$
Σύνολο (pooled sample)	$d_{j,g}$	$r_{j,g} - d_{j,g}$	$r_{j,g}$

Όπως και στην ενότητα 4.6, ακολουθούμε παρόμοια διαδικασία και υποθέτουμε ότι τα περιθώρια αθροίσματα $d_{j,g}$, $r_{j,g}-d_{j,g}$, $r_{1j,g}$, $r_{2j,g}$,..., $r_{kj,g}$ είναι σταθερά και ότι ισχύει η μηδενική υπόθεση, δηλαδή ανεξαρτησία «Ομάδας» και «Κατάστασης». Επομένως, τα εσωτερικά κελιά του πίνακα μπορούν να πάρουν οποιαδήποτε τυχαία τιμή, εφόσον ικανοποιούνται τα περιθώρια αθροίσματα. Για την χρονική στιγμή $t_{j,g}$ το διάνυσμα τυχαίων μεταβλητών $\mathbf{D}_{j,g}=\left(d_{1j,g},d_{2j,g},...,d_{mj,g}\right)'$, που αφορά το στρώμα g, ακολουθεί την πολυδιάστατη υπεργεωμετρική κατανομή με παραμέτρους $v=d_{j,g}$, $a_1=r_{1j,g}$, $a_2=r_{2j,g}$,..., $a_m=r_{mj,g}$, δηλαδή $\mathbf{D}_{j,g}\sim MultHg\left(d_{j,g},r_{1j,g},r_{2j,g},...,r_{mj,g}\right)$. Τελικά, για $1\le j\le k_g$, $1\le i\le m$ και $1\le g\le G$ ορίζονται οι παρακάτω ποσότητες

$$E\left(d_{ij,g}\right)=E_{ij}^{g}=r_{ij,g}\frac{d_{j,g}}{r_{j,g}}$$

$$V\left(d_{ij,g}\right)=V_{ij}^{g}=\frac{r_{ij,g}\left(r_{i,g}-r_{ij,g}\right)d_{j,g}\left(r_{i,g}-d_{j,g}\right)}{r_{j,g}^{2}\left(r_{j,g}-1\right)}$$

$$Cov\left(d_{ij,g},d_{sj,g}\right)=V_{is,j}^{g}=-\frac{r_{ij,g}r_{sj,g}d_{j,g}\left(r_{j,g}-d_{j,g}\right)}{r_{j,g}^{2}\left(r_{j,g}-1\right)}\quad \text{για } 1\le i\ne s\le m$$

$$Cov\left(d_{ij,g},d_{sj^{*},g}\right)=V_{ij,sj^{*}}^{g}=0 \text{ για } 1\le i\ne s\le m \text{ και } 1\le j\ne j^{*}\le k \qquad (4.8.1)$$

Θα χρησιμοποιήσουμε την στατιστική συνάρτηση

$$\mathbf{D}_g=\sum_{j=1}^{k_g}\left(\vec{d}_{j,g}-\vec{E}_{j,g}\right)=\left(D_{1,g},D_{2,g},...,D_{m,g}\right)'$$

Υποθέτουμε ότι οι τυχαίες μεταβλητές $d_{ij,g}$ είναι ανεξάρτητες, οπότε σύμφωνα με τη σχέση $(4.8.1)$ συμπεραίνουμε ότι

$$E\left(\mathbf{D}_g\right)=0 \qquad (4.8.2)$$

$$\sigma_{ii,g}=Cov\left(D_{i,g},D_{i,g}\right)=V\left(D_{i,g}\right)=\sum_{j=1}^{k_g}V_{ij}^{g}\quad 1\le i\le m \qquad (4.8.3)$$

$$\sigma_{is,g}=Cov\left(D_{i,g},D_{s,g}\right)=\sum_{j=1}^{k_g}V_{is,j}^{g}\quad 1\le i\ne s\le m \qquad (4.8.4)$$

Τα στοιχεία του διανύσματος $\mathbf{D}_g$ είναι εξαρτημένα αφού $\sum_{i=1}^{m}D_{i,g}=0$.

Τελικά, ο ολικός έλεγχος για την ισότητα των m συναρτήσεων επιβίωσης στο στρώμα g πραγματοποιείται με την εφαρμογή της στατιστικής συνάρτησης

$$Q_g = \mathbf{D}_g^{'} \mathbf{V}_g^{-1} \mathbf{D}_g$$

η οποία ακολουθεί προσεγγιστικά χ^2 κατανομή με $m-1$ βαθμούς ελευθερίας, όπου έχουμε παραλείψει την i ομάδα από τον υπολογισμό των $\mathbf{D}_g$ και $\mathbf{V}_g$, όπως και στο κεφάλαιο 4.6. Η ποσότητα $\mathbf{V}_g = \left(\sigma_{is,g}\right)_{(m-1)\times(m-1)}$ είναι ο πίνακας διακυμάνσεων των στοιχείων του τυχαίου διανύσματος $\left(D_{1,g}, D_{2,g}, ..., D_{m-1,g}\right)'$. Η μηδενική υπόθεση απορρίπτεται για το στρώμα g, όταν ικανοποιείται η ανισότητα $Q_g > \chi^2_{m-1,a}$ σε προκαθορισμένο επίπεδο σημαντικότητας a.
(Αντζουλάκος, 2009)

Κεφάλαιο 5

Το μοντέλο αναλογικού κινδύνου του Cox

5.1 Εισαγωγή

Στο κεφάλαιο 4 μελετήσαμε την μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης. Στο κεφάλαιο 5 θα μελετήσουμε την ημιπαραμετρική εκτίμηση της συνάρτησης επιβίωσης χρησιμοποιώντας το ημιπαραμετρικό μοντέλο παλινδρόμησης του Cox ή μοντέλο αναλογικού κινδύνου του Cox. Τόσο στην περίπτωση που υπάρχουν δεσμοί όσο και στην περίπτωση που απουσιάζουν θα παρουσιάσουμε την συνάρτηση μερικής πιθανοφάνειας των Breslow και Efron για την εκτίμηση και την εξέταση ελέγχων υποθέσεων του διανύσματος παραμέτρων b, όπως επίσης και την μεθοδολογία εκτίμησης της αναφορικής συνάρτησης κινδύνου και επιβίωσης. Η ανάλυση του μοντέλου αναλογικού κινδύνου του Cox ολοκληρώνεται με την αξιολόγησή του μέσω της ανάλυσης υπολοίπων. Βασική προϋπόθεση για την μελέτη του μοντέλου αναλογικού κινδύνου του Cox είναι η επιλογή του βέλτιστου μοντέλου που προσαρμόζεται στα δεδομένα που μελετάμε. Η διαδικασία επιλογής αυτού του μοντέλου περιγράφεται αναλυτικά βήμα-βήμα. Τέλος, παρουσιάζουμε και περιγράφουμε συνοπτικά ορισμένες μεθόδους που αφορούν την μείωση των δεδομένων υψηλής διάστασης, ώστε να μπορεί να εφαρμοστεί σωστά η Ανάλυση Επιβίωσης.

5.2 Ημιπαραμετρικό μοντέλο παλινδρόμησης του Cox

Τα τελευταία χρόνια στην ιατρική στατιστική βρίσκονται στο επίκεντρο της προσοχής οι μέθοδοι εντοπισμού και αξιολόγησης των ανεξάρτητων συμμεταβλητών $\mathbf{X} = (X_1, X_2, ..., X_k)'$ που επηρεάζουν τον χρόνο επιβίωσης T. Αυτές οι

συμμεταβλητές μπορεί να είναι το φύλο των ασθενών, το στάδιο της ασθένειας ή μία δίτιμη μεταβλητή (0/1) που αντιπροσωπεύει την ομάδα ελέγχου (*control*) έναντι της ομάδας θεραπείας (*treatment*). Το μοντέλο παλινδρόμησης του Cox υποθέτει ότι η συνάρτηση κινδύνου της συνεχούς τυχαίας μεταβλητής T δίνεται από την σχέση

$$h(t/\mathbf{X}) = h_0(t)e^{\mathbf{b'X}} \quad t \geq 0$$

όπου $\mathbf{b} = (b_1, b_2, ..., b_k)'$ είναι ένα $k \times 1$ διάνυσμα αγνώστων παραμέτρων που αντανακλούν τις επιδράσεις του διανύσματος συμμεταβλητών $\mathbf{X}$ στην επιβίωση των ασθενών και $h_0(t)$ είναι μία αυθαίρετη συνάρτηση του χρόνου. Μία επέκταση αυτού του μοντέλου παρουσιάστηκε από τον Kalbfleisch (Kay, 2006), όπου χρησιμοποιεί μεταβλητή στρωματοποίησης. Το μοντέλο αυτό θα μελετηθεί στη συνέχεια.

5.2.1 Το μοντέλο αναλογικού κινδύνου του Cox

Το πρόβλημα σύγκρισης των δύο επιπέδων μιας μεταβλητής μπορεί να γενικευτεί αν θεωρήσουμε αυτά τα επίπεδα ως δύο ψευδομεταβλητές (*dummy variables*), οι οποίες ορίζονται σύμφωνα με τον παρακάτω πίνακα

Πίνακας 5.2.1

Επίπεδα μεταβλητής X	X_1	X_2
1	1	0
2	0	1

Η συνάρτηση κινδύνου για το πρώτο επίπεδο ορίζεται ως $h_1(t) = h(t/X=0) = h_0(t)$, $t \geq 0$, ενώ για το δεύτερο επίπεδο ορίζεται ως $h_2(t) = h(t/X=1)$, $t \geq 0$. Τότε σύμφωνα με το μοντέλο του αναλογικού κινδύνου θα ισχύει ότι

$$\frac{h_2(t)}{h_1(t)} = \frac{h(t/X=1)}{h(t/X=0)} = c = e^b \quad t \geq 0$$

αν θεωρήσουμε ότι $b = \log c \Leftrightarrow c = e^b$ όπου $-\infty < b < \infty$. Δηλαδή, ο κίνδυνος για το δεύτερο επίπεδο είναι e^b επί τον κίνδυνο του πρώτου επιπέδου για κάθε $t \geq 0$. Συνεπώς, το μοντέλο

$$h(t/X) = h_0(t)e^{bX} \text{ για κάθε } t \geq 0 \text{ όπου } X = \left(X_1, X_2\right)'$$

αποτελεί το ημιπαραμετρικό μοντέλο παλινδρόμησης του Cox (*Cox semi-parametric regression model*), το οποίο συχνά αναφέρεται και ως μοντέλο αναλογικού κινδύνου (*proportional hazard model*) ή μοντέλο PH.

Το μοντέλο μπορεί να γενικευθεί για $k-$επίπεδα ενός παράγοντα ή για ένα διάνυσμα k συμμεταβλητών $\mathbf{X} = \left(X_1, X_2, ..., X_k\right)'$, οπότε παίρνει την γενική μορφή

$$h(t/\mathbf{X}) = h_0(t)e^{\mathbf{b'X}} = h_0(t)\exp\left(\sum_{p=1}^{k} b_p X_p\right) \text{ για κάθε } t \geq 0 \qquad (5.2.1)$$

όπου ισχύουν οι παρακάτω περιορισμοί και υποθέσεις:

- η ποσότητα $h_0(t)$, η οποία ονομάζεται αναφορική συνάρτηση κινδύνου (*baseline hazard function*) ή πιο σύντομα ως ΑΣΚ, είναι η συνάρτηση κινδύνου ενός ατόμου για την οποία ξέρουμε ότι ισχύει $h_0(t) \geq 0$.

- το διάνυσμα παραμέτρων $\mathbf{b} = \left(b_1, b_2, ..., b_k\right)'$ είναι σταθερό για κάθε χρονική στιγμή, διότι θεωρούμε ότι ο λόγος των κινδύνων δεν εξαρτάται από τον χρόνο. Συνεπώς, οι συμμεταβλητές που εξετάζουμε δεν εξαρτώνται από τον χρόνο (*fixed covariates*), οπότε η παρατηρούμε τιμή του διανύσματος $\mathbf{X}_j = \left(X_{1j}, X_{2j}, ..., X_{kj}\right)'$ για το j άτομο έχει καταγραφεί στην αρχή της έρευνα, δηλαδή στον χρόνο $t = 0$ και δεν αλλάζει με την πάροδο του χρόνου.
- δεν κάνουμε καμία παραμετρική υπόθεση για την αναφορική συνάρτηση κινδύνου. Μόνο το διάνυσμα των συμμεταβλητών $\mathbf{X}$ έχει παραμετρική μορφή στο μοντέλο αυτό, γι΄ αυτό θεωρείται ημιπαραμετρικό μοντέλο.
- η ποσότητα $\exp\left(\mathbf{b'X}\right)$ είναι η συνάρτηση που πρότεινε ο Cox για την συνάρτηση γενικής μορφής $C\left(\mathbf{b'X}\right)$, η οποία θεωρείται γνωστή.

(Αντζουλάκος, 2009)

Στην περίπτωση ενός παράγοντα με k επίπεδα ο πίνακας 5.2.1 παίρνει την εξής μορφή

Πίνακας 5.2.2

Επίπεδα μεταβλητής X	X_1	X_2	**...**	X_k
1	1	0	...	0
2	0	1	...	0
...	...	...	...	...
k	0	0	...	1

οπότε το μοντέλο παλινδρόμησης του Cox ορίζεται ως εξής

$$h(t/\mathbf{X}) = h_0(t)e^{\mathbf{b'x}} = h_0(t)\exp\left(\sum_{p=2}^{k} b_p X_p\right)$$

όπου θεωρούμε το επίπεδο X_1 ως επίπεδο αναφοράς με το οποίο συγκρίνουμε τα υπόλοιπα επίπεδα. Στην περίπτωση αυτή η αναφορική συνάρτηση κινδύνου ορίζεται ως $h_0(t) = h(t/X_2 = 0,...,X_k = 0)$.

Για να συγκρίνουμε τις συναρτήσεις κινδύνου δύο οποιωνδήποτε ατόμων με διανύσματα συμμεταβλητών $\mathbf{X}$ και $\mathbf{X}^*$ αντίστοιχα χρησιμοποιούμε το μοντέλο αναλογικού κινδύνου λαμβάνοντας υπόψη το μοντέλο $(5.2.1)$

$$RHR(t) = \frac{h(t/\mathbf{X})}{h(t/\mathbf{X}^*)} = \exp\left(\mathbf{b'X} - \mathbf{b'X}^*\right) = \exp\left(\mathbf{b'}(\mathbf{X} - \mathbf{X}^*)\right) = \exp\left(\sum_{p=1}^{k} b_p\left(X_p - X_p^*\right)\right)$$

Η ποσότητα $RHR(t)$ ονομάζεται σχετικός λόγος κινδύνων (*relative hazard ratio*) και παρατηρούμε ότι παραμένει σταθερή για κάθε χρονική στιγμή t. Στον λόγο $RHR(t)$ αποδίδει το μοντέλο $(5.2.1)$ το εναλλακτικό του όνομα «μοντέλο αναλογικού κινδύνου του Cox».

Με το μοντέλο παλινδρόμησης του Cox στόχος είναι η μοντελοποίηση του κινδύνου, έτσι ώστε να καθορίσουμε τις συμμεταβλητές που επηρεάζουν τη συνάρτηση κινδύνου και να εκτιμήσουμε την συνάρτηση κινδύνου, οπότε και την συνάρτηση επιβίωσης, διότι $S(t/\mathbf{X}) = \exp\left(-H\left(t/\mathbf{X}\right)\right) = \exp\left(-\int_0^t h\left(u/\mathbf{X}\right)du\right)$, οπότε

$$S(t/\mathbf{X}) = \exp\left(-\int_0^t h_0(u)\exp(\mathbf{b}'\mathbf{X})\,du\right) = \left(\exp\left(-\int_0^t h_0(u)\,du\right)\right)^{\exp(\mathbf{b}'\mathbf{X})} = \left(S_0(t)\right)^{\exp(\mathbf{b}'\mathbf{X})}$$

για κάθε $t \geq 0$.

(Αντζουλάκος, 2009)

5.2.2 Η συνάρτηση μερικής ποθανοφάνειας του Cox

Το 1972 ο Cox πρότεινε την προσαρμογή του μοντέλου μέσω της μεγιστοποίησης μιας ιδιαίτερης συνάρτησης πιθανοφάνειας. Συνήθως, στα δεδομένα επιβίωσης μελετώνται οι λογοκριμένοι χρόνοι από δεξιά, δηλαδή κάποιοι ασθενείς δεν «αποτυγχάνουν» μέχρι το προκαθορισμένο τέλος της έρευνας. Επιπλέον, υποθέτουμε ότι οι μηχανισμοί παραγωγής λογοκριμένων και πλήρων χρόνων είναι ανεξάρτητοι μεταξύ τους.

Αρχικά ας θεωρήσουμε ένα δείγμα με n ασθενείς και ένα διάνυσμα ανεξάρτητων συμμεταβλητών $\mathbf{X}_j = \left(X_{1j}, X_{2j}, ..., X_{kj}\right)'$ για τον j ασθενή. Έστω οι διατεταγμένοι πλήρεις χρόνοι ζωής $t_{(1)} < t_{(2)} < ... < t_{(d)}$ στους οποίους συμβαίνει μόνο ένα ενδεχόμενο (έστω θάνατος), δηλαδή θεωρούμε ότι απουσιάζουν οι δεσμοί. Επίσης, θεωρούμε το σύνολο R_i, όπου περιέχει τους ασθενείς που βρίσκονται σε κίνδυνο τη χρονική στιγμή $t_{(i)}$. Αξιωματικά, λοιπόν, ορίζουμε ότι στην αρχή της έρευνας θα ισχύει $R_0 = \{1, 2, ..., n\}$. Σύμφωνα με τους Prentice και Kalbfleisch (Kay, 2005) έστω A_i είναι το ενδεχόμενο ότι κάποιο άτομο πεθαίνει τη χρονική στιγμή $t_{(i)}$ και B_i είναι το ενδεχόμενο να συμβεί ένας θάνατος ή μια διαφυγή τη στιγμή $t_{(i)}$.

Η υπό συνθήκη πιθανότητα ότι κάποιο άτομο θα πεθάνει τη χρονική στιγμή $t_{(i)}$ δοθέντος του συνόλου κινδύνου R_i ορίζεται ως (Kay, 2005)

$$P(A_i/B_i) = \frac{h\left(t_{(i)}/\mathbf{X}_{(i)}\right)}{\sum_{j\in R_i} h\left(t_j/\mathbf{X}_j\right)} = \frac{h_0\left(t_{(i)}\right)\exp\left(\mathbf{b}'\mathbf{X}_{(i)}\right)}{\sum_{j\in R_i} h_0\left(t_{(i)}\right)\exp\left(\mathbf{b}'\mathbf{X}_j\right)}$$

όπου ο αριθμητής του κλάσματος περιλαμβάνει τη συνάρτηση κινδύνου του ατόμου που πεθαίνει τη χρονική στιγμή t_i, ενώ ο παρονομαστής περιλαμβάνει τη συνάρτηση κινδύνου των ατόμων που είναι σε κίνδυνο τη χρονική στιγμή t_i. Τελικά, απλοποιούμε την αθροιστική συνάρτηση κινδύνου και συμπεραίνουμε ότι η υπό

συνθήκη πιθανότητα δεν εξαρτάται από την αθροιστική συνάρτηση κινδύνου, δηλαδή (Kay, 2005)

$$\frac{\exp\left(\mathbf{b}'\mathbf{X}_{(i)}\right)}{\sum_{j\in R_i}\exp\left(\mathbf{b}'\mathbf{X}_j\right)}$$

Λαμβάνοντας υπόψη όλους τους πλήρους χρόνους, τότε βρίσκουμε την μερική πιθανοφάνεια, δηλαδή (Kay, 2005)

$$L(\mathbf{b})=\prod_{i=1}^{d}\frac{\exp\left(\mathbf{b}'\mathbf{X}_{(i)}\right)}{\sum_{j\in R_i}\exp\left(\mathbf{b}'\mathbf{X}_j\right)}=\prod_{i=1}^{d}\frac{\exp\left(\sum_{p=1}^{k}b_pX_{(i)p}\right)}{\sum_{j\in R_i}\exp\left(\sum_{p=1}^{k}b_pX_{jp}\right)} \qquad (5.2.2)$$

Ο Cox ονόμασε αυτή την ποσότητα ως «υπό συνθήκη πιθανοφάνεια», διότι είναι αποτέλεσμα γινομένου υπό συνθήκη πιθανοτήτων που αντιστοιχούν σε όλους τους πλήρεις χρόνους του προβλήματος. Ωστόσο, αργότερα απέρριψε αυτό το όνομα, διότι ουσιαστικά η συνάρτηση L δεν είναι από μόνη της μία υπό συνθήκη πιθανότητα (Rodriquez, 2005).

Οι Prentice και Kalbfleisch θεώρησαν την περίπτωση όπου οι συμμεταβλητές είναι σταθερές (*fixed*) καθ' όλη τη διάρκεια της έρευνας και απέδειξαν ότι η συνάρτηση L είναι η περιθώρια πιθανοφάνεια (*marginal likelihood*) των τάξεων των παρατηρούμενων χρόνων, σύμφωνα με τη σειρά που πέθαναν οι ασθενείς και όχι σύμφωνα με τον πραγματικό χρόνο που πέθαναν (Rodriquez, 2005).

Ο Cox μέσω της ιδέας της μερικής πιθανοφάνειας δείχνει ότι η συμπερασματολογία γύρω από το διάνυσμα **b** χρησιμοποιώντας τις μεθόδους πιθανοφάνειας που αφορούν μεγάλα δείγματα, μπορεί να βασιστεί στη σχέση $(5.2.2)$. Όταν δεν υπάρχει λογοκρισία, τότε η σχέση $(5.2.2)$ είναι η οριακή κατανομή των ranks των πλήρων χρόνων (Kay, 2005).

5.2.3 Εκτιμητής Μεγίστης Πιθανοφάνειας του διανύσματος b

Στην συνέχεια μπορούμε να υπολογίσουμε τον λογάριθμο της μερικής πιθανοφάνειας του Cox, δηλαδή (Rodriquez, 2005)

$$\log L=\sum_{i=1}^{d}\sum_{p=1}^{k}b_pX_{(i)p}-\sum_{i=1}^{d}\log\left[\sum_{j\in R_i}\exp\left(\sum_{p=1}^{k}b_pX_{jp}\right)\right]=l(b) \qquad (5.2.3)$$

Η εύρεση του ΕΜΠ $\hat{\mathbf{b}}$ επιτυγχάνεται με την μεγιστοποίηση ως προς $\mathbf{b}$ της παραπάνω ποσότητας, οπότε ορίζουμε ένα σύστημα εξισώσεων όπου ουσιαστικά είναι ένα σύστημα μερικών παραγώγων ως προς το διάνυσμα $\mathbf{b}$ εξισωμένες με το μηδέν, δηλαδή (Rodriquez, 2005)

$$\mathbf{U}(\mathbf{b}) = \frac{\partial l(\mathbf{b})}{\partial \mathbf{b}} = \sum_{i=1}^{d} \left(\mathbf{X}_{(i)} - \frac{\sum_{j \in R_i} \mathbf{X}_j \exp\left(\mathbf{b}'\mathbf{X}_j\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)} \right) = 0 \qquad (5.2.4)$$

όπου $\mathbf{U}(\mathbf{b}) = \left(U_1(\mathbf{b}), U_2(\mathbf{b}), ..., U_p(\mathbf{b})\right)'$ είναι το διάνυσμα των σκορ (*score vector*) και

$$U_r(\mathbf{b}) = \frac{\partial l(\mathbf{b})}{\partial b_r} = \sum_{i=1}^{d} \left(X_{(i)r} - \frac{\sum_{j \in R_i} X_{jr} \exp\left(\sum_{p=1}^{k} b_p X_{jp}\right)}{\sum_{j \in R_i} \exp\left(\sum_{p=1}^{k} b_p X_{jp}\right)} \right) = \sum_{i=1}^{d} U_{(i)r}(\mathbf{b}) = 0 \quad 1 \le r \le p$$

Ουσιαστικά η ποσότητα

$$\frac{\sum_{j \in R_i} X_{jr} \exp\left(\sum_{p=1}^{k} b_p X_{jp}\right)}{\sum_{j \in R_i} \exp\left(\sum_{p=1}^{k} b_p X_{jp}\right)}$$

είναι ο σταθμισμένος μέσος της συμμεταβλητής $\mathbf{X}_r$ για το σύνολο κινδύνου R_i με σταθμίσεις τους σχετικούς λόγους κινδύνων $\exp\left(\mathbf{b}'\mathbf{X}_j\right)$. Οπότε μπορούμε να γράψουμε τα σκορ ως εξής (Rodriquez, 2005)

$$U_r(\mathbf{b}) = \frac{\partial l(\mathbf{b})}{\partial b_r} = \sum_{i=1}^{d} \left(X_{(i)r} - A_{(i)r}(\mathbf{b}) \right) \quad 1 \le r \le p$$

όπου $A_{(i)r}(\mathbf{b})$ είναι ο σταθμισμένος μέσος της συμμεταβλητής $\mathbf{X}_r$ για το σύνολο κινδύνου R_i.

Υπολογίζουμε τη δεύτερη μερική παράγωγο της $(5.2.3)$ και πολλαπλασιάζουμε με το -1, οπότε βρίσκουμε τον παρατηρούμενο πίνακα πληροφορίας του Fisher (*Fisher's observed information matrix*) (Αντζουλάκος, 2009)

$$\mathbf{I}_0(\mathbf{b}) = \left[I_{0,rs}(\mathbf{b}) \right]_{k\times k} = \left(-\frac{\partial^2}{\partial b_r \partial b_s} l(\mathbf{b}) \right)_{k\times k} = \begin{pmatrix} I_{0,11} & I_{0,12} & \cdots & I_{0,1k} \\ I_{0,21} & I_{0,21} & \cdots & I_{0,2k} \\ \vdots & \vdots & \ddots & \vdots \\ I_{0,k1} & I_{0,k1} & \cdots & I_{0,kk} \end{pmatrix}$$

όπου

$$-\frac{\partial^2}{\partial b_r \partial b_s} l(\mathbf{b}) = \sum_{i=1}^{d} \left(\frac{\sum_{j\in R_i} X_{jr} X_{js} \exp\left(\sum_{p=1}^{k} b_p X_{jp} \right)}{\sum_{j\in R_i} \exp\left(\sum_{p=1}^{k} b_p X_{jp} \right)} - A_{(i)r}(\mathbf{b}) \cdot A_{(i)s}(\mathbf{b}) \right) = \sum_{i=1}^{d} U_{(i),rs}(\mathbf{b}) \qquad (5.2.5)$$

$$A_{(i)r}(\mathbf{b}) = \frac{\sum_{j\in R_i} X_{jr} \exp\left(\sum_{p=1}^{k} b_p X_{jp} \right)}{\sum_{j\in R_i} \exp\left(\sum_{p=1}^{k} b_p X_{jp} \right)} \quad \text{και} \quad A_{(i)s}(\mathbf{b}) = \frac{\sum_{j\in R_i} X_{js} \exp\left(\sum_{p=1}^{k} b_p X_{jp} \right)}{\sum_{j\in R_i} \exp\left(\sum_{p=1}^{k} b_p X_{jp} \right)}$$

Η ποσότητα στην παρένθεση της σχέσης $(5.2.5)$ θυμίζει συνδιακύμανση δύο μεταβλητών, οπότε εναλλακτικά μπορούμε να γράψουμε τη σχέση $(5.2.4)$ ως εξής

$$\mathbf{I}_0(\mathbf{b}) = \left(-\frac{\partial^2}{\partial b_r \partial b_s} l(\mathbf{b}) \right)_{p\times p} = \left(\sum_{i=1}^{d} C_{i,rs}(\mathbf{b}) \right)_{p\times p}$$

όπου η $C_{i,rs}(\mathbf{b})$ δηλώνει τη συνδιακύμανση των σταθμισμένων συμμεταβλητών $\mathbf{X}_r$ και $\mathbf{X}_s$ για το σύνολο κινδύνου R_i με σταθμίσεις τους σχετικούς λόγους κινδύνων $\exp\left(\mathbf{b}'\mathbf{X}_j\right)$ (Rodriquez, 2005).

Με την χρήση της μεθόδου *Newton-Raphson* μπορούμε να επιλύσουμε το σύστημα $(5.2.4)$ για την εύρεση του ΕΜΠ του διανύσματος $\mathbf{b}$, αρκεί ο πίνακας των δευτέρων μερικών παραγώγων

$$\frac{\partial^2}{\partial \mathbf{b}^2} l\left(\mathbf{b}\right) = \left(\frac{\partial^2}{\partial b_r \partial b_s} l(\mathbf{b}) \right)_{p\times p} = \left(-I_{0,rs}(\mathbf{b}) \right)_{p\times p} = -\mathbf{I}_0(\mathbf{b})$$

υπολογισμένος στη θέση $\hat{\mathbf{b}}$ να είναι αρνητικά ορισμένος (Αντζουλάκος, 2009 και Rodriquez, 2005).

5.2.4 Ύπαρξη δεσμών *(ties)*

Στην περίπτωση των δεσμών υποθέτουμε ότι τη χρονική στιγμή $t_{(i)}$ συμβαίνει τουλάχιστον ένας θάνατος. Πριν αναλύσουμε την περίπτωση των δεσμών αξίζει να σημειώσουμε ότι σύμφωνα με το άρθρο του Rodriguez (2005), οι χρόνοι ζωής μπορεί να είναι:

- διακριτοί, οπότε υπάρχει μία πιθανότητα αποτυχίας (δηλαδή, να συμβεί ο θάνατος) τη χρονική στιγμή $t_{(i)}$, δηλαδή $P\left(T = t_{(i)}\right) = f\left(t_{(i)}\right)$. Σε αυτή την περίπτωση εφαρμόζουμε ένα διακριτό μοντέλο.
- συνεχείς, αλλά ομαδοποιημένοι, οπότε η μεταβλητή d_i αντιπροσωπεύει το πλήθος θανάτων σε κάποιο διάστημα που περιέχει τη χρονική στιγμή $t_{(i)}$. Στην περίπτωση αυτή μπορούμε να εκτιμήσουμε μία παράμετρο για κάθε διάστημα χρόνων ζωής χωριστά χρησιμοποιώντας complementary log-log διωνυμικό μοντέλο ή ένα Poisson μοντέλο.
- συνεχείς, αλλά μη ομαδοποιημένοι με αρκετούς δεσμούς. Στην περίπτωση αυτή χρησιμοποιούμε μία επέκταση της συνάρτησης μερικής πιθανοφάνειας που έχουμε περιγράψει ήδη.

Έστω D_i είναι το σύνολο με τους ασθενείς που πεθαίνουν τη χρονική στιγμή $t_{(i)}$. Η πιθανότητα ότι τα d_i άτομα που πράγματι αποτυγχάνουν ανήκουν στο σύνολο D_i δοθέντος του συνόλου κινδύνου R_i λαμβάνοντας υπόψη όλους τους δυνατούς συνδυασμούς των d_i ατόμων που αποτυγχάνουν τη χρονική στιγμή $t_{(i)}$, ορίζεται ως

$$\frac{\prod_{j \in D_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}{\sum_{g=1}^{|P_i|} \prod_{h \in P_{i(g)}} \exp\left(\mathbf{b}'\mathbf{X}_h\right)} \qquad (5.2.6)$$

όπου ο αριθμητής περιέχει την πιθανότητα αποτυχίας των ατόμων που ανήκουν στο σύνολο D_i, ενώ ο παρονομαστής περιέχει το άθροισμα των πιθανοτήτων αποτυχίας για όλους τους δυνατούς συνδυασμούς των $d_i \left(= |D_i|\right)$ ατόμων που αποτυγχάνουν την χρονική στιγμή $t_{(i)}$ (Rodriguez, 2005). Σύμφωνα με ένα παράδειγμα από το άρθρο του Rodriguez (2005), έστω ότι εξετάζουμε το διάνυσμα συμμεταβλητών $\mathbf{X}$ και ότι τη χρονική στιγμή $t_{(i)}$ βρίσκονται σε κίνδυνο τέσσερα άτομα, οπότε το σύνολο

κινδύνου ορίζεται ως $R_i = \{1,2,3,4\}$ και πεθαίνουν συνολικά δύο, τα οποία ορίζουν το σύνολο $D_i = \{2,4\}$. Το σύνολο με τους δυνατούς συνδυασμούς δύο ατόμων που πεθαίνουν τη χρονική στιγμή $t_{(i)}$ ορίζεται ως $P_i = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$. Ο αριθμητής του τύπου $(5.2.6)$ γράφεται ως $e^{\mathbf{b}'\mathbf{X}_2+\mathbf{b}'\mathbf{X}_4} = e^{\mathbf{b}'(\mathbf{X}_2+\mathbf{X}_4)}$, ενώ ο παρονομαστής γράφεται ως

$$\sum_{g=1}^{|P_i|=6} \prod_{h\in P_{i(g)}} \exp(\mathbf{b}'\mathbf{X}_h) = \prod_{h\in P_{i[(1,2)]}} \exp(\mathbf{b}'\mathbf{X}_h) + \prod_{h\in P_{i[(1,3)]}} \exp(\mathbf{b}'\mathbf{X}_h) + \ldots + \prod_{h\in P_{i[(3,4)]}} \exp(\mathbf{b}'\mathbf{X}_h)$$

δηλαδή

$$\exp(\mathbf{b}'(\mathbf{X}_1+\mathbf{X}_2)) + \exp(\mathbf{b}'(\mathbf{X}_1+\mathbf{X}_3)) + \ldots + \exp(\mathbf{b}'(\mathbf{X}_3+\mathbf{X}_4))$$

Τελικά, λαμβάνοντας υπόψη όλους τους πλήρεις χρόνους, τότε βρίσκουμε την συνάρτηση μερικής πιθανοφάνειας όταν υπάρχουν δεσμοί , δηλαδή

$$L(\mathbf{b}) = \prod_{i=1}^{d} \frac{\prod_{j\in D_i} \exp(\mathbf{b}'\mathbf{X}_j)}{\sum_{g=1}^{|P_i|} \prod_{h\in P_{i(g)}} \exp(\mathbf{b}'\mathbf{X}_h)}$$

Ο αριθμητής υπολογίζεται ιδιαίτερα εύκολα και μάλιστα απλοποιείται στην μορφή

$$\prod_{j\in D_i} \exp(\mathbf{b}'\mathbf{X}_j) = \exp\left(\mathbf{b}'\sum_{j\in D_i}\mathbf{X}_j\right) = \exp(\mathbf{b}'\mathbf{S}_i)$$

Ωστόσο, ο παρονομαστής είναι υπολογιστικά χρονοβόρος όταν ο αριθμός των μεταθέσεων (*permutations*) είναι αρκετά μεγάλος (Rodriguez, 2005).

Οι Peto και Breslow πρότειναν μία προσέγγιση του παρονομαστή υπολογίζοντας το άθροισμα $\sum \exp(\mathbf{b}'\mathbf{X}_j)$ όπου $j \in R_i$. Η συνάρτηση μερικής πιθανοφάνειας των Peto-Breslow δίνεται παρακάτω

$$L(\mathbf{b}) = \prod_{i=1}^{d} \frac{\exp(\mathbf{b}'\mathbf{S}_i)}{\left(\sum_{j\in R_i} \exp(\mathbf{b}'\mathbf{X}_j)\right)^{d_i}}$$

Δίνει ικανοποιητικά αποτελέσματα όταν το d_i είναι σημαντικά μικρότερο από το $r_i\left(=|R_i|\right)$, δηλαδή έχουμε μικρό πλήθος δεσμών και λόγω της ευκολίας στην εφαρμογή της θεωρείται ιδιαίτερα γνωστή συνάρτηση μερικής πιθανοφάνειας στην περίπτωση δεσμών (Αντζουλάκος, 2009).

Ο Efron (1977) πρότεινε την παρακάτω συνάρτηση μερικής πιθανοφάνειας, όταν υπάρχουν δεσμοί

$$L(\mathbf{b}) = \prod_{i=1}^{d} \frac{\exp(\mathbf{b}'\mathbf{S}_i)}{\prod_{r=1}^{d_i}\left(\sum_{j\in R_i}\exp(\mathbf{b}'\mathbf{X}_j) - \frac{r-1}{d_i}\sum_{j\in D_i} exp(\mathbf{b}'\mathbf{X}_j)\right)}$$

Στην περίπτωση που το d_i είναι αρκετά μεγάλο, δηλαδή έχουμε μεγάλο πλήθος δεσμών, τότε ο τύπος του Efron δίνει καλύτερα αποτελέσματα. Ωστόσο και οι δύο τύποι συμπίπτουν όταν δεν υπάρχουν δεσμοί, δηλαδή $d_i = 1$ (Αντζουλάκος, 2009).

5.2.5 Έλεγχοι υποθέσεων για το διάνυσμα παραμέτρων **b**

Υπάρχουν τρεις προσεγγίσεις για να ελέγξουμε υποθέσεις για το διάνυσμα παραμέτρων $\mathbf{b}$, όπως (Rodriguez, 2005):

• το *Likelihood Ratio τεστ*, σύμφωνα με το οποίο για δύο εμφωλευμένα (*nested*) μοντέλα (δηλαδή το ένα μοντέλο είναι υποσύνολο του άλλου ως προς τις μεταβλητές που περιέχει) βρίσκουμε το διπλάσιο της διαφοράς μεταξύ της log-μερικής πιθανοφάνειας στη θέση $\mathbf{b} = \hat{\mathbf{b}}$ και της log-μερικής πιθανοφάνειας στη θέση $\mathbf{b} = \mathbf{b}_0$, δηλαδή

$$X^2_{LR} = -2\log\left(\frac{L(\mathbf{b}_0)}{L(\hat{\mathbf{b}})}\right) = 2\left(l(\hat{\mathbf{b}}) - l(\mathbf{b}_0)\right) \overset{\alpha}{\sim} \chi^2_p$$

όπου p είναι το πλήθος των παραμέτρων.

• το *Wald* τεστ, όπου δίνει αξιόλογα αποτελέσματα όταν το δείγμα χρόνων ζωής είναι μεγάλο. Συγκεκριμένα, η στατιστική συνάρτηση $\hat{\mathbf{b}}$ ακολουθεί προσεγγιστικά την πολυδιάστατη κανονική κατανομή με μέση τιμή $\mathbf{b}$ και πίνακα διακυμάνσεων-συνδιακυμάνσεων $\mathbf{I}_0^{-1}(\mathbf{b})$. Επομένως, υπό την μηδενική υπόθεση $H_0 : \mathbf{b} = \mathbf{b}_0$ η στατιστική συνάρτηση

$$X^2_W = \left(\hat{\mathbf{b}} - \mathbf{b}_0\right)' \mathbf{I}_0\left(\hat{\mathbf{b}}\right)\left(\hat{\mathbf{b}} - \mathbf{b}_0\right) \overset{\alpha}{\sim} \chi^2_p$$

Στην μονοδιάστατη περίπτωση το *Wald* τεστ ορίζεται από τη στατιστική συνάρτηση

$$Z = \frac{\hat{b} - b_0}{\sqrt{\left(I_0\left(\hat{b}\right)\right)^{-1}}} \overset{\alpha}{\sim} \mathrm{N}(0,1)$$

όπου η στατιστική συνάρτηση $\hat{b}$ ακολουθεί προσεγγιστικά την κανονική κατανομή με μέση τιμή b και διακύμανση $\left(I_0(b)\right)^{-1}$ (Αντζουλάκος, 2009).

• το *Score* τεστ, στο οποίο δεν χρησιμοποιούμε τον ΕΜΠ $\hat{\mathbf{b}}$, σε αντίθεση με τους προηγούμενους ελέγχους. Το διάνυσμα σκορ $\mathbf{U}(\mathbf{b})$ ακολουθεί προσεγγιστικά την πολυδιάστατη κανονική κατανομή με μέση τιμή 0 και πίνακα διακυμάνσεων-συνδιακυμάνσεων $\mathbf{I}_0(\mathbf{b})$, όταν μελετάμε μεγάλο δείγμα. Επομένως, για τον έλεγχο της μηδενική υπόθεση $H_0 : \mathbf{b} = \mathbf{b}_0$ χρησιμοποιείται η στατιστική συνάρτηση

$$X^2_{Sc} = \left(\mathbf{U}(\mathbf{b}_0) - 0\right)' \mathbf{I}_0^{-1}(\mathbf{b}_0)\left(\mathbf{U}(\mathbf{b}_0) - 0\right) \overset{\alpha}{\sim} \chi^2_p$$

Στην μονοδιάστατη περίπτωση το *Score* τεστ ορίζεται από τη στατιστική συνάρτηση

$$Z = \frac{U(b_0) - 0}{\sqrt{I_0(b_0)}} \overset{\alpha}{\sim} \mathrm{N}(0,1)$$

όπου η στατιστική συνάρτηση $U(b)$ ακολουθεί προσεγγιστικά την κανονική κατανομή με μέση τιμή 0 και διακύμανση $I_0(b)$ (Αντζουλάκος, 2009).

Αξίζει να σημειώσουμε ότι το *Score* τεστ για τον έλεγχο της $H_0 : \mathbf{b} = 0$ χρησιμοποιώντας το μοντέλο παλινδρόμησης του Cox ισοδυναμεί με το Mantel-Haenszel log-rank τεστ, όταν μελετάμε την ισότητα των συναρτήσεων επιβίωσης k ομάδων, δηλαδή την μηδενική υπόθεση $H_0 : S_1(t) = S_2(t) = ... = S_k(t)$, $t \geq 0$. Παρακάτω δίνεται σύντομα η απόδειξη αυτού του συμπεράσματος (Rodriguez, 2005):

Έστω η πιο απλή περίπτωση, όπου απουσιάζουν οι δεσμοί, δηλαδή $d_i = 1 = \sum_{p=1}^{k} d_{pi}$.

Εάν $\mathbf{b} = 0$, τότε οι σταθμίσεις $\exp\left(\mathbf{b}'\mathbf{X}_j\right)$ που χρησιμοποιήσαμε για να υπολογίσουμε τις συντεταγμένες του διανύσματος

$$\mathbf{A}_{(i)}(\mathbf{b}) = \left(A_{(i)1}(\mathbf{b}), A_{(i)2}(\mathbf{b}), ..., A_{(i)p}(\mathbf{b})\right)'$$

και τα στοιχεία του πίνακα

$$\mathbf{C}_i(\mathbf{b}) = \begin{pmatrix} C_{i,11}(\mathbf{b}) & C_{i,12}(\mathbf{b}) & \cdots & C_{i,1p}(\mathbf{b}) \\ C_{i,21}(\mathbf{b}) & C_{i,22}(\mathbf{b}) & \cdots & C_{i,2p}(\mathbf{b}) \\ \vdots & \vdots & \ddots & \vdots \\ C_{i,p1}(\mathbf{b}) & C_{i,p2}(\mathbf{b}) & \cdots & C_{i,pp}(\mathbf{b}) \end{pmatrix}$$

όπου $C_{i,rs}(\mathbf{b}) = \dfrac{\sum_{j \in R_i} X_{jr} X_{js} \exp\left(\sum_{p=1}^{k} b_p X_{jp}\right)}{\sum_{j \in R_i} \exp\left(\sum_{p=1}^{k} b_p X_{jp}\right)} - A_{(i)r}(\mathbf{b}) \cdot A_{(i)s}(\mathbf{b}) \quad 1 \le r, s \le p$

ισούται με 1, όπου $A_{(i)r}(\mathbf{b})$ είναι το ποσοστό ατόμων σε κίνδυνο για το σύνολο κινδύνου R_i που αφορά την $r-$οστή συμμεταβλητή (δηλαδή, ισοδυναμεί με το αναμενόμενο πλήθος θανάτων σε αυτή τη συμμεταβλητή τη χρονική στιγμή $t_{(i)}$) και $\mathbf{C}_i(\mathbf{b})$ είναι ο διωνυμικός πίνακας διακυμάνσεων-συνδιακυμάνσεων (*binomial variance-covariance matrix*). Στην περίπτωση που υπάρχουν δεσμοί, το μοντέλο του Cox οδηγεί στον έλεγχο $Q = \mathbf{D}'\mathbf{V}^{-1}\mathbf{D}$ που περιγράψαμε στην ενότητα 4.7. Χρησιμοποιούμε την συνάρτηση μερικής πιθανοφάνειας των Peto-Breslow, που ισοδυναμεί με τον πίνακα διακυμάνσεων-συνδιακυμάνσεων $\mathbf{V}$, όπου έχουμε παραλείψει την ποσότητα $(r_i - d_i)/(r_i - 1)$ (Rodriquez, 2005).

Και οι τρεις παραπάνω έλεγχοι είναι ασυμπτωτικά ισοδύναμοι. Η ποιότητα της κανονικής προσέγγισης εξαρτάται από το μέγεθος του δείγματος, την κατανομή των παρατηρούμενων χρόνων ζωής στον χώρο των συμμεταβλητών και την ένταση της λογοκρισίας (Rodriquez, 2005).

Παρατήρηση

Συχνά, θέλουμε να ελέγξουμε ένα υποσύνολο παραμέτρων του διανύσματος b. Αυτός ο έλεγχος ονομάζεται **τοπικός έλεγχος** (*local test*). Συγκεκριμένα, έστω $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2)'$ ένα $p \times 1$ διάνυσμα παραμέτρων, όπου $\mathbf{b}_1 = (\mathrm{b}_1, \mathrm{b}_2, ..., \mathrm{b}_{p_1})'$ και $\mathbf{b}_2 = (\mathrm{b}_{p_1+1}, \mathrm{b}_{p_1+2}, ..., \mathrm{b}_p)'$ είναι ένα $p_1 \times 1$ και $(p - p_1) \times 1 = p_2 \times 1$ διάνυσμα παραμέτρων αντίστοιχα. Για τον έλεγχο της μηδενικής υπόθεσης $H_0 : \mathbf{b}_1 = \mathbf{b}_{10}$ οι έλεγχοι που περιγράψαμε παραπάνω ορίζονται ως εξής (Αντζουλάκος, 2009)

- το *Likelihood Ratio* τεστ ορίζεται από τη στατιστική συνάρτηση

$$X^2_{LR} = 2\left(l(\hat{\mathbf{b}}) - l(\tilde{\mathbf{b}})\right) \overset{a}{\sim} \chi^2_{p_1}$$

- το *Wald* τεστ ορίζεται από τη στατιστική συνάρτηση

$$X_W^2 = \left(\hat{\mathbf{b}} - \mathbf{b}_{10}\right)' \left[\mathbf{I}_0^{11}\left(\hat{\mathbf{b}}\right)\right]^{-1} \left(\hat{\mathbf{b}} - \mathbf{b}_{10}\right) \overset{\alpha}{\sim} \chi^2_{p_1}$$

• το *Score* τεστ ορίζεται από τη στατιστική συνάρτηση

$$X_{Sc}^2 = \left(\mathbf{U}_1\left(\tilde{\mathbf{b}}\right) - 0\right)' \mathbf{I}_0^{11}\left(\tilde{\mathbf{b}}\right)\left(\mathbf{U}_1\left(\tilde{\mathbf{b}}\right) - 0\right) \overset{\alpha}{\sim} \chi^2_{p_1}$$

όπου $U\left(\mathbf{b}\right) = \left(U_1\left(\mathbf{b}\right), U_2\left(\mathbf{b}\right)\right)$, $\mathbf{U}_1\left(\mathbf{b}\right) = \left(\mathrm{U}_1\left(\mathbf{b}\right), \mathrm{U}_2\left(\mathbf{b}\right), ..., \mathrm{U}_{p_1}\left(\mathbf{b}\right)\right)'$, $\tilde{\mathbf{b}} = \left(\mathbf{b}_{10}, \tilde{\mathbf{b}}_2\right)$ και $\tilde{\mathbf{b}}_2 = \tilde{\mathbf{b}}_2\left(\mathbf{b}_{10}\right)$ είναι ο ΕΜΠ της παραμέτρου $\mathbf{b}_2$ όπως προκύπτει από τη συνάρτηση πιθανοφάνειας $L\left(\mathbf{b}_{10}, \mathbf{b}_2\right)$. Επιπλέον, ο πίνακας $\mathbf{I}_0^{11}\left(\mathbf{b}\right)$ είναι τμήμα του αντιστρόφου του παρατηρούμενου πίνακα πληροφορίας

$$\mathbf{I}_0^{-1}\left(\mathbf{b}\right) = \left[I_0^{ij}\left(\mathbf{b}\right)\right]_{p\times p} = \begin{pmatrix} \mathbf{I}_0^{11}\left(\mathbf{b}\right) & \mathbf{I}_0^{12}\left(\mathbf{b}\right) \\ \mathbf{I}_0^{21}\left(\mathbf{b}\right) & \mathbf{I}_0^{22}\left(\mathbf{b}\right) \end{pmatrix}_{p\times p}$$

όπου $\left[\mathbf{I}_0^{11}\left(\mathbf{b}\right)\right]_{p_1\times p_1}$, $\left[\mathbf{I}_0^{22}\left(\mathbf{b}\right)\right]_{p_2\times p_2}$, $\left[\mathbf{I}_0^{12}\left(\mathbf{b}\right)\right]_{p_1\times p_2}$ και $\left[\mathbf{I}_0^{21}\left(\mathbf{b}\right)\right]_{p_2\times p_1}$.

5.2.6 Η εκτίμηση της αναφορικής συνάρτησης κινδύνου και επιβίωσης

Εκτός από την εκτίμηση του διανύσματος συντελεστών παλινδρόμησης $\mathbf{b}$, ιδιαίτερο ενδιαφέρον παρουσιάζει και η εκτίμηση της αναφορικής συνάρτησης κινδύνου $h_0(t)$, η οποία δεν επηρεάζει την μερική πιθανοφάνεια $L(\mathbf{b})$, αφού δεν περιέχεται στον τύπο της. Σύμφωνα με τους Kalbfleisch και Prentice (Rodriguez, 2005) θεωρούμε την ποσότητα π_i που δηλώνει την υπό συνθήκη πιθανότητα επιβίωσης τη χρονική στιγμή $t_{(i)}$. Λαμβάνοντας υπόψη και το διάνυσμα συμμεταβλητών, τότε στο π_i υψώνουμε την ποσότητα $\exp\left(\mathbf{b}'\mathbf{X}\right)$. Η συνάρτηση πιθανοφάνειας γράφεται στην μορφή

$$\begin{aligned} L(\mathbf{b}) &= \left(\prod_{j\in D} f\left(t_j / \mathbf{X}_j\right)\right)\left(\prod_{j\in C} S\left(t_j / \mathbf{X}_j\right)\right) \\ &= \prod_{i=1}^{d}\prod_{j\in D_i}\left(1-\pi_i\right)^{\exp\left(\mathbf{b}'\mathbf{X}_j\right)} \prod_{j\in\left(R_i - D_i\right)} \pi_i^{\exp\left(\mathbf{b}'\mathbf{X}_j\right)} \end{aligned}$$

Ο Meier πρότεινε την μεγιστοποίηση αυτής της πιθανοφάνειας λαμβάνοντας υπόψη τις παραμέτρους π_i και $\mathbf{b}$. Μία απλή προσέγγιση είναι να αντικαταστήσουμε

το $\mathbf{b}$ με τον ΕΜΠ $\hat{\mathbf{b}}$ που βρήκαμε στο κεφάλαιο 5.2.3 μέσω της συνάρτησης μερικής πιθανοφάνειας και να μεγιστοποιήσουμε ως προς π_i (Rodriguez, 2005).

Στην περίπτωση που δεν υπάρχουν δεσμοί βρίσκουμε ότι ο ΕΜΠ της π_i είναι

$$\hat{\pi}_i = \left(1 - \frac{\exp\left(\mathbf{b}'\mathbf{X}_{(i)j}\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}\right)^{\exp\left(-\mathbf{b}'\mathbf{X}_{(i)j}\right)}$$

Χωρίς συμμεταβλητές η εκτίμηση της π_i είναι $\hat{\pi}_i = 1 - d_i / r_i = 1 - 1/r_i$, όπου είναι ο εκτιμητής Kaplan-Meier για $t_{(i)} < t \le t_{(i+1)}$ (λόγω απουσίας δεσμών ισχύει ότι $d_i = 1$) (Rodriguez, 2005).

Στην περίπτωση που υπάρχουν δεσμοί θα πρέπει να λύσουμε επαναληπτικά ως προς π_i την παρακάτω σχέση

$$\sum_{j \in D_i} \frac{\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)}{1 - \pi_i^{\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)}} = \sum_{j \in R_i} \exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)$$

Μία κατάλληλη αρχική τιμή είναι

$$\log \pi_i = -\frac{d_i}{\sum_{j \in R_i} \exp\left(-\hat{\mathbf{b}}'\mathbf{X}_j\right)}$$

Η εκτίμηση της αναφορικής συνάρτησης επιβίωσης είναι ουσιαστικά μία σκαλωτή συνάρτηση και δίνεται παρακάτω

$$\hat{S}_0(t) = \prod_{i: t_{(i)} < t} \hat{\pi}_i$$

Οι Cox και Oakes (Rodriquez, 2005) περιγράφουν μία απλούστερη διαδικασία εκτίμησης της αναφορικής συνάρτησης επιβίωσης που αποτελεί επέκταση του εκτιμητή Nelson-Aalen της αθροιστικής συνάρτησης κατανομής,

$$\hat{H}_{NA}(t) = \sum_{j: t_j < t} \frac{d_j}{r_j}$$

Έστω ότι η αναφορική συνάρτηση κατανομής ισούται με μηδέν στους χρόνους που δεν συμβαίνει αποτυχία. Ο αναμενόμενος αριθμός θανάτων την χρονική στιγμή $t_{(i)}$ ισούται με το άθροισμα των συναρτήσεων κατανομής που αντιστοιχούν στο σύνολο κινδύνου R_i, δηλαδή

$$E\left(d_i\right)=\sum_{j\in R_i}h_0\left(t_{(i)}\right)\exp\left(\mathbf{b}'\mathbf{X}_j\right)$$

Εξισώνουμε το παρατηρούμενο με το αναμενόμενο πλήθος θανάτων για τη χρονική στιγμή $t_{(i)}$ και βρίσκουμε την εκτίμηση της αναφορικής συνάρτησης κατανομής, δηλαδή

$$\begin{aligned} d_i=E\left(d_i\right)&\Leftrightarrow d_i=\sum_{j\in R_i}h_0\left(t_{(i)}\right)\exp\left(\mathbf{b}'\mathbf{X}_j\right)\\ &\Leftrightarrow d_i=h_0\left(t_{(i)}\right)\sum_{j\in R_i}\exp\left(\mathbf{b}'\mathbf{X}_j\right)\\ &\Leftrightarrow h_0\left(t_{(i)}\right)=\frac{d_i}{\sum_{j\in R_i}\exp\left(\mathbf{b}'\mathbf{X}_j\right)}\end{aligned}$$

Συνεπώς, βρήκαμε ότι η εκτίμηση της αναφορικής συνάρτησης κατανομής στη χρονική στιγμή $t_{(i)}$ δίνεται από τη σχέση

$$\hat{h}_0\left(t_{(i)}\right)=\frac{d_i}{\sum_{j\in R_i}\exp\left(\mathbf{b}'\mathbf{X}_j\right)}=\hat{h}_i$$

Τελικά η εκτίμηση της αναφορικής αθροιστικής συνάρτησης κατανομής και της αναφορικής συνάρτησης επιβίωσης δίνονται αντίστοιχα από τις σχέσεις

$$\hat{H}_0\left(t\right)=\sum_{i:t_{(i)}<t}\hat{h}_i\quad\text{και}\quad\hat{S}_0\left(t\right)=\exp\left(-\hat{H}_0\left(t\right)\right)=\exp\left(-\sum_{i:t_{(i)}<t}\hat{h}_i\right)\qquad\left(5.2.7\right)$$

Εάν δεν υπάρχουν συμμεταβλητές, τότε οι ποσότητες στην $\left(5.2.7\right)$ συμπίπτουν με τον Nelson-Aalen και Kaplan-Meier εκτιμητή αντίστοιχα (Αντζουλάκος, 2009).

Έχοντας εκτιμήσει την αναφορική συνάρτηση κινδύνου και επιβίωσης, μπορούμε να προσαρμόσουμε το μοντέλο παλινδρόμησης του Cox για οποιαδήποτε τιμή του διανύσματος συμμεταβλητών $\mathbf{X}$. Αυτή η διαδικασία είναι εύκολη στην περίπτωση των χρονο-ανεξάρτητων συμμεταβλητών (*time-fixed covariates*), διότι το μόνο που μπορούμε να κάνουμε είναι να πολλαπλασιάσουμε την αναφορική συνάρτηση κινδύνου με τον σχετικό κίνδυνο, δηλαδή $h_0\left(t\right)\exp\left(\mathbf{b}'\mathbf{X}\right)=h\left(t/\mathbf{X}\right)$ ή να υψώσουμε την αναφορική συνάρτηση επιβίωσης στον σχετικό κίνδυνο, δηλαδή $\left[S_0\left(t\right)\right]^{\exp(\mathbf{b}'\mathbf{X})}=S\left(t/\mathbf{X}\right)$. Ωστόσο, στην περίπτωση των χρονο-εξαρτημένων συμμεταβλητών (*time-varying covariates*) η διαδικασία γίνεται πολύπλοκη, καθώς πρέπει να επιλέξουμε την κατάλληλη συνάρτηση κινδύνου για κάθε χρόνο αποτυχίας

λαμβάνοντας υπόψη τις τιμές των συμμεταβλητών σε αυτόν τον χρόνο (Rodriguez, 2005).

5.3 Αξιολόγηση μοντέλου αναλογικού κινδύνου μέσω ανάλυσης υπολοίπων

Το μοντέλο παλινδρόμησης του Cox χρησιμοποιείται ευρέως για την ανάλυση των επιδράσεων των συμμεταβλητών στην συνάρτηση κινδύνου χρησιμοποιώντας δείγμα από πλήρεις και λογοκριμένους χρόνους επιβίωσης. Ωστόσο, το μοντέλο αυτό δεν ικανοποιεί πάντα την υπόθεση του αναλογικού κινδύνου. Για να εξετάσουμε αν οι χρονικά ανεξάρτητες (σταθερές ή *fixed*) μεταβλητές του μοντέλου ικανοποιούν την υπόθεση του αναλογικού κινδύνου χρησιμοποιούμε κάποια υπόλοιπα του μοντέλου που περιγράφονται παρακάτω.

5.3.1 Χρήση των Schoenfeld residuals (ή partial residuals)

Ο Schoenfeld χρησιμοποίησε τα υπόλοιπα (*residuals*) για να ελέγξει την υπόθεση του αναλογικού κινδύνου και οι Grambsch & Therneau πρότειναν την τυποποίηση (*scaling*) και «εξομάλυνση» (*smoothing*) αυτών των υπολοίπων (Winnett and Sasieni, 2001).

Γνωρίζουμε ότι το μοντέλο αναλογικού κινδύνου του Cox υποθέτει μία συνάρτηση κινδύνου με διάνυσμα συμμεταβλητών $\mathbf{X}$, δηλαδή

$$h(t/\mathbf{X}) = h_0(t)\exp(\mathbf{b}'\mathbf{X}) = h_0(t)\exp\left(\sum_{p=1}^{k} b_p X_p\right) \quad t \geq 0$$

Το μοντέλο αυτό συνεπάγεται ότι ο λόγος κίνδυνων δύο οποιωνδήποτε ατόμων είναι ανεξάρτητος του χρόνου. Έστω ότι εισάγουμε στο μοντέλο την (ορισμένη) μεταβλητή $\mathbf{g}(t)'\mathbf{X}$, όπου $\mathbf{g}(t)$ είναι ένα διάνυσμα άγνωστων συναρτήσεων του χρόνου. Τότε το μοντέλο παίρνει την παρακάτω μορφή

$$h(t/\mathbf{X}) = h_0(t)\exp\left[\left(\mathbf{b}+\mathbf{g}(t)\right)'\mathbf{X}\right] = h_0(t)\exp\left[\sum_{p=1}^{k}\left(b_p + g_p(t)\right)X_p\right]$$

οπότε η συνάρτηση κινδύνου αλλάζει με την πάροδο του χρόνου. Η υπόθεση του αναλογικού κινδύνου αξιολογείται από την απόφασή μας για τον αν η $\mathbf{g}(t)$ ισούται ή όχι με μηδέν.

Έστω $T_1, T_2, ..., T_n$ ένα δείγμα από πλήρεις και λογοκριμένους χρόνους. Στο δείγμα αυτό υπάρχουν d διαφορετικοί διατεταγμένοι πλήρεις χρόνοι. Επιπλέον, έστω $\delta_1, \delta_2, ..., \delta_n$ οι δείκτες λογοκρισίας που αντιστοιχούν στους χρόνους ζωής του δείγματος, όπου $\delta_j = 1$ εάν T_j είναι πλήρης χρόνος και $\delta_j = 0$ εάν T_j είναι λογοκριμένος χρόνος για $1 \le j \le n$ και $\mathbf{X} = \left(X_1, X_2, ..., X_k\right)'$ ένα $k-$διάνυσμα σταθερών συμμεταβλητών. Η μερική πιθανοφάνεια του μοντέλου του Cox δίνεται παρακάτω

$$L(\mathbf{b}) = \prod_{i=1}^{d} L_{(i)}(\mathbf{b}) = \prod_{i=1}^{d} \frac{\exp\left(\mathbf{b}'\mathbf{X}_{(i)}\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)} = \prod_{i=1}^{n} \left(\frac{\exp\left(\mathbf{b}'\mathbf{X}_i\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)} \right)^{\delta_i} = \prod_{i=1}^{n} L_i(\mathbf{b})$$

οπότε τελικά ο λογάριθμος της μερικής πιθανοφάνειας είναι

$$l(\mathbf{b}) = \log L(\mathbf{b}) = \sum_{i=1}^{n} \delta_i \left(\mathbf{b}'\mathbf{X}_i - \log\left(\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right) \right) \right)$$

όπου μεγιστοποιώντας ως προς $\mathbf{b}$ τον λογάριθμο της μερικής πιθανοφάνειας βρίσκουμε τον Εκτιμητή Μεγίστης Πιθανοφάνειας $\hat{\mathbf{b}}$ του διανύσματος παραμέτρων $\mathbf{b}$ (Winnett and Sasieni, 2001). Οι εξισώσεις score δίνονται παρακάτω

$$U_g(\mathbf{b}) = \frac{\partial}{\partial b_g} l(\mathbf{b}) = \sum_{i=1}^{n} \delta_i \left(X_{ig} - \frac{\sum_{j \in R_i} X_{jg} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)} \right) = \sum_{i=1}^{n} \delta_i \left(X_{ig} - A_{ig} \right) = \sum_{i=1}^{n} r_{ig} = 0$$

Τελικά για το i άτομο με συμμεταβλητή X_g, όπου $1 \le i \le n$ και $1 \le g \le k$ τα υπόλοιπα του Schoenfeld ορίζονται ως εξής

$$\hat{r}_{ig} = \delta_i \left(X_{ig} - \frac{\sum_{j \in R_i} X_{jg} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)} \right) \qquad (5.3.1)$$

Από τη σχέση είναι προφανές ότι σε περίπτωση λογοκριμένου χρόνου για το i άτομο τα Schoenfeld υπόλοιπα ισούται με μηδέν για κάθε μεταβλητή X_g, $1 \le g \le k$ λόγω του $\delta_i = 0$. Σε περίπτωση πλήρη χρόνου για το i άτομο τα Schoenfeld υπόλοιπα ισούνται με

$$\hat{r}_{ig} = X_{ig} - \frac{\sum_{j \in R_i} X_{jg} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)} = X_{ig} - A_{ig}$$

για κάθε μεταβλητή $X_g, 1 \le g \le k$ λόγω του $\delta_i = 1$, όπου A_{ig} θυμίζουμε ότι είναι ο σταθμισμένος μέσος όρος των τιμών της g μεταβλητής για το i άτομο με σταθμίσεις $\exp\left(\mathbf{b}'\mathbf{X}_j\right)$ που αντιστοιχούν στο σύνολο κινδύνου R_i για την χρονική στιγμή t_i. Να σημειώσουμε ότι σε περίπτωση δεσμών το $\hat{r}_{ig}$ είναι το ίδιο για τα άτομα του συνόλου D_i.

Λόγω της μηδενικής τιμής των Schoenfeld υπολοίπων στους λογοκριμένους χρόνους, η μελέτη των υπολοίπων αυτών περιορίζεται μόνο για τους πλήρεις χρόνους. Συνεπώς, μελετάμε το παρακάτω διάνυσμα Schoenfeld υπολοίπων

$$\hat{\mathbf{r}}_{(i)} = \left(\hat{r}_{(i)1}, \hat{r}_{(i)2}, ..., \hat{r}_{(i)k}\right)' \quad 1 \le i \le d$$

που αντιστοιχεί στον διατεταγμένο πλήρη χρόνο $t_{(i)}$. Επιπλέον, στους πλήρεις χρόνους ισχύουν οι παρακάτω σχέσεις, ως αποτέλεσμα της διαδικασίας υπολογισμού των Schoenfeld υπολοίπων που περιγράφηκε ήδη

$$\mathbf{r}_{(i)} = \mathbf{r}_{(i)}(\mathbf{b}) = \left(\frac{\partial}{\partial \mathbf{b}} \log L_{(i)}(\mathbf{b})\right)_{k \times k} = \left(U_{(i)1}(\mathbf{b}), U_{(i)2}(\mathbf{b}), ..., U_{(i)k}(\mathbf{b})\right)' = \mathbf{Q}_{(i)}(\mathbf{b})$$

όπου $r_{(i)g} = r_{(i)g}(b) = \frac{\partial}{\partial b_g} \log L_{(i)}(b) = U_{(i)g}(b)$

οπότε λαμβάνοντας υπόψη τη μέση τιμή και τον πίνακα διακυμάνσεων-συνδιακυμάνσεων των διανυσμάτων $\mathbf{Q}_{(i)}(\mathbf{b})$ και $\mathbf{U}(\mathbf{b})$ συμπεραίνουμε ότι

$E\left(\mathbf{r}_{(i)}(\mathbf{b})\right) = 0$,

$$V_{(i)}(\mathbf{b}) = V\left(\mathbf{r}_{(i)}(\mathbf{b})\right) = E\left(\mathbf{r}_{(i)}(\mathbf{b})\mathbf{r}_{(i)}'(\mathbf{b})\right) = -\left(E\frac{\partial^2}{\partial b_h \partial b_m} l_{(i)}(\mathbf{b})\right)_{k \times k}, \quad 1 \le h, m \le k$$

$\sum_{i=1}^{d} V_{(i)}(\mathbf{b}) = 0$,

$$\left(\sum_{i=1}^{d} U_{(i)1}(\mathbf{b}), \sum_{i=1}^{d} U_{(i)2}(\mathbf{b}), ..., \sum_{i=1}^{d} U_{(i)k}(\mathbf{b})\right)' = \left(U_1(\mathbf{b}), U_2(\mathbf{b}), ..., U_k(\mathbf{b})\right)' = \mathbf{U}(\mathbf{b})$$

οπότε $\mathbf{U}(\mathbf{b}) = \sum_{i=1}^{d} \mathbf{r}_{(i)}(\mathbf{b})$

Οι Grambsch & Therneau (1994) πρότειναν τα *scaled Schoenfeld residuals* που προκύπτουν από την τυποποίηση των *Schoenfeld residuals* και δίνονται από την σχέση

$$\hat{\mathbf{r}}^{*}_{(i)} = {}_0\hat{\mathbf{V}}^{-1}_{(i)} \cdot \hat{\mathbf{r}}_{(i)} \cong d \cdot \mathbf{I}_0^{-1}\left(\hat{\mathbf{b}}\right) \cdot \hat{\mathbf{r}}_{(i)}, 1 \le i \le d \qquad (5.3.2)$$

όπου ${}_0\hat{\mathbf{V}}_{(i)}\left(\hat{\mathbf{b}}\right) = \left(-\frac{\partial}{\partial b_h \partial b_m} l_{(i)}\left(\hat{\mathbf{b}}\right) \right)_{k\times k} = \begin{pmatrix} {}_0\hat{V}_{(i),11} & {}_0\hat{V}_{(i),12} & \cdots & {}_0\hat{V}_{(i),1k} \\ {}_0\hat{V}_{(i),21} & {}_0\hat{V}_{(i),22} & \cdots & {}_0\hat{V}_{(i),2k} \\ \vdots & \vdots & \ddots & \vdots \\ {}_0\hat{V}_{(i),k1} & {}_0\hat{V}_{(i),k2} & \cdots & {}_0\hat{V}_{(i),kk} \end{pmatrix} = \mathbf{U}_{(i)}\left(\hat{\mathbf{b}}\right)$

με ${}_0\hat{V}_{(i),gg} = \sum_{j\in R_i} \frac{\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)}{\sum_{l\in R_i}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_l\right)} \left(X_{jg} - \frac{\sum_{j\in R_i} X_{ig}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)}{\sum_{j\in R_i}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)} \right)^2 = \sum_{j\in R_i} \hat{w}_{(i)j}\left(X_{jg} - \hat{A}_{(i)g}\right)^2$

και ${}_0\hat{V}_{(i),hm} = \sum_{j\in R_i} \frac{\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)}{\sum_{l\in R_i}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_l\right)} \left(X_{jh} - \frac{\sum_{j\in R_i} X_{ih}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)}{\sum_{j\in R_i}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)} \right)\left(X_{jm} - \frac{\sum_{j\in R_i} X_{im}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)}{\sum_{j\in R_i}\exp\left(\hat{\mathbf{b}}'\mathbf{X}_j\right)} \right)$

$$= \sum_{j\in R_i} \hat{w}_{(i)j}\left(X_{ih} - \hat{A}_{(i)h}\right)\left(X_{jm} - \hat{A}_{(i)m}\right),\ 1 \le h \ne m \le k \text{ και } 1 \le g \le k$$

είναι ο εκτιμημένος παρατηρούμενος πίνακας διακυμάνσεων-συνδιακυμάνσεων και αποτελεί την ασυμπτωτική εκτίμηση του πίνακα διακυμάνσεων-συνδιακυμάνσεων του διανύσματος $\mathbf{r}_{(i)}\left(\mathbf{b}\right)$.

(Αντζουλάκος, 2009)

Παρατήρηση

Η προσέγγιση στην σχέση $(5.3.2)$ προκύπτει από την αντικατάσταση κάθε πίνακα ${}_0\hat{\mathbf{V}}_{(i)}$ με τον πίνακα $\hat{\bar{\mathbf{V}}}$, όπου

$$\hat{\bar{\mathbf{V}}} = \frac{\sum_{i=1}^{d} {}_0\hat{\mathbf{V}}_{(i)}}{d} = \frac{\mathbf{I}_0\left(\hat{\mathbf{b}}\right)}{d} \Rightarrow \hat{\bar{\mathbf{V}}}^{-1} = d \cdot \mathbf{I}_0^{-1}\left(\hat{\mathbf{b}}\right)$$

Η προσέγγιση αυτή βασίζεται στην υπόθεση ότι οι πίνακες ${}_0\hat{\mathbf{V}}_{(i)}$ είναι κατά προσέγγιση ισοδύναμοι. Με την προσέγγιση στοχεύουμε στην μείωση των υπολογισμών, αφού αντικαθιστούμε τον υπολογισμό και την αντιστροφή των d

πινάκων ${}_0\hat{\mathbf{V}}_{(i)}$ με τον υπολογισμό και την αντιστροφή μόνο του πίνακα $\hat{\bar{\mathbf{V}}}$ (Winnett and Sasieni, 2001).

Στην πράξη προτιμάται ο πίνακας $\hat{\bar{\mathbf{V}}}$ αντί του ${}_0\hat{\mathbf{V}}_{(i)}$, ο οποίος γνωρίζουμε ότι είναι η σταθμισμένη διακύμανση των συμμεταβλητών για τα άτομα που βρίσκονται σε κίνδυνο τη στιγμή $t_{(i)}$ με αποτέλεσμα η τιμή του ${}_0\hat{\mathbf{V}}_{(i)}$ να ποικίλει σημαντικά εάν αλλάζει η διακύμανση των τιμών των συμμεταβλητών για τα άτομα του συνόλου κινδύνου R_i, επειδή τα άτομα που πεθαίνουν ή λογοκρίνονται αφαιρούνται από το R_i και έτσι η τιμή της ${}_0\hat{\mathbf{V}}_{(i)}^{-1}$ γίνεται πολύ μικρή. Συνεπώς με την χρήση του $\hat{\bar{\mathbf{V}}}$ αντί του ${}_0\hat{\mathbf{V}}_{(i)}$ περιορίζεται αυτή η ανισορροπία (*instability*). Γενικά, όταν δεν έχουμε λογοκρισία, η διακύμανση των τιμών των συμμεταβλητών των ατόμων που βρίσκονται στο σύνολο κινδύνου μιας χρονικής στιγμής θα τείνει να μειώνεται καθώς το πλήθος των ατόμων σε κίνδυνο μειώνεται και οι συμμεταβλητές είναι σημαντικοί προγνωστικοί παράγοντες του μοντέλου. Ωστόσο, στην περίπτωση που δεν είναι σημαντική η επίδραση κάποιας συμμεταβλητής, τότε καθώς θα μειώνεται το πλήθος των ατόμων σε κίνδυνο, η τιμή της ${}_0\hat{\mathbf{V}}_{(i)}$ θα αλλάζει ελάχιστα με αποτέλεσμα να είναι κοντά στην τιμή του $\hat{\bar{\mathbf{V}}}$ (Winnett et al., 2001).

Με την γραφική αναπαράσταση των Schoenfeld υπολοίπων έναντι του χρόνου ζωής t για κάθε συμμεταβλητή μπορούμε να αξιολογήσουμε την υπόθεση του αναλογικού κινδύνου για κάθε συμμεταβλητή χωριστά ελέγχοντας την υπόθεση $b_g(t) = b_g$, $1 \le g \le k$, δηλαδή η παράμετρος b_g παραμένει σταθερή στον χρόνο, οπότε η αντίστοιχη συμμεταβλητή δεν εξαρτάται από τον χρόνο. Εισάγοντας στο γράφημα μία *smoothing curve* μπορούμε να αξιολογήσουμε την υπόθεση του αναλογικού κινδύνου για κάθε συμμεταβλητή χωριστά ελέγχοντας την υπόθεση της μηδενικής κλίσης της *smoothing curve*, δηλαδή η παράμετρος b_g παραμένει σταθερή στον χρόνο, $b_g(t) = b_g$, οπότε ικανοποιείται η υπόθεση του αναλογικού κινδύνου γι' αυτή την μεταβλητή. Να σημειώσουμε ότι οι δύο αυτοί έλεγχοι είναι ισοδύναμοι (Αντζουλάκος, 2009).

Για να ελέγξουμε αν οι συμμεταβλητές ικανοποιούν ταυτόχρονα την υπόθεση του αναλογικού κινδύνου, δηλαδή να ελέγξουμε την ολική επάρκεια του μοντέλου

αναλογικού κινδύνου (*overall (global) model fit)* προτείνεται από τους Grambsch & Therneau (1994) η παρακάτω στατιστική

$$T=\left(\sum_{i=1}^{d}\left(g\left(t_{(i)}\right)-\overline{g}\right)\hat{\mathbf{r}}'_{(i)}\right)\frac{\hat{\mathbf{r}}^{*}_{(i)}}{\hat{\mathbf{r}}_{(i)}}\frac{1}{\sum_{i=1}^{d}\left(g\left(t_{(i)}\right)-\overline{g}\right)^{2}}\left(\sum_{i=1}^{d}\left(g\left(t_{(i)}\right)-\overline{g}\right)\hat{\mathbf{r}}_{(i)}\right)\overset{\alpha}{\sim}\chi^{2}_{k}$$

με μηδενική υπόθεση $H_0:b_1(t)=b_1,b_2(t)=b_2,...,b_k(t)=b_k$ και συνήθεις επιλογές της συνάρτησης $g(t)$ τις $g(t)=t$ και $g(t)=\ln t$.

Για τους ατομικούς ελέγχους (*individual tests*) που αναφέραμε παραπάνω οι Grambsch & Therneau (1994) πρότειναν την στατιστική συνάρτηση

$$T_g=\frac{\left(\sum_{i=1}^{d}\left(g\left(t_{(i)}\right)-\overline{g}\right)\hat{R}_{(i)g}\right)^{2}}{d\cdot\mathbf{I}_0^{gg}\left(\hat{\mathbf{b}}\right)\cdot\sum_{i=1}^{d}\left(g\left(t_{(i)}\right)-\overline{g}\right)^{2}}\overset{\alpha}{\sim}\chi^{2}_{1}\ ,\ 1\le g\le k$$

με μηδενική υπόθεση $H_0:b_g(t)=b_g$,

όπου $\hat{\mathbf{R}}_{(i)}=\left(\hat{R}_{(i)1},\hat{R}_{(i)2},...,\hat{R}_{(i)k}\right)'=\left(\hat{b}_1+\hat{r}^{*}_{(i)1},\hat{b}_2+\hat{r}^{*}_{(i)2},...,\hat{b}_k+\hat{r}^{*}_{(i)k}\right)'=\hat{\mathbf{b}}+\hat{\mathbf{r}}^{*}_{(i)}$, $1\le i\le d$

είναι το διάνυσμα με τα *rescaled Schoenfeld residuals* που όρισαν οι Grambsch & Therneau (1994). Παρατήρησαν ότι

$$E\left(\hat{\mathbf{R}}_{(i)}\right)=E\left(\hat{\mathbf{b}}+\hat{\mathbf{r}}^{*}_{(i)}\right)\cong\mathbf{b}=\left(b_1,b_2,...,b_k\right)' \qquad (5.3.3)$$

διότι $E\left(\hat{\mathbf{r}}^{*}_{(i)}\right)=E\left(d\cdot I_0^{-1}\left(\hat{\mathbf{b}}\right)\cdot\hat{\mathbf{r}}_{(i)}\right)=d\cdot I_0^{-1}\left(\hat{\mathbf{b}}\right)\cdot E\left(\hat{\mathbf{r}}_{(i)}\right)=d\cdot I_0^{-1}\left(\hat{\mathbf{b}}\right)\cdot 0=0$

οπότε από την σχέση $(5.3.3)$ συμπεραίνουμε ότι

$$E\left(\hat{R}_{(i)g}\right)=b_g\ ,\ 1\le g\le k$$

(Αντζουλάκος, 2009)

5.3.2 Χρήση των διαφορών Δέλτα-βήτα ή scaled score residuals

Με τις Δέλτα-Βήτα διαφορές μπορούμε να εξετάσουμε πόσο επηρεάζει κάθε παρατήρηση την εκτίμηση του διανύσματος παραμέτρων $\mathbf{b}$. Συγκεκριμένα για να εξετάσουμε αυτήν την επιρροή της i παρατήρησης, όπου $1\le i\le n$ αρκεί αρχικά να εκτιμήσουμε το διάνυσμα παραμέτρων $\mathbf{b}$ χρησιμοποιώντας όλες τις παρατηρήσεις (χρόνους ζωής) και στην συνέχεια να εξαιρέσουμε την i παρατήρηση και να

εκτιμήσουμε το διάνυσμα παραμέτρων **b** χρησιμοποιώντας τις υπόλοιπες παρατηρήσεις. Η ποσότητα

$$\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(i)} = \left(\hat{b}_1 - \hat{b}_{(i)1}, \hat{b}_2 - \hat{b}_{(i)2}, ..., \hat{b}_k - \hat{b}_{(i)k} \right)' \quad 1 \le i \le n$$

καλείται Δέλτα-βήτα διαφορά για την i παρατήρηση (Αντζουλάκος, 2009). Γενικά χρησιμοποιείται ο παρακάτω συμβολισμός

$$\Delta_{(i)}\hat{\mathbf{b}} = \hat{\mathbf{b}} - \hat{\mathbf{b}}_{(i)} = \left(\Delta_{(i)1}\hat{b}, \Delta_{(i)2}\hat{b}, ..., \Delta_{(i)k}\hat{b} \right)' = \left(\hat{b}_1 - \hat{b}_{(i)1}, \hat{b}_2 - \hat{b}_{(i)2}, ..., \hat{b}_k - \hat{b}_{(i)k} \right)' \quad 1 \le i \le n$$

Είναι προφανές ότι αν η ποσότητα $\Delta_{(i)}\hat{\mathbf{b}}$ είναι κοντά στο μηδέν, τότε η επιρροή της i παρατήρησης στην εκτίμηση του διανύσματος **b** είναι αμελητέα, ενώ αν η ποσότητα $\Delta_{(i)}\hat{\mathbf{b}}$ είναι μεγάλη, τότε αντίστοιχα μεγάλη θεωρείται και η επιρροή της i παρατήρησης στην εκτίμηση του διανύσματος **b**. Συνολικά πρέπει να εφαρμόσουμε $n+1$ μοντέλα του Cox για να υπολογίσουμε τις ποσότητες $\Delta_{(i)}\hat{\mathbf{b}}$, δηλαδή όσο είναι το πλήθος των παρατηρήσεων για την εκτίμηση των διανυσμάτων $\mathbf{b}_{(i)}$, $1 \le i \le n$ εξαιρώντας κάθε φορά από μία παρατήρηση από το σετ των δεδομένων, συν ένα επιπλέον μοντέλο για την εκτίμηση του διανύσματος **b**, όπου λαμβάνουμε υπόψη όλες τις παρατηρήσεις.

Στην περίπτωση μικρού δείγματος παρατηρήσεων μπορούμε εύκολα να υπολογίσουμε τις Δέλτα-Βήτα διαφορές. Ωστόσο, το ίδιο δεν συμβαίνει στην περίπτωση μεγάλου δείγματος γιατί ο υπολογισμός των ποσοτήτων $\Delta_{(i)}\hat{b}$ είναι ιδιαίτερα χρονοβόρος. Αυτό το πρόβλημα αντιμετωπίζεται με την προσέγγιση των Δέλτα-βήτα διαφορών ως εξής

$$\Delta_{(i)}\mathbf{b} = \hat{\mathbf{b}} - \hat{\mathbf{b}}_{(i)} \cong \mathbf{I}_0^{-1}\left(\hat{\mathbf{b}}\right) \cdot \hat{\mathbf{L}}_i \text{ όπου } \hat{\mathbf{L}}_i = \left(\hat{L}_{i1}, \hat{L}_{i2}, ..., \hat{L}_{ik} \right)' \text{ και } 1 \le i \le n$$

Το διάνυσμα $\hat{\mathbf{L}}_i$ αποτελεί το διάνυσμα των *score residuals* για το i άτομο με διάνυσμα συμμεταβλητών $\mathbf{X} = \left(X_1, X_2, ..., X_k \right)'$. Παρατηρούμε ότι και στα *score residuals* αντιστοιχεί ένα υπόλοιπο σε κάθε συμμεταβλητή για κάθε άτομο. Τα *score residuals* ορίζονται ως εξής

$$L_{ig} = r_{(i)g} - X_{ig} \cdot H\left(t_i / \mathbf{X}_i\right) + \exp\left(\mathbf{b}'\mathbf{X}_i\right) \cdot \sum_{j:t_j \le t_i} \frac{\sum_{l \in R_j} X_{lg} \exp\left(\mathbf{b}'\mathbf{X}_l\right)}{\sum_{l \in R_j} \exp\left(\mathbf{b}'\mathbf{X}_l\right)} \cdot \frac{\delta_j}{\sum_{l \in R_j} \exp\left(\mathbf{b}'\mathbf{X}_l\right)}$$

$$= r_{(i)g} - X_{ig} \cdot H\left(t_i / \mathbf{X}_i\right) + \exp\left(\mathbf{b}'\mathbf{X}_i\right) \cdot \sum_{j:t_j \le t_i} A_{jg} \frac{\delta_j}{\sum_{l \in R_j} \exp\left(\mathbf{b}'\mathbf{X}_l\right)} \quad 1 \le i \le d \quad 1 \le g \le k$$

$$(5.3.4)$$

Συγκεκριμένα για τον υπολογισμό των *score residuals* L_{ig} βασιστήκαμε στο σκορ της g συμμεταβλητής

$$U_g\left(\mathbf{b}\right) = \frac{\partial}{\partial b_g} l\left(\mathbf{b}\right) = \sum_{i=1}^{d} \delta_i \left(X_{ig} - \frac{\sum_{j \in R_i} X_{jg} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)} \right)$$

$$= \sum_{i=1}^{d} \delta_i X_{ig} - \sum_{i=1}^{d} \delta_i \frac{\sum_{j \in R_i} X_{jg} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}$$

$$= \sum_{i=1}^{d} \delta_i X_{ig} - \sum_{i=1}^{d} X_{ig} \exp\left(\mathbf{b}'\mathbf{X}_i\right) \sum_{i:t_i \le t} \frac{\delta_i}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}$$

$$= \sum_{i=1}^{d} X_{ig} \left(\delta_i - \exp\left(\mathbf{b}'\mathbf{X}_i\right) \cdot H_0\left(t_i\right) \right)$$

$$= \sum_{i=1}^{d} X_{ig} \left(\delta_i - H\left(t_i / \mathbf{X}_i\right) \right) = \sum_{i=1}^{d} L_{ig}$$

και στην αναφορική αθροιστική συνάρτηση κινδύνου κατά Breslow, δηλαδή

$$H_0\left(t\right) = \sum_{i:t_i \le t} h_0\left(t\right) = \sum_{i:t_i \le t} \frac{\delta_i}{\sum_{j \in R_i} \exp\left(\mathbf{b}'\mathbf{X}_j\right)}$$

Έπειτα από αρκετές πράξεις καταλήξαμε στην σχέση $(5.3.4)$.

(Αντζουλάκος, 2009)

Παρατήρηση

Για την αξιολόγηση της υπόθεσης του αναλογικού κινδύνου εξετάζονται και άλλα υπόλοιπα σύμφωνα με το βιβλίο του Αντζουλάκου (2009):

1. Cox-Snell residuals

$$\hat{r}_{c_j} = \hat{H}\left(t_j / \mathbf{X}_j\right) = \hat{H}_0\left(t_j\right)\exp\left(\sum_{p=1}^{k}\hat{b}_p X_{jp}\right) \in [0,\infty) \quad 1 \le j \le n$$

όπου δεν λαμβάνονται υπόψη οι λογοκριμένοι χρόνοι στην γραφική παράσταση των $\left(\hat{r}_{c_j}, \delta_j\right)$. Χρησιμοποιούνται για τον έλεγχο της ολικής επάρκειας του μοντέλου αναλογικού κινδύνου.

2. Modified Cox-Snell residuals

$$\hat{r}'_{c_j} = 1 - \delta_j + \hat{r}_{c_j} = \begin{cases} \hat{r}_{c_j} & , \ \delta_j = 1 \\ 1 + \hat{r}_{c_j} & , \ \delta_j = 0 \end{cases} \quad \text{όπου} \quad \hat{r}'_{c_j} = \hat{r}_{c_j} \in [0,\infty) \ \text{ενώ} \ \ \hat{r}'_{c_j} = 1 + \hat{r}_{c_j} \in [1,\infty)$$

Να σημειώσουμε ότι τα *Modified Cox-Snell residuals* είναι τροποποίηση των *Cox-Snell residuals* για να λαμβάνονται υπόψη οι λογοκριμένοι χρόνοι στην γραφική παράσταση τους.

3. Martingale residuals

$$\hat{r}_{M_j} = \delta_j - \hat{r}_{c_j} = \begin{cases} 1 - \hat{r}_{c_j} & , \ \delta_j = 1 \\ -\hat{r}_{c_j} & , \ \delta_j = 0 \end{cases}$$

Ισχύει ότι $\sum_{j=1}^{n}\hat{r}_{M_j} = 0$ όταν εκτιμάμε την $H_0\left(t_j\right)$ με Breslow και επιπλέον ως προς τις τιμές που παίρνουν αυτά τα υπόλοιπα ισχύει ότι

$$\hat{r}_{M_j} = 1 - \hat{r}_{c_j} \in (-\infty, 1] \ \text{ενώ} \ \ \hat{r}_{M_j} = -\hat{r}_{c_j} \in (-\infty, 0]$$

δηλαδή τα *Martingale residuals* που αντιστοιχούν σε λογοκριμένους χρόνους είναι αρνητικά. Χρησιμοποιούνται κυρίως για την εύρεση συναρτησιακής μορφής μίας ερμηνευτικής μεταβλητής, αλλά και για τον εντοπισμό ακραίων τιμών (*outliers*).

4. Deviance residuals

$$\hat{r}_{D_j} = \begin{cases} sign\left(1-\hat{r}_{c_j}\right)\sqrt{-2\left[1-\hat{r}_{c_j}+\log\left(\hat{r}_{c_j}\right)\right]} & , \delta_j = 1 \\ sign\left(-\hat{r}_{c_j}\right)\sqrt{2\hat{r}_{c_j}} & , \delta_j = 0 \end{cases}$$

$$= sign\left(\hat{r}_{M_j}\right)\sqrt{-2\left[\hat{r}_{M_j}+\delta_j \log\left(\delta_j-\hat{r}_{M_j}\right)\right]} \quad \text{όπου} \quad \hat{r}_{D_j} \in \left(-\infty,\infty\right)$$

Για τα υπόλοιπα *deviance ισχύει ότι,* όταν $\hat{r}_{M_j} \in \left(-\infty,0\right)$, τότε το $\hat{r}_{D_j}$ κατευθύνεται προς το μηδέν, ενώ όταν $\hat{r}_{M_j} \in \left(0,1\right)$, τότε το $\hat{r}_{D_j}$ κατευθύνεται προς το $+\infty$. Επειδή κατανέμονται περισσότερο συμμετρικά γύρω από το μηδέν σε σχέση με τα *Martingale residuals*, εντοπίζουν πιο εύκολα τις ακραίες τιμές.

Ωστόσο, για τον υπολογισμό αυτών των υπολοίπων επιβάλλεται η εκτίμηση της αθροιστικής συνάρτησης κινδύνου, από την οποία εξαρτώνται. Επιπλέον, για κάθε άτομο υπολογίζεται μόνο ένα υπόλοιπο χωρίς να λαμβάνουμε υπόψη τις συμμεταβλητές του μοντέλου. Δηλαδή, η τιμή του υπολοίπου που βρήκαμε για κάποιο άτομο θα είναι ίδια για όλες τις συμμεταβλητές που αντιστοιχούν στο άτομο αυτό. Αυτά τα μειονεκτήματα αντιμετωπίζονται με την χρήση των Schoefeld υπολοίπων, διότι ούτε χρειάζεται η εκτίμηση της αθροιστικής συνάρτησης κινδύνου για τον υπολογισμό τους και επιπλέον, για κάθε συμμεταβλητή που αντιστοιχεί σ' ένα άτομο ορίζεται και από ένα υπόλοιπο (Αντζουλάκος, 2009).

5.4 Γραφική αξιολόγηση υπόθεσης αναλογικού κινδύνου

Η σύγκριση γραφικών μεθόδων θεωρείται αρκετά αυθαίρετη, αφού δεν υπάρχουν συγκεκριμένες οδηγίες για το πώς θα ερμηνεύσουμε τα γραφήματα. Τα συμπεράσματα εξαρτώνται κυρίως από την εμπειρία του ερευνητή. Συνεπώς, για να μπορούμε να αξιολογήσουμε τα αποτελέσματα των διαφορετικών μεθόδων, μπορούμε να χρησιμοποιήσουμε κάποιους ελέγχους καλής προσαρμογής που έχουν προταθεί για τον εντοπισμό παραβιάσεων της υπόθεση του αναλογικού κινδύνου (ή PH υπόθεσης). Ωστόσο, κανένας από αυτούς τους ελέγχους δεν χρησιμοποιήθηκε ευρέως κυρίως λόγω μικρής στατιστικής ισχύος. Επιπλέον, οι περισσότεροι από

αυτούς τους ελέγχους βασίζονται στην διαίρεση του χρόνου ζωής σε διαστήματα με αποτέλεσμα να υπάρχει προβληματισμός στην επιλογή του πλήθους και του περιεχομένου των χρονικών αυτών διαστημάτων (Hess, 1995).

Υπάρχουν αρκετές γραφικές μέθοδοι που προτείνονται για την αξιολόγηση της PH υπόθεσης. Παρακάτω παρουσιάζονται κάποιες από αυτές:

1. Γράφημα των αθροιστικών συναρτήσεων κινδύνου έναντι του χρόνου και έλεγχος σταθερότητας του λόγου (*constant ratio*).
2. Γράφημα των αθροιστικών συναρτήσεων κινδύνου μεταξύ τους και έλεγχος σταθερής κλίσης (*constant slope*).
3. Γράφημα των λογαρίθμων των αθροιστικών συναρτήσεων κινδύνου έναντι του χρόνου και έλεγχος παραλληλότητας.
4. Γράφημα διαφορών μεταξύ των λογαρίθμων των αθροιστικών συναρτήσεων κινδύνου έναντι του χρόνου και έλεγχος σταθερότητας (*constancy*).
5. Γράφημα συναρτήσεων επιβίωσης βασισμένων στο μοντέλο παλινδρόμησης του Cox μαζί με Kaplan-Meier συναρτήσεις επιβίωσης.
6. Διαίρεση του άξονα χρόνων ζωής και προσαρμογή μοντέλων χωριστά σε κάθε χρονικό διάστημα.
7. Εισαγωγή όρου αλληλεπίδρασης στο μοντέλο που εξαρτάται από τον χρόνο (*time-by-covariate interaction terms*) και εκτίμηση του λογαρίθμου της συνάρτησης κινδύνου.

Να σημειώσουμε ότι χαρακτηριστικό των μεθόδων αυτών είναι ότι η εκτίμηση των συναρτήσεων επιβίωσης και των αθροιστικών συναρτήσεων κινδύνου που προκύπτει με τη μέθοδο Kaplan-Meier και Nelson-Aalen αντίστοιχα, όπου δεν χρησιμοποιείται η υπόθεση του αναλογικού κινδύνου. Γι' αυτό το λόγο, οι αντίστοιχες γραφικές παραστάσεις χαρακτηρίζονται γενικά ως «εμπειρικές».

Μέθοδος 1: Γράφημα των λογαρίθμων των αθροιστικών συναρτήσεων κινδύνου έναντι του χρόνου και έλεγχος παραλληλότητας.

Η υπόθεση PH υπονοεί ότι οι ο λόγος δύο αθροιστικών συναρτήσεων κινδύνου είναι σταθερός για κάθε t, δηλαδή ισχύει ότι

$$H\left(t/\mathbf{X}\right)=H_0\left(t\right)\exp\left(\mathbf{b}'\mathbf{X}\right)\Rightarrow\frac{H\left(t/\mathbf{X}\right)}{H\left(t/\mathbf{X}^*\right)}=\frac{H_0\left(t\right)\exp\left(\mathbf{b}'\mathbf{X}\right)}{H_0\left(t\right)\exp\left(\mathbf{b}'\mathbf{X}^*\right)}=\mathrm{e}^{\mathbf{b}'\left(\mathbf{X}-\mathbf{X}^*\right)}\quad t\geq 0$$

Εάν οι καμπύλες τέμνονται, τότε αυτό είναι ένδειξη παραβίασης της υπόθεσης PH. Αφού ισχύει $H(t) = -\log S(t)$, τότε μπορούμε να χρησιμοποιήσουμε τον $-\log$ μετασχηματισμό της Kaplan-Meier εκτίμησης γι' αυτήν την αξιολόγηση. Συνεπώς, η PH υπόθεση υπονοεί ότι

$$\log H(t/\mathbf{X}) = \log H_0(t) = \exp(\mathbf{b}'\mathbf{X}) \quad t \geq 0$$

ή αλλιώς $\log\left[-\log S(t/\mathbf{X})\right] = \log\left[-\log S_0(t)\right] + \mathbf{b}'\mathbf{X} \quad t \geq 0$

Επομένως, υπό την PH υπόθεση, γραφήματα των $\log\left[-\log \hat{S}_i(t/\mathbf{X})\right]$ ή ισοδύναμα γραφήματα των $\log \hat{H}_i(t/\mathbf{X})$ έναντι του χρόνου θα πρέπει να είναι «παράλληλα» μεταξύ τους. Στην περίπτωση μίας δίτιμης συμμεταβλητής, μπορούμε να εφαρμόσουμε τους $\log(-\log)$ μετασχηματισμούς των Kaplan-Meier εκτιμήσεων και να ελέγξουμε αν οι καμπύλες ισαπέχουν μεταξύ τους.

Μέθοδος 2: Γράφημα των αθροιστικών συναρτήσεων κινδύνου μεταξύ τους και έλεγχος σταθερής κλίσης (*constant slope*).

Στην περίπτωση μίας δίτιμης συμμεταβλητής αρκετοί συγγραφείς προτείνουν το γράφημα της $H_1(t) = H(t/X=1)$ έναντι της $H_0(t) = H(t/X=0)$ για κάθε $t \geq 0$, γνωστό ως $H-H$ γράφημα (Hess, 1995). Στο γράφημα αυτό ορίζουμε στον άξονα x την ποσότητα $\hat{H}_0(t)$ και στον άξονα y την ποσότητα $\hat{H}_1(t)$, οι οποίες είναι Nelson-Aalen εκτιμήσεις. Υπό την PH υπόθεση ο (σχετικός) λόγος κινδύνων δύο οποιωνδήποτε ατόμων γνωρίζουμε ότι ορίζεται ως $H_1(t)/H_0(t) = e^b = \theta$ και είναι σταθερός για κάθε $t \geq 0$. Ουσιαστικά, παριστάνουμε γραφικά την απλή γραμμική παλινδρόμηση $H_1(t) = \theta \cdot H_0(t)$, $t \geq 0$ με κλίση θ και τεταγμένη μηδέν. Η PH υπόθεση ικανοποιείται όταν η καμπύλη του γραφήματος τείνει να συμπέσει στην $y = x$.

Μέθοδος 3: Γράφημα των αθροιστικών συναρτήσεων κινδύνου έναντι του χρόνου και έλεγχος σταθερότητας του λόγου (*constant ratio*).

Ο Muenz προτείνει την γραφική αναπαράσταση της συνάρτησης $\hat{\omega}(t)$ έναντι του χρόνου ζωής t, όπου $\hat{\omega}(t) = H_1(t)/H_0(t)$ για δίτιμη συμμεταβλητή (Hess, 1995). Ουσιαστικά παριστάνουμε γραφικά τον σχετικό λόγο κινδύνων έναντι του χρόνου ζωής, ο οποίος παραμένει σταθερός με την πάροδο του χρόνου, όταν ισχύει η PH υπόθεση. Για να ικανοποιείται η PH υπόθεση θα πρέπει η καμπύλη του γραφήματος να είναι μία ευθεία γραμμή παράλληλη στον άξονα x για κάθε t.

Μέθοδος 4: Γράφημα διαφορών μεταξύ των λογαρίθμων των αθροιστικών συναρτήσεων κινδύνου έναντι του χρόνου και έλεγχος σταθερότητας (*constancy*).

Υπό την PH υπόθεση, ο Schumacher προτείνει την γραφική αναπαράσταση της συνάρτησης $\hat{\gamma}(t)$ έναντι του χρόνου επιβίωσης t (Hess, 1995), όπου

$$\hat{\gamma}(t) = \log\left[-\log \hat{S}_1(t)\right] - \log\left[-\log \hat{S}_0(t)\right] \quad t \geq 0$$

Εξάλλου γνωρίζουμε ότι $H(t) = -\log S(t)$, οπότε συμπεραίνουμε άμεσα ότι

$$\hat{\gamma}(t) = \log \hat{H}_1(t) - \log \hat{H}_0(t) = \log\left(\hat{H}_1(t)/\hat{H}_0(t)\right) = \log \hat{\omega}(t)$$

Όταν ισχύει η PH υπόθεση η συνάρτηση $\gamma(t) = \log\theta = b$ είναι σταθερή για κάθε t, οπότε οι καμπύλες θα εμφανίζονται παράλληλες μεταξύ τους και θα έχουν απόσταση ίση με b.

Σύμφωνα με Dabrowska (Hess, 1995) προτείνεται το γράφημα της συνάρτησης $\hat{\phi}(t)$ έναντι του t, όπου

$$\hat{\phi}(t) = \left(\hat{H}_1(t) - \hat{H}_0(t)\right)/\hat{H}_0(t) = \hat{H}_1(t)/\hat{H}_0(t) - 1$$

Δηλαδή, υπό την PH υπόθεση ισχύει ότι

$$\hat{\phi}(t) = 1 - \hat{\omega}(t) \quad t \geq 0$$

Οι συναρτήσεις $\omega(t)$, $\gamma(t) = \log\omega(t)$ και $\phi(t) = 1 - \omega(t)$ συνδέονται μεταξύ τους και έτσι δημιουργούν παρόμοια γραφήματα. Λαμβάνοντας υπόψη την ευρύτατη χρήση των $\log\left[-\log\right]$ γραφημάτων επιβίωσης και το γεγονός ότι

$$\gamma(t) = \log\left[-\log S_1(t)\right] - \log\left[-\log S_0(t)\right]$$

τότε συμπεραίνουμε ότι η γραφική αναπαράσταση της συνάρτησης $\hat{\gamma}(t)$ έναντι του t φαίνεται η πιο λογική επιλογή (Hess, 1995). Αυτό το γράφημα δίνει μία άμεση αξιολόγηση της PH υπόθεσης. Αντί να αξιολογήσουμε την παραλληλότητα δύο καμπυλών, αξιολογούμε την σταθερότητα μίας καμπύλης. Υπό την PH υπόθεση, αυτό το γράφημα είναι σταθερό για κάθε t και κεντρισμένο γύρω από το $\hat{b} = \log\left(\hat{H}_1(t)/\hat{H}_0(t)\right)$.

Παρατηρήσεις

1. Στο γράφημα συναρτήσεων επιβίωσης βασισμένων στο μοντέλο παλινδρόμησης του Cox μαζί με Kaplan-Meier συναρτήσεις επιβίωσης δεν είναι πάντα δυνατόν να κρίνουμε αν οι διαφορές μεταξύ των προβλεπόμενων (*predicted, model-based*) και των παρατηρούμενων (*observed, non-model-based*) εκτιμήσεων επιβίωσης είναι αποτέλεσμα της μεταβλητότητας του δείγματος ή πράγματι υφίσταται αυτή η τάση. Ο υπολογισμός διαστημάτων εμπιστοσύνης ίσως βοηθήσει. Ο υπολογισμός διαστημάτων εμπιστοσύνης για τις εκτιμήσεις Kaplan-Meier είναι εύκολη διαδικασία. Ωστόσο, διαστήματα εμπιστοσύνης για τις εκτιμήσεις από το μοντέλο του Cox απαιτούν περίπλοκους υπολογισμούς. Ακόμα και αν ο υπολογισμός των διαστημάτων εμπιστοσύνης ήταν γρήγορος, ωστόσο, η παρουσίαση σε κοινό γράφημα τόσο των δύο καμπυλών όσο και των αντίστοιχων διαστημάτων εμπιστοσύνης θα οδηγούσε σε ένα γράφημα δυσνόητο οπτικά.

2. Το γράφημα που βασίζεται στη διαίρεση του άξονα χρόνων ζωής και την προσαρμογή μοντέλων χωριστά σε κάθε χρονικό διάστημα μειονεκτεί στο γεγονός ότι το αποτέλεσμα εξαρτάται από το πλήθος και την σύσταση των χρονικών διαστημάτων. Για να περιοριστεί αυτή η αδυναμία προτείνεται να υπάρχει ένα συγκρίσιμο πλήθος γεγονότων σε κάθε διάστημα. Για να συμβεί αυτό προτείνεται να επιλέγουμε να χρησιμοποιούμε τα ποσοστημόρια των χρόνων αποτυχιών (*death times*) ως «οδηγό» (*breakpoints*) για τη δημιουργία χρονικών διαστημάτων. Ωστόσο, σε αυτήν την περίπτωση το αποτέλεσμα εξαρτάται άμεσα από τον μηχανισμό παραγωγής λογοκριμένων χρόνων, κάτι το οποίο θεωρείται ανεπιθύμητο. Μία εναλλακτική λύση είναι να επιλέξουμε ως *breakpoints* τα ποσοστημόρια των παρατηρούμενων χρόνων ζωής, οι οποίοι περιλαμβάνουν λογοκριμένους και πλήρεις χρόνους. Όμως, και σε αυτήν την περίπτωση πάλι μένει άλυτο το πρόβλημα επιλογής

αυτών των ποστοστημορίων. Επιπλέον, το πλήθος και η θέση των *breakpoints* εξαρτάται από το μέγεθος του δείγματος χρόνων που μελετάμε.

3. Το γράφημα που αφορά εισαγωγή όρου αλληλεπίδρασης στο μοντέλο που εξαρτάται από τον χρόνο (*time-by-covariate interaction terms*) και εκτίμηση του λογαρίθμου της συνάρτησης κινδύνου βασίζεται σε μία περίπλοκη μεθοδολογία που απαιτεί την εκτίμηση αρκετών παραμέτρων. Επιπλέον, βασίζεται στην μοντελοποίηση μεταβλητών που εξαρτώνται όχι μονότονα από τον χρόνο (*modelling non-monotonic time-dependence*), που αποτελεί έναν όρο για τον οποίον υπάρχουν ελάχιστες αναφορές στην βιβλιογραφία.

Η μεθοδολογία και των τριών παραπάνω γραφικών μεθόδων αξιολόγησης της PH υπόθεσης περιγράφεται ικανοποιητικά από τον Hess (1995).

4. Για να μπορούμε να εξετάσουμε αν μία συνεχής μεταβλητή ικανοποιεί την PH υπόθεση, επιβάλλεται η κατάλληλη κατηγοριοποίησή της σε μικρό αριθμό κατηγοριών (συνήθως δύο ή τρεις), οι οποίες έχουν λογική ερμηνεία (Αντζουλάκος, 2009).

5. Στις μεθόδους που παρουσιάσαμε εξετάσαμε την περίπτωση μίας μεταβλητής. Η αξιολόγηση της PH υπόθεσης για περισσότερες από μία μεταβλητές ταυτόχρονα συνήθως αποφεύγεται διότι δύσκολα ερμηνεύονται τα αποτελέσματα του γραφήματος και επιπλέον υπάρχει το ενδεχόμενο να μην μπορεί να παρουσιαστεί γραφικά κάποιος συνδυασμός επιπέδων όταν δεν υπάρχουν πλήρεις χρόνοι για τον συνδυασμό αυτό, διότι δεν μπορεί να υπολογιστεί ο Kaplan-Meier εκτιμητής (Αντζουλάκος, 2009).

6. Όταν θέλουμε να αξιολογήσουμε την PH υπόθεση για μία μεταβλητή λαμβάνοντας υπόψη (*adjusting for*) κάποιες μεταβλητές για τις οποίες γνωρίζουμε ήδη ότι ικανοποιείται η PH υπόθεση, τότε παρουσιάζουμε γραφικά το στρωματοποιημένο μοντέλο του Cox, το οποίο σε κάθε στρώμα της υπό εξέταση μεταβλητής περιέχει τις μεταβλητές αυτές που δεν εξαρτώνται από τον χρόνο. Στο γράφημα θα εμφανίζονται οι καμπύλες που αντιστοιχούν στα στρώματα της υπό εξέταση μεταβλητής. Αν είναι μεταξύ τους παράλληλες, τότε η μεταβλητή αυτή ικανοποιεί την PH υπόθεση (Αντζουλάκος, 2009).

5.5 Επιλογή μοντέλου (*model selection*)

Η επιλογή του κατάλληλου μοντέλου που θα προσαρμοστεί στα δεδομένα που μελετάμε κρίνεται αναγκαία στην περίπτωση που εξετάζουμε μεγάλο πλήθος μεταβλητών, έτσι ώστε να αποφύγουμε την υπερπροσαρμογή (*overfitting*) και την εκτίμηση ενός κορεσμένου μοντέλου (*saturated*) που θεωρείται υπολογιστικά σημαντικά χρονοβόρο όσο πιο μεγάλο είναι το πλήθος των μεταβλητών που εξετάζουμε. Ωστόσο δεν υπάρχει κάποια «μαγική» μέθοδος που θα μας οδηγήσει με απόλυτη ακρίβεια στο «άριστο» μοντέλο (Yi, 2010).

Συνήθως χρησιμοποιείται το *Likelihood Ratio* τεστ σε συνδυασμό με το κριτήριο πληροφορίας του *Akaike* (*Akaike's information criterion, AIC*), ώστε να επιλέξουμε τις κύριες επιδράσεις και αλληλεπιδράσεις που θα εισάγουμε στο μοντέλο (Αντζουλάκος, 2009).

Οι πιο γνωστές μέθοδοι επιλογής μοντέλου είναι (Yi, 2010):

- η *stepwise regression*, όπου ξεκινάμε από το μοντέλο που δεν περιέχει καμία μεταβλητή, εισάγουμε και εξετάζουμε μία-μία τις μεταβλητές και βήμα-βήμα ελέγχουμε αν πρέπει να καταργήσουμε κάποια μεταβλητή. Σταματάμε, όταν δεν χρειάζεται να καταργηθεί κάποια μεταβλητή.
- η *backward elimination*, όπου ξεκινάμε από το πλήρες μοντέλο και βήμα-βήμα απορρίπτουμε μία-μία τις μεταβλητές έως ότου να μην απορρίπτεται καμία μεταβλητή.
- η *forward selection*, η οποία εφαρμόζεται όπως η *stepwise regression* μόνο που δεν καταργούμε μεταβλητές.

Μειονέκτημα των παραπάνω μεθόδων είναι ότι πρέπει να εξετάζουμε μία μεταβλητή την φορά και αυτό κάνει τη διαδικασία χρονοβόρα.

5.5.1 Πρώτη διαδικασία επιλογής μεταβλητών

Στην συνέχεια περιγράφουμε την προσέγγιση του Collett (2003) για την επιλογή του κατάλληλου μοντέλου (Yi, 2010). Στην προσέγγιση αυτή θεωρούμε ότι όλες οι μεταβλητές έχουν την ίδια «τύχη» να εισαχθούν στο μοντέλο και υποθέτουμε ότι δεν υπάρχει κάποιος *a priori* λόγος να περιλάβουμε στο μοντέλο συγκεκριμένες μεταβλητές, όπως η θεραπεία.

ΒΗΜΑ 1: Προσαρμόζουμε ένα μονομεταβλητό μοντέλο (*univariate model*) για κάθε μεταβλητή και καθορίζουμε ποιες μεταβλητές θεωρούνται σημαντικές, άρα πρέπει να εισαχθούν στο μοντέλο, χρησιμοποιώντας ένα προκαθορισμένο επίπεδο αναφοράς, έστω $a_1 = 0.20$.

ΒΗΜΑ 2: Προσαρμόζουμε ένα πολυμεταβλητό μοντέλο (*multivariate model*) με όλες τις μεταβλητές που κρίθηκαν σημαντικές στο προηγούμενο βήμα. Χρησιμοποιούμε *backward elimination* για να καταργήσουμε τις μη σημαντικές μεταβλητές σε επίπεδο σημαντικότητας $a_2 = 0.10$.

ΒΗΜΑ 3: Προσαρμόζουμε το μοντέλο που βρήκαμε στο προηγούμενο βήμα και χρησιμοποιούμε *forward selection* για να ελέγχουμε αν πρέπει να εισαχθούν κάποιες από τις μεταβλητές που καταργήσαμε στο **ΒΗΜΑ 1**. Χρησιμοποιούμε επίπεδο σημαντικότητας $a_3 = 0.10$.

ΒΗΜΑ 4: Προσαρμόζουμε το μοντέλο που βρήκαμε στο προηγούμενο βήμα και εφαρμόζουμε *stepwise regression* για να εξετάζουμε αν πρέπει να καταργήσουμε τις μεταβλητές που θεωρούμε μη σημαντικές και να εισάγουμε κάποιες μεταβλητές που θεωρούμε σημαντικές σε επίπεδο σημαντικότητας $a_4 = 0.10$. Σε αυτό το βήμα εξετάζουμε αν μπορούν να εισαχθούν αλληλεπιδράσεις που ορίζονται μεταξύ των κύριων επιδράσεων που υπάρχουν ήδη στο μοντέλο σύμφωνα με την ιεραρχική αρχή (*hierarchical principle*).

Ο Collett προτείνει την χρήση του *Likelihood Ratio* τεστ για την επιλογή και την κατάργηση των μεταβλητών σε κάθε βήμα. Ωστόσο, μία επιπρόσθετη βοήθεια στην επιλογή των μεταβλητών δίνει το κριτήριο πληροφορίας *AIC*, το οποίο ορίζεται από την σχέση $AIC = -2\log L + k \cdot p$, όπου p είναι το πλήθος των μεταβλητών που υπάρχουν στο μοντέλο μέσω του οποίου υπολογίσαμε την συνάρτηση πιθανοφάνειας L και k είναι μία προκαθορισμένη σταθερά με τιμές $2 \le k \le 6$. Συνήθως χρησιμοποιείται η σταθερά $k = 3$. Σύμφωνα με το κριτήριο AIC καθώς εισάγουμε μεταβλητές στο μοντέλο η τιμή του θα μειώνεται. Από ένα σημείο και μετά μόλις η τιμή του αρχίζει να αυξάνεται, τότε αυτό είναι ένδειξη ότι δεν πρέπει να εισαχθούν άλλες μεταβλητές στο μοντέλο. Προφανώς, επιλέγουμε το μοντέλο που αντιστοιχεί στην μικρότερη τιμή του AIC (Αντζουλάκος, 2009).

5.5.2 Δεύτερη διαδικασία επιλογής μεταβλητών

Στην περίπτωση αυτή έχουμε αποφασίσει *a priori* ποιες μεταβλητές θα εισάγουμε στο μοντέλο οπωσδήποτε. Συνεπώς, σκοπός μας είναι να ελέγξουμε ποιες από τις υπόλοιπες μεταβλητές μπορούμε να εισάγουμε στο μοντέλο. Ακολουθούμε τα βήματα 1-4 και μόλις εισάγουμε και τις προεπιλεγμένες μεταβλητές, τότε εξετάζουμε αν μπορούμε να εισάγουμε αλληλεπιδράσεις σύμφωνα πάντα με την ιεραρχική αρχή (Αντζουλάκος, 2009).

Να σημειώσουμε ότι μπορούν να προκύψουν αλληλεπιδράσεις μεταξύ παραγόντων και μίας μεταβλητής που δεν είναι παράγοντας σχηματίζοντας μικτούς όρους (*mixed terms*). Δεν μπορούν να προκύψουν αλληλεπιδράσεις από μεταβλητές που δεν είναι παράγοντες. Επιπλέον, οι αλληλεπιδράσεις τουλάχιστον τριών μεταβλητών είναι δύσκολο να ερμηνευτούν. Συνεπώς, όσο πιο μικρής τάξης αλληλεπιδράσεις υπάρχουν στο μοντέλο τόσο πιο κατανοητή είναι η ερμηνεία του. Προφανώς, οι αλληλεπιδράσεις δύο μεταβλητών παρέχουν την πιο κατανοητή ερμηνεία, καθώς εκφράζουν πως επηρεάζει ο ένας παράγοντας την μεταβλητή απόκρισης για κάθε επίπεδο του άλλου παράγοντα (Αντζουλάκος, 2009).

Γενικές παρατηρήσεις

Στην συνέχεια δίνονται κάποιες γενικές παρατηρήσεις σχετικά με όσα αναφέρθηκαν στην παρούσα ενότητα σύμφωνα με το βιβλίο του Αντζουλάκου (2009).

1. Όταν ένας παράγοντας Α με a επίπεδα αλληλεπιδρά με έναν άλλο παράγοντα Β με b επίπεδα, τότε στο μοντέλο εισάγουμε συνολικά $(a-1)(b-1)$ αλληλεπιδράσεις και όχι $a \cdot b$, διότι σε κάθε μεταβλητή θεωρούμε ένα επίπεδό της (συνήθως το πρώτο) ως επίπεδο αναφοράς.

2. Το μέγεθος τόσο της εκτίμησης των συντελεστών παλινδρόμησης όσο και του τυπικού σφάλματος της εκτίμησης καθορίζουν αν το μοντέλο μας είναι επαρκές ή πλεονάζον ως προς τις μεταβλητές που περιέχει. Συγκεκριμένα, ένας συντελεστής παλινδρόμησης με ιδιαίτερα μεγάλη εκτίμηση κατ' απόλυτη τιμή και μεγάλο τυπικό σφάλμα είναι ένδειξη ότι στο μοντέλο έχει γίνει υπερβολική προσαρμογή (*overfitting*), οπότε αυτή η μεταβλητή δεν πρέπει να εισαχθεί στο μοντέλο.

3. Είναι δυνατόν στο μοντέλο να συναντήσουμε ισχυρά συσχετισμένες μεταβλητές. Ισχυρά συσχετισμένες θεωρούνται οι μεταβλητές οι οποίες φαίνονται μη σημαντικές όταν ελέγχονται χωριστά, ενώ φαίνονται σημαντικές όταν ελέγχονται ταυτόχρονα στο μοντέλο.

5.6 Μοντέλο αναλογικού κινδύνου του Cox σε δεδομένα υψηλής διάστασης

Το μοντέλο αναλογικού κινδύνου του Cox εφαρμόζεται παραδοσιακά στην περίπτωση που το πλήθος των δειγμάτων n υπερέχει σημαντικά του πλήθους των γονιδίων p. Στην αντίθετη περίπτωση, το σετ δεδομένων θεωρείται υψηλής διάστασης (*high-dimensionality*) και χρησιμοποιείται ο συμβολισμός $p \gg n$. Ιδιαίτερο χαρακτηριστικό αυτών των δεδομένων είναι ότι οι γονιδιακές εκφράσεις είναι συχνά υψηλά συσχετισμένες μεταξύ τους με αποτέλεσμα να αυξάνεται το ενδεχόμενο πολυσυγγραμικότητας (*collinearity*) μεταξύ των ερμηνευτικών μεταβλητών. Στόχος είναι να ληφθεί υπόψη η λογοκρισία και να μειωθεί η διάσταση των δεδομένων, έτσι ώστε να εφαρμοστεί κατάλληλη μέθοδος της Ανάλυσης Επιβίωσης, όπως το πολυμεταβλητό μοντέλο του Cox (Wieringen et al. 2009).

Για τη μείωση της διάστασης των δεδομένων εφαρμόζεται συνήθως η Ανάλυση σε Κύριες Συνιστώσες (*Principal Component Analysis, PCA*) ή η Αποσύνθεση Ιδιόμορφων Τιμών (*Singular* Value *Decomposition, SVD*) (Shoemaker and Lin, 2005). Ωστόσο, και οι δύο αυτές μέθοδοι έχουν δύο κύρια μειονεκτήματα (Shoemaker and Lin, 2005):

- Δύσκολα μπορούμε να ορίσουμε κάποιο όνομα ή ερμηνεία στις κύριες συνιστώσες ή στα ιδιόμορφα διανύσματα, όπου είναι δυνατόν να μην συσχετίζονται υψηλά με την μεταβλητή απόκρισης.
- Οι συντελεστές των κυρίων συνιστωσών ή των ιδιόμορφων διανυσμάτων που χρησιμοποιούνται στον εντοπισμό των γονιδίων που συνεισφέρουν περισσότερο στη συνολική διακύμανση των δεδομένων, μερικές φορές δεν μπορούν να εντοπίσουν τα πιο «κυρίαρχα» γονίδια.

Αν σκοπός της μελέτης είναι να κατασκευάσουμε ένα μοντέλο ακριβούς πρόβλεψης αδιαφορώντας για την ερμηνευτικότητά του, τότε οι παραπάνω μέθοδοι μπορούν να εφαρμοστούν και να δώσουν πολύ καλά αποτελέσματα (Shoemaker and Lin, 2005).

Υπάρχουν και πιο περίπλοκες μέθοδοι για τη μείωση της διάστασης των δεδομένων, όπου κάποιες από αυτές χρησιμοποιούν και χρόνους ζωής και επιδιώκουν την ομαδοποίηση γονιδίων παρόμοιας έκφρασης, ώστε να δημιουργήσουν «μεταγονίδια» (*metagenes*) ή αλλιώς «υπεργονίδια» (*supergenes*) (Shoemaker and Lin, 2005). Στη συνέχεια περιγράφονται ορισμένες μέθοδοι που λαμβάνουν υπόψη την υψηλή διάσταση των δεδομένων:

5.6.1 Συστάδες ως προβλεπτικοί παράγοντες

Αρχικά ομαδοποιούμε τα δείγματα (δηλαδή, το προφίλ εκφράσεων κάθε ασθενή) σε συστάδες σύμφωνα με κάποια ιεραρχική μέθοδο ομαδοποίησης και χρησιμοποιούμε αυτήν την ομαδοποίηση ως προγνωστικό παράγοντα στο μοντέλο αναλογικού κινδύνου του Cox. Το όνομα (*label*) κάθε συστάδας ουσιαστικά ανακεφαλαιώνει την πληροφορία πρόβλεψης που παρέχουν οι εκφράσεις γονιδίων (δηλαδή, συστάδα υψηλής, μέτριας ή χαμηλής πρόγνωσης) (Wieringen et al., 2009).

- **Μειονεκτήματα**

Ωστόσο, αυτό το μοντέλο δεν είναι κατάλληλο για πρόβλεψη της επιβίωσης, διότι με την ομαδοποίηση των ασθενών χάνουμε πολύτιμη πληροφορία για το προφίλ τους, η οποία «καλύπτεται» από το όνομα της συστάδας στην οποία ανήκουν και πάντα υπάρχει ο κίνδυνος σχηματισμού συστάδων μικρής εκτός-συστάδας μεταβλητότητας και μεγάλης εντός-συστάδας μεταβλητότητας. Δηλαδή, η μετατροπή συνεχών μεταβλητών σε παράγοντες έχει αποδειχθεί ως μη κατάλληλη (Wieringen et al., 2009). Επιπλέον, το πλήθος συστάδων που θα επιλέξουμε και ο αλγόριθμος ομαδοποίησης που θα χρησιμοποιήσουμε δεν είναι πάντα προφανή.

Παρόλα αυτά, με το μοντέλο αυτό μπορούμε να τοποθετούμε κάθε νέο ασθενή στην κατάλληλη συστάδα σύμφωνα με την εκτιμώμενη Kaplan-Meier καμπύλη κάθε συστάδας, όπου μας δίνει τον εκτιμώμενο χρόνο επιβίωσης κάθε ασθενή.

5.6.2 Υπό επιτήρηση επιλογή συστάδων γονιδιακών εκφράσεων με μορφή δέντρων (Supervised harvesting of expression trees)

Αρχικά ομαδοποιούμε τα p γονίδια σύμφωνα με κάποια ιεραρχική μέθοδο ομαδοποίησης σε $p + p - 1 = 2p - 1$ συστάδες. Η ομαδοποίηση αυτή θα είναι ο προγνωστικός παράγοντας της επιβίωσης. Σε κάθε συστάδα υπολογίζουμε το μέσο προφίλ έκφρασης κάθε ασθενή για τα γονίδια που περιέχει, οπότε προκύπτουν $2p-1$ νέοι προγνωστικοί παράγοντες. Χρησιμοποιούμε τους νέους αυτούς $2p-1$ προγνωστικούς παράγοντες για να δημιουργήσουμε το μοντέλο αναλογικού κινδύνου του Cox περιλαμβάνοντας και αλληλεπιδράσεις πρώτης τάξης. Προτείνεται να αποφεύγεται η εισαγωγή αλληλεπιδράσεων μεγαλύτερης τάξης, διότι ο υπολογισμός τους είναι χρονοβόρος και η ερμηνεία τους δύσκολη (Wieringen et al., 2009). Στην συνέχεια επιλέγουμε τις k συστάδες που θα εισάγουμε στο τελικό μοντέλο αναλογικού κινδύνου χρησιμοποιώντας την *forward selection* ή *backward elimination* μέθοδο. Έπειτα, το τελικό μοντέλο επιλέγεται με την μέθοδο της «$k-$φορές διασταυρούμενης αξιολόγησης» ($k-$*fold cross-validation)*.

➢ **Μειονεκτήματα**

Η συγκεκριμένη μέθοδος έχει δεχτεί έντονη κριτική. Οι συγγραφείς της μεθόδου προτείνουν την εφαρμογή της σ' ένα μεγάλο δείγμα ασθενών, ώστε να μπορούν να εντοπιστούν οι αλληλεπιδράσεις (Wieringen et al., 2009). Άλλα δύο μειονεκτήματα που παρουσιάζει αυτή η μέθοδος είναι η ευαισθησία του μοντέλου αναλογικού κινδύνου του Cox στην μέθοδο ομαδοποίησης που εφαρμόστηκε στο πρώτο βήμα και η έντονη ετερογένεια που μπορεί να χαρακτηρίσει τις συστάδες (Wieringen et al., 2009).

5.6.3 Μονομεταβλητή επιλογή γονιδίων (Univariate gene selection)

Σύμφωνα με τη συγκεκριμένη μέθοδο λαμβάνουμε υπόψη το p-value κάθε γονιδίου μέσω του στατιστικού του Wald στο απλό μοντέλο παλινδρόμησης του Cox. Στατιστικά σημαντικά θεωρούνται τα γονίδια με p-value μικρότερο από το προκαθορισμένο επίπεδο σημαντικότητας, οπότε και επιλέγονται. Στην συνέχεια

χρησιμοποιούμε τα επιλεγμένα γονίδια στο πολλαπλό μοντέλο αναλογικού κινδύνου του Cox (Wieringen et al., 2009).

➢ **Μειονεκτήματα**

Η συγκεκριμένη μέθοδος δεν λαμβάνει υπόψη τη συσχέτιση μεταξύ των γονιδίων με αποτέλεσμα να υπάρχει το ενδεχόμενο να επιλεγούν υψηλά συσχετισμένα γονίδια. Συνέπεια αυτής της υψηλής συσχέτισης είναι ότι πολλά γονίδια που στην απλή παλινδρόμηση του Cox βρέθηκαν στατιστικά σημαντικά, είναι δυνατόν να μην είναι σημαντικά στο πολλαπλό μοντέλο αναλογικού κινδύνου (Wieringen et al., 2009).

5.6.4 Υπό επιτήρηση Ανάλυση σε Κύριες Συνιστώσες (Supervised Principal Component Analysis, SuperPC)

Ουσιαστικά η SuperPC είναι η τροποποίηση της γνωστής PCA, διότι αντιμετωπίζει την υψηλή διάσταση ενός συνόλου δεδομένων γονιδιακών εκφράσεων χρησιμοποιώντας όχι όλα τα γονίδια, αλλά μόνο εκείνα που εκτιμάται ότι έχουν υψηλή συσχέτιση με τον χρόνο επιβίωσης. Αυτά τα γονίδια-συνιστώσες χρησιμοποιούνται στο μοντέλο αναλογικού κινδύνου του Cox. Αρχικά η SuperPC υπολογίζει το (απόλυτο) σκορ στατιστικό του Cox (*Cox absolute score statistic*), το οποίο εκφράζει το βαθμό της μονομεταβλητής σχέσης μεταξύ κάθε γονιδίου και του χρόνου επιβίωσης. Έπειτα, μέσω *cross-validation* καθορίζεται το επίπεδο αναφοράς (*threshold*) αυτών των σκορ. Στη συνέχεια κατασκευάζουμε έναν μειωμένο πίνακα εκφράσεων που περιέχει μόνο εκείνα τα γονίδια με Cox-σκορ μεγαλύτερο του επιπέδου αναφοράς. Εφαρμόζουμε *PCA* σ' αυτόν τον μειωμένο πίνακα για να ορίσουμε τους προγνωστικούς παράγοντες που θα περιλάβουμε στο μοντέλο αναλογικού κινδύνου (Wieringen et al., 2009).

➢ **Μειονεκτήματα**

Να σημειώσουμε ότι οι κύριες συνιστώσες είναι ένας σταθμισμένος μέσος (*weighted average*) των αρχικών προφίλ εκφράσεων και ερμηνεύονται ως «ιδιογονίδια» (*eigengenes*), «υπεργονίδια» (*supergenes*) ή ««μεταγονίδια» (*metagenes*). Αυτή η ερμηνεία των κυρίων συνιστωσών δεν έχει κάποιο ουσιαστικό

περιεχόμενο (*content*), αφού δεν έχει κάποια θεωρητική βάση ούτε κάποιο βιολογικό χαρακτηριστικό. Συνήθως, οι συνιστώσες δεν έχουν κάποια ιδιαίτερη ερμηνεία, κυρίως εάν είναι μεγάλος ο αριθμός των γονιδίων που συντελούν τη συνιστώσα (Wieringen et al., 2009).

Πέρα από την ερμηνευτική δυσκολία των συνιστωσών, η SuperPC μέθοδος μπορεί να αποτελέσει άριστη μέθοδο πρόβλεψης της επιβίωσης, διότι όχι μόνο μειώνει ένα σετ δεδομένων υψηλής διάστασης στους πιο σημαντικούς παράγοντες, αλλά επιπλέον επιτρέπει να παραστήσουμε γραφικά αυτούς τους παράγοντες (Wieringen et al., 2009). Στην R χρησιμοποιείται η εντολή *superpc*( ) γι' αυτή τη μέθοδο (Tibshirani, 2004).

5.6.5 Μερική παλινδρόμηση του Cox (Partial Cox regression)

Οι Nguyen & Rocke (2002) πρότειναν την εφαρμογή του αλγορίθμου των μερικών ελαχίστων τετραγώνων (*Partial Least Squares, PLS*) για την πρόβλεψη του χρόνου επιβίωσης σε δεδομένα που προέρχονται από μικροσυστάδες. Ωστόσο, αυτός ο αλγόριθμος δεν χειριζόταν ικανοποιητικά τα λογοκριμένα δεδομένα με αποτέλεσμα να μελετηθούν κάποιες τροποποιήσεις του *PLS* αλγόριθμου, ώστε να χρησιμοποιεί σωστά τα λογοκριμένα δεδομένα. Μεταξύ αυτών των τροποποιήσεων ξεχωρίζει η μέθοδος του Bastien, διότι οι ιδιότητές του είναι κατανοητά διατυπωμένες και επιπλέον ο αλγόριθμος της μεθόδου δεν περιέχει κάποιο άριστο επίπεδο επανάληψης (*iterative optimization step)* σε αντίθεση με την μέθοδο του Park (Wieringen et al., 2009).

Συνοπτικά, αναφέρουμε ότι η *PLS* είναι μία τεχνική μείωσης της διάστασης ενός συνόλου δεδομένων με τη βοήθεια της επιτήρησης, όπου μπορεί να χρησιμοποιηθεί στο να συνδέσουμε μία μεταβλητή απόκρισης με τις ερμηνευτικές μεταβλητές. Όταν εφαρμόζουμε αυτή τη μέθοδο σε προβλήματα παλινδρόμησης, που δεν περιέχουν λογοκριμένα δεδομένα, τότε δημιουργούμε νέες συνιστώσες (*components)* που είναι ορθογώνιοι γραμμικοί συνδυασμοί των μεταβλητών που έχουν μέγιστη συνδιακύμανση με την μεταβλητή απόκρισης. Να σημειώσουμε ότι η *PLS* διαφέρει από την *PCA*, διότι οι συνιστώσες που προκύπτουν σύμφωνα με την *PLS* έχουν την μέγιστη (*maximal*) συνδιακύμανση αντί διακύμανση με την μεταβλητή απόκρισης.

Επιπλέον, σε αντίθεση με την *PCA* είναι μία μέθοδος μείωσης της διάστασης με επιτήρηση (*Supervised dimension reduction method)*. Το «βάρος» (*loading)* που δίνεται σε κάθε ερμηνευτική μεταβλητή στο ν' αποτελέσει μέρος μίας συνιστώσας, είναι μία μη-γραμμική συνάρτηση της ερμηνευτικής μεταβλητής και της μεταβλητής απόκρισης και υπολογίζεται επαρκώς μέσω ενός επαναληπτικού αλγόριθμου (Wieringen et al., 2009).

Σύμφωνα με την μέθοδο του Bastien τροποποιούμε το κλασικό *PLS* μοντέλο αντικαθιστώντας τη γραμμική παλινδρόμηση με την Cox παλινδρόμηση για να βρούμε τους *PLS* συντελεστές. Η πρώτη *PLS* συνιστώσα είναι ένα άθροισμα σταθμισμένων κεντραρισμένων γονιδιακών εκφράσεων, όπου ως j στάθμιση ορίζεται η συνδιακύμανση της κεντραρισμένης αποκριτικής μεταβλητής με τις κεντραρισμένες εκφράσεις του j γονιδίου, όπου $j = 1, 2, ..., p$. Αυτές οι σταθμίσεις συμπίπτουν με τους αντίστοιχους συντελεστές παλινδρόμησης του μονομεταβλητού (ή απλού) μοντέλου αναλογικού κινδύνου. Στην συνέχεια σχηματίζουμε p μοντέλα παλινδρόμησης όπου το γονίδιο αποτελεί την αποκριτική μεταβλητή, ενώ η πρώτη συνιστώσα αποτελεί την ερμηνευτική μεταβλητή. Με αυτά τα μοντέλα παλινδρόμησης κατασκευάζουμε την επόμενη συνιστώσα, η οποία είναι ένα άθροισμα σταθμισμένων υπολοίπων (*residuals)*, όπου ως j στάθμιση ορίζεται η συνδιακύμανση του υπολοίπου y_1 (αυτό το υπόλοιπο αφορά το μοντέλο παλινδρόμησης της κεντραρισμένης αποκριτικής μεταβλητής y έναντι της πρώτης συνιστώσας, δηλαδή $y = c_1 t_1 + y_1$) με το j υπόλοιπο x_{1j}, $j = 1, 2, ..., p$. Αυτή η διαδικασία συνεχίζεται έως ότου σχηματιστούν k συνιστώσες, όπου το k καθορίζεται μέσω *cross-validation* (Bastien et al., 2005).

➢ **Μειονεκτήματα**

Η *PLS* έχει λάβει κριτική για κάποιες ανεπιθύμητες ιδιότητες συρρίκνωσης (*shrinkage properties)* σύμφωνα με Butler. Ωστόσο, δεν είναι ξεκάθαρο εάν αυτή η κριτική αφορά τη γενικευμένη *PLS* μέθοδο. Όπως και με την *SuperPC* μέθοδο, οι συνιστώσες δύσκολα ερμηνεύονται (Wieringen et al., 2009).

Παρά τα μειονεκτήματα που παρουσιάζει αποτελεί μία άριστη μέθοδο μείωσης της διάστασης, γραφικής αναπαράστασης των συνιστωσών και πρόβλεψης του χρόνου επιβίωσης (Wieringen et al., 2009).

Κεφάλαιο 6

Εφαρμογή της Ανάλυσης Επιβίωσης

6.1 Εισαγωγή: Τα τερματικά σημεία και οι κλινικοπαθολογικές και ιστολογικές παράμετροι της μελέτης

Μετά την χειρουργική αφαίρεση του καρκινώματος στο στήθος, οι γυναίκες εισέρχονται στην έρευνα και παρακολουθείται η πορεία της υγείας τους. Για την χειρουργική αφαίρεση των όγκων εφαρμόστηκε τροποποιημένη ριζική μαστεκτομή (*modified radical mastectomy*) ή κλασική χειρουργική επέμβαση (*breast conversing surgery*). Η παρακολούθηση των γυναικών αυτών ξεκίνησε τον Οκτώβριο του 1997 και κράτησε περίπου δέκα χρόνια. Να σημειωθεί ότι η παρακολούθηση κάθε ασθενούς δεν συμπίπτει με την χρονική έναρξη της έρευνας.

Για κάθε ασθενή καταγράφηκε ο συνολικός χρόνος ζωής της (*overall survival, OS*), δηλαδή το διάστημα από την εισαγωγή της στην έρευνα έως τον θάνατό της από οποιαδήποτε αιτία. Επιπλέον καταγράφηκε και το χρονικό διάστημα χωρίς ασθένεια (*disease-free survival, DFS*), δηλαδή το διάστημα από την εισαγωγή της ασθενούς στην έρευνα έως την εμφάνιση δεύτερου καρκινώματος όχι απαραίτητα στο στήθος (*secondary malignancy*) ή την επανεμφάνιση καρκίνου στο μαστό (*relapse*) ή τον θάνατο λόγω καρκίνου στο μαστό ή τον θάνατο από οποιαδήποτε αιτία μη σχετιζόμενη με τον καρκίνο στον μαστό, οποιοδήποτε από τα παραπάνω γεγονότα συμβεί πρώτο. Λόγω της μη σύμπτωσης της χρονικής έναρξης της έρευνας με την χρονική στιγμή εισαγωγής κάθε γυναίκας στην έρευνα αυτή αντιστοιχούν σε κάθε ασθενή συγκεκριμένες ημερομηνίες εξέτασής της (*dates of contact*).

Στην συνέχεια εφαρμόστηκε τυχαιοποιημένη δοκιμή για να καθοριστεί η ομάδα των ασθενών στην οποία χορηγείται χημειοθεραπεία με πακλιταξέλη (*adjuvant chemotherapy with paclitaxel, E-T-CMF*) και η ομάδων ασθενών όπου χορηγείται χημειοθεραπεία χωρίς πακλιταξέλη (*adjuvant chemotherapy without paclitaxel, E-*

CMF) με σκοπό την εξάλειψη ενδεχόμενης μετάστασης της ασθένειας. Για την μελέτη της επίδρασης της χημειοθεραπείας στην επιβίωση των ασθενών συμπεριλήφθηκαν αρκετές κλινικοπαθολογικές και ιστολογικές παράμετροι (*clinicopathological parameters*), όπως το μέγεθος του όγκου σε cm (*tumor size*), η ταξινόμηση του όγκου ανάλογα με την δομή και ανάπτυξη των κυττάρων του (*tumor grade*), το πλήθος θετικών αδενωμάτων (*nodes*), η πρωτεΐνη VEGF, η κατάσταση των βιοδεικτών (*biomarkers*) ER-mRNA (*estrogen receptor*), PgR-mRNA (*progesterone receptor*) και HER2 (*Human Epidermal growth factor Receptor-type 2*), αλλά και ένα σύνολο εκφράσεων 37 γονιδίων. Η κατάσταση των βιοδεικτών ER-mRNA και PgR-mRNA διακρίνεται σε θετική (*over-expressed*) και αρνητική (*under-expressed*), ενώ του βιοδείκτη HER2 διακρίνεται σε υπερέκφραση (*over-expression*) και μη-υπερέκφραση (*no over-expression*). Η αρνητική κατάσταση των ER-mRNA και PgR-mRNA, αλλά και η μη-υπερέκφραση του HER2 αποτελούν ένδειξη δυσοίωνης πρόγνωσης για την επιβίωση των ασθενών.

Στην ανάλυση επιβίωσης των 316 συνολικά ασθενών μας ενδιαφέρει να μελετήσουμε την προγνωστική δύναμη των βιοδεικτών στον χρόνο ζωής των ασθενών που έχουν λάβει είτε E-T-CMF είτε E-CMF λαμβάνοντας υπόψη παράλληλα τις παραπάνω κλινικοπαθολογικές και ιστολογικές παραμέτρους.

6.2 Κατανομή των ασθενών στους βιοδείκτες της μελέτης με βάση τα τερματικά τους σημεία

Μελετώντας τον συνολικό χρόνο επιβίωσης και τον χρόνο επιβίωσης χωρίς ασθένεια παρατηρούμε ότι 83 ασθενείς απεβίωσαν από οποιαδήποτε αιτία (δηλαδή το 26.3% του συνόλου των ασθενών) και 118 ασθενείς (37.3%) υποτροπίασαν ή εμφάνισαν δεύτερο όγκο ή απεβίωσαν από την ασθένεια ή από οποιαδήποτε άλλη αιτία αντίστοιχα. Δηλαδή, οι επιζώντες με ασθένεια είναι συνολικά 118-83=35 (11%).

Ο πίνακας 6.1 περιέχει το πλήθος και ποσοστό ασθενών που απεβίωσαν ή επιβίωσαν με ασθένεια για κάθε βιοδείκτη.

Πίνακας 6.1

	ER-mRNA			PgR-mRNA		
	Relapse	**Death**	**Total**	**Relapse**	**Death**	**Total**
Positive	7.7% (24)	17.3% (54)	**25% (78)**	7.8% (24)	14.2% (44)	**22% (68)**
Negative	3.5% (11)	8.7% (27)	**12.2% (38)**	3.5% (11)	12% (37)	**15.5% (48)**
Total	**11.2% (35)**	**26% (81)**	**37.2% (116)**	**11.3% (35)**	**26.2% (81)**	**37.5% (116)**

	HER2		
	Relapse	**Death**	**Total**
Overexpression	4.1% (12)	11.4% (33)	**15.5% (45)**
No Overexpression	7.6% (22)	14.9% (43)	**22.5% (65)**
Total	**11.7% (34)**	**26.3% (76)**	**38% (110)**

Από τις 78 ασθενείς με θετικό ER-mRNA το 7.7% παρουσίασε υποτροπή και το 17.3% απεβίωσε σε σχέση με 3.5% και 8.7% αντίστοιχα των ασθενών με αρνητικό ER-mRNA. Παρόμοια είναι η κατάσταση και στην περίπτωση του PgR-mRNA. Συγκεκριμένα, από τις 68 ασθενείς με θετικό Pgr-mRNA το 7.8% παρουσίασε υποτροπή και το 14.2% απεβίωσε σε σχέση με 3.5% και 12% αντίστοιχα των ασθενών με αρνητικό PgR-mRNA. Δηλαδή, στις ασθενείς με θετικό ER-mRNA ή/ και θετικό PgR-mRNA το ποσοστό θανάτων ή επιβίωσης με ασθένεια υπερέχει σε σχέση με τις ασθενείς με αρνητικό ER-mRNA ή/ και αρνητικό PgR-mRNA Ωστόσο, η κατάσταση διαφοροποιείται στην περίπτωση του HER2, διότι από τις 45 ασθενείς με υπερέκφραση του HER2 το 4.1% παρουσίασε υποτροπή και το 11.4% απεβίωσε σε σχέση με 7.6% και 14.9% αντίστοιχα των ασθενών με μη-υπερέκφραση του HER2.

Να σημειώσουμε ότι το 70% του συνόλου των ασθενών εμφάνισε θετικό ER-mRNA και το 29% εμφάνισε αρνητικό ER-mRNA, ενώ για το 1,2% περίπου των ασθενών δεν προσδιορίστηκε η κατάσταση του ER-mRNA. Παρόμοια είναι τα αντίστοιχα ποσοστά στην περίπτωση του PgR-mRNA, όπου στο 61.7% του συνόλου των ασθενών είναι θετικό, στο 36% των ασθενών είναι αρνητικό ,ενώ στο 8.2% περίπου των ασθενών δεν έχει προσδιοριστεί. Τέλος, το HER2 παρουσιάζει αντίθετη εικόνα στα αντίστοιχα ποσοστά, καθώς το 27% του συνόλου των ασθενών εμφάνισε υπερέκφραση και το 64% εμφάνισε μη-υπερέκφραση του HER2, ενώ για το 8.5% περίπου των ασθενών δεν προσδιορίστηκε η κατάσταση του HER2.

Στον πίνακα 6.2 ανακεφαλαιώνεται το πλήθος και ποσοστό ασθενών για κάθε κατάσταση των υπό μελέτη βιοδεικτών.

Πίνακας 6.2

	Biomarkers			
	ER-mRNA	**PgR-mRNA**	**HER2**	
Positive	220 (70%)	195 (61.7%)	**Overexpression**	87 (27%)
Negative	92 (29%)	114 (36%)	**No overexpression**	202 (64%)
Missing Values	4 (1.2%)	7 (2.2%)	**Missing Values**	27 (8.5%)
Total	**316**	**316**	**Total**	**316**

6.3 Μελέτη επίδρασης κάθε βιοδείκτη στον κίνδυνο θανάτου και στην επιβίωση χωρίς ασθένεια λαμβάνοντας υπόψη τη χημειοθεραπεία

Εφαρμόζουμε το (μονομεταβλητό) μοντέλο παλινδρόμησης του Cox (*univariate Cox regression model*) για να μελετήσουμε την επίδραση κάθε προγνωστικού παράγοντα στον κίνδυνο θανάτου και στον κίνδυνο επιβίωσης με ασθένεια

«κρατώντας σταθερό» (*adjusting for*) το είδος χημειοθεραπείας (παράγοντας *group*) που χορηγήθηκε στις ασθενείς.

Στην περίπτωση του ER-mRNA μελετάμε το παρακάτω μοντέλο αναλογικού κινδύνου χωριστά για την OS και για την DFS

$$h(t/Z) = h_0(t)\exp(b_1 \cdot ER + b_2 \cdot group) \qquad (6.1)$$

Πίνακας 6.3a

Συντελεστές παλινδρόμησης του μοντέλου (6.1) για OS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
ER	-0.176	0.838	0.236	-0.743	0.457	1.192	0.528	1.333
group	0.163	1.177	0.224	0.728	0.467	0.849	0.758	1.828

Πίνακας 6.3b

Συντελεστές παλινδρόμησης του μοντέλου (6.1) για DFS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
ER	-0.183	0.832	0.198	-0.925	0.355	1.201	0.564	1.228
group	0.026	1.026	0.186	0.140	0.888	0.974	0.712	1.480

Σύμφωνα με τον (σχετικό) λόγο κινδύνων υπάρχει η τάση να μειώνεται παρόμοια τόσο ο κίνδυνος θανάτου όσο και ο κίνδυνος επιβίωσης με ασθένεια στις ασθενείς με θετικό ER-mRNA σε σχέση με τις ασθενείς με αρνητικό ER-mRNA, αλλά αυτή η τάση δεν είναι (στατιστικά) σημαντική (HR=0.84, 95% CI: 0.53-1.33, Wald p-value=0.457 για την OS και HR=0.83, 95% CI: 0.56-1.23, Wald p-value=0.355 για την DFS).

Εξετάζουμε την περίπτωση του PgR-mRNA μελετώντας το παρακάτω μοντέλο αναλογικού κινδύνου

$$h(t/Z) = h_0(t)\exp(b_1 \cdot PgR + b_2 \cdot group) \qquad (6.2)$$

Πίνακας 6.4a

Συντελεστές παλινδρόμησης του μοντέλου (6.2) για OS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
PgR	-0.452	0.636	0.223	-2.026	0.043	1.572	0.411	0.985
group	0.203	1.225	0.225	0.900	0.368	0.816	0.788	1.904

Πίνακας 6.4b

Συντελεστές παλινδρόμησης του μοντέλου (6.2) για DFS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
PgR	-0.259	0.771	0.188	-1.375	0.169	1.296	0.533	1.117
group	0.058	1.059	0.186	0.310	0.757	0.944	0.735	1.528

Ενώ υπάρχει η τάση να μειώνεται σημαντικά ο κίνδυνος θανάτου στις ασθενείς με θετικό PgR-mRNA σε σχέση με τις ασθενείς με αρνητικό PgR-mRNA, δεν μειώνεται σημαντικά ο κίνδυνος επιβίωσης με ασθένεια στις ασθενείς με θετικό PgR-mRNA σε σχέση με τις ασθενείς με αρνητικό PgR-mRNA (HR=0.64, 95% CI: 0.41-0.98, Wald p-value=0.0428 και HR=0.77, 95% CI: 0.53-1.12, Wald p-value=0.169 για την OS και DFS αντίστοιχα).

Τέλος, μελετάμε το παρακάτω μοντέλο αναλογικού κινδύνου στην περίπτωση του HER2

$$h(t/Z) = h_0(t)\exp(b_1 \cdot HER2 + b_2 \cdot group) \qquad (6.3)$$

Πίνακας 6.5a

Συντελεστές παλινδρόμησης του μοντέλου (6.3) για OS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
HER2	0.680	1.974	0.232	2.930	0.003	0.507	1.252	3.110
group	0.119	1.126	0.231	0.515	0.607	0.888	0.716	1.772

Πίνακας 6.5b

Συντελεστές παλινδρόμησης του μοντέλου (6.3) για DFS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
*HER*2	0.670	1.954	0.196	3.419	0.0006	0.512	1.331	2.869
group	0.047	1.048	0.193	0.246	0.806	0.954	0.718	1.530

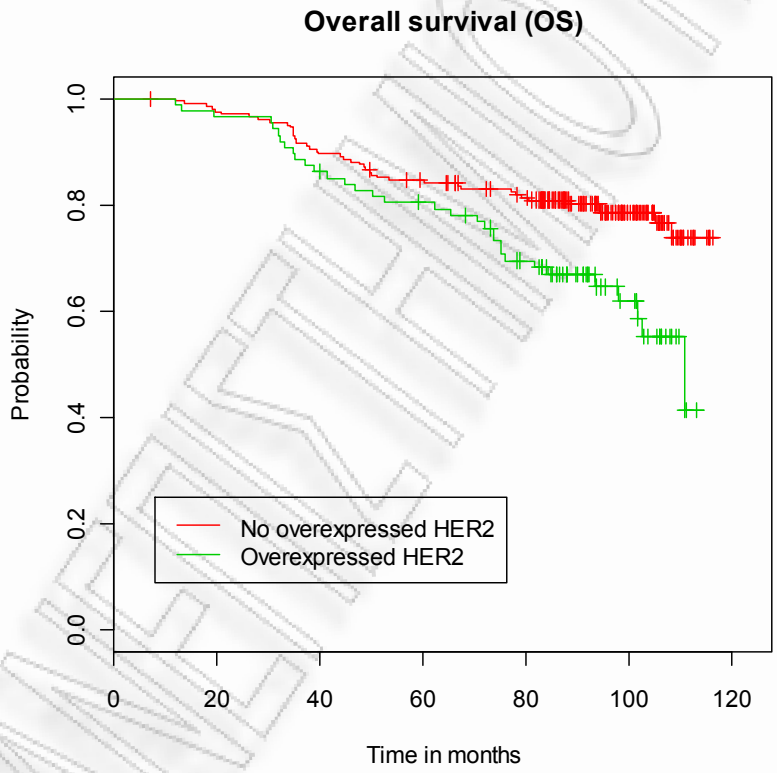
Σύμφωνα με τον (σχετικό) λόγο κινδύνων υπάρχει η τάση να αυξάνεται παρόμοια τόσο ο κίνδυνος θανάτου όσο και ο κίνδυνος επιβίωσης με ασθένεια στις ασθενείς με υπερέκφραση του HER2 σε σχέση με τις ασθενείς με μη-υπερέκφραση του HER2, και μάλιστα αυτή η τάση είναι (στατιστικά) σημαντική (HR=1.97, 95% CI: 1.25-3.11, Wald p-value=0.0034 και HR=1.95, 95% CI: 1.33-2.87, Wald p-value=0.00063 για OS και DFS αντίστοιχα).

Παρακάτω παρουσιάζονται γραφικά η επίδραση των σημαντικών βιοδεικτών στην συνολική και/ ή χωρίς ασθένεια επιβίωση. Συγκεκριμένα στο γράφημα 6.1 φαίνεται η διαφοροποίηση της συνολικής επιβίωσης μεταξύ των ασθενών με υπερέκφραση του βιοδείκτη HER2 και των ασθενών με μη- υπερέκφραση αυτού του βιοδείκτη. Οι ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2 υπερέχουν ελάχιστα από τις ασθενείς με υπερέκφραση αυτού του βιοδείκτη για τους πρώτους 70 μήνες περίπου, ενώ στην συνέχεια η υπεροχή αυτή γίνεται πιο έντονη έως το τέλος της έρευνας. Για τις ασθενείς με μη-υπερέκφραση του HER2 η πιθανότητα συνολικής επιβίωσης φτάνει περίπου το 74% με μέγιστο (πλήρη) χρόνο τους 108.492 μήνες, ενώ για τις ασθενείς με υπερέκφραση του HER2 η πιθανότητα συνολικής επιβίωσης φτάνει περίπου το 41% με μέγιστο χρόνο τους 110.852 μήνες. Η διαφοροποίηση της πιθανότητας συνολικής επιβίωσης μεταξύ των ασθενών με μη-υπερέκφραση του βιοδείκτη και των ασθενών με υπερέκφραση του βιοδείκτη αποδεικνύεται και από τον logrank έλεγχο όπου η στατιστική συνάρτηση είναι $LR = 8.8$ με $p - value = 0.00304$. Συνεπώς, απορρίπτουμε την μηδενική υπόθεση περί μη διαφοροποίησης των συναρτήσεων συνολικής επιβίωσης στις δύο καταστάσεις του βιοδείκτη HER2, $H_0 : S_1(t) = S_2(t)$, $t \le \tau$ σε επίπεδο σημαντικότητας $a = 0.05$.

Πίνακας 6.6

Logrank έλεγχος επίδρασης του HER2 στην OS					
	N	Observed	Expected	χ^2	Q
No-overexpression	202	43	54.6	2.47	8.78
Overexpression	87	33	21.4	6.29	8.78
$Chisq = 8.8$ με 1 βαθμό ελευθερίας, $p - value = 0.00304$					

Γράφημα 6.1



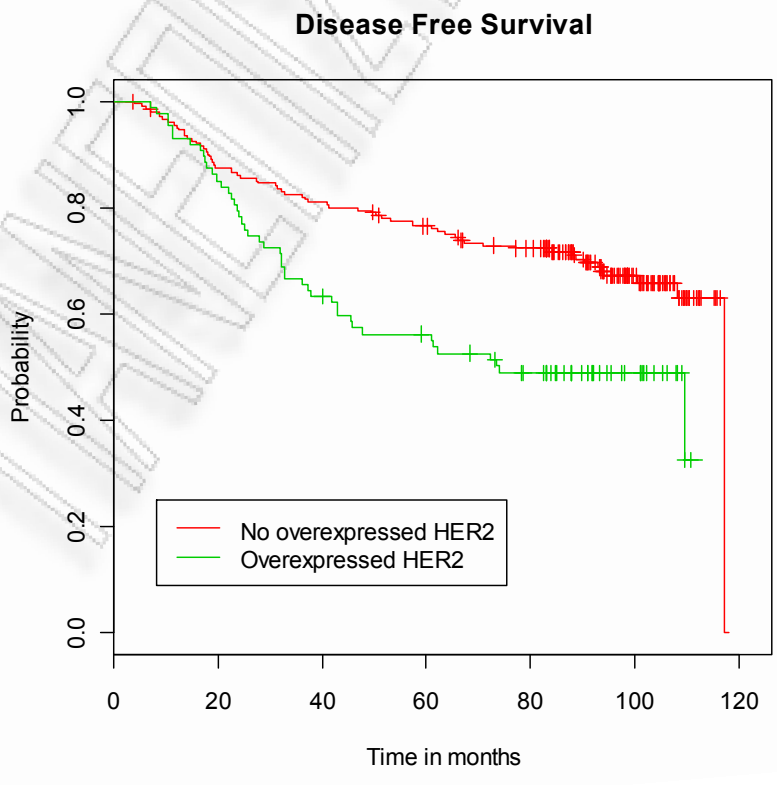
Στο γράφημα 6.2 μελετάμε την διαφοροποίηση της επιβίωσης χωρίς ασθένεια στα δύο επίπεδα (ή καταστάσεις, *status*) του δείκτη HER2. Οι ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2 υπερέχουν σημαντικά από τις ασθενείς με υπερέκφραση αυτού του βιοδείκτη μετά τους 20 μήνες περίπου και ως το τέλος της έρευνας. Οι ασθενείς με μη-υπερέκφραση του HER2 φτάνουν τους 121.475 μήνες ως μέγιστο χρόνο ζωής, όπου αποβιώνει και η τελευταία ασθενής της ομάδας αυτής, ενώ

για τις ασθενείς με υπερέκφραση του HER2 η πιθανότητα επιβίωσης χωρίς ασθένεια φτάνει περίπου το 33% με μέγιστο χρόνο τους 109.672 μήνες. Η διαφοροποίηση της πιθανότητας επιβίωσης χωρίς ασθένεια μεταξύ των ασθενών αυτών των ομάδων επιβεβαιώνεται και από τον logrank έλεγχο όπου η στατιστική συνάρτηση είναι $LR = 12.1$ με $p - value = 0.000516$. Συνεπώς, απορρίπτουμε την μηδενική υπόθεση περί μη διαφοροποίησης των συναρτήσεων επιβίωσης χωρίς ασθένεια στις δύο καταστάσεις του βιοδείκτη HER2, $H_0 : S_1(t) = S_2(t)$, $t \le \tau$ σε επίπεδο σημαντικότητας $a = 0.05$.

Πίνακας 6.7

Logrank έλεγχος επίδρασης του HER2 στην DFS					
	N	Observed	Expected	χ^2	Q
No-overexpression	202	65	81	3.15	12.1
Overexpression	87	45	29	8.79	12.1
$Chisq = 12.1$ με 1 βαθμό ελευθερίας, $p - value = 0.000516$					

Γράφημα 6.2

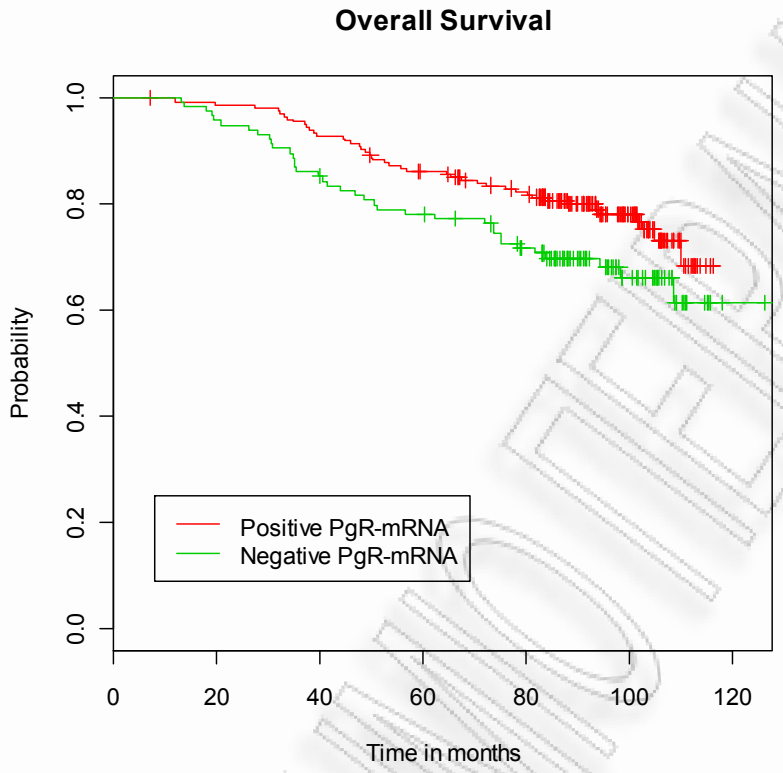


Τέλος, στο γράφημα 6.3 συγκρίνουμε τις ασθενείς με θετικό PgR-mRNA με τις ασθενείς με αρνητικό PgR-mRNA ως προς την συνολική τους επιβίωση. Συγκεκριμένα οι ασθενείς με θετικό PgR-mRNA φαίνονται να υπερέχουν ελάχιστα από τις ασθενείς με αρνητικό PgR-mRNA ως το τέλος της έρευνας. Οι ασθενείς με θετικό PgR-mRNA φτάνουν τους 110.852 μήνες ως μέγιστο χρόνο ζωής με πιθανότητα συνολικής επιβίωσης 68%, ενώ για τις ασθενείς με αρνητικό PgR-mRNA η πιθανότητα συνολικής επιβίωσης φτάνει περίπου το 61% με μέγιστο χρόνο τους 108.787 μήνες. Αυτή η οριακή διαφοροποίηση της πιθανότητας συνολικής επιβίωσης μεταξύ αυτών των ομάδων ασθενών επαληθεύεται και από τον logrank έλεγχο όπου η στατιστική συνάρτηση είναι $LR = 4.1$ με $p - value = 0.0419$. Συνεπώς, απορρίπτουμε οριακά την μηδενική υπόθεση περί μη διαφοροποίησης των συναρτήσεων συνολικής επιβίωσης στις δύο καταστάσεις του βιοδείκτη PgR-mRNA, $H_0 : S_1(t) = S_2(t)$, $t \le \tau$ σε επίπεδο σημαντικότητας $a = 0.05$.

Πίνακας 6.8

Logrank έλεγχος επίδρασης του PgR στην OS					
	N	Observed	Expected	χ^2	Q
Negative	114	37	28.3	2.69	4.14
Positive	195	44	52.7	1.44	4.14
$Chisq = 4.1$ με 1 βαθμό ελευθερίας, $p - value = 0.0419$					

Γράφημα 6.3



6.4 Μελέτη σημαντικότητας της αλληλεπίδρασης κάθε σημαντικού βιοδείκτη με την χημειοθεραπεία στον κίνδυνο θανάτου και στην επιβίωση χωρίς ασθένεια

Στην συνέχεια λαμβάνουμε υπόψη την αλληλεπίδραση κάθε σημαντικού βιοδείκτη με την χημειοθεραπεία για να ελέγξουμε αν ο συνδυασμός αυτών των δύο παραγόντων επηρεάζει σημαντικά τον κίνδυνο θανάτου ή/ και τον κίνδυνο επιβίωσης με ασθένεια.

Στην περίπτωση του PgR-mRNA το μοντέλο αναλογικού κινδύνου για την OS είναι το εξής

$$h(t/Z) = h_0(t)\exp\left(b_1 \cdot PgR + b_2 \cdot group + b_3 \cdot \left[PgR \times group\right]\right) \qquad (6.4)$$

Πίνακας 6.9

Συντελεστές παλινδρόμησης του μοντέλου (6.4.1) για OS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
PgR	0.438	1.550	0.764	0.573	0.566	0.645	0.347	6.927
group	0.513	1.671	0.345	1.490	0.136	0.598	0.850	3.283
PgR × *group*	-0.561	0.571	0.458	-1.225	0.220	1.753	0.232	1.400

Συμπεραίνουμε ότι η αλληλεπίδραση του PgR-mRNA με την χημειοθεραπεία δεν επηρεάζει σημαντικά τον κίνδυνο θανάτου, καθώς τόσο η χορήγηση E-T-CMF όσο και η χορήγηση E-CMF διαφοροποίησε μεν τον κίνδυνο μεταξύ των ασθενών με θετικό PgR-mRNA και αρνητικό PgR-mRNA, διότι η πρώτη αύξησε τον κίνδυνο θανάτου (HR=1.55), ενώ η δεύτερη μείωσε τον κίνδυνο θανάτου (HR=0.88) στις πρώτες σε σχέση με τις δεύτερες, αλλά όχι σημαντικά. Συνεπώς, η αλληλεπίδραση δεν είναι στατιστικά σημαντική για να περιληφθεί στο μοντέλο.
(για την αλληλεπίδραση έχουμε Wald p-value=0.22 και 95% CI: 0.23-1.40)

Μελετάμε την περίπτωση του HER2 χωριστά για την OS και DFS σύμφωνα με το παρακάτω μοντέλο αναλογικού κινδύνου

$$h(t/Z) = h_0(t)\exp\left(b_1 \cdot HER2 + b_2 \cdot group + b_3 \cdot [HER2 \times group]\right) \qquad (6.5)$$

Πίνακας 6.10a

Συντελεστές παλινδρόμησης του μοντέλου (6.5) για OS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
HER2	0.125	1.133	0.763	0.163	0.870	0.883	0.254	5.059
group	-0.394	0.674	0.706	-0.558	0.577	1.482	0.169	2.691
HER2 × *group*	0.357	1.429	0.467	0.766	0.444	0.699	0.573	3.569

Πίνακας 6.10b

Συντελεστές παλινδρόμησης του μοντέλου (6.5) για DFS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
HER2	-0.169	0.844	0.640	-0.265	0.791	1.184	0.241	2.959
group	-0.722	0.486	0.585	-1.235	0.217	2.058	0.154	1.528
HER2 × *group*	0.545	1.724	0.392	1.388	0.165	0.580	0.799	3.721

Συνολικά συμπεραίνουμε ότι ούτε η αλληλεπίδραση του HER2 με την χημειοθεραπεία φαίνεται να επηρεάζει σημαντικά τόσο τον κίνδυνο θανάτου όσο και τον κίνδυνο επιβίωσης με ασθένεια. Συγκεκριμένα, η χορήγηση E-CMF αύξησε τόσο τον κίνδυνο θανάτου (HR=1.62) όσο και τον κίνδυνο επιβίωσης με ασθένεια (HR=1.45) στις ασθενείς με υπερέκφραση του HER2 σε σχέση με τις ασθενείς με μη-υπερέκφραση του HER2, αλλά όχι σημαντικά. Η χορήγηση E-T-CMF δεν διαφοροποίησε τον κίνδυνο θανάτου μεταξύ των ασθενών με υπερέκφραση και με μη-υπερέκφραση του HER2, ωστόσο μείωσε τον κίνδυνο επιβίωσης με ασθένεια (HR=0.84) στις ασθενείς με υπερέκφραση του HER2 σε σχέση με τις ασθενείς με μη-υπερέκφραση του HER2, αλλά όχι σημαντικά. Οπότε ούτε αυτή η αλληλεπίδραση χρειάζεται να συμπεριληφθεί στο μοντέλο.
(για την αλληλεπίδραση έχουμε Wald p-value=0.44 και 95% CI: 0.57-3.57 για την OS και Wald p-value=0.16 και 95% CI: 0.80-3.72 για την DFS)

Η επίδραση της χημειοθεραπείας στην συνολική επιβίωση (*OS*) και στην επιβίωση χωρίς ασθένεια (DFS) στις ασθενείς με υπερέκφραση του HER2 και στις ασθενείς με μη-υπερέκφραση του HER2 παρουσιάζεται αντίστοιχα στα γραφήματα 6.4 και 6.5.

Συγκεκριμένα σύμφωνα με το γράφημα 6.4 παρατηρούμε ότι στις ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2 οι καμπύλες των χημειοθεραπειών E-T-CMF και E-CMF τείνουν να συμπέσουν και αυτό είναι ένδειξη ότι η πιθανότητα επιβίωσης δεν διαφοροποείται μεταξύ των ασθενών που πήραν E-T-CMF και E-CMF. Ωστόσο, στις ασθενείς με υπερέκφραση του βιοδείκτη HER2 οι καμπύλες των χημειοθεραπειών E-T-CMF και E-CMF τέμνονται σε αρκετά σημεία μόνο για τους πρώτους 40 μήνες περίπου και αυτό είναι ένδειξη ότι η πιθανότητα επιβίωσης διαφοροποείται ελάχιστα

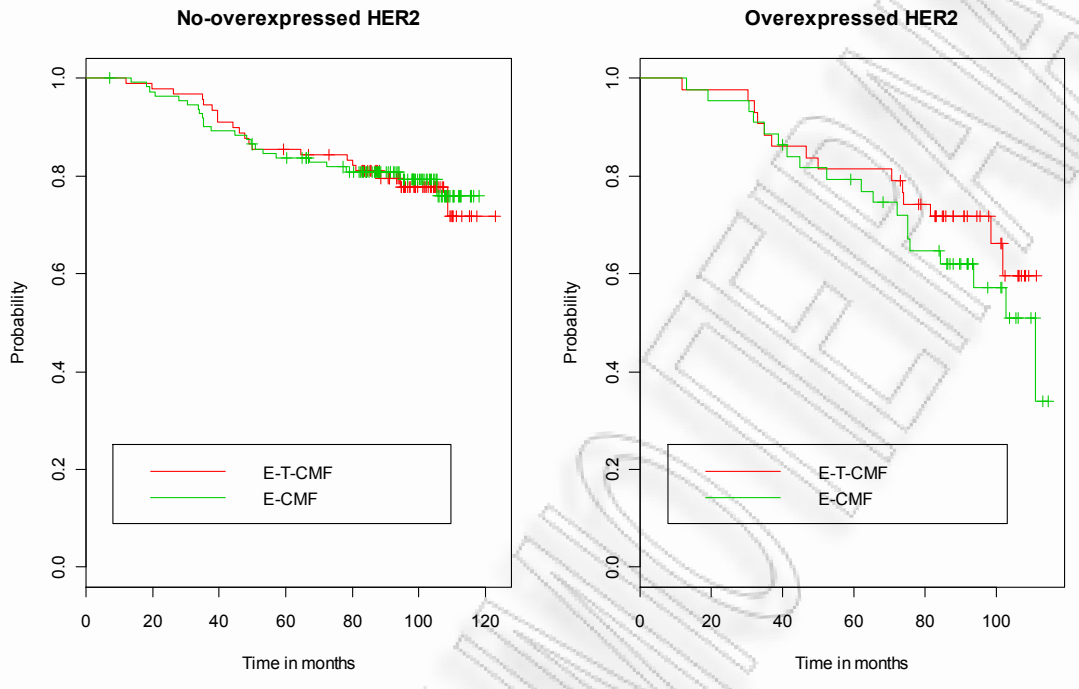
(μη σημαντικά) μεταξύ των ασθενών που πήραν E-T-CMF και E-CMF, καθώς οι πρώτες φαίνεται να υπερέχουν ελάχιστα από τιςς δεύτερες. Τα συμπεράσματα αυτά επαληθεύονται και από τον Gehan-Breslow έλεγχο, όπου για την περίπτωση της μη-υπερέκφρασης και της υπερέκφρασης του HER2 οι στατιστικές συναρτήσεις είναι $Gehan = 0$ και $Gehan = 0.7$ με αντίστοιχα $p - value = 0.96$ και $p - value = 0.7$. Δηλαδή, η μηδενική υπόθεση περί μη διαφοροποίησης των συναρτήσεων συνολικής επιβίωσης των δύο ομάδων χημειοθεραπείας, $H_0 : S_1(t) = S_2(t)$, $t \le \tau$ γίνεται αποδεκτή σε επίπεδο σημαντικότητας $a = 0.05$ και για τις δύο καταστάσεις του βιοδείκτη HER2.

Πίνακας 6.11

Gehan-Wilcoxon έλεγχοι για OS					
No-overexpressed HER2					
	N	Observed	Expected	χ^2	Q
E-T-CMF	90	17.6	17.5	0.00142	0.0029
E-CMF	112	20.8	21.0	0.00118	0.0029
$Chisq = 0.5$ με 1 βαθμό ελευθερίας, $p - value = 0.484$					
Overexpressed HER2					
	N	Observed	Expected	χ^2	Q
E-T-CMF	43	11.5	13.4	0.272	0.665
E-CMF	44	15.2	13.3	0.275	0.665
$Chisq = 0.7$ με 1 βαθμό ελευθερίας, $p - value = 0.415$					

Γράφημα 6.4

Συνολική επιβίωση (OS)



Στο γράφημα 6.5 τα συμπεράσματα για την επίδραση της χημειοθεραπείας στην πιθανότητα επιβίωσης χωρίς ασθένεια τόσο για τΙς ασθενείς με υπερέκφραση του βιοδείκτη HER2 όσο και για τΙς ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2 δεν φαίνονται να διαφοροποιούνται από τα συμπεράσματα στο γράφημα 6.4. Να σημειώσουμε ότι μόνο στις ασθενείς μη-υπερέκφραση του βιοδείκτη HER2 στις οποίες χορηγήθηκε E-T-CMF ο μέγιστος χρόνος επιβίωσης ($t = 121.48$) είναι πλήρης, δηλαδή απεβίωσαν όλες οι ασθενείς αυτής της ομάδας. Χρησιμοποιούμε τον Gehan-Breslow έλεγχο για να ελέγξουμε την αξιοπιστία αυτών των αποτελεσμάτων. Για την περίπτωση της μη-υπερέκφρασης και της υπερέκφρασης του HER2 οι στατιστικές συναρτήσεις είναι $Gehan = 0.5$ και $Gehan = 1.1$ με αντίστοιχα $p - value = 0.48$ και $p - value = 0.29$. Συνεπώς, η μηδενική υπόθεση περί μη διαφοροποίησης των συναρτήσεων επιβίωσης χωρίς ασθένεια των δύο ομάδων χημειοθεραπείας, $H_0 : S_1(t) = S_2(t)$, $t \le \tau$ γίνεται αποδεκτή σε επίπεδο σημαντικότητας $a = 0.05$ και για τις δύο καταστάσεις του βιοδείκτη HER2.

Πίνακας 6.12

Gehan-Wilcoxon έλεγχοι για DFS					
No-overexpressed HER2					
	N	Observed	Expected	χ^2	Q
E-T-CMF	90	26.7	24.3	0.227	0.489
E-CMF	112	27.8	30.1	0.184	0.489
$Chisq = 0.5$ με 1 βαθμό ελευθερίας, $p - value = 0.484$					
Overexpressed HER2					
	N	Observed	Expected	χ^2	Q
E-T-CMF	43	14.4	17.1	0.421	1.12
E-CMF	44	19.2	16.5	0.435	1.12
$Chisq = 1.1$ με 1 βαθμό ελευθερίας, $p - value = 0.29$					

Μελετάμε το γράφημα 6.6 στο οποίο παρουσιάζεται ο βαθμός διαφοροποίησης της πιθανότητας συνολικής επιβίωσης μεταξύ ασθενών με E-T-CMF και E-CMF περιορίζοντας την μελέτη στις ασθενείς με θετικό βιοδείκτη PgR-mRNA και στις ασθενείς με αρνητικό βιοδείκτη PgR-mRNA. Συγκεκριμένα, στις ασθενείς με θετικό βιοδείκτη PgR-mRNA η πιθανότητα συνολικής επιβίωσης δεν φαίνεται να διαφοροποιείται μεταξύ ασθενών με E-T-CMF και E-CMF, αφού η καμπύλες φαίνεται να έχουν την τάση να συμπέσουν μεταξύ τους, ενώ στις ασθενείς με αρνητικό βιοδείκτη PgR-mRNA η πιθανότητα συνολικής επιβίωσης φαίνεται να διαφοροποιείται αρκετά μεταξύ ασθενών με E-T-CMF και E-CMF μόνο στους πρώτους 110 μήνες περίπου. Για τις ασθενείς με αρνητικό PgR-mRNA η στατιστική του συνάρτηση του Gehan-Breslow ελέγχου είναι $Gehan = 2.8$ με $p - value = 0.09$, οπότε αποδεχόμαστε οριακά την μηδενική υπόθεση λόγω σύμπτωσης των καμπυλών μόλις ενάμιση χρόνο περίπου πριν την λήξη της έρευνας. Δηλαδή, αν δεν συνέβαινε αυτή η σύμπτωση των καμπυλών, τότε η διαφοροποίηση της συνολικής πιθανότητας επιβίωσης μεταξύ των δύο ομάδων χημειοθεραπείας θα ήταν σημαντική. Τέλος, για τις ασθενείς με θετικό PgR-mRNA η στατιστική του συνάρτηση του Gehan-Breslow ελέγχου είναι $Gehan = 0.1$ με $p - value = 0.77$, οπότε αποδεχόμαστε την μηδενική υπόθεση ότι η πιθανότητα συνολικής επιβίωσης δεν διαφοροποιείται σημαντικά μεταξύ των ασθενών με E-T-CMF και E-CMF.

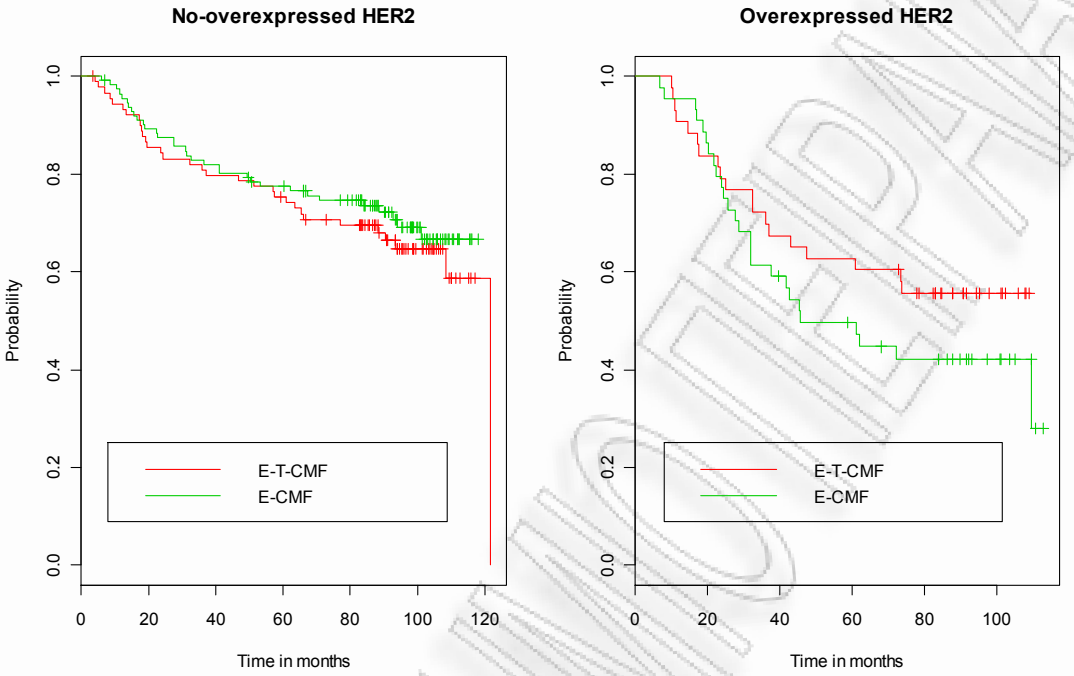
Πίνακας 6.13

Gehan-Wilcoxon έλεγχοι για OS					
Negative PgR					
	N	Observed	Expected	χ^2	Q
E-T-CMF	49	10.3	14.6	1.28	2.83
E-CMF	65	20.8	16.5	1.13	2.83
$Chisq = 2.8$ με 1 βαθμό ελευθερίας, $p - value = 0.0926$					
Positive PgR					
	N	Observed	Expected	χ^2	Q
E-T-CMF	92	18.9	18.1	0.0393	0.0824
E-CMF	103	20.0	20.9	0.0341	0.0824
$Chisq = 0.1$ με 1 βαθμό ελευθερίας, $p - value = 0.774$					

Συμπεραίνουμε, λοιπόν, ότι η προβλεπτική δύναμη της χημειοθεραπείας, είτε με πακλιταξέλη είτε χωρίς, είναι μη σημαντικός προβλεπτικός παράγοντας για την συνολική επιβίωση και στην επιβίωση χωρίς ασθένειας στις ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2, ενώ είναι ιδιαίτερα χαμηλή στην συνολική επιβίωση και στην επιβίωση χωρίς ασθένειας στις ασθενείς με υπερέκφραση αυτού του βιοδείκτη. Επίσης, η ασήμαντη προβλεπτική δύναμη της χημειοθεραπείας στην συνολική επιβίωση παρατηρείται στις ασθενείς με θετικό βιοδείκτη PgR-mRNA, ενώ στις ασθενείς με αρνητικό βιοδείκτη PgR-mRNA η χημειοθεραπεία διαφοροποιεί μεν την συνολική επιβίωση των ασθενών αυτών, αλλά αυτή η διαφοροποίηση είναι οριακά μη σημαντική.

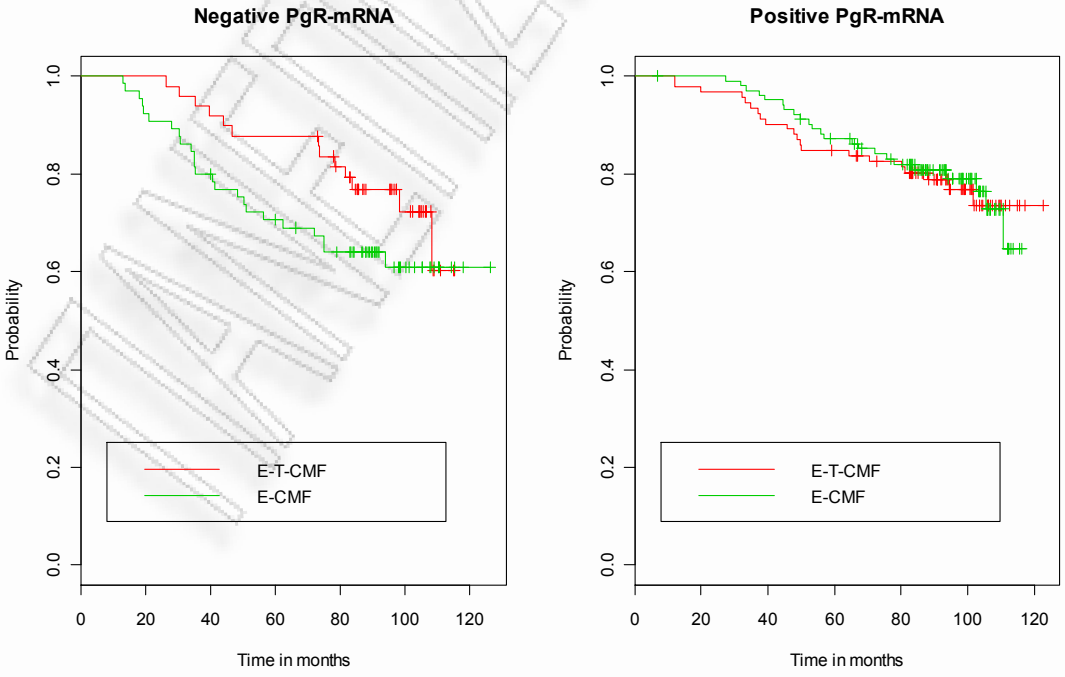
Γράφημα 6.5

Επιβίωση χωρίς ασθένεια (DFS)



Γράφημα 6.6

Συνολική επιβίωση (OS)



6.5 Μελέτη επίδρασης της χημειοθεραπείας, των βιοδεικτών, των κλινικοπαθολογικών παραμέτρων και των γονιδίων στον κίνδυνο θανάτου και στην επιβίωση χωρίς ασθένεια

Σύμφωνα με το πολυμεταβλητό μοντέλο παλινδρόμησης του Cox (πίνακας 6.16), στο οποίο συμπεριλάβαμε τους βιοδείκτες ER-mRNA, PgR-mRNA και HER2, τις τρεις ομάδες γονιδίων, την πρωτεΐνη VEGF, τις κλινικοπαθολογικές παραμέτρους, όπως το μέγεθος του όγκου (*tumor size*), την ταξινόμηση του όγκου (*tumor grade*) και το πλήθος θετικών αδενωμάτων (*nodes*) και την χημειοθεραπεία με ή χωρίς πακλιταξέλη (E-T-CFM ή E-CFM αντίστοιχα) συμπεραίνουμε ότι το πλήθος θετικών αδενωμάτων, οι δύο πρώτες ομάδες γονιδίων και ο βιοδείκτης HER2 επηρεάζουν σημαντικά την συνολική επιβίωση των ασθενών, ενώ το πλήθος θετικών αδενωμάτων και ο βιοδείκτης HER2 επηρεάζουν σημαντικά την επιβίωση των ασθενών χωρίς ασθένεια. Το είδος χημειοθεραπείας και ο βιοδείκτης PgR-mRNA δεν επηρεάζουν σημαντικά την συνολική επιβίωση ούτε την επιβίωση χωρίς ασθένεια, ενώ η δεύτερη ομάδα γονιδίων δεν επηρεάζει σημαντικά την επιβίωση χωρίς ασθένεια. Τέλος, ο βιοδείκτης ER-mRNA, η πρωτεΐνη VEGF, το μέγεθος και η ταξινόμηση του όγκου δεν έχουν καμία προγνωστική δύναμη τόσο στην συνολική όσο και στην χωρίς ασθένεια επιβίωση.

Συγκεκριμένα, η ύπαρξη ενός μέχρι τριών θετικών αδενωμάτων αυξάνει 3.56 φορές τον κίνδυνο θανάτου (HR=3.56, 95% Cl: 1.447-8.780, Wald-P=0.0057) και 2.63 φορές τον κίνδυνο υποτροπής (HR=2.63, 95% Cl: 1.286-5.375, Wald-P=0.0081) σε σχέση με την απουσία αδενωμάτων, ενώ η ύπαρξη τεσσάρων μέχρι εννέα θετικών αδενωμάτων αυξάνει 4.32 φορές τον κίνδυνο θανάτου (HR=4.32, 95% Cl: 1.780-10.474, Wald-P=0.0012) και 4.11 φορές το κίνδυνο υποτροπής (HR=4.11, 95% Cl: 2.057-8.228, Wald-P=6.35e-05) σε σχέση με την απουσία θετικών αδενωμάτων.

Ως προς τον βιοδείκτη HER2 παρατηρούμε ότι μία ασθενής με υπερέκφραση του βιοδείκτη HER2 παρουσιάζει 2.27 φορές μεγαλύτερο κίνδυνο θανάτου (HR=2.27, 95% Cl: 1.283-4.029, Wald-P=0.0049) και 1.67 φορές μεγαλύτερο κίνδυνο υποτροπής (HR=1.67, 95% Cl: 1.066-2.601, Wald-P=0.0025) σε σχέση με κάποια άλλη ασθενή με μη-υπερέκφραση αυτού του βιοδείκτη.

Επιπλέον, ως προς τον βιοδείκτη PgR-mRNA συμπεραίνουμε ότι ασθενής με θετικό βιοδείκτη PgR-mRNA παρουσιάζει 4.61 φορές μεγαλύτερο κίνδυνο θανάτου (HR=4.61, 95% Cl: 0.876-24.298, Wald-P=0.0712) και 2.43 φορές μεγαλύτερο

κίνδυνο υποτροπής (HR=2.43, 95% Cl: 0.617-9.549, Wald-P=0.204) σε σχέση με κάποια άλλη ασθενή με αρνητικό βιοδείκτη PgR-mRNA. Ωστόσο, αυτή η επίδραση δεν είναι (στατιστικά) σημαντική.

Εξετάζοντας τις ομάδες γονιδίων παρατηρούμε ότι όταν σε κάποια ασθενή η έκφραση των γονιδίων της πρώτης ομάδας αυξάνεται κατά ένα βαθμό, τότε ο κίνδυνος θανάτου μειώνεται (HR=0.58, 95% Cl: 0.408-0.826, Wald-P=0.0025) σε σχέση με κάποια άλλη ασθενή όπου η έκφραση των γονιδίων της ομάδας αυτής παραμένει σταθερή. Όταν σε κάποια ασθενή η έκφραση των γονιδίων της δεύτερης ομάδας γονιδίων αυξάνεται κατά ένα βαθμό, τότε ο κίνδυνος θανάτου αυξάνεται 1.89 φορές (HR=1.89, 95% Cl: 1.235-2.889, Wald-P=0.0028) και ο κίνδυνος υποτροπής αυξάνεται 1.33 φορές (HR=1.33, 95% Cl: 0.982-1.816, Wald-P=0.065) σε σχέση με κάποια άλλη ασθενή στην οποία η έκφραση των γονιδίων της ομάδας αυτής παραμένει σταθερή. Ωστόσο, η επίδραση της δεύτερης ομάδας γονιδίων στην επιβίωση χωρίς ασθένεια δεν είναι σημαντική.

Τέλος, είτε χορηγήσουμε E-T-CMF είτε E-CMF ο βιοδείκτης PgR-mRNA επιδρά σημαντικά στην συνολική επιβίωση των ασθενών (έλεγχος για την αλληλεπίδραση: HR:0.3412, 95% Cl: 0.124-0.942, Wald-P=0.038), αλλά όχι στην επιβίωση χωρίς ασθένεια (έλεγχος για την αλληλεπίδραση: HR:0.50, 95% Cl: 0.218-1.159, Wald-P=0.1064). Συγκεκριμένα, όταν χορηγείται E-T-CMF, τότε ασθενής με θετικό βιοδείκτη PgR-mRNA παρουσιάζει μειωμένο κίνδυνο θανάτου σε σχέση κάποια ασθενή με αρνητικό βιοδείκτη, ενώ όταν χορηγείται E-CMF, τότε ασθενής με θετικό βιοδείκτη PgR-mRNA παρουσιάζει 1.57 φορές μεγαλύτερο κίνδυνο θανάτου σε σχέση με κάποια ασθενή με αρνητικό βιοδείκτη. Παρόμοιος είναι ο σχετικός κίνδυνος θανάτου όταν γίνεται χορήγηση της E-CMF σε σχέση με την χορήγηση της E-T-CMF σε ασθενή με αρνητικό και θετικό PgR-mRNA αντίστοιχα.

Το πολυμεταβλητό μοντέλο του Cox όταν μελετάμε το συνολικό χρόνο επιβίωσης είναι το εξής

$$h(t/Z) = h_0(t)\exp(b_1 \cdot VEGF\_ich + b_2 \cdot nodes13 + b_3 \cdot nodes49 + b_4 \cdot genes1 + b_5 \cdot erIhc + \\ + b_6 \cdot pgrIhc + b_7 \cdot her2 + b_8 \cdot genes2 + b_9 \cdot group + b_{10} \cdot [group \times pgrIhc]) \qquad (6.6)$$

Πίνακας 6.14

Συντελεστές παλινδρόμησης του μοντέλου (6.6) για OS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
VEGF	0.505	1.657	0.355	1.422	0.155	0.604	0.826	3.323
*nodes*13	1.271	3.564	0.460	2.763	0.006	0.281	1.447	8.780
*nodes*49	1.463	4.318	0.452	3.235	0.001	0.232	1.780	10.474
*genes*1	-0.543	0.581	0.179	-3.024	0.002	1.721	0.408	0.826
*genes*2	0.640	1.897	0.215	2.982	0.003	0.527	1.245	2.889
ER	0.118	1.126	0.312	0.379	0.705	0.888	0.610	2.076
PgR	1.529	4.614	0.847	1.804	0.071	0.217	0.876	24.298
*HER*2	0.821	2.274	0.292	2.813	0.005	0.439	1.283	4.029
group	0.675	1.963	0.385	1.753	0.079	0.509	0.923	4.175
$PgR \times group$	-1.075	0.341	0.518	-2.075	0.038	2.931	0.124	0.942

Όταν μελετάμε τον χρόνο επιβίωσης χωρίς ασθένεια, το πολυμεταβλητό μοντέλο είναι το εξής

$$h(t/Z) = h_0(t)\exp(b_1 \cdot VEGF\_ich + b_2 \cdot nodes13 + b_3 \cdot nodes49 + b_4 \cdot genes2 + b_5 \cdot erIhc +$$

$$+ b_6 \cdot pgrIhc + b_7 \cdot her2 + b_8 \cdot group + b_9 \cdot [group \times pgrIhc]) \qquad (6.7)$$

Πίνακας 6.15

Συντελεστές παλινδρόμησης του μοντέλου (6.7) για DFS								
	$\hat{b}$	$\exp(\hat{b})$	$se(\hat{b})$	z	$P(>\|z\|)$	$\exp(-\hat{b})$	lower .95	upper .09
VEGF	0.167	1.182	0.262	0.638	0.523	0.846	0.707	1.977
*nodes*13	0.967	2.629	0.365	2.649	0.008	0.380	1.286	5.375
*nodes*49	1.414	4.114	0.354	3.999	6.3e-05	0.243	2.057	8.228
*genes*2	0.289	1.336	0.157	1.845	0.065	0.748	0.982	1.816
ER	-0.150	0.861	0.254	-0.590	0.555	1.162	0.523	1.416
PgR	0.887	2.427	0.698	1.269	0.204	0.412	0.617	9.549
*HER*2	0.510	1.665	0.227	2.242	0.025	0.600	1.066	2.601
group	0.472	1.603	0.323	1.462	0.144	0.624	0.852	3.017
$PgR \times group$	-0.688	0.503	0.426	-1.615	0.106	1.989	0.218	1.159

Πίνακας 6.16

Πολυμεταβλητή ανάλυση παλινδρόμησης του Cox

	HR	**95% Cl**	**Wald P-value**
Overall Survival			
Treatment group			
E-T-CMF	1		
E-CMF	1.96	0.923-4.175	0.079
Positive nodes			
0	1		
1-3	3.56	1.447-8.780	0.006
4-9	4.32	1.780-10.474	0.001
ER-mRNA levels			
negative	1		
positive	1.13	0.610-2.076	0.705
PgR-mRNA levels			
negative	1		

positive	4.61	0.876-24.298	0.071
HER2 levels			
no over-expression	1		
over-expression	2.27	1.283-4.029	0.005
Genes group			
group1	0.58	0.408-0.826	0.002
group2	1.96	1.245-2.889	0.003
Disease-Free Survival			
Treatment group			
E-T-CMF	1		
E-CMF	1.60	0.852-3.017	0.144
Positive nodes			
0	1		
1-3	2.63	1.286-5.375	0.008
4-9	4.11	2.057-8.228	6.35e-05
ER-mRNA levels			
negative	1		
positive	0.86	0.523-1.416	0.555
PgR-mRNA levels			
negative	1		
positive	2.43	0.617-9.549	0.204
HER2 levels			
no over-expression	1		
over-expression	1.66	1.066-2.601	0.025
Genes group			
group2	1.34	0.982-1.816	0.065

6.6 Ανάλυση των scaled Schoefeld και scaled score υπολοίπων

Μελετάμε τα scaled Schoenfeld και τα scaled score υπόλοιπα των πολυμεταβλητών μοντέλων του Cox (6.6) και (6.7), έτσι ώστε να ελέγξουμε ποιες συμμεταβλητές ικανοποιούν την υπόθεση του αναλογικού κινδύνου. Προτιμάμε την μελέτη των scaled Schoenfeld υπολοίπων αντί των Schoenfeld υπολοίπων, διότι με τα πρώτα μπορούμε να ελέγξουμε με στατιστικό έλεγχο αν ικανοποιείται η υπόθεση του αναλογικού κινδύνου τόσο για κάθε συμμεταβλητή ξεχωριστά (ατομικοί έλεγχοι) όσο και για όλες τις μεταβλητές ταυτόχρονα (ολικός έλεγχος). Ελέγχοντας αν οι συμμεταβλητές ικανοποιούν ταυτόχρονα την υπόθεση του αναλογικού κινδύνου μπορούμε παράλληλα να συμπερασματολογήσουμε για την ολική επάρκεια του μοντέλου (*overall model fit*).

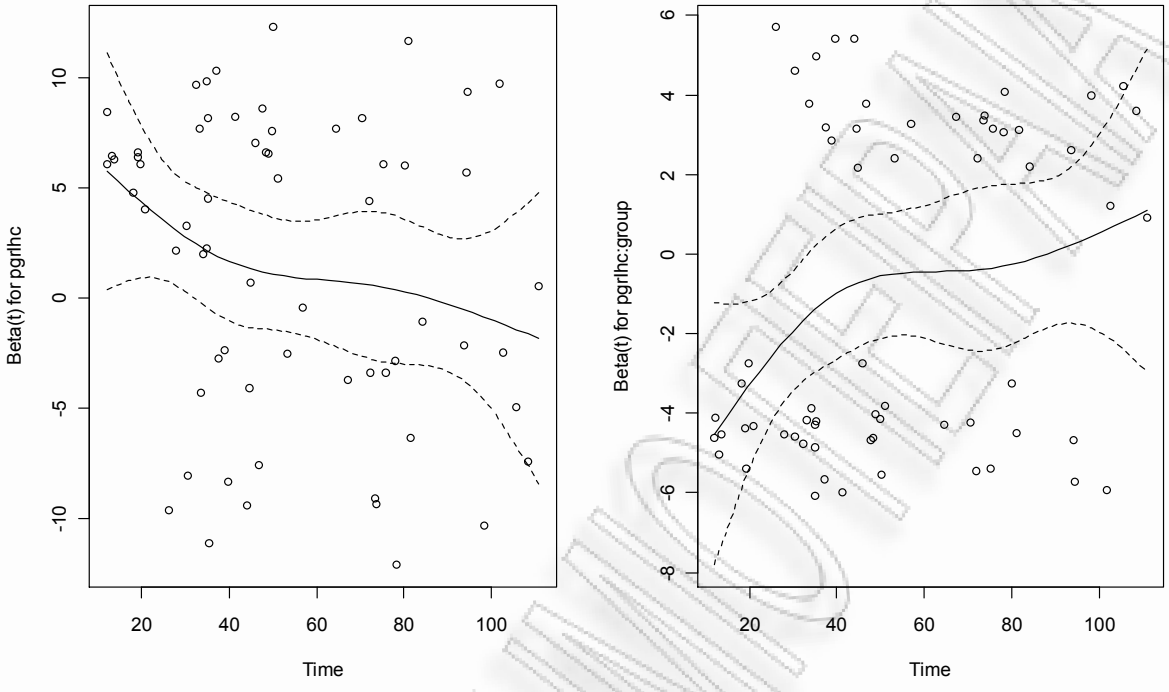
Σύμφωνα με τον πίνακα 6.17 για το μοντέλο (6.6) ο βιοδείκτης PgR-mRNA και η αλληλεπίδραση μεταξύ αυτού του βιοδείκτη και της θεραπείας δεν ικανοποιούν την υπόθεση του αναλογικού κινδύνου, οπότε σε επίπεδο σημαντικότητας $a = 0.05$ απορρίπτουμε τις μηδενικές υποθέσεις $H_0 : b_5(t) = b_5$ και $H_0 : b_{10}(t) = b_{10}$ αντίστοιχα. Συμπεραίνουμε ότι με την πάροδο του χρόνου ο βιοδείκτης PgR-mRNA τείνει να επηρεάζει διαφορετικά τον κίνδυνο θανάτου, αφού φαίνεται να εξαρτάται από τον χρόνο (αν και οριακά λόγω $p-value = 0.0475$). Επιπλέον, τόσο η επίδραση της χημειοθεραπείας στον κίνδυνο θανάτου για κάθε επίπεδο του βιοδείκτη PgR-mRNA όσο και η επίδραση του βιοδείκτη PgR-mRNA στον κίνδυνο θανάτου για κάθε είδος χημειοθεραπείας φαίνεται να διαφοροποιείται με την πάροδο του χρόνου λόγω εξάρτησης αυτής της αλληλεπίδρασης από τον χρόνο. Λαμβάνοντας υπόψη το αποτέλεσμα του ολικού ελέγχου συμπεραίνουμε ότι το μοντέλο δεν φαίνεται να είναι επαρκές (οριακά λόγω $p-value = 0.0487$), δηλαδή οι συμμεταβλητές δεν ικανοποιούν ταυτόχρονα την υπόθεση του αναλογικού κινδύνου. Εξάλλου, ήδη από τους τοπικούς ελέγχους βρήκαμε ότι ο βιοδείκτης PgR-mRNA και η αλληλεπίδραση μεταξύ αυτού του βιοδείκτη και της θεραπείας δεν ικανοποιούν την υπόθεση του αναλογικού κινδύνου.

Πίνακας 6.17

Ολικός έλεγχος και τοπικοί έλεγχοι για το μοντέλο (6.6)			
	Pearson's rho	chi-square statistic	$p-value$
VEGF	0.000959	6.12e-05	0.9938
*nodes*13	-0.059890	2.35e-01	0.6281
*nodes*49	-0.173503	2.07e+00	0.1507
*genes*1	0.043725	1.16e-01	0.7334
*genes*2	0.138693	1.19e+00	0.2744
ER	0.193988	2.40e+00	0.1214
PgR	-0.249315	3.93e+00	0.0475
*HER*2	0.178116	2.34e+00	0.1260
group	-0.203522	2.75e+00	0.0970
$PgR \times group$	0.288591	5.16e+00	0.0231
Global test	NA	1.84e+01	0.0487

Η μη ικανοποίηση της υπόθεσης του αναλογικού κινδύνου τόσο από τον βιοδείκτη PgR-mRNA όσο και από την αλληλεπίδραση μεταξύ αυτού του βιοδείκτη και της θεραπείας μπορεί να αποδειχθεί και από την γραφική αναπαράσταση των αντίστοιχων scaled Schoenfeld υπολοίπων έναντι του χρόνου. Σύμφωνα με τα γραφήματα 6.7 η καμπύλη «εξομάλυνσης» (*smoothing curve*) των scaled Schoenfeld υπολοίπων για τον βιοδείκτη PgR-mRNA σχεδόν τείνει να γίνει ευθεία, δηλαδή να έχει μηδενική κλίση, αποδεικνύοντας ότι αυτός ο βιοδείκτης οριακά δεν ικανοποιεί την υπόθεση του αναλογικού κινδύνου με την πάροδο του χρόνου. Η καμπύλη «εξομάλυνσης» (*smoothing curve*) των scaled Schoenfeld υπολοίπων για την αλληλεπίδραση μεταξύ αυτού του βιοδείκτη και της θεραπείας δεν φαίνεται να έχει μηδενική κλίση, οπότε συμπεραίνουμε ότι αυτή η αλληλεπίδραση δεν ικανοποιεί την υπόθεση αναλογικού κινδύνου με την πάροδο του χρόνου.

Γραφήματα 6.7



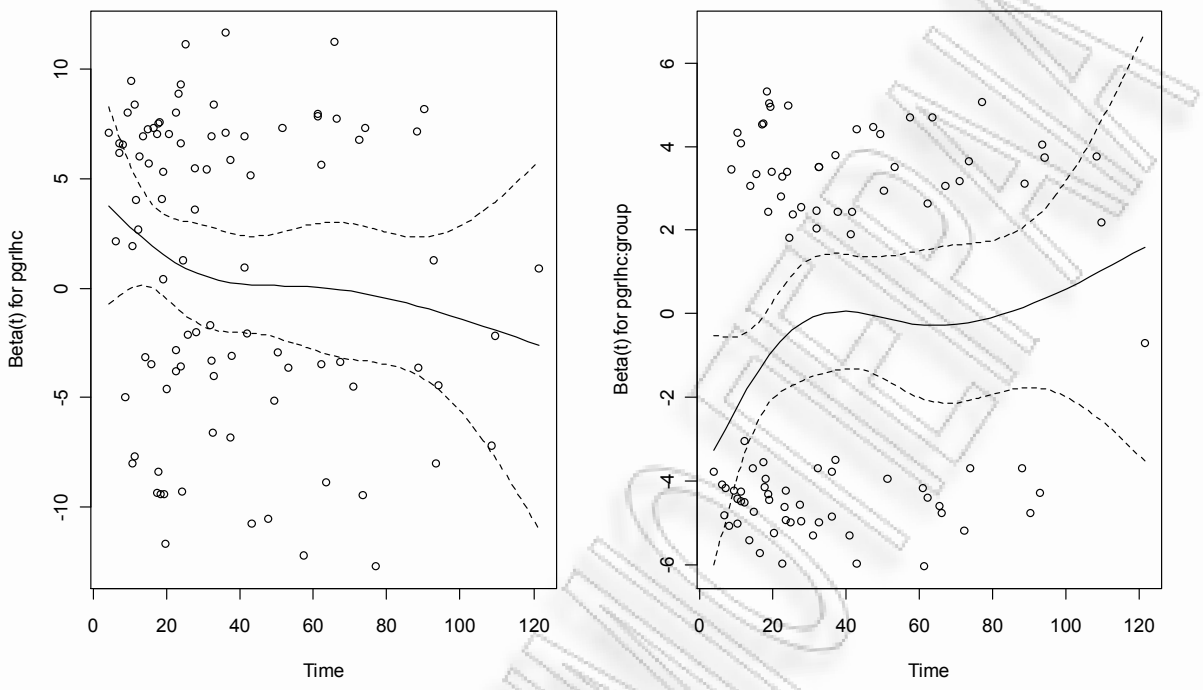
Στον πίνακα 6.18 παρουσιάζονται οι τοπικοί έλεγχοι και ο ολικός έλεγχος για το πολυμεταβλητό μοντέλου του Cox (6.7). Συγκεκριμένα, δεν φαίνεται να υπάρχει κάποια συμμεταβλητή που να μην ικανοποιεί την υπόθεση του αναλογικού κινδύνου και επιπλέον αυτό το μοντέλο θεωρείται επαρκές, οπότε η υπόθεση του αναλογικού κινδύνου ικανοποιείται ταυτόχρονα από τις συμμεταβλητές του μοντέλου αυτού. Δηλαδή, σε επίπεδο σημαντικότητας $a = 0.05$ αποδεχόμαστε τις μηδενικές υποθέσεις $H_0 : b_g(t) = b_g$, $g = 1,2,...,9$ για τους τοπικούς ελέγχους, αλλά και την μηδενική υπόθεση $H_0 : b_1(t) = b_1, b_2(t) = b_2, ..., b_9(t) = b_9$ για τον ολικό έλεγχο.

Πίνακας 6.18

Ολικός έλεγχος και τοπικοί έλεγχοι για το μοντέλο (6.7)			
	Pearson's rho	chi-square statistic	$p-value$
VEGF	4.26e-02	1.82e-01	0.6700
*nodes*13	-6.43e-02	4.07e-01	0.5235
*nodes*49	-9.32e-02	8.70e-01	0.3510
*genes*2	1.31e-01	2.98e-07	0.9996
ER	-1.72e-01	1.44e+00	0.2298
PgR	-4.49e-02	2.76e+00	0.0965
*HER*2	6.33e-05	1.96e-01	0.6583
group	-1.66e-01	2.59e+00	0.1073
$PgR \times group$	1.77e-01	2.90e+00	0.0887
Global test	NA	5.83e+01	0.7566

Επειδή η υπόθεση του αναλογικού κινδύνου φαίνεται να ικανοποιείται από τις συμμεταβλητές του μοντέλου (6.7) σύμφωνα με τον ολικό έλεγχο και τους τοπικούς ελέγχους, τότε η καμπύλη «εξομάλυνσης» θα τείνει να έχει μηδενική κλήση στα γραφήματα των scaled Schoenfeld υπολοίπων για κάθε συμμεταβλητή του μοντέλου. Μάλιστα όσο πιο μεγάλη είναι η τιμή του $p-value$ τόσο πιο μεγάλη θα είναι η «τάση» της καμπύλης «εξομάλυνσης» να έχει μηδενική κλίση. Έστω στα γραφήματα 6.8 παρουσιάζουμε τα scaled Schoenfeld υπολοίπων του βιοδείκτη PgR-mRNA και της αλληλεπίδραση μεταξύ αυτού του βιοδείκτη και της θεραπείας. Οι τοπικοί έλεγχοι αυτών των συμμεταβλητών έχουν μεν $p-value$ μεγαλύτερο του επιπέδου σημαντικότητας $a=0.05$, αλλά όχι σημαντικά μεγάλο για να παρουσιάζει η αντίστοιχη καμπύλη «εξομάλυνσης» μεγάλη τάση για μηδενική κλίση. Δηλαδή, αυτές οι συμμεταβλητές φαίνονται να ικανοποιούν οριακά την υπόθεση του αναλογικού κινδύνου.

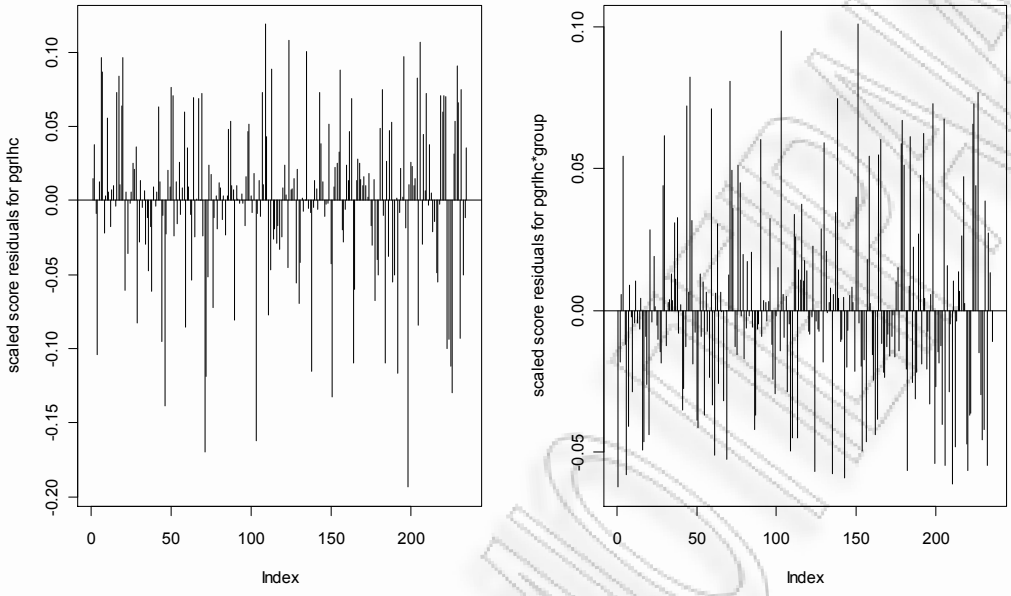
Γραφήματα 6.8



Ενδιαφέρον παρουσιάζει να μελετήσουμε γραφικά τα scaled score υπόλοιπα έστω πάλι για τον βιοδείκτη PgR-mRNA και την αλληλεπίδραση μεταξύ αυτού του βιοδείκτη και της θεραπείας για να δούμε πόσο επηρεάζουν οι ασθενείς στην εκτίμηση των παραμέτρων αυτών των συμμεταβλητών. Δηλαδή, «μετράμε» την επιρροή κάθε ασθενούς στο μοντέλο (6.6) και (6.7). Η μέθοδος αυτή είναι γνωστή με τον όρο μόχλευση (leverage). Στα γραφήματα 6.9 παρουσιάζουμε τα scaled score υπόλοιπα για τον βιοδείκτη PgR-mRNA και την αλληλεπίδραση μεταξύ αυτού του βιοδείκτη και της θεραπείας του μοντέλου (6.6). Παρατηρούμε ότι και στις δύο συμμεταβλητές υπάρχει αρκετή διαφορετικότητα στην μόχλευση κάθε ασθενούς, διότι υπάρχουν ασθενείς με σημαντικά μεγάλη μόχλευση, ασθενείς με μέτρια μόχλευση και ασθενής με μικρή ως σχεδόν μηδενική μόχλευση. Σε παρόμοια συμπεράσματα καταλήγουμε και για scaled score υπόλοιπα των συμμεταβλητών αυτών του μοντέλου (6.7) σύμφωνα με τα γραφήματα 6.10.

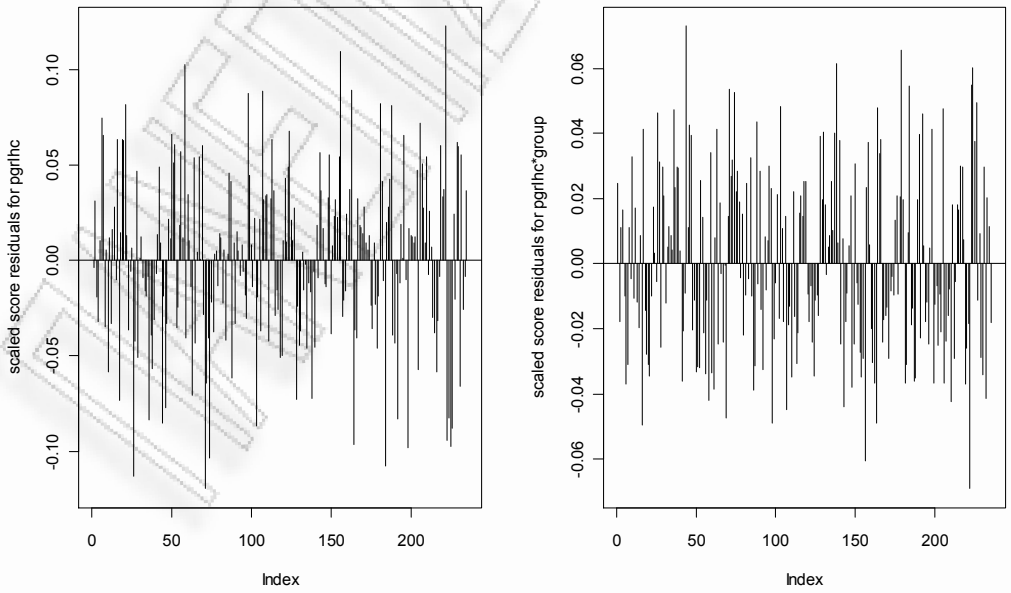
Γραφήματα 6.9

Scaled score υπόλοιπα για το πολυμεταβλητό μοντέλο του Cox για την OS



Γραφήματα 6.10

Scaled score υπόλοιπα για το πολυμεταβλητό μοντέλο του Cox για την DFS



Κεφάλαιο 7

Συμπεράσματα

Για την ομαδοποίηση των γονιδίων σε τρεις συστάδες επιλέξαμε την ward linkage μεταξύ των ιεραρχικών μεθόδων ομαδοποίησης average, complete και ward linkage, ενώ μεταξύ των μη ιεραρχικών μεθόδων ομαδοποίησης $k-$means, PAM και SOM επιλέξαμε την $k-$means ως τις πιο κατάλληλες σύμφωνα με τα μέτρα αξιολόγησης (εσωτερικά, σταθερότητας και βιολογικά).

Συγκεκριμένα, μελετώντας τις ιεραρχικές μεθόδους ομαδοποίησης συμπεράναμε ότι μόνο η complete και η average linkage δημιουργούν μεν ευδιάκριτες συστάδες, αλλά η απόσταση τους είναι μικρή όταν μετακινούμαστε στο επόμενο επίπεδο συνένωσης, δηλαδή αυτές οι συστάδες δεν διαφέρουν σημαντικά μεταξύ τους ως προς τα γονίδια που περιέχουν με αποτέλεσμα αυτές οι δύο μέθοδοι ομαδοποίησης να απορρίπτονται.

Ως προς τις μη ιεραρχικές μεθόδους ομαδοποίησης μελετώντας τα μέτρα αξιολόγησης παρατηρούμε ότι γενικά μόνο η $k-$means και η PAM δεν φαίνονται να διαφέρουν σημαντικά μεταξύ τους. Η SOM υπερέχει πολύ ελάχιστα έναντι της PAM και σχετικά αρκετά έναντι της $k-$means μόνο στον δείκτη βιολογικής ομοιογένειας (*BHI*), δηλαδή λαμβάνοντας υπόψη τη βιολογική λειτουργία των γονιδίων μόνο η SOM δημιουργεί συστάδες που περιέχουν γονίδια με συγκριτικά πιο ικανοποιητική παρόμοια βιολογική λειτουργία. Ωστόσο, γενικά και οι τρεις αυτές μέθοδοι ομαδοποίησης δημιουργούν συστάδες που χαρακτηρίζονται από υψηλή εντός-συστάδας μεταβλητότητα και χαμηλή εκτός-συστάδας μεταβλητότητα. Δηλαδή, παρόμοια γονίδια δεν βρίσκονται στην ίδια συστάδα με αποτέλεσμα τον σχηματισμό συστάδων παρόμοιας χαμηλής ομοιογένειας. Αυτή η αδυναμία φαίνεται να μετριάζεται συνολικά στην $k-$means μέθοδο, γι' αυτό προτιμάται μεταξύ των άλλων.

Στην τελική απόφαση για την μέθοδο ομαδοποίησης σε τρεις συστάδες, που θα εφαρμοστεί στα γονίδια, κρίνεται σημαντική η σύγκριση της ward linkage με την $k-$means μέθοδο χρησιμοποιώντας πάλι τα εσωτερικά, σταθερότητας και βιολογικά μέτρα αξιολόγησης της ομαδοποίησης. Αυτές οι δύο μέθοδοι ομαδοποίησης διαφέρουν πολύ ελάχιστα και στα τρία μέτρα αξιολόγησης με αποτέλεσμα να μην μπορεί να ξεχωρίσει σημαντικά κάποια από αυτές τις δύο μεθόδους. Ωστόσο, η $k-$means φαίνεται να υπερέχει, αν και ελάχιστα, έναντι της ward linkage μόνο στα εσωτερικά μέτρα και στα μέτρα σταθερότητας με αποτέλεσμα να κρίνεται γενικά ως η πιο κατάλληλη για την ομαδοποίηση των 37 γονιδίων της μελέτης σε τρεις συστάδες.

Η ανάλυση των δεδομένων ολοκληρώνεται με την εφαρμογή της ανάλυσης επιβίωσης. Σύμφωνα με τα αποτελέσματα του μονομεταβλητού μοντέλου παλινδρόμησης του Cox συμπεραίνουμε ότι ο βιοδείκτης ER-mRNA δεν έχει προγνωστική δύναμη τόσο στην συνολική όσο και στην χωρίς ασθένεια επιβίωση των ασθενών. Ωστόσο , ο βιοδείκτης HER2 επηρεάζει σημαντικά τόσο την συνολική όσο και την χωρίς ασθένεια επιβίωση των ασθενών, διότι και στις δύο περιπτώσεις οι ασθενείς με υπερέκφραση αυτού του βιοδείκτη παρουσιάζουν μικρότερη πιθανότητα επιβίωσης σε σχέση με τις ασθενείς με μη-υπερέκφραση αυτού του βιοδείκτη. Ο βιοδείκτης PgR-mRNA επηρεάζει σημαντικά μόνο την συνολική επιβίωση των ασθενών και συγκεκριμένα οι ασθενείς με θετικό βιοδείκτη PgR-mRNA έχουν μεγαλύτερη πιθανότητα επιβίωσης σε σχέση με τις ασθενείς που έχουν αρνητικό βιοδείκτη PgR-mRNA.

Μελετώντας την επίδραση των βιοδεικτών PgR-mRNA και HER2 στην συνολική και χωρίς ασθένεια επιβίωση για κάθε είδος χημειοθεραπείας συμπεραίνουμε πως όταν χορηγήσουμε E-T-CMF ο βιοδείκτης PgR-mRNA αυξάνει την συνολική επιβίωση, αλλά όχι σημαντικά, ενώ όταν χορηγήσουμε E-CMF ο βιοδείκτης αυτός μειώνει την συνολική επιβίωση, αλλά πάλι όχι σημαντικά. Στην περίπτωση του βιοδείκτη HER2, όταν χορηγήσουμε E-T-CMF τότε δεν διαφοροποιείται η συνολική επιβίωση μεταξύ των επιπέδων του βιοδείκτη, ενώ αυξάνεται η πιθανότητα επιβίωσης χωρίς κίνδυνο για τις ασθενείς με υπερέκφραση του βιοδείκτη HER2 σε σχέση με τις ασθενείς με μη-υπερέκφραση αυτού του βιοδείκτη. Τέλος, όταν χορηγήσουμε E-CMF, τότε η πιθανότητα τόσο της συνολικής όσο και της χωρίς ασθένειας επιβίωσης μειώνεται για τις ασθενείς με υπερέκφραση του βιοδείκτη HER2 σε σχέση με τις ασθενείς με μη-υπερέκφραση αυτού του βιοδείκτη. Επιπλέον, μελετώντας την

επίδραση της χημειοθεραπείας στην συνολική και χωρίς ασθένεια επιβίωση για κάθε επίπεδο των βιοδεικτών PgR-mRNA και HER2 παρατηρούμε ότι στις ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2 η χορήγηση της E-CMF δεν διαφοροποιεί την συνολική ούτε την χωρίς ασθένεια επιβίωση σε σχέση με την χορήγηση της E-T-CMF, ενώ στις ασθενείς με υπερέκφραση του βιοδείκτη HER2 η χορήγηση της E-CMF δεν αυξάνει τόσο την συνολική όσο και την χωρίς ασθένεια επιβίωση σε σχέση με την χορήγηση της E-T-CMF, αλλά όχι σημαντικά. Τέλος, στις ασθενείς με θετικό βιοδείκτη PgR-mRNA η χορήγηση της E-CMF δεν διαφοροποιεί την συνολική επιβίωση σε σχέση με την χορήγηση της E-T-CMF, ενώ στις ασθενείς με αρνητικό βιοδείκτη PgR-mRNA η χορήγηση της E-CMF μειώνει την συνολική επιβίωση σε σχέση με την χορήγηση της E-T-CMF, αλλά αυτή η μείωση είναι οριακά ασήμαντη.

Συνεπώς, τόσο η αλληλεπίδραση της χημειοθεραπείας τόσο με το βιοδείκτη HER2 όσο και με τον βιοδείκτη PgR-mRNA δεν αποτελεί σημαντικό προγνωστικό παράγοντα τόσο για την συνολική όσο και για την χωρίς ασθένεια επιβίωση.

Εξετάζουμε τα αποτελέσματα του πολυμεταβλητού μοντέλου παλινδρόμησης του Cox και συμπεραίνουμε πως όταν λαμβάνουμε υπόψη κάποιους κλινικοπαθολογικούς παράγοντες αλλά και τις ομάδες των γονιδίων, τότε ο βιοδείκτης HER2 μειώνει την πιθανότητα τόσο της συνολικής όσο και της χωρίς ασθένεια επιβίωσης στις ασθενείς με υπερέκφραση σε σχέση με τις ασθενείς με μη-υπερέκφραση αυτού του βιοδείκτη. Επιπλέον, ο βιοδείκτης PgR-mRNA μειώνει την πιθανότητα τόσο της συνολικής όσο και της χωρίς ασθένεια επιβίωσης στις ασθενείς με θετική σε σχέση με τις ασθενείς με αρνητική έκφραση αυτού του βιοδείκτη, αλλά η μείωση αυτή δεν είναι σημαντική. Τέλος, ο βιοδείκτης ER-mRNA δεν έχει ισχυρή προγνωστική δύναμη στην συνολική επιβίωση ούτε στην επιβίωση χωρίς ασθένεια όταν συνυπολογίζουμε τους παραπάνω παράγοντες.

Από τους κλινικοπαθολογικούς παράγοντες και τις ομάδες των γονιδίων παρατηρούμε ότι η ύπαρξη θετικών αδενωμάτων μειώνει την πιθανότητα συνολικής και χωρίς ασθένεια επιβίωσης σε σχέση με την απουσία αδενωμάτων και μάλιστα όσο πιο πολλά θετικά αδενώματα υπάρχουν τόσο πιο πολύ μειώνεται η πιθανότητα επιβίωσης. Οι υπόλοιποι κλινικοπαθολογικοί παράγοντες δεν φαίνεται να έχουν σημαντική προγνωστική δύναμη τόσο στην συνολική όσο και στην χωρίς ασθένεια επιβίωση. Επιπλέον, από τις τρεις ομάδες γονιδίων μόνο οι δύο πρώτες αποτελούν σημαντικούς προγνωστικούς παράγοντες της επιβίωσης, όπου συγκεκριμένα η πρώτη αυξάνει σημαντικά την πιθανότητα μόνο της συνολικής επιβίωσης και η δεύτερη

μειώνει σημαντικά την πιθανότητα της συνολικής επιβίωσης των ασθενών, ενώ μειώνει οριακά όχι σημαντικά την πιθανότητα επιβίωσης χωρίς ασθένεια.

Τέλος, από τους βιοδείκτες που μελετήσαμε μόνο ο PgR-mRNA αλληλεπιδρά σημαντικά με την χημειοθεραπεία και μόνο στην συνολική επιβίωση των ασθενών. Συγκεκριμένα, η χορήγηση E-T-CMF αυξάνει την πιθανότητα συνολικής επιβίωσης των ασθενών με θετικό βιοδείκτη PgR-mRNA σε σχέση με τις ασθενείς με αρνητικό βιοδείκτη, ενώ η χορήγηση E-CMF μειώνει την πιθανότητα συνολικής επιβίωσης των ασθενών με θετικό βιοδείκτη PgR-mRNA σε σχέση με τις ασθενείς με αρνητικό βιοδείκτη. Επιπλέον, στις ασθενείς με αρνητικό βιοδείκτη PgR-mRNA η χορήγηση της E-CMF αυξάνει την πιθανότητα συνολικής επιβίωσης σε σχέση με την χορήγηση της E-T-CMF, ενώ στις ασθενείς με θετικό βιοδείκτη PgR-mRNA η χορήγηση της E-CMF μειώνει την πιθανότητα συνολικής επιβίωσης σε σχέση με την χορήγηση της E-T-CMF.

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα Α

Τεχνολογία μικροσυστάδων

Παράρτημα Β1

Εφαρμογή στην ομαδοποίηση των εκφράσεων γονιδίων με εντολές του προγράμματος R

Παράρτημα Β2

Παρουσίαση των εντολών του προγράμματος R που χρησιμοποιήθηκαν στην ομαδοποίηση των εκφράσεων γονιδίων

Παράρτημα Γ1

Εφαρμογή στην Ανάλυση Επιβίωσης με εντολές του προγράμματος R

Παράρτημα Γ2

Παρουσίαση των εντολών του προγράμματος R που χρησιμοποιήθηκαν στην Ανάλυση Επιβίωσης

ΠΑΡΑΡΤΗΜΑ Α

Τεχνολογία μικροσυστάδων

Πίνακας 1.7.1

			Gene			
Array	**Dye**	**Variety**	**1**	**2**	…	**p**
1	**1**	**1**	y_{1111}	y_{1112}	…	y_{111p}
		…	…	…		…
		v	y_{11v1}	y_{11v2}		y_{11vp}
	…	…	…	…	…	…
	d	**1**	y_{1d11}	y_{1d12}		y_{1d12}
		…	…	…		…
		v	y_{1dv1}	y_{1dv2}		y_{1dvp}
…	…	…	…	…	…	…
α	**1**	**1**	y_{a111}	y_{a112}	…	y_{a11p}
		…	…	…		…
		v	y_{a1v1}	y_{a1v2}		y_{a1vp}
	…	…	…	…	…	…
	d	**1**	y_{ad11}	y_{ad12}		y_{ad12}
		…	…	…		…
		v	y_{adv1}	y_{1dv2}		y_{advp}

Πίνακας 1.7.2

Source of variability	***Sum of Square***	**degree of freedom**	**Mean of Square**	**F**
Array	SSA	$a-1$	$MSA=\frac{SSA}{a-1}$	$\frac{MSA}{MSE}$
Dye	SSD	$d-1$	$MSD=\frac{SSD}{d-1}$	$\frac{MSD}{MSE}$
Variety	SSV	$v-1$	$MSV=\frac{SSV}{v-1}$	$\frac{MSV}{MSE}$
Gene	SSG	$p-1$	$MSG=\frac{SSG}{p-1}$	$\frac{MSG}{MSE}$
Array × Dye	SSAD	$(a-1)(d-1)$	$MSAD=\frac{SSAD}{(a-1)(d-1)}$	$\frac{MSAD}{MSE}$
Array × Variety	SSAV	$(a-1)(v-1)$	$MSAV=\frac{SSAV}{(a-1)(v-1)}$	$\frac{MSAV}{MSE}$
Array × Gene	SSAG	$(a-1)(p-1)$	$MSAG=\frac{SSAG}{(a-1)(p-1)}$	$\frac{MSAG}{MSE}$
Dye × Variety	SSDV	$(d-1)(v-1)$	$MSDV=\frac{SSDV}{(d-1)(v-1)}$	$\frac{MSDV}{MSE}$
Dye × Gene	SSDG	$(d-1)(p-1)$	$MSDG=\frac{SSDG}{(d-1)(p-1)}$	$\frac{MSDG}{MSE}$
Variety × Gene	SSGV	$(v-1)(p-1)$	$MSVG=\frac{SSVG}{(v-1)(p-1)}$	$\frac{MSVG}{MSE}$
Error	SSE	$adv(p-1)$	$MSE=\frac{SSE}{adv(p-1)}$	-
Total	SSTO	$advp-1$	-	-

Παράρτημα Β1

Εφαρμογή στην ομαδοποίηση εκφράσεων γονιδίων με εντολές του προγράμματος R

Αρχικά θα «φορτώσουμε» (*load*) τα δεδομένα στην R κονσόλα (*R console*) χρησιμοποιώντας την εντολή read.spss(). Για την ομαδοποίηση θα βασιστούμε μόνο στα 37 γονίδια-μεταβλητές που ορίζουν οι στήλες 35 ως 71 για το δείγμα ασθενών που προκύπτει όταν διαγράψουμε εκείνες τις ασθενείς που δεν έχουν καμία τιμή για όλα τα γονίδια. Διαγράφουμε συνολικά 51 ασθενείς. Οι εντολές ομαδοποίησης που ακολουθούν αναλύουν δεδομένα σε πίνακες, γι' αυτό επιβάλλεται να περιλάβουμε αυτό το υποσύνολο δεδομένων στην εντολή as.matrix().

```
a=read.spss("C:/ProgramFiles/SPSSInc/SPSS16/HeCOG_10_97_V3_RPL37A.sav",
use.value.labels=TRUE, to.data.frame=TRUE);a
```

Εκτιμάμε τις ελλείπουσες τιμές χρησιμοποιώντας την $k-$NN μέθοδο.

```
imp <- impute.knn(t(a), 20, rowmax=0.5, colmax=0.9)
```

Για πλήθος γονιδίων από $k=10$ ως 20 η μέθοδος αυτή εκτιμά πιο σωστά τις ελλείπουσες τιμές. Εξετάζοντας το δενδρόγραμμα στο θερμικό χάρτη συσχετίσεων των γονιδίων για το σετ δεδομένων που προκύπτει μετά την εκτίμηση των ελλειπούσων τιμών για πλήθος γειτονικών γονιδίων $k=10$ ως 20, παρατηρούμε ότι τόσο το δενδρόγραμμα όσο και ο θερμικός χάρτης αλλάζουν αμελητέα, οπότε επιλέγουμε να «κρατήσουμε» έστω το $k=20$ στην $k-$NN μέθοδο. Από τα αντικείμενα της εντολής impute.knn() θα χρησιμοποιούμε το $data ως το σετ

δεδομένων που θα αναλύσουμε στη συνέχεια. Έστω, λοιπόν, το σετ δεδομένων με το όνομα dat = imp$data.

Για να υπολογίσουμε τον πίνακα αποστάσεων του Pearson χρησιμοποιούμε την παρακάτω εντολή χρησιμοποιώντας ανεστραμμένο το data frame δεδομένων με το όνομα d = t(dat), ώστε τα γονίδια να βρίσκονται στις γραμμές του data frame.

```
p=as.dist(1-cor(d,method="pearson"))
```

Η εντολή 1-cor() υπολογίζει τον πίνακα συσχετίσεων του Pearson με τιμές στο διάστημα $[0,2]$, ενώ η εντολή as.dist() μετατρέπει τον πίνακα αυτό σε πίνακα αποστάσεων έτσι ώστε να μπορεί να χρησιμοποιηθεί από την εντολή hclust() για την ιεραρχική ομαδοποίηση, χωρίς να παρουσιαστεί σφάλμα, καθώς στα μέτρα απόστασης που υπολογίζει δεν περιέχεται η απόσταση του Pearson. Για την κατασκευή του θερμικού χάρτη των συσχετίσεων του Pearson χρησιμοποιούμε την εντολή heatmap.2(), αφού πρώτα τοποθετήσουμε την εντολή cor(d,method="pearson") στην εντολή as.matrix(), διότι για την χρήση της εντολής heatmap.2() επιβάλλεται το αντικείμενο που θα εισάγουμε προς ανάλυση να έχει οριστεί αυστηρά ως πίνακας δεδομένων :

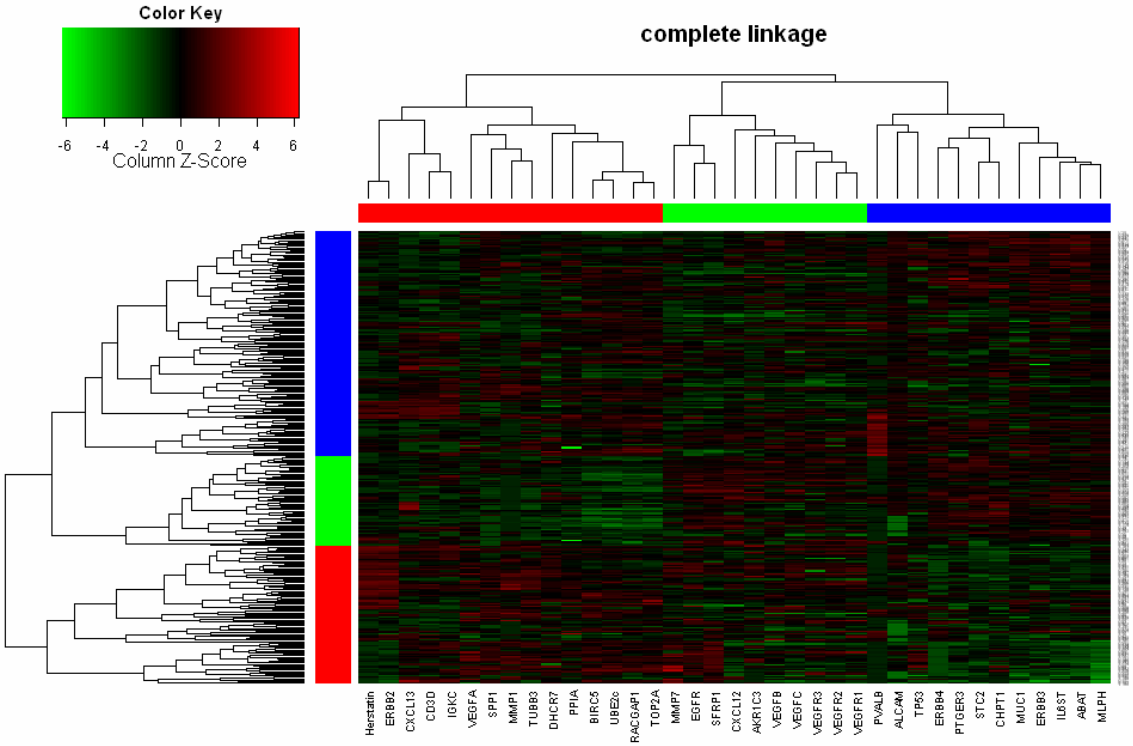
```
p2= as.matrix(cor(d ,method="pearson"))
heatmap.2(p2,col=greenred(75), trace="none", density.info="none", scale="none",
margins=c(7,7),symkey=T)
```

Στη συνέχεια θα εφαρμόσουμε μία σειρά από εντολές για να πετύχουμε την κατασκευή θερμικού χάρτη για τις μεθόδους ιεραρχικής ομαδοποίησης single, complete, average και ward linkage χρησιμοποιώντας ταυτόχρονα πίνακα αποστάσεων (συσχετίσεων) του Pearson. Προτιμάται η εντολή heatmap.2() έναντι της hmap(), διότι δίνει μεγάλη ποικιλία επιλογών για την επιθυμητή διαμόρφωση του χάρτη. Έστω η κατασκευή θερμικού χάρτη για την complete linkage:

```
hclust=function(d, method="complete")
hclust(d, method=method)
dist=function(d)
as.dist(1-cor(t(d), method="pearson"))

heatmap.2(d, distfun=dist, hclustfun=hclust, scale="column", trace="none",
density.info="none", col=greenred(75), cexRow=0.4, main="single linkage",
margins=c(7,7))
```

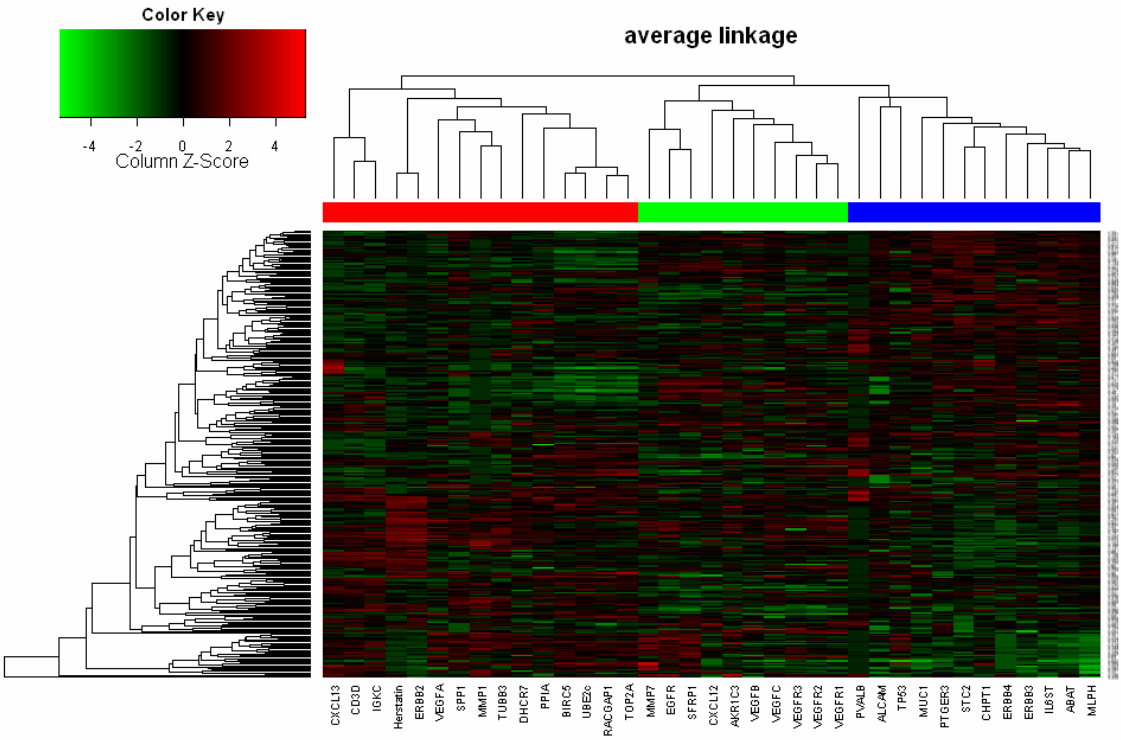
Με τον ίδιο τρόπο ορίζεται ο θερμικός χάρτη για τις υπόλοιπες μεθόδους. Παρακάτω παρουσιάζουμε το θερμικό χάρτη για την complete και average linkage, τις οποίες συγκρίναμε μεταξύ τους και με την ward linkage, έτσι ώστε να καταλήξουμε στην πιο αντιπροσωπευτική συσσωρευτική μέθοδο για το σετ δεδομένων που αναλύουμε. Και στις τρεις μεθόδους είναι προφανές ότι οι βέλτιστες συστάδες γονιδίων είναι τρεις στο πλήθος λαμβάνοντας υπόψη εκείνο το σημείο του δενδρογράμματος που παρατηρείται η μεγαλύτερη μεταβολή στην απόσταση όταν μετακινούμαστε στο επόμενο επίπεδο συνένωσης.

Γράφημα Β1.1



Μελετάμε την complete linkage (Γράφημα Β1.1) και παρατηρούμε ότι λαμβάνοντας υπόψη πόσο μεγαλώνει η απόσταση μεταξύ των συστάδων όταν πηγαίνουμε στο επόμενο στάδιο συνένωσης των συστάδων, μπορούμε να διακρίνουμε εύκολα τρεις συστάδες γονιδίων μικρής ανισορροπίας ως προς το μέγεθός τους και τέσσερις συστάδες ασθενών μεγάλης ανισορροπίας ως προς το μέγεθός τους. Οι αποστάσεις μεταξύ των συστάδων γονιδίων, όπως προκύπτουν στα δύο τελευταία στάδια της ομαδοποίησης, δεν θεωρούνται ιδιαίτερα μεγάλες και αυτό αποτελεί ένδειξη ότι αυτές οι συστάδες δεν διαφέρουν σημαντικά μεταξύ τους, άρα δεν είναι ιδιαίτερα ανομοιογενείς, κάτι το οποίο δεν θεωρείται ιδανικό στην ομαδοποίηση. Αξίζει να σημειώσουμε ότι γενικά οι ευδιάκριτες συστάδες των ασθενών και των γονιδίων βοηθούν στο να ερμηνεύσουμε σωστά τον θερμικό χάρτη.

Γράφημα Β1.2



Στη συνέχεια μελετάμε την average linkage (Γράφημα Β1.2) και μπορούμε να διακρίνουμε με ευκολία τρεις συστάδες γονιδίων χωρίς σημαντική ανισορροπία ως προς το μέγεθός τους. Ωστόσο, η ομαδοποίηση των ασθενών είναι κάθε άλλο παρά ικανοποιητική με αποτέλεσμα ο θερμικός χάρτης να θεωρείται ακατάλληλος για ερμηνεία. Επιπλέον, οι αποστάσεις μεταξύ των συστάδων των γονιδίων που

σχηματίζονται στο τελικό στάδιο της συνένωσης είναι ιδιαίτερα μικρές και αυτό συνεπάγεται μικρή ανομοιογένεια μεταξύ των συστάδων, με αποτέλεσμα η ομαδοποίηση αυτή να μην προτιμάται.

Ωστόσο, η εντολή για τον θερμικό χάρτη με τη μέθοδο UPGMA θα οριστεί διαφορετικά, διότι η εντολή heatmap.2() δεν μπορεί να κατασκευάσει δενδρόγραμμα χρησιμοποιώντας μία συνάρτηση της οικογένειας phylo, γι' αυτό θα ορίσουμε as.hclust() την εντολή upgma(), η οποία πραγματοποιεί ομαδοποίηση με την μέθοδο UPGMA.

```
hclust2=function(d,method="average")
as.hclust(upgma(d,method=method))
dist2=function(d)
as.dist(1-cor(t(d),method="pearson"))

heatmap.2(d,distfun=dist7, hclustfun=hclust7,scale="column", trace="none",
density.info="none", col=greenred(75), cexRow=0.4, main="upgma",
margins=c(7,7))
```

Για να επιλέξουμε την κατάλληλη ιεραρχική μέθοδο μεταξύ της complete, average και ward linkage χρησιμοποιώντας εσωτερικά μέτρα και μέτρα σταθερότητας χρησιμοποιούμε την εντολή clvalid(). Έστω ότι υπολογίζουμε τα εσωτερικά μέτρα για την ward linkage για τρεις συστάδες:

```
v=clValid(t(d), 3:3, clMethods = "hierarchical", validation ="internal",
metric ="correlation", method ="ward")
summary(v)
```

```
Clustering Methods: hierarchical
Cluster sizes: 3

Validation Measures:
                                  3
hierarchical  Connectivity   17.9210
              Dunn            0.5106
              Silhouette      0.2108
```

```
Optimal Scores:
             Score   Method       Clusters
Connectivity 17.9210 hierarchical  3
Dunn          0.5106 hierarchical  3
Silhouette    0.2108 hierarchical  3
```

Αναλόγως εργαζόμαστε και για τις μεθόδους average και complete. Για να υπολογίσουμε τα μέτρα σταθερότητας αντικαθιστούμε την επιλογή "internal" με την επιλογή "stability" . Για τα βιολογικά μέτρα πρέπει να ορίσουμε λίγο διαφορετικά την εντολή clvalid():

```
d=t(dat)
d=as.data.frame(d)
genes=names(d)            ## (1) ##
fc=tapply(genes, FC, c)  ## (2) ##

bio=clValid(t(d), 3:3, clMethods = "hierarchical", validation ="biological",
metric ="correlation", method ="ward", annotation = fc)
summary(bio)
```

(1) Ορίζουμε ως data frame τον πίνακα d, έτσι ώστε να μπορούμε να «καλέσουμε» τα ονόματα των γονιδίων με την εντολή names().

(2) Για κάθε επίπεδο του παράγοντα FC, ο οποίος περιέχει τις λειτουργικές ομάδες, παρουσιάζει ποια γονίδια περιέχονται. Αν στην εντολή tapply() δεν περιλάβουμε την παράμετρο c, τότε το αποτέλεσμα θα είναι ένα διάνυσμα μήκους όσο το πλήθος των γονιδίων, που περιέχει ακέραιους αριθμούς που αντιπροσωπεύουν το id της λειτουργικής ομάδας (όπως θα εμφανιζόταν αν είχαμε περιλάβει την παράμετρο c) που δηλώνουν σε ποια λειτουργική ομάδα ανήκει το κάθε γονίδιο.

Τέλος, για το υπολογισμό του συσσωρευτικού συντελεστή και για τις τρεις παραπάνω μεθόδους ιεραρχικής ομαδοποίησης χρησιμοποιούμε την εντολή coef.hclust(). Στην συγκεκριμένη εντολή εισάγουμε το αντικείμενο της εντολής hclust() και όχι της ακολουθίας εντολών

```
{  hclust=function(d, method="complete");
   hclust(d, method=method);
```

dist=function(d);

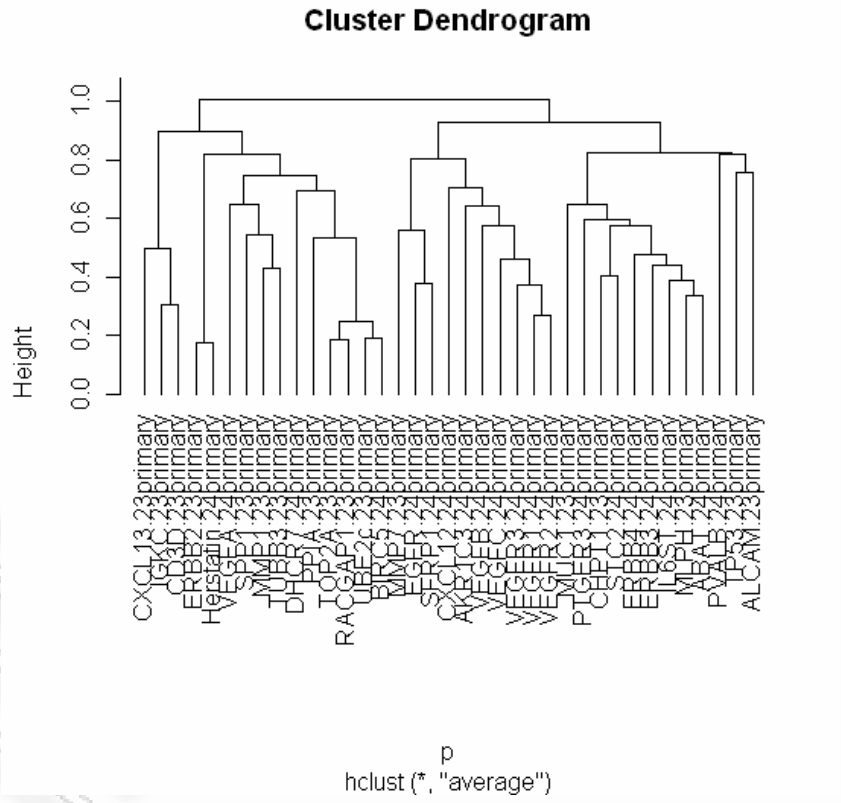
as.dist(1-cor(t(d), method="pearson")) }

που παρουσιάσαμε παραπάνω, διότι στην περίπτωση αυτή εμφανίζεται σφάλμα (*warning*).

Έστω πραγματοποιούμε ομαδοποίηση με την average linkage με απόσταση Pearson (Γράφημα Β1.3):

```
p3= as.dist(1-cor(t(d), method="pearson"))
h=hclust(p3,"average")
plot(h, hang=-1)
```

Γράφημα Β1.3



Στην συνέχεια εισάγουμε το αντικείμενο h στην εντολή coef.hclust() και παίρνουμε το παρακάτω αποτέλεσμα

```
agglcoef=coef.hclust(h); agglcoef
[1] 1.243610
```

Αντί να κατασκευάσουμε πίνακα με τις συστάδες γονιδίων που προκύπτουν από την ward linkage, μπορούμε κάτω από το δενδρόγραμμα των γονιδίων που κατασκευάσαμε με το heatmap.2() να προσδιορίσουμε αυτές τις συστάδες χρησιμοποιώντας μία μπάρα τριών χρωμάτων, όπου το καθένα αντιπροσωπεύει την αντίστοιχη συστάδα.

```
p3= as.dist(1-cor(t(d), method="pearson"))
h=hclust(p3,"ward")

plot(h, hang=-1)
k=h$order;k
srt=seq(1,37)

v=c(1,2,3);v
r=rep(v,c(15,11,11));r
d1=cbind(t(d),srt);d1
d1[1:37,317]=factor(d1[1:37,317],levels=k)
o_n=as.matrix(d1[order(d1[1:37,317]),]);o_n      ## (1) ##
d2=o_n[1:37,1:316];d2
d3=cbind(d2,r);d3                                ## (2) ##
kk=seq(1,37)                                     ## (3) ##
d5=cbind(d3,k);d5                                ## (4) ##
d5[1:37,318]=factor(d5[1:37,318],levels=kk)
o_nn=as.matrix(d5[order(d5[1:37,318]),]);o_nn   ## (5) ##
d6=o_nn[1:37,1:317]                              ## (6) ##
d7=d6[1:37,1:317]                                ## (7) ##
d8=t(d7);d8

hclust_n=function(d8,method="ward")
hclust(d8,method=method)
dist_n=function(d8)
as.dist(1-cor(t(d8),method="pearson"))

color.map=function(r){if(r= ="1"){"#FF0000"}else  ## (8) ##
{if(r= ="2"){"#00FF00"} else "#0000FF"}}
groupscolors=unlist(lapply(d6[,317], color.map))     ## (9) ##

heatmap.2(d8,distfun=dist_n, hclustfun=hclust_n, scale="column", trace="none",
density.info="none", col=greenred(75), cexRow=0.4, main="ward linkage",
margins=c(7,7), ColSideColors=groupscolors)
```

(1) Διατάσσουμε τη στήλη srt του πίνακα d1 σύμφωνα με το διάνυσμα k, οπότε ταυτόχρονα διατάσσονται και τα γονίδια όμοια με τη στήλη srt.

(2) Ο πίνακας d3 περιέχει τα 37 γονίδια και για τα 316 δείγματα και μία επιπλέον στήλη την r που δείχνει την ομάδα που ανήκει κάθε γονίδιο και παίρνει τις τιμές 1, 2, 3.

(3) Έστω μία ακολουθία αριθμών από το 1 ως το 37.

(4) Στον πίνακα d3 τοποθετούμε και άλλη μία στήλη, την k.

(5) Διατάσσουμε τη στήλη k του πίνακα d5 σύμφωνα με το διάνυσμα kk, οπότε ταυτόχρονα διατάσσονται και τα γονίδια όμοια με τη στήλη k.

(6) Από τον πίνακα o_nn εξαιρούμε την στήλη kk.

(7) Από τον πίνακα d6 εξαιρούμε τη στήλη k.

(8) Καθορίζουμε το χρώμα της μπάρα που αντιστοιχεί σε κάθε ομάδα

(9) Η εντολή αυτή αντιστοιχίζεται με την παράμετρο ColSideColors της εντολής heatmap.2() και αφορά τα χρώματα που εμφανίζονται στην μπάρα πάνω από τις στήλες του θερμικού χάρτη. Για τα χρώματα που εμφανίζονται στην μπάρα που τοποθετείται παράλληλα με τις γραμμές του θερμικού χάρτη χρησιμοποιούμε την παράμετρο RowSideColors.

Να σημειώσουμε ότι οι επιλογές “#FF0000”, “#00FF00” και “#0000FF” αντιστοιχούν στο κόκκινο, πράσινο και μπλε χρώμα.

Παρατηρήσεις

1. Το δενδρόγραμμα που προκύπτει από την εντολή hclust() διαφέρει με το δενδρόγραμμα του θερμικού χάρτη μόνο ως προς τη διάταξη των στοιχείων και όχι ως προς τις αποστάσεις και τα στοιχεία που περιέχουν οι συστάδες που δημιουργούνται. Συγκεκριμένα, με την εντολή hclust() η αναδιάταξη κάθε συστάδας γίνεται ως εξής:

Αν μία συστάδα που πρόκειται να συγχωνεύσουμε στη συστάδα προηγούμενου βήματος αντιστοιχεί σε μεγαλύτερη τιμή της κλίμακας αποστάσεων, τότε τοποθετείται δεξιά της τελευταίας. Η συστάδα του προηγούμενου βήματος θεωρείται «ισχυρότερη» (*tighter*) γι’ αυτό βρίσκεται αριστερά του γραφήματος. Τα στοιχεία που πρόκειται να συγχωνευτούν σε μία συστάδα θεωρούνται πάντα «ισχυρά», οπότε

τοποθετούνται αριστερά της συστάδας. Εντός κάθε συστάδας η διάταξη των στοιχείων ακολουθεί τη διάταξή τους στον πίνακα δεδομένων.

2. Για να εμφανίσουμε τις χρωματιστές μπάρες που ορίζουν τις συστάδες των ασθενών κάτω από το αντίστοιχο δενδρόγραμμα θα πρέπει αρχικά να εντοπίσουμε το πλήθος των ασθενών σε αυτές τις συστάδες, κάτι το οποίο είναι ιδιαίτερα δύσκολο, διότι είναι αδύνατο απλά να παρατηρήσουμε τις ασθενείς που αντιστοιχούν στις συστάδες αυτές μέσω του δενδρογράμματος, καθώς λόγω του σημαντικά μεγάλου πλήθους ασθενών οι «ετικέτες» των ασθενών στο δενδρόγραμμα εμφανίζονται η μία πάνω στην άλλη.

Το πρόβλημα αυτό αντιμετωπίζεται με την χρήση της εντολής identify(). Σύμφωνα με την εντολή αυτή μπορούμε να μαρκάρουμε πάνω στο δενδρόγραμμα τη συστάδα που μας ενδιαφέρει και να εμφανίσουμε στο output τόσο τις «ετικέτες» των ασθενών που ανήκουν σε αυτή τη συστάδα όσο και το πλήθος τους. Συγκεκριμένα, γι' αυτόν τον σκοπό ακολουθούμε την παρακάτω σειρά εντολών

```
pp=hclust(p,"complete")
plot(pp,hang=-1)
(x1=identify(pp)); x1; as.matrix(x1)

(x2=identify(pp)); x2; as.matrix(x2)
(x3=identify(pp)); x3 ;as.matrix(x3)
```

Στην συνέχεια χρησιμοποιούμε το μέγεθος των συστάδων που πήραμε από τις παραπάνω εντολές στις εντολές που αναλύσαμε παραπάνω για την περίπτωση των γονιδίων, ώστε να εμφανίσουμε στον θερμικό χάρτη τις χρωματιστές μπάρες κάτω από το δενδρόγραμμα των ασθενών. Λαμβάνουμε υπόψη την **παρατήρηση 1** για την διαφορετική διάταξη των συστάδων στο δενδρόγραμμα που δίνουν οι εντολές hclust() και heatmap.2(), έτσι ώστε τα μεγέθη των συστάδων που βρήκαμε μέσω του δενδρογράμματος της hclust() να τα ορίσουμε στις αντίστοιχες συστάδες του δενδρογράμματος της εντολής heatmap.2().

Έχοντας καταλήξει στο πλήθος $\mathrm{k}(=3)$ των συστάδων που βρήκαμε με την ward linkage, μπορούμε να εφαρμόσουμε την μη ιεραρχική $k-$means μέθοδο

ομαδοποίησης και να μελετήσουμε τις συστάδες γονιδίων που προκύπτουν. Πρώτα θα προσδιορίσουμε τα αρχικά κέντρα βάρους των συστάδων χρησιμοποιώντας την εντολή initial.Centers() στην οποία εισάγουμε τον ανεστραμμένο πίνακα δεδομένων (διότι τόσο η συγκεκριμένη εντολή όσο και η εντολή kmeans() θεωρούν ως μεταβλητές προς ομαδοποίηση τις γραμμές του πίνακα) και το πλήθος των συστάδων ως εξής:

```
ic=initial.Centers(dat,3)
print(ic)
centrs=dat[ic,];centrs
```

```
> print(ic)
[1] 1   2   3
> centrs=dat[ic,];centrs
                          1         2         3         4
PPIA.23primary     39.30506  39.78587  40.28778  40.07722
UBE2c.23primary    32.56824  33.57381  35.75368  32.08713
MMP1.23primary     34.78235  30.61283  28.77205  30.52551
                       . . .     . . .     . . .     . . .
                         365       366       367
PPIA.23primary     40.37573  40.69244  38.08120
UBE2c.23primary    35.41882  34.98105  32.48556
MMP1.23primary     33.35316  35.30509  29.24583
```

Σύμφωνα με την εντολή initial.Centers(dat,3) εμφανίζεται στο output το id των αντίστοιχων γονιδίων στον πίνακα δεδομένων. Αυτά τα γονίδια είναι το αρχικό κέντρο βάρους της πρώτης, δεύτερης και τρίτης συστάδας αντίστοιχα. Με την εντολή dat[ic,] επιλέγουμε από το πίνακα δεδομένων και εμφανίζουμε το σχέδιο έκφρασης αυτών των γονιδίων-κέντρων βάρους.

Για να εφαρμόσουμε την $k-$means μέθοδο ομαδοποίησης χρησιμοποιούμε την εντολή kmeans() στην οποία εισάγουμε ανεστραμμένο τον πίνακα δεδομένων και τα κέντρα βάρους που έχουμε βρει ήδη:

```
cl=kmeans(dat,centrs);cl
```

```
K-means clustering with 3 clusters of sizes 8, 10, 19
Cluster means:
        1        2        3        4        5        6
1 35.82234 36.39291 36.82811 37.21600 36.19897 37.55209
2 33.94567 35.15041 33.19529 34.48281 33.61893 34.19734
3 32.60368 31.90273 31.86683 32.20333 30.98915 31.37808
      ...      ...      ...      ...      ...      ...
        311      312      313      314      315      316
1 36.92987 37.26308 37.65825 35.34468 35.33018 35.37487
2 34.12897 34.38575 33.49941 34.92633 34.85497 33.57193
3 31.82545 31.66580 32.0237  31.64480 32.30443 32.71613

Clustering vector:
   PPIA.23primary        UBE2c.23primary        MMP1.23primary
                1                      2                      3
   IGKC.23primary        TP53.23primary         MLPH.23primary
                3                      2                      1
              ...                    ...                    ...
VEGFR1.24primary VEGFR2.24primary VEGFR3.24primary
                3                3                3

Within cluster sum of squares by cluster:
[1]  7898.57    7955.79     18838.73

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

Το output της εντολής αυτής περιέχει το πλήθος γονιδίων κάθε συστάδας, τα τελικά κέντρα βάρους κάθε συστάδας ως διανύσματα μεγέθους ίσο με το πλήθος των

ασθενών όπου περιέχει το μέσο επίπεδο έκφρασης των γονιδίων της αντίστοιχης συστάδας για κάθε ασθενή, την συστάδα στην οποία ανήκει κάθε γονίδιο, την εντός-συστάδας διακύμανση για κάθε συστάδα και τέλος τα «ονόματα» των παραπάνω αποτελεσμάτων για να τα «καλέσουμε» ατομικά χρησιμοποιώντας τον σύνδεσμο cl$.

Γραφικά μπορούμε να παρουσιάσουμε τις συστάδες που δημιουργούνται με διαφορετικά χρώματα και να εμφανίσουμε το κέντρο βάρους κάθε συστάδας. Αποφεύγουμε να εμφανίσουμε στο γράφημα το όνομα που αντιστοιχεί σε κάθε σημείο-γονίδιο, γιατί λόγω του αρκετά μεγάλου πλήθους των γονιδίων και της μέτριας διασποράς ορισμένων σημείων θα συγχωνευτούν τα αντίστοιχα ονόματα δημιουργώντας μία «ταραχή» στο γράφημα και το αποτέλεσμα δεν θα είναι ικανοποιητικό οπτικά.

```
plot(dat, col=cl$cluster, xlab="ασθενής 1", ylab="ασθενής 2",
main="Ομαδοποίηση σε 3 συστάδες")
text(32.4, 32, label="Κέντρο συστάδας 3",col=3)
text(34.3, 34.6, label="Κέντρο συστάδας 2",col=2)
text(35.8, 37, label="Κέντρο συστάδας 2",col=1)
points(cl$centers, col=1:3, pch=11)
legend(28.8, 38.6, c("Συστάδα 1","Συστάδα 2","Συστάδα 3"), col=cl$cluster, lty=1
```

Για την εφαρμογή της PAM ομαδοποίησης για k=3 συστάδες χρησιμοποιούμε την εντολή pam() στην οποία εισάγουμε τον ανεστραμμένο πίνακα δεδομένων, z, το πλήθος των συστάδων, το μέτρο απόστασης των γονιδίων από τα medoids και ορίζουμε κάποιες επιλογές της εντολής αυτής ως εξής

```
pam=pam(dat, k=3, metric="euclidean", medoids=NULL, diss=FALSE,
stand=TRUE, cluster.only=FALSE, do.swap=TRUE);pam
```

```
Medoids:
                      ID        1         2         3         4
PPIA.23primary         1 39.30506  39.78587  40.28778  40.07722
VEGFR1.24primary      35 32.86094  32.16772  32.93975  32.02527
VEGFB.24primary       33 35.59921  35.41613  35.34674  35.36149
                            . . .     . . .     . . .     . . .
                            365       366       367
PPIA.23primary          40.37573  40.69244  38.08120
VEGFR1.24primary        31.43399  32.13230  31.87168
VEGFB.24primary         34.84790  34.64712  34.94631

Clustering vector:
    PPIA.23primary      UBE2c.23primary      MMP1.23primary
                 1                    2                   2
    IGKC.23primary      TP53.23primary       MLPH.23primary
                 2                    3                   3
             . . .                . . .               . . .
  VEGFR1.24primary   VEGFR2.24primary   VEGFR3.24primary
                 2                    2                   2

Objective function:
   build        swap
14.77640    14.76685

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"
 [5] "isolation"  "clusinfo"   "silinfo"    "diss"
 [9] "call"       "data"
```

Σύμφωνα με το παραπάνω output μπορούμε να πληροφορηθούμε για το μέσο επίπεδο έκφρασης για κάθε συνδυασμό medoid και δείγματος, την συστάδα στην

οποία ανήκει κάθε γονίδιο, την αντικειμενική συνάρτηση του build-step και swap-step, τα «ονόματα» των παραπάνω αποτελεσμάτων για να τα «καλέσουμε» ατομικά χρησιμοποιώντας τον σύνδεσμο pam$, αλλά και κάποιων επιπλέον αποτελεσμάτων, όπως:

- "isolation", δηλαδή καθορίζει αν κάποια συστάδα είναι απομονωμένη (οπότε θα είναι $L-$ ή L^*- συστάδα) ή όχι .

```
> pam$isolation
 1  2  3
no no no
Levels: no L L*
```

- "clusinfo", δηλαδή παρουσιάζει σε έναν πίνακα το μέγεθος (size) κάθε συστάδας, την μέγιστη (max_diss) και την μέση απόσταση (av_diss) των γονιδίων κάθε συστάδας από το medoid της συστάδας, τη διάμετρο (diameter) και την διαίρεση (separation) κάθε συστάδας.

```
> pam$clusinfo
     size  max_diss   av_diss   diameter  separation
[1,]    1   0.00000   0.00000   0.00000    30.23114
[2,]   23  26.47336  15.32555  35.82881    13.47912
[3,]   13  23.71683  14.91430  37.18188    13.47912
```

- "silinfo", δηλαδή εμφανίζει έναν πίνακα με τα silhouette widths των γονιδίων, τις συστάδες που ανήκουν αλλά και τις γειτονικές τους συστάδες. Επιπλέον, κάτω από τον πίνακα δίνεται ο μέσος όρος των silhouette widths κάθε συστάδας και ο συνολικός μέσος όρος των silhouette widths. Να σημειώσουμε ότι τα silhouette widths εμφανίζονται κατά φθίνουσα διάταξη σε κάθε συστάδα.

```
> pam$silinfo
$widths
                      cluster  neighbor   sil_width
PPIA.23primary          1        3        0.000000000
AKR1C3.24primary        2        3        0.505652469
EGFR.24primary          2        3        0.486957014
VEGFR2.24primary        2        3        0.474741282
          . . .         . . .    . . .        . . .
STC2.24primary          3        2        0.200424419
ALCAM.23primary         3        2        0.179763703
SFRP1.24primary         3        2        0.092171820

$clus.avg.widths
[1] 0.0000000  0.3722739  0.3361764
$avg.width
[1] 0.3495296
```

- "diss", δηλαδή μας δίνει τον πίνακα αποστάσεων και το μέτρο απόστασης. Στη συγκεκριμένη περίπτωση έχουμε εφαρμόσει ευκλείδεια απόσταση (metric="euclidean") σε τυποποιημένα (stand=TRUE) δεδομένα.

- "data", όπου εμφανίζει τον πίνακα δεδομένων, dat.

Μπορούμε να παρουσιάσουμε γραφικά τις συστάδες που προκύπτουν και τα αντίστοιχα medoids τους χρησιμοποιώντας διαφορετικό χρώμα. Παρακάτω δίνονται οι κατάλληλες εντολές:

```
plot(dat, col=pam$clustering, xlab="ασθενής 1", ylab="ασθενής 2",
main="PAM ομαδοποίηση σε 3 συστάδες")
points(pam$medoids,col=1:3,pch=11)
text(32.5, 32.5, label="VEGFR1")
text(35.8, 35.8, label="VEGFB")
text(39, 39.5, label="PPIA")
legend(28.8,38.5,c("Συστάδα 1","Συστάδα 2","Συστάδα 3"),col=1:3,lty=1)
```

Η ομαδοποίηση τόσο με την μέθοδο $k-$means όσο και με την μέθοδο PAM για συγκεκριμένο πλήθος συστάδων μπορούν να αξιολογηθούν μέσω των silhouette widths, τα οποία μπορούμε να παρουσιάσουμε γραφικά μέσω του silhouette plot για καλύτερα συμπεράσματα ως προς την καταλληλότητα της συγκεκριμένης μεθόδου ομαδοποίησης. Θα μελετήσουμε την κατασκευή silhouette plot ξεχωριστά για την $k-$means και PAM μέθοδο.

Για να υπολογίσουμε τα silhouette widths των ομαδοποιημένων γονιδίων σύμφωνα με την $k-$means μέθοδο, αρκεί να εισάγουμε στην εντολή silhouette() το αντικείμενο cl$cluster που μας δίνει η εντολή cl=kmeans();cl και το μέτρο απόστασης ως εξής:

```
diss <- dist(dat)
sk <- silhouette(cl$cluster,diss)
rownames(sk)=names(Groups_of_genes)
sortSilhouette(sk)
```

Οι εντολές αυτές μας δίνουν σε μορφή πίνακα τα silhouette widths των γονιδίων διατεταγμένα σε φθίνουσα διάταξη σε κάθε συστάδα λόγω της εντολής sortSilhouette(sk), την συστάδα που ανήκει κάθε γονίδιο καθώς και τη γειτονική του συστάδα.

```
                      cluster     neighbor      sil_width
ABAT.24primary          1           2          0.39050484
PPIA.23primary          1           2          0.32968981
CXCL13.23primary        1           2          0.29052509
      . . .           . . .       . . .          . . .
VEGFA.24primary         3           2          0.12406152
ERBB3.24primary         3           2          0.10000069
IGKC.23primary          3           2         -0.03360509

attr(,"call")
silhouette.default(x = cl$cluster, dist = diss)
attr(,"class")
[1] "silhouette"
attr(,"iOrd")
 [1]  26 1  21 11 6  13  9  33  5  22  2  24 14  7 29 30  8  31
[19] 19 17 23  3 16 37 36 25 10 28 35 32  4 12 27 34 18 15
[37] 20
attr(,"Ordered")
[1] TRUE
```

Τα τελευταία τέσσερα αντικείμενα αφορούν αντίστοιχα την εντολή που εφαρμόσαμε για τον υπολογισμό των silhouette widths, την κατηγορία στην οποία ανήκει αυτή η εντολή, το αντίστοιχο id αριθμό στον πίνακα δεδομένων που έχουν τα γονίδια με την σειρά που εμφανίζονται στον παραπάνω πίνακα και τέλος αν διατάξαμε ή όχι κατά φθίνουσα τάξη τα silhouette widths (συνεπώς και τα γονίδια) εντός κάθε συστάδας.

Παρατήρηση

Παρακάτω δίνεται μία ακολουθία εντολών για να μπορούμε να εμφανίζουμε τα ονόματα των γονιδίων στο silhouette plot και το αποτέλεσμα της εντολής silhouette() μέσω της εντολής names(Groups_of_genes):

```
mat=as.factor(cl$cluster);mat
Groups_of_genes=sort(mat, decreasing = FALSE)
as.data.frame(Groups_of_genes)
name=names(Groups_of_genes)
```

Με την εντολή summary(sk) παίρνουμε το μέγεθος και τον μέσο όρο των silhouette widths κάθε συστάδας, αλλά και τα περιγραφικά στατιστικά των silhouette widths:

```
Silhouette of 37 units in 3 clusters from
silhouette.default(x = cl$cluster, dist = diss) :

Cluster sizes and average silhouette widths:
      8          10          19
0.2418807  0.2170527  0.2350883

Individual silhouette widths:
    Min.      1st Qu.   Median    Mean      3rd Qu.   Max.
-0.03361   0.18670   0.24960   0.23170   0.29130   0.39290
```

Για την γραφική απεικόνιση των silhouette widths των γονιδίων κάθε συστάδας εφαρμόζουμε την παρακάτω εντολή:

```
plot(sk,cex.names=0.8,do.n.k=T, do.clus.stat=T, main="Silhouette plot")
```

Εκτός από την γραφική αναπαράσταση των silhouette widths των γονιδίων, μπορούμε να δούμε τον συνολικό (*overall*) μέσο όρο των silhouette widths, όπως επίσης το μέγεθος και τον μέσο όρο των silhouette widths κάθε συστάδας. Για να βρούμε τον δείκτη $SC = \max \bar{s}(k)$ για διάφορα k αρκεί να εφαρμόσουμε τις παρακάτω εντολές αλλάζοντας κάθε φορά την τιμή της παραμέτρου k στην εντολή initial.Centers(dat, k):

```
ic=initial.Centers(dat, k )
centers=dat[ic,]
cl=kmeans(dat,centers)
diss <- dist(dat)
sk <- silhouette(cl$cluster,diss)
plot(sk,cex.names=0.8,do.n.k=T, do.clus.stat=T, main="Silhouette plot")
```

Για να μελετήσουμε τα silhouette widths των γονιδίων που έχουν ομαδοποιηθεί σύμφωνα με την μέθοδο PAM αρκεί να εφαρμόσουμε τις παρακάτω εντολές

```
pam=pam(dat, k=3,metric="euclidean", medoids=NULL, diss=FALSE,
cluster.only=FALSE, do.swap=TRUE)
diss <- dist(dat)
sk2 <- silhouette(pam,diss)
rownames(sk2)=names(Groups_of_genes)
summary(sk2)
```

Ουσιαστικά εφαρμόζουμε την εντολή pam = pam() και στην εντολή silhouette() τοποθετούμε το όνομα pam της παραπάνω εντολής στη θέση του cl$cluster, που χρησιμοποιήσαμε προηγουμένως για να μελετήσουμε τα silhouette widths των γονιδίων, όπως ομαδοποιήθηκαν με την $k-$means μέθοδο. Όπως παρουσιάσαμε γραφικά τα silhouette widths για την $k-$means μέθοδο, μπορούμε με τον ίδιο τρόπο να παρουσιάσουμε γραφικά τα silhouette widths για την PAM μέθοδο.

Παράρτημα Β2

Παρουσίαση των εντολών του προγράμματος R που χρησιμοποιήθηκαν στην ομαδοποίηση εκφράσεων γονιδίων

Σε αυτό το παράρτημα παρουσιάζουμε και αναπτύσσουμε τις εντολές της **R** που εφαρμόστηκαν για τα αποτελέσματα που αναλύσαμε στο κεφάλαιο 3.

Impute Missing Values with k-nearest neighbour (impute.knn {impute})

Αποδίδει τα missing values ενός σετ δεδομένων χρησιμοποιώντας το μοντέλο της Principal Component Analysis (PCA).

impute.knn(x, k, rowmax, colmax)

x: πίνακας ή data.frame, όπου οι γραμμές ορίζουν τις μεταβλητές που περιέχουν συνεχείς ελλείπουσες τιμές..

k: αριθμός γειτόνων που θα χρησιμοποιηθούν στην απόδοση των ελλειπουσών τιμών. Ως προεπιλογή ορίζεται $k = 10$.

rowmax: το μέγιστο ποσοστό ελλειπουσών τιμών που επιτρέπονται σε κάθε γραμμή. Ως προεπιλογή ορίζουμε **rowmax = 0.5**. Οι ελλείπουσες τιμές αποδίδονται

χρησιμοποιώντας τον ολικό μέσο ανά δείγμα αν και μόνο αν **rowmax > 0.5** για κάποιες γραμμές.

colmax: το μέγιστο ποσοστό ελλειπουσών τιμών που επιτρέπονται σε κάθε στήλη. Ως προεπιλογή ορίζουμε **colmax = 0.8**. Εάν κάποιες στήλες έχουν **colmax > 0.8**, τότε η εντολή **impute.knn()** δεν δίνει αποτέλεσμα. Στην περίπτωση αυτή καλύτερα να ορίσουμε **colmax = 1**.

Παρατήρησεις

1. Η μέθοδος του $k-$κοντινότερου γείτονα θεωρείται η πιο αξιόλογη μέθοδος απόδοσης των ελλειπουσών τιμών (Troyanskaya, 2001). Επίσης, ως άριστο θεωρείται το πλήθος γειτόνων k που κυμαίνεται από το 10 ως το 20 (Hastie et al. 1999).

2. Ουσιαστικά παίρνουμε τον αρχικό πίνακα, όπου τώρα οι ελλείπουσες τιμές έχουν συμπληρωθεί σύμφωνα με την μέθοδο του $k-$κοντινότερου γείτονα. Ο πίνακας αυτός ορίζεται από το αντικείμενο **$data**.

Correlation Matrices (cor {stats})

Υπολογίζει την συσχέτιση των μεταβλητών x και y αν αυτές είναι διανύσματα. Αν είναι πίνακες τότε η συσχέτιση υπολογίζεται μεταξύ των στηλών του πίνακα x και των στηλών του πίνακα y. Στην πρώτη περίπτωση το αποτέλεσμα είναι μία τιμή, ενώ στην δεύτερη περίπτωση το αποτέλεσμα είναι ένας πίνακας συσχέτισης.

cor(x, y, method=c(“pearson”,”kendall”,”spearman”))

x: αριθμητικό διάνυσμα ή πίνακας ή data.frame.

y: ως προεπιλογή ορίζεται **NULL,** αν ισχύει ότι $y = x$. Διαφορετικά μπορεί να είναι ένα διάνυσμα ή ένας πίνακας ή ένα data.frame με διάσταση όπως του x.

method: ως προεπιλογή ορίζεται **"pearson",** διαφορετικά ορίζουμε **"kendall"** ή **"spearman"**.

Enhanced Heat Map (heatmap.2 {gplots})

Κατασκευάζει θερμικό χάρτη με ή χωρίς δενδρόγραμμα στην πάνω πλευρά ή / και στην αριστερά πλευρά του θερμικού χάρτη. Η συγκεκριμένη συνάρτηση έχει περισσότερες επιλογές από την συνάρτηση **hmap (hmap {seriation})** κυρίως ως προς τα χρώματα που θα χρησιμοποιήσουμε.

heatmap.2(x, distfun = dist, hclustfun = hclust, Rowv ,scale, key, symkey, col = redgreen(75), density.info, trace, cexRow, symm, main, margins = c(·,·))

x: ένας πίνακας δεδομένων ορισμένος ως **as.matrix()**.

distfun: συνάρτηση που υπολογίζει την απόσταση τόσο μεταξύ των γραμμών όσο και μεταξύ των στηλών. Ως προεπιλογή ορίζεται **dist**.

hclustfun: συνάρτηση που εφαρμόζει ιεραρχική ομαδοποίηση. Ως προεπιλογή ορίζεται **hclust**.

Rowv: στην συνάρτηση μπορούμε να εισάγουμε την επιλογή **Rowv**, όπου καθορίζει πως θα υπολογιστεί το δενδρόγραμμα των γραμμών, αν θα γίνει αναδιάταξη των γραμμών και πως. Αν παραλείψουμε αυτή την επιλογή τότε η αναδιάταξη των γραμμών γίνεται σύμφωνα με τον μέσο τους και πάντα κατ' αύξουσα σειρά. Αν ορίσουμε **Rowv = NULL**, τότε δεν γίνεται αναδιάταξη γραμμών. Ανάλογα ορίζεται και η επιλογή **Colv**. Αν ο πίνακας είναι συμμετρικός, οπότε έχουμε ορίσει **symkey=TRUE**, τότε ορίζουμε **Colv = "Rowv"**, οπότε η αναδιάταξη των στηλών γίνεται σύμφωνα με την αναδιάταξη των γραμμών.

scale: αν ο πίνακας **x** είναι συμμετρικός τότε ορίζουμε **"none"**, διαφορετικά ορίζουμε **"row"** ή **"column"** αν θέλουμε να τυποποιήσουμε ως προς τις γραμμές ή τις στήλες του πίνακα αντίστοιχα.

key: αν ορίσουμε **TRUE** τότε εμφανίζεται πάνω-αριστερά στο output ένα «υπόμνημα» με κλιμακωμένες αποχρώσεις της παλέτας χρωμάτων που έχουμε επιλέξει, ενώ αν ορίσουμε **FALSE** τότε δεν εμφανίζεται αυτό το «υπόμνημα».

symkey: εάν ο πίνακας **x** περιέχει αρνητικές τιμές, τότε ορίζουμε **TRUE**. Με αυτόν τον τρόπο τα χρώματα στο «υπόμνημα» εμφανίζονται συμμετρικά γύρω από το μηδέν.

col: με την επιλογή **redgreen(75)** επιλέγουμε την παλέτα κόκκινου και πράσινου χρώματος με 75 συνολικά αποχρώσεις.

density.info: με την επιλογή **"none"** δεν εμφανίζεται στο «υπόμνημα» χρωμάτων ένα ιστόγραμμα.

trace: με την επιλογή **"none"** δηλώνουμε ο θερμικός χάρτης να μην εμφανίζεται με διαχωρισμένες τις στήλες ή τις γραμμές, όπου θα ξεχώριζαν μεταξύ τους οι αντίστοιχες μεταβλητές.

cexRow: ρυθμίζουμε το μέγεθος των ονομάτων των γραμμών. Μέχρι το 0.5 η γραμματοσειρά είναι πολύ μικρή, οπότε προτιμάται όταν το πλήθος των ονομάτων είναι μεγάλο, διότι έτσι το ένα όνομα δεν καλύπτει το άλλο. Αντίστοιχα, ορίζουμε την επιλογή **cexCol** που αφορά τις στήλες.

symm: αν ο πίνακας είναι συμμετρικός ορίζουμε **TRUE**, διαφορετικά ορίζουμε **FALSE.**

margins: είναι ένα αριθμητικό διάνυσμα μήκους ίσο με 2, όπου περιέχει τα όρια για τα ονόματα των γραμμών και των στηλών αντίστοιχα.

Αντικείμενα του Output Επιλεκτικά

$colInd: εμφανίζει το id που έχουν οι στήλες στον πίνακα δεδομένων, οι οποίες είναι διατεταγμένες στον θερμικό χάρτη σύμφωνα με το δενδρόγραμμα. Ανάλογα ερμηνεύεται το αντικείμενο **$rowInd** που αφορά τις γραμμές.

$carpet: περιέχει τις αναδιατεταγμένες γραμμές και στήλες με τυποποιημένες τιμές κατά στήλες ή γραμμές, όπου δημιουργούν τον θερμικό χάρτη.

$colMeans: εμφανίζεται αν ορίσουμε την επιλογή **scale = "column",** οπότε για κάθε στήλη εμφανίζει τη μέση τιμή. Ανάλογα ερμηνεύεται το αντικείμενο **$rowMeans** που αφορά τις γραμμές.

$colSDs: εμφανίζεται, επίσης, αν ορίσουμε την επιλογή **scale = "column",** οπότε για κάθε στήλη εμφανίζει την τυπική απόκλιση. Ανάλογα ερμηνεύεται το αντικείμενο **$rowSDs** που αφορά τις γραμμές.

Παρατήρηση

Στο **heatmap.2()** η συσσωρευτική μέθοδος ομαδοποίησης που ορίζεται ως προεπιλογή είναι η complete linkage.

Hierarchical Clustering (hclust {stats})

Εφαρμόζει ιεραρχική ομαδοποίηση σ' έναν πίνακα δεδομένων σύμφωνα με το μέτρο απόστασης μεταξύ στοιχείων και μεταξύ συστάδων ή συστάδων και στοιχείων που θα ορίσουμε.

```
hclust( d, method = "·" )
plot( x, labels, hang = −1 )
```

d: ο πίνακας αποστάσεων, όπου συνήθως παράγεται από την συνάρτηση **dist( ).**

method: η μέθοδος υπολογισμού των αποστάσεων μεταξύ στοιχείων και συστάδων ή μεταξύ συστάδων, όπως **"ward"**, **"single"**, **"complete"**, **"average"**, **"median"** και **"centroid".**

x: εισάγω το όνομα της συνάρτησης **hclust( ).**

labels: αν ορίσουμε **NULL**, τότε στα φύλλα του δενδρογράμματος θα εμφανίζονται τα ονόματα των στοιχείων που ομαδοποιήσαμε, ενώ αν ορίσουμε **FALSE** τότε θα εμφανίζεται μόνο ο αριθμός που αντιστοιχεί στην θέση τους στον πίνακα δεδομένων.

hang: καθορίζουμε από ποιο σημείο του άξονα των αποστάσεων του δενδρογράμματος θέλουμε να «κρέμονται» τα ονόματα των στοιχείων. Οποιαδήποτε αρνητική τιμή ορίζει να «κρέμονται» από το μηδέν, οπότε το δενδρόγραμμα είναι πιο εμφανίσιμο.

Αντικείμενα του Output Επιλεκτικά

$merge: πρόκειται για έναν πίνακα $(n-1)\times 2$, όπου n είναι το πλήθος των στοιχείων που ομαδοποιούμε. Κάθε γραμμή αντιπροσωπεύει και από ένα στάδιο στο οποίο συγχωνεύονται δύο στοιχεία ή δύο συστάδες ή ένα στοιχείο και μία συστάδα. Έστω j το στοιχείο σε μία γραμμή αυτού του πίνακα το οποίο έχει αρνητική τιμή. Τότε θα συγχωνευτεί με συστάδα αν η τιμή στην δίπλα στήλη είναι θετική, αλλιώς θα συγχωνευτεί με στοιχείο. Αν σε μία γραμμή και οι δύο τιμές είναι θετικές, τότε έχουμε συγχώνευση δύο συστάδων σε μία.

$height: είναι ένα διάνυσμα μήκους $n-1$ το οποίο περιέχει κατ' αύξουσα σειρά τις αποστάσεις μεταξύ στοιχείων, συστάδων ή συστάδων και στοιχείων (συνολικά πραγματοποιούνται $n-1$ συγχωνεύσεις).

$order: είναι ένα διάνυσμα μήκους n το οποίο περιέχει τον αριθμό κάθε ομαδοποιημένου στοιχείου που αντιστοιχεί στην θέση του στον πίνακα δεδομένων.

$labels: πρόκειται για ένα διάνυσμα μήκους n το οποίο περιέχει τα ονόματα των στοιχείων που ομαδοποιήσαμε.

$method: εμφανίζει την συσσωρευτική μέθοδο που ορίσαμε .

Distance Matrix Computation (dist {stats})

Υπολογίζει τον πίνακα αποστάσεων μεταξύ των γραμμών σύμφωνα με το μέτρο αποστάσεων που ορίσαμε.

dist(y, method = "·", diag, upper)

y: ο πίνακας δεδομένων ή το data.frame που θα αναλύσουμε. Ως μεταβλητές ορίζονται οι γραμμές.

method: ορίζουμε την μέθοδο αποστάσεων μεταξύ των στοιχείων. Επιλέγουμε μεταξύ των **"euclidean"**, **"manhattan"**, **"maximum"**, **"canberra"**, **"binary"** και **"minkowski".**

diag: αν επιλέξουμε **FALSE**, τότε στον πίνακα αποστάσεων δεν θα εμφανίζεται η διαγώνιος $diag(0,0,...,0)$.

upper: αν επιλέξουμε **FALSE**, τότε το άνω τρίγωνο (χωρίς τη διαγώνιο) του πίνακα αποστάσεων δεν θα εμφανίζεται.

Παρατήρηση

Η απόσταση του Pearson δεν μπορεί να οριστεί στην συνάρτηση **dist().** Η δυσκολία αυτή αντιμετωπίζεται αν εφαρμόσουμε την συνάρτηση **1-cor(),** την εισάγουμε στην εντολή **as.dist()** και ορίσουμε με αυτόν τον τρόπο την επιλογή **d** στην συνάρτηση **hclust().**

UPGMA Tree (upgma {phangorn})

Εφαρμόζει ιεραρχική ομαδοποίηση χρησιμοποιώντας τη μέθοδο UPGMA και παρουσιάζει ένα δενδρόγραμμα

```
upgma( d, names, method=”ave” )
```

d: ο πίνακας αποστάσεων που υπολογίσαμε.

names: τα ονόματα (labels) των στοιχείων που ομαδοποιούμε.

method: η συσσωρευτική μέθοδος που επιλέξαμε. Στην συγκεκριμένη περίπτωση ορίσαμε την **average** μέθοδο.

Validate Cluster Results (clValid {clValid})

Αξιολογεί τις μεθόδους ομαδοποίησης που εφαρμόσαμε, το εύρος πλήθους των συστάδων που επιλέξαμε και καταλήγει στην άριστη μέθοδο ομαδοποίησης και πλήθος συστάδων χρησιμοποιώντας εσωτερικά μέτρα, μέτρα σταθερότητας και βιολογικά μέτρα.

clValid(obj, nClust, clMethods=$c(\text{"·"},...)$**, validation, maxitems, metric, method)**

obj: ο πίνακας ή data.frame δεδομένων. Τα στοιχεία που θέλουμε να ομαδοποιήσουμε βρίσκονται στις γραμμές.

nClust: αριθμητικό διάνυσμα σε μορφή εύρους, δηλαδή $a:b\ (a \geq 2,\ b \leq n-1)$ που δίνει το πλήθος των συστάδων που θέλουμε να εξετάσουμε.

clMethods: οι μέθοδοι ομαδοποίησης (ιεραρχικής και μη) που θέλουμε να αξιολογήσουμε. Ορίζεται ως διάνυσμα με μήκος τουλάχιστον ίσο με μονάδα με δυνατές επιλογές τις μεθόδους **"hierarchical"**, **"kmeans"**, **"diana"**, **"fanny"**, **"som"**, **"model"**, **"sota"**, **"pam"**, **"clara"**, και **"agnes"**.

validation: επιλέγουμε κάποια από τρία μέτρα αξιολόγησης, δηλαδή **"internal"**, **"stability"** και **"biological"**.

maxitems: ορίζουμε το μέγιστο πλήθος στοιχείων που μπορούν να ομαδοποιηθούν, δηλαδή ίσο με n.

metric: αφορά το μέτρο απόστασης που θέλουμε να χρησιμοποιήσουμε. Μπορούμε να επιλέξουμε μεταξύ των δυνατών επιλογών **"euclidean"**, **"correlation"** και **"manhattan"**.

method: η συσσωρευτική μέθοδος ιεραρχικής ομαδοποίηση με δυνατές επιλογές την **"ward"**, **"single"**, **"complete"** και **"average"**. Αφορά μόνο τις συναρτήσεις **hclust** και **agnes**.

Παρατηρήσεις

1. Με την εντολή **plot()** στην οποία έχουμε εισάγει την συνάρτηση **clValid** παίρνουμε διαφορετικά γραφήματα ένα για κάθε μέτρο, όπου συγκρίνονται για κάθε μέγεθος συστάδας που έχουμε επιλέξει οι μέθοδοι ομαδοποίησης που μας ενδιαφέρουν.

2. Για να υπολογίσουμε τα βιολογικά μέτρα χρειαζόμαστε τα ονόματα των λειτουργικών ομάδων (*functional classes, FC*) στις οποίες ταξινομούνται τα γονίδια και επιπλέον την παράμετρο **annotation** στην εντολή **clValid()**, στην οποία ορίζουμε το όνομα της μεταβλητής-παράγοντα που για κάθε γονίδιο έχει αντιστοιχίσει την λειτουργική ομάδα στην οποία ανήκει. Οπότε η εντολή που χρησιμοποιούμε για το υπολογισμό των βιολογικών μέτρων ορίζεται ως

clValid(obj, nClust, clMethods=$c("\cdot",...)$**, validation="biological", maxitems, metric, method, annotation)**

Calculation of Initial Cluster Centers for k-Means (initial.Centers {clusterSim})

Η συγκεκριμένη συνάρτηση προσδιορίζει τα αρχικά κέντρα βάρους, όπως το SPSS, ως εξής:

- Επιλέγει τα πρώτα k διανύσματα-γραμμές ως κέντρα βάρους (ή μητρικά σημεία).
- Υπολογίζει τις αποστάσεις των υπολοίπων $p-k$ διανυσμάτων-γραμμών από τα κέντρα αυτά, όπως επίσης και τις αποστάσεις μεταξύ των κέντρων αυτών χρησιμοποιώντας την ευκλείδεια απόσταση.
- Εντοπίζει το διάνυσμα-γραμμή και το κέντρο με την μικρότερη μεταξύ τους απόσταση. Αν η απόσταση μεταξύ παρατήρησης και κέντρου είναι μεγαλύτερη από την απόσταση των δύο κοντινότερων κέντρων, τότε με το διάνυσμα αυτό αντικαθιστάται εκείνο το κέντρο που απέχει ελάχιστα από το διάνυσμα και έτσι ορίζεται το νέο κέντρο.
- Η διαδικασία αυτή επαναλαμβάνεται $p-k$ φορές έως ότου καταλήξουμε στα κέντρα που θα χρησιμοποιήσουμε στην k – means μέθοδο.

initial.Centers(x, k)

x: πίνακας ή data.frame δεδομένων, όπου οι γραμμές ορίζουν τα στοιχεία που θέλουμε να ομαδοποιήσουμε.

k: πλήθος κέντρων βάρους.

Η τιμή αυτής της συνάρτησης είναι η θέση των κέντρων αυτών στον πίνακα δεδομένων που μελετάμε.

k-Means Clustering (kmeans {stats})

Εφαρμόζει την $k-$means μέθοδο ομαδοποίησης για συγκεκριμένο πλήθος μητρικών σημείων και συγκεκριμένο αλγόριθμο.

```
kmeans( x, centers, iter.max, algorithm)
plot( x, col=$cluster)
```

x: πίνακας ή data.frame αριθμητικών δεδομένων, όπου οι γραμμές ορίζουν τα στοιχεία που θέλουμε να ομαδοποιήσουμε.

centers: αν δεν γνωρίζουμε εκ των προτέρων τα κέντρα βάρους, τότε ορίζουμε απλά το πλήθος τους. Ωστόσο, αν τα γνωρίζουμε εκ των προτέρων τότε ορίζουμε το αντίστοιχο διάνυσμα-γραμμών, που αποτελεί τμήμα του πίνακα που μελετάμε.

iter.max: ορίζουμε τον μέγιστο αριθμό επαναλήψεων του αλγόριθμου.

algorithm: προσδιορίζουμε τον αλγόριθμο που θέλουμε να εφαρμόσουμε. Ως προεπιλογή χρησιμοποιείται ο **"Hartigan-Wong"** αλγόριθμος. Ωστόσο, μπορούμε να επιλέξουμε τον **"Lloyd"**, **"Forgy"** ή **"MacQueen"** αλγόριθμο.

Αντικείμενα του Output

$centers: δίνει τα κέντρα βάρους των συστάδων που προέκυψαν.

$cluster: ουσιαστικά είναι ένα διάνυσμα που περιέχει τα ονόματα των στοιχείων που ομαδοποιήσαμε μαζί και την συστάδα στην οποία ανήκουν.

$withinss: για κάθε συστάδα δίνει το εντός-συστάδας άθροισμα (*within-cluster variation*).

$size: για κάθε συστάδα δίνει το μέγεθός της.

Παρατηρήσεις

1. Λαμβάνοντας υπόψη το αποτέλεσμα της συνάρτησης **initial.Centers()**, η παράμετρος **centers** ορίζεται ως **centers** = **x[ic,]**, όπου **ic** = **initial.Centers(x, k)**.

2. Για να εμφανιστούν τα κέντρα βάρους στο scatterplot της $k-$means ομαδοποίησης εφαρμόζουμε κάτω από την συνάρτηση **plot(x, col=$cluster)** την συνάρτηση (χαμηλού επιπέδου) **points($centers, col=1:k, pch=8, cex=3)**, όπου με την παράμετρο **pch=8** επιλέγουμε το κέντρο κάθε συστάδας να εμφανίζεται με το σύμβολο $*$, ενώ με την παράμετρο **cex=3** προσδιορίσαμε το μέγεθος του συμβόλου $*$. Κάθε συστάδα θα έχει το δικό της χρώμα και αυτό το προσδιορίζουμε με την παράμετρο **col=1:k**, όπου k είναι το πλήθος τω νσυστάδων.

Partitioning Around Medoids (PAM) Object (pam.object {cluster})

Πραγματοποιεί ομαδοποίηση των στοιχείων με την μέθοδο PAM για συγκεκριμένα medoids και γνωστό πλήθος συστάδων.

pam(x, k, diss, metric, medoids, cluster.only, do.swap)

x: πίνακας ή data.frame αριθμητικών δεδομένων (οι γραμμές ορίζουν τα στοιχεία που θέλουμε να ομαδοποιήσουμε) ή πίνακας αποστάσεων.

k: θετικός ακέραιος αριθμός που προσδιορίζει το πλήθος των συστάδων. Επιβάλλεται να ισχύει ότι $1 \le k \le n-1$.

diss: αν ορίσουμε **TRUE**, τότε το **x** θα θεωρηθεί ως πίνακας αποστάσεων (*dissimilar matrix*), διαφορετικά θα θεωρηθεί ως πίνακας δεδομένων (*data matrix*).

metric: καθορίζει το μέτρο απόστασης. Επιλέγουμε **"euclidean"** ή **"manhattan"**, όπου συνηθίζονται στην PAM μέθοδο ομαδοποίησης. Ωστόσο, εάν το **x** είναι πίνακας αποστάσεων, τότε παραλείπεται αυτή η παράμετρος.

medoids: αν δεν γνωρίζουμε εκ των προτέρων τα αρχικά medoids επιλέγουμε **NULL**, οπότε εφαρμόζεται το build-step για τον προσδιορισμό τους. Εάν τα γνωρίζουμε ήδη, τότε κατασκευάζουμε ένα διάνυσμα μήκους k που περιέχει τη θέση των στοιχείων αυτών ως αρχικά medoids στον πίνακα δεδομένων που μελετάμε.

cluster.only: αν επιλέξουμε **TRUE**, τότε στο output εμφανίζονται μόνο τα ονόματα των στοιχείων που ομαδοποιούμε και η συστάδα στην οποία ανήκουν. Ωστόσο, προτιμάται η επιλογή **FALSE**, διότι έτσι εμφανίζονται στο output περισσότερα αποτελέσματα.

do.swap: αν επιλέξουμε **TRUE**, τότε πραγματοποιείται το swap-step. Για μεγάλο πλήθος στοιχείων η διαδικασία αυτή θεωρείται υπολογιστικά χρονοβόρα, οπότε συνιστάται η επιλογή **FALSE** και περιοριζόμαστε στο build-step.

Αντικείμενα του Output

$medoids: εάν έχουμε δηλώσει πίνακα αποστάσεων, τότε δίνεται ένα διάνυσμα με τα ονόματα των στοιχείων που αποτελούν τα medoids. Διαφορετικά, δίνεται ένας πίνακας όπου οι γραμμές του ορίζουν τα medoids.

$id.med: διάνυσμα με ακέραιες τιμές, όπου κάθε τιμή ορίζει τη θέση του στοιχείου-medoid στον πίνακα δεδομένων.

$clustering: δίνει ένα διάνυσμα με τα ονόματα των στοιχείων και τη συστάδα στην οποία ανήκουν.

$objective: δίνει την αντικειμενική συνάρτηση (*Objective Function, OF*) του build-step και του swap-step.

$isolation: πρόκειται για ένα διάνυσμα με μήκος όσο το πλήθος των συστάδων, όπου καθορίζει ποιες συστάδες είναι $L-$ ή L^*- απομονωμένες (*isolated*) και ποιες όχι. Να σημειώσουμε ότι όταν μία συστάδα είναι L^*- isolated, τότε είναι και $L-$ isolated.

$clusinfo: δίνει έναν πίνακα διαστάσεων $k \times 5$, όπου k είναι το πλήθος των συστάδων και οι στήλες αντιπροσωπεύουν αντίστοιχα το μέγεθος κάθε συστάδας (*size*), την μέγιστη απόσταση των στοιχείων από το medoid της συστάδας που ανήκουν (*max_diss*), την μέση απόσταση των στοιχείων από το medoid της συστάδας που ανήκουν (*av_diss*), τη διάμετρο (*diameter*), δηλαδή την μέγιστη απόσταση μεταξύ δύο στοιχείων που ανήκουν στην ίδια συστάδα και τη διαίρεση (*separation*), δηλαδή την ελάχιστη απόσταση μεταξύ δύο στοιχείων που ανήκουν σε διαφορετικές συστάδες.

$silinfo: δίνει έναν πίνακα διαστάσεων $p \times 3$, όπου p είναι το πλήθος των στοιχείων που ομαδοποιούμε και οι στήλες αντιπροσωπεύουν αντίστοιχα τη συστάδα που ανήκει κάθε στοιχείο (*cluster*), την γειτονική συστάδα (*neighbor*) και το silhouette width κάθε στοιχείου (*sil_width*). Επιπλέον, το συγκεκριμένο αντικείμενο περιέχει το μέσο silhouette width κάθε συστάδας ($clus.avg.widths) και το συνολικό silhouette width της ομαδοποίησης ($avg.width).

Παρατήρηση

Το scatterplot της ομαδοποίησης με τη μέθοδο PAM, ορίζεται όπως και της ομαδοποίησης με τη μέθοδο k – means, που αναλύσαμε παραπάνω.

<u>Compute & Extract Silhouette Information from Clustering</u> (silhouette {cluster})

Υπολογίζει τα silhouette widths, average silhouette widths για κάθε συστάδα και κάποια περιγραφικά στατιστικά για τα silhouette widths της μεθόδου ομαδοποίησης που εφαρμόσαμε, έτσι ώστε να την αξιολογήσουμε ως προς το πόσο σωστά έχει κατανείμει τα στοιχεία στις συστάδες που ανήκουν. Συνήθως, εφαρμόζεται για την k – means και PAM μέθοδο ομαδοποίησης.

silhouette(obj, diss)

obj: εισάγουμε τη συνάρτηση **pam()** με το όνομα που της έχουμε ορίσει ή το αντικείμενο **$cluster** της συνάρτησης **kmeans()** με το όνομα που της έχουμε ορίσει, επίσης.

diss: εισάγουμε τη συνάρτηση **dist()** με το όνομα που της έχουμε δώσει.

Παρατήρηση

Με την συνάρτηση **summary()** μπορούμε να μελετήσουμε τα περιγραφικά στατιστικά των silhouette widths, όπως ελάχιστο και μέγιστο silhouette width, το πρώτο και τρίτο τεταρτημόριο, το μέσο και διάμεσο silhouette width.

Για την γραφική απεικόνιση των silhouette widths χρησιμοποιούμε την παρακάτω συνάρτηση με τις παραμέτρους που περιέχει

```
plot( obj, cex.names=0.8, do.n.k, do.clus.stat, border,
main="Silhouette plot")
```

obj: εισάγουμε τη συνάρτηση **silhouette()** με το όνομα που της έχουμε ορίσει.

cex.names: προσδιορίζει το μέγεθος των ονομάτων των στοιχείων στον άξονα y.

do.n.k: αν επιλέξουμε **TRUE**, τότε πάνω δεξιά και αριστερά στο γράφημα θα εμφανίζονται αντίστοιχα το πλήθος των συστάδων και το πλήθος των στοιχείων που ομαδοποιήσαμε.

do.clus.stat: αν επιλέξουμε **TRUE**, τότε σε κάθε συστάδα θα εμφανίζεται το μέγεθός της και το average silhouette width.

border: αν επιλέξουμε **TRUE**, τότε τα silhouette width bars θα διακρίνονται μεταξύ των στοιχείων.

main: αν παραλείψουμε την συγκεκριμένη παράμετρο, τότε ως τίτλος του γραφήματος θα εμφανίζεται η συνάρτηση ομαδοποίησης που ορίσαμε ως αντικείμενο της συνάρτησης **silhouette()**. Για παράδειγμα, αν μελετάμε τα silhouette widths της

PAM μεθόδου, τότε ο τίτλος του silhouette plot θα είναι **"pam(x, k, diss, metric, medoids, cluster.only, do.swap)"** με τις παραμέτρους που έχουμε ορίσει. Για να αποφύγουμε αυτόν τον τίτλο περιλαμβάνουμε την παράμετρο **main** με έναν δικό μας τίτλο.

Παράρτημα Γ1

Εφαρμογή στην Ανάλυση Επιβίωσης με εντολές του προγράμματος R

Για την εκτίμηση της Kaplan-Meier συνάρτησης επιβίωσης χωρίς να λάβουμε υπόψη κάποια συμμεταβλητή χρησιμοποιούμε την εντολή **survfit()**, όπως παρακάτω

```
kaplan1=survfit(Surv(survival, death)~1, conf. type="none")
summary(kaplan1)
```

time	n.risk	n.event	survival	std.err
12.0	315	1	0.997	0.00317
12.1	314	1	0.994	0.00448
13.0	313	1	0.990	0.00547
13.6	312	1	0.987	0.00631
. . .	. . .	. . .	. . .	. . .
117.2	315	1	0.997	0.00317
118.0	314	1	0.994	0.00448
122.8	313	1	0.990	0.00547
126.3	312	1	0.987	0.00631

Εμφανίζει τους διαφορετικούς πλήρεις χρόνους διατεταγμένους (*time*), το πλήθος ατόμων σε κίνδυνο τη χρονική στιγμή t_j (*n.risk*), το πλήθος αποτυχιών τη στιγμή t_j

(*n.event*), την πιθανότητα επιβίωσης τουλάχιστον πέρα από τη χρονική στιγμή t_j (*survival*) και την τυπική απόκλιση της επιβίωσης τύπου Greenwood (*std.error*), όπου ισχύει

$$\sqrt{\hat{V}\left(\hat{S}(t)\right)} = \sqrt{\left\{\hat{S}(t)\right\}^2 \sum_{j:t_j<t} \frac{d_j}{r_j\left(r_j - d_j\right)}}$$

Με την επιλογή **conf.type="none"** δεν εμφανίζει διάστημα εμπιστοσύνης για την συνάρτηση επιβίωσης για κάθε $t \in \left(t_j, t_{j+1}\right]$.

Στην περίπτωση που θέλουμε να μελετήσουμε τον πίνακα επιβίωσης για κάθε επίπεδο μιας συμμεταβλητής, τότε αντί της μονάδας εισάγουμε το όνομα της συμμεταβλητής

```
kaplan2=survfit(Surv(survival, death)~her2, conf. type="none")
summary(kaplan2)
```

```
                her2=0
 time  n.risk  n.event  survival  std.err
 12.1   201       1       0.995   0.00496
 13.6   200       1       0.990   0.00700
 18.1   199       1       0.985   0.00855
  . . .   . . .   . . .    . . .      . . .
108.1    27       1       0.737   0.04393
118.0   201       1       0.995   0.00496
122.8   200       1       0.990   0.00700


                her2=1
 time  n.risk n.event  survival  std.err
 12.0    87      1       0.989   0.0114
 13.0    86      1       0.977   0.0161
 19.2    85      1       0.966   0.0196
  . . .   . . .   . . .    . . .     . . .
102.4    17      1       0.552   0.0688
110.8     4      1       0.414   0.1301
114.5    87      1       0.989   0.0114
```

Η υπό εξέταση συμμεταβλητή HER2 (βιοδείκτης) έχει δύο επίπεδα: το her2=0, δηλαδή η μη-υπερέκφραση (no over-expression) του βιοδείκτη και το her2=1, δηλαδή η υπερέκφραση του βιοδείκτη (over-expression). Μπορούμε να δούμε πόσο διαφέρει η συνάρτηση επιβίωσης μεταξύ αυτών των δύο επιπέδων μέσω γραφικής παράστασης της συνάρτησης επιβίωσης των ομάδων αυτών.

```
plot(kaplan2, xlab="Time in months", ylab="Probability",main="Overall Survival",
lty=1,col=2:3)
legend(8,.25,c("No overexpression","Overexpression"), lty=1:1, col=2:3)
```

Στο γράφημα Γ1.1 παρουσιάζεται η καμπύλη επιβίωσης κάθε επιπέδου του βιοδείκτη. Για να συγκρίνουμε με μεγαλύτερη αξιοπιστία τις συναρτήσεις επιβίωσης αυτών των δύο επιπέδων εφαρμόζουμε στατιστικό έλεγχο, γνωστό ως Mantel-Haenszel έλεγχο για την υπόθεση

$$H_0 : S_1(t) = S_2(t),\ t \le \tau \text{ έναντι } H_1 : S_1(t) \ne S_2(t),\ t \le \tau$$

όπου τ είναι ο μέγιστος πλήρης χρόνος. Η επιλογή μεταξύ Logrank και Gehan test εξαρτάται από το αν οι καμπύλες του γραφήματος Γ1.1 είναι παράλληλες ή όχι. Στην πρώτη περίπτωση ικανοποιείται η υπόθεση του αναλογικού κινδύνου, οπότε εφαρμόζουμε το Logrank test που δίνει ίσο βάρος ($w_j = 1$) και στα δύο άκρα των πλήρων χρόνων ζωής, ενώ στη δεύτερη περίπτωση δεν ικανοποιείται η υπόθεση αναλογικού κινδύνου, οπότε εφαρμόζουμε το Gehan test που δίνει περισσότερο βάρος στο αριστερό άκρο των πλήρων χρόνων ζωής ($w_j = r_j$) όπου υπάρχουν περισσότερα άτομα υπό μελέτη και συνεπώς περισσότερη πληροφορία για τις συναρτήσεις επιβίωσης των δύο επιπέδων. Οι καμπύλες δεν φαίνονται να τέμνονται παρά μόνο στους πρώτους χρόνους. Στη συνέχεια εφαρμόζουμε τον Logrank έλεγχο, όπου χρησιμοποιούμε την εντολή **survdiff()**.

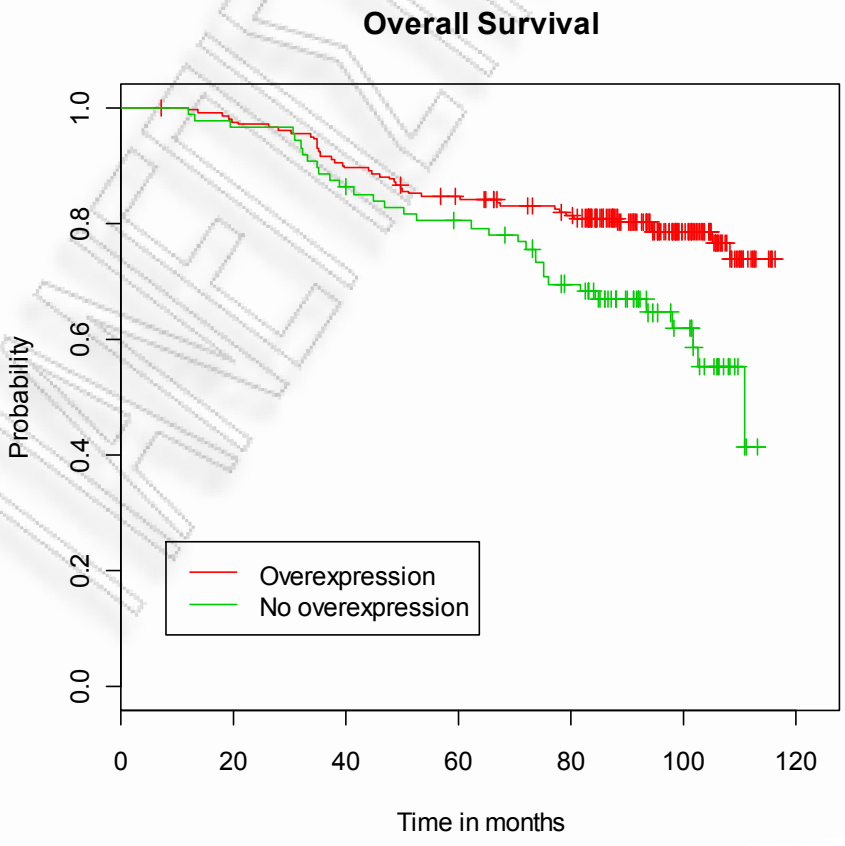
```
comp1=survdiff(Surv(survival, death)~her2, rho=0);comp1
```

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
her2=1	202	43	54.6	2.47	8.78
her2=2	87	33	21.4	6.29	8.78

Chisq= 8.8 on 1 degrees of freedom, p= 0.00304

Ο τύπος (O-E)^2/V αφορά την στατιστική συνάρτηση Q, ενώ ο τύπος (O-E)^2/E είναι ουσιαστικά ο κλασικός χ^2 έλεγχος, όπου προσεγγίζει την ποσότητα Q όταν το δείγμα είναι αρκετά μεγάλο. Επειδή $p-value=0.00304$ απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας $a=5\%$ και συμπεραίνουμε ότι η συνολική επιβίωση διαφέρει μεταξύ των ασθενών με μη-υπερέκφραση και των ασθενών με υπερέκφραση του βιοδείκτη και συγκεκριμένα οι δεύτεροι έχουν μεγαλύτερη πιθανότητα επιβίωσης από τους πρώτους κυρίως μετά τους 60 πρώτους μήνες περίπου (σύμφωνα με γράφημα Γ1.1).

Γράφημα Γ1.1



Παρατηρήσεις

1. Για να εφαρμόσουμε Gehan έλεγχο, τότε στην ρουτίνα **survdiff()**, εξισώνουμε την παράμετρο **rho** με τη μονάδα.
2. Είναι προφανές ότι η συμμεταβλητή που μελετάμε τόσο στην εντολή **survfit()** όσο και στην εντολή **survdiff()** επιβάλλεται να είναι κατηγορική. Στην περίπτωση συνεχούς μεταβλητής πρέπει να κατηγοριοποιηθεί κατάλληλα.
3. Οι συναρτήσεις επιβίωσης για κάθε επίπεδο μιας συμμεταβλητής μπορούν να βρεθούν εναλλακτικά ως εξής:

```
kaplan3=survfit(Surv(survival, death)~1, subset=her2= =0, conf. type="none")
summary(kaplan3)
kaplan4=survfit(Surv(survival, death)~1, subset=her2= =1, conf. type="none")
summary(kaplan4)
```

όπου με την παράμετρο **subset** επιλέγουμε από το σετ δεδομένων εκείνες τις γραμμές που αντιστοιχούν στο επίπεδο της συμμεταβλητής που μας ενδιαφέρει. Στην περίπτωση αυτή η γραφική απεικόνιση των καμπυλών των επιπέδων της συμμεταβλητής γίνεται ως εξής:

```
plot(kaplan3, xlab="Time in months",ylab="Probability",main="Overall Survival",
lty=1,col=2:3)
lines(kaplan4, lty=1, col=3)
legend(8, .25, c("No overexpression","Overexpression"), lty=1:1, col=2:3)
```

Όταν θέλουμε να μελετήσουμε τον χρόνο επιβίωσης στα επίπεδα μιας συμμεταβλητής λαμβάνοντας υπόψη κάποιον άλλο παράγοντα, τότε ουσιαστικά συγκρίνουμε τα επίπεδα της συμμεταβλητής αυτής σε κάθε στρώμα του παράγοντα. Έστω ότι ενδιαφερόμαστε να συγκρίνουμε την επιβίωση των ασθενών στους οποίους χορηγήθηκε χημειοθεραπεία με πακλιταξέλη (E-T-CMF) με την επιβίωση των ασθενών στους οποίους χορηγήθηκε χημειοθεραπεία χωρίς πακλιταξέλη (E-CMF) για

κάθε επίπεδο του βιοδείκτη HER2. Δηλαδή μας ενδιαφέρει ο τοπικός έλεγχος της υπόθεσης

$H_0 : S_{1h}(t) = S_{2h}(t),\ t \le \tau$ έναντι $H_1 : S_{1h}(t) \neq S_{2h}(t),\ t \le \tau$ για το $h\ (1 \le h \le 2)$ στρώμα του βιοδείκτη HER2. Τότε εφαρμόζουμε την παρακάτω εντολή

```
comp2=survdiff(Surv(survival, death)~group, subset=her2==0, rho=1);comp2
comp3=survdiff(Surv(survival, death)~group, subset=her2==1, rho=1);comp3
```

Για την μη-υπερέκφραση του βιοδείκτη (her2=0) το αποτέλεσμα είναι

```
             N  Observed  Expected  (O-E)^2/E  (O-E)^2/V
group= 1    90     17.6      17.5    0.00142     0.0029
group= 2   112     20.8      21.0    0.00118     0.0029
 Chisq= 0  on 1 degrees of freedom, p= 0.957
```

Για την υπερέκφραση του βιοδείκτη (her2=1) το αποτέλεσμα είναι

```
             N  Observed  Expected  (O-E)^2/E  (O-E)^2/V
group= 1    43     11.5      13.4      0.272      0.665
group= 2    44     15.2      13.3      0.275      0.665
 Chisq= 0.7  on 1 degrees of freedom, p= 0.415
```

Και στα δύο στρώματα αποδεχόμαστε τη μηδενική υπόθεση γιατί το $p-value$ είναι σημαντικά μεγαλύτερο από το επίπεδο σημαντικότητας και συμπεραίνουμε ότι η πιθανότητα συνολικής επιβίωσης των ασθενών δεν διαφέρει σημαντικά είτε χορηγηθεί E-T-CMF είτε E-CMF. Ιδιαίτερα στις ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2 λόγω του ότι το $p-value$ είναι πολύ κοντά στη μονάδα συμπεραίνουμε ότι τόσο οι χρόνοι επιβίωσης όσο και οι αντίστοιχες πιθανότητες επιβίωσης σχεδόν ταυτίζονται.

Με τις παρακάτω εντολές μπορούμε να κατασκευάσουμε για κάθε επίπεδο έκφρασης του βιοδείκτη HER2 τους πίνακες επιβίωσης για τις ασθενείς με πακλιταξέλη (group=0) και για τις ασθενείς χωρίς πακλιταξέλη (group=1).

Για την ομάδα των ασθενών με μη-υπερέκφραση (her2=0)

```
kaplan5=survfit(Surv(survival, death)~group, subset=her2= =0, conf. type="none")
summary(kaplan5)
```

```
                group=0
 time  n.risk  n.event  survival  std.err
 12.1     90        1     0.989   0.0110
 19.8     89        1     0.978   0.0155
 26.2     88        1     0.967   0.0189
  . . .   . . .   . . .    . . .    . . .
 88.3     54        1     0.795   0.0434
 94.5     45        1     0.777   0.0459
108.5     13        1     0.717   0.0714

                group=1
 time  n.risk  n.event  survival  std.err
 13.6    111        1     0.991   0.00897
 18.1    110        1     0.982   0.01263
 19.0    109        1     0.973   0.01539
  . . .   . . .   . . .    . . .    . . .
 78.0     84        1     0.809   0.03762
 94.2     52        1     0.793   0.03998
105.5     23        1     0.759   0.05099
```

Για την ομάδα των ασθενών με υπερέκφραση (her2=1)

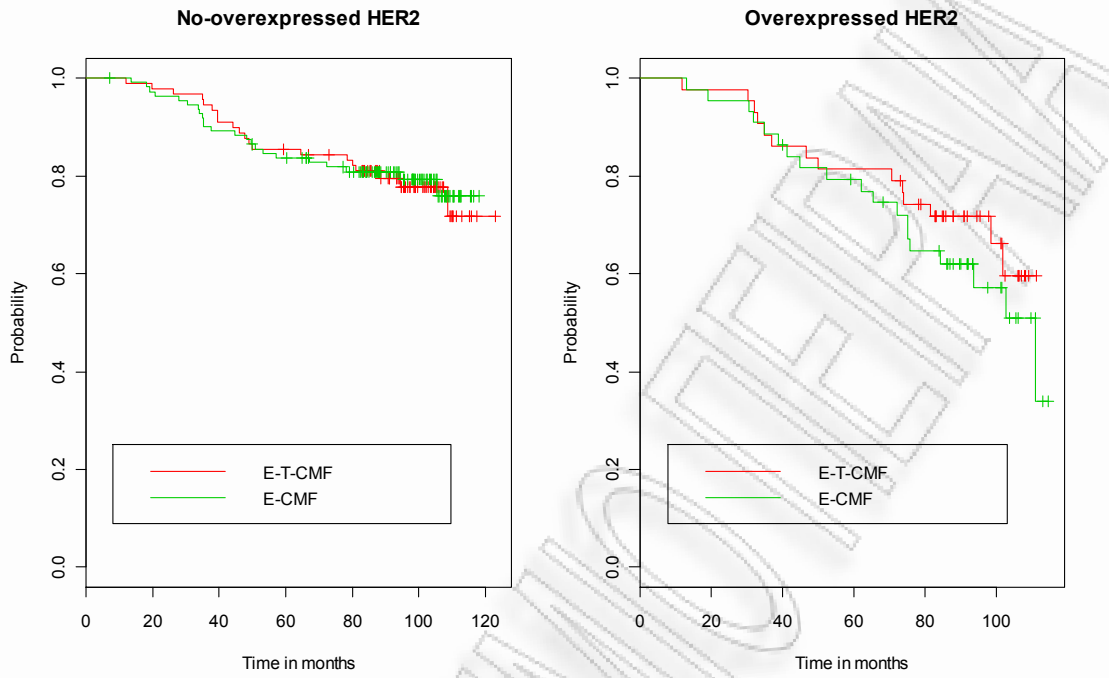
```
kaplan6=survfit(Surv(survival, death)~group, subset=her2= =2, conf. type="none")
summary(kaplan6)
```

```
                group=0
  time n.risk n.event survival std.err
  12.0     43       1    0.977  0.0230
  30.4     42       1    0.953  0.0321
  32.3     41       1    0.930  0.0388
   ...    ...     ...      ...     ...
  81.6     29       1    0.717  0.0693
  98.3     13       1    0.662  0.0831
 101.7     10       1    0.596  0.0977
                group=1
  time n.risk n.event survival std.err
  13.0     44       1    0.977  0.0225
  19.2     43       1    0.955  0.0314
  30.7     42       1    0.932  0.0380
   ...    ...     ...      ...     ...
  93.6     13       1    0.573  0.0832
 102.7      9       1    0.509  0.0952
 110.9      3       1    0.339  0.1524
```

Τα αποτελέσματα από τους τοπικούς στρωματοποιημένους ελέγχους αποδεικνύονται και γραφικά για κάθε επίπεδο του βιοδείκτη χρησιμοποιώντας τις παρακάτω εντολές

```
par(mfcol=c(1,2))
plot(kaplan5, xlab="Time in months",ylab="Probability", main="No-overexpressed HER2", lty=1:1, col=2:3)
legend(8, .25 ,c("E-T-CMF","E-CMF"), lty=1:1, col=2:3)
plot(kaplan6, xlab="Time in months",ylab="Probability", main="Overexpressed HER2", lty=1:1, col=2:3)
legend(8, .25 ,c("E-T-CMF","E-CMF"), lty=1:1, col=2:3)
par(mfcol=c(1,1))
```

Γράφημα Γ1.2



Λόγω του γεγονότος ότι οι καμπύλες τέμνονται και στα δύο γραφήματα εφαρμόσαμε τον Gehan έλεγχο, δηλαδή **rho=1**.

Για τον ολικό στρωματοποιημένο έλεγχο δεν συγκρίνουμε τις συναρτήσεις επιβίωσης των δύο ομάδων ασθενών για κάθε επίπεδο του βιοδείκτη, αλλά λαμβάνοντας υπόψη τον βιοδείκτη αυτό. Συγκεκριμένα ελέγχουμε την υπόθεση

$$H_0 : S_{1h}(t) = S_{2h}(t),\ t \le \tau \ \text{έναντι}\ \ H_1 : S_{1h}(t) \neq S_{2h}(t),\ t \le \tau,\ 1 \le h \le 2$$

χρησιμοποιώντας την παρακάτω εντολή

```
comp4=survdiff(Surv(survival, death)~group+strata(her2), rho=1);comp4
```

```
            N  Observed  Expected  (O-E)^2/E  (O-E)^2/V
group= 1  133      29.2      30.9     0.0993      0.219
group= 2  156      36.0      34.2     0.0897      0.219
 Chisq= 0.2  on 1 degrees of freedom, p= 0.64
```

Λόγω $p-value$ συμπεραίνουμε ότι λαμβάνοντας υπόψη τον βιοδείκτη HER2 οι συναρτήσεις επιβίωσης αυτών των δύο ομάδων δεν διαφέρουν σημαντικά. Στο ίδιο συμπέρασμα καταλήξαμε και με τους τοπικούς ελέγχους. Ωστόσο, πρέπει να έχουμε υπόψη ότι το αποτέλεσμα του ολικού ελέγχου δεν συμπίπτει πάντα με τα αποτελέσματα των τοπικών ελέγχων. Για παράδειγμα, θα μπορούσαν οι τοπικοί έλεγχοι να έδειχναν ότι οι συναρτήσεις επιβίωσης διαφέρουν στα επίπεδα του βιοδείκτη, αλλά ο ολικός έλεγχος να μην καταγράφει αυτή τη διαφορά.

Για να εφαρμόσουμε το μοντέλο παλινδρόμησης του Cox χρησιμοποιούμε την εντολή **coxph()**. Έστω ότι θέλουμε να εξετάσουμε αν ο βιοδείκτης HER2 επηρεάζει και πόσο τη συνάρτηση (συνολικού) κινδύνου ή αντίστοιχα την συνάρτηση συνολικής επιβίωσης. Εφαρμόζουμε την παρακάτω εντολή

```
cox1=coxph(Surv(survival,death)~her2, method="breslow")
summary(cox1)
```

```
n=289 (27 observations deleted due to missingness)
        coef  exp(coef) se(coef)    z      Pr(>|z|)    exp(-coef)  lower .95  upper .95
her2   0.6742  1.9624    0.2318  2.908  0.00363 **  0.5096      1.246       3.091
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Rsquare= 0.028   (max possible= 0.941 )
Likelihood ratio test  =  8.07  on 1 df,   p=0.004489
Wald test              =  8.46  on 1 df,   p=0.003633
Score (logrank) test   =  8.78  on 1 df,   p=0.003043
```

Σύμφωνα με τους ολικούς ελέγχους (*Likelihood ratio test, Wald test, Score test*) ο βιοδείκτης HER2 φαίνεται να επηρεάζει την συνάρτηση συνολικού κινδύνου, διότι και τα τρία $p-value$ είναι πολύ μικρότερα του επιπέδου σημαντικότητας $a=5\%$ με αποτέλεσμα να απορρίπτουμε την μηδενική υπόθεση $H_0: b=0$ για το μοντέλο

$$h(t/Z)=h_0(t)\exp(b\cdot HER2),\ t\geq 0 \text{ όπου } Z=\{HER2\}$$

Λόγω του ότι στο μοντέλο υπάρχει μόνο μία συμμεταβλητή ο ολικός *Wald* έλεγχος συμπίπτει με τον τοπικό *Wald* έλεγχο τόσο ως προς την στατιστική συνάρτηση όσο και ως προς το $p-value$. Η στατιστική συνάρτηση του ολικού *Wald* ελέγχου ακολουθεί χ^2 κατανομή με 1 βαθμό ελευθερίας, ενώ η στατιστική συνάρτηση του τοπικού *Wald* ελέγχου ακολουθεί τυπική κανονική κατανομή και γνωρίζουμε ότι το τετράγωνο της τυπικής κανονικής κατανομής προσεγγίζει την χ^2 κατανομή με 1 βαθμό ελευθερίας. Επιπλέον, ο *Score* έλεγχος θα συμπίπτει με τον *Logrank* έλεγχο λόγω της μίας μεταβλητής που υπάρχει στο μοντέλο παλινδρόμησης του Cox.

Σύμφωνα με το πρόσημο του συντελεστή παλινδρόμησης $\hat{b}$ συμπεραίνουμε ότι ο βιοδείκτης HER2 επηρεάζει θετικά, ενώ λαμβάνοντας υπόψη το γεγονός ότι η τιμή του νεπέριου αυτού του συντελεστή παλινδρόμησης είναι μεγαλύτερη της μονάδας καταλαβαίνουμε ότι όταν «ανεβαίνουμε» επίπεδο στον βιοδείκτη HER2, τότε αυξάνεται ο κίνδυνος θανάτου του ασθενή. Πιο συγκεκριμένα ισχύει ότι

$$\widehat{HR}(t)=\frac{\hat{h}(t/HER2=1)}{\hat{h}(t/HER2=0)}=e^{\hat{b}}=1.962,\ t\geq 0$$

δηλαδή ασθενείς με υπερέκφραση του βιοδείκτη HER2 έχουν περίπου 2 φορές μεγαλύτερο κίνδυνο θανάτου από τις ασθενείς με μη-υπερέκφραση του βιοδείκτη αυτού.

Παρατήρηση

Λαμβάνοντας υπόψη το αποτέλεσμα του αντικειμένου **kaplan1** δεν υπάρχουν δεσμοί, διότι συμβαίνει μόνο μία αποτυχία σε κάθε χρονική στιγμή, οπότε μπορούμε να εφαρμόσουμε είτε Breslow είτε Efron και να πάρουμε το ίδιο αποτέλεσμα.

Έστω ότι θέλουμε να μελετήσουμε την επίδραση του βιοδείκτη HER2 στη συνάρτηση κινδύνου λαμβάνοντας υπόψη και την επίδραση της χημειοθεραπείας. Δηλαδή, μελετάμε το μοντέλο παλινδρόμησης του Cox

$$h(t/Z)=h_0(t)\exp(b_1\cdot HER2+b_2\cdot group),\ t\geq 0 \text{ όπου } Z=\{HER2, group\}$$

Εφαρμόζουμε τις παρακάτω εντολές για το μοντέλο αυτό

```
cox2=coxph(Surv(survival,death)~her2+group, method="breslow")
summary(cox2)
```

```
n=289 (27 observations deleted due to missingness)

        coef  exp(coef) se(coef)    z Pr(>|z|)    exp(-coef) lower .95 upper .95
her2   0.6799   1.9736   0.2321 2.930  0.00339 **  0.5067     1.2523     3.110
group  0.1190   1.1263   0.2311 0.515  0.60663      0.8878     0.7161     1.772
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Rsquare= 0.028   (max possible= 0.941 )
Likelihood ratio test  =  8.34  on 2 df,   p=0.01545
Wald test              =  8.72  on 2 df,   p=0.01275
Score (logrank) test =  9.05  on 2 df,   p=0.01085
```

Από τα $p-value$ των ολικών ελέγχων συμπεραίνουμε ότι απορρίπτεται η μηδενική υπόθεση $H_0 : b = (b_1, b_2)' = 0$, οπότε τουλάχιστον μία από τις παραμέτρους είναι διαφορετική του μηδενός και αυτό συνεπάγεται ότι τουλάχιστον μία από τις συμμεταβλητές του μοντέλου επηρεάζει τη συνάρτηση κινδύνου. Σύμφωνα με τον πίνακα συντελεστών παλινδρόμησης μόνο ο βιοδείκτης φαίνεται να επηρεάζει σημαντικά την συνάρτηση κινδύνου $(p-value = 0.00339)$ και συγκεκριμένα ασθενείς με υπερέκφραση του βιοδείκτη HER2 έχουν περίπου 2 φορές μεγαλύτερο κίνδυνο θανάτου από τις ασθενείς με μη-υπερέκφραση του βιοδείκτη αυτού είτε χορηγήσουμε E-T-CMF είτε E-CMF, δηλαδή «κρατώντας» σταθερό το είδος χημειοθεραπείας (*adjusted for group*). Ο αντίστοιχοι σχετικοί λόγοι κινδύνων υπολογίζονται ως εξής

$$\widehat{HR}(t) = \frac{\hat{h}(t / HER2 = 1, group = 0)}{\hat{h}(t / HER2 = 0, group = 0)} = e^{\hat{b}_1} = 1.9736 = \frac{\hat{h}(t / HER2 = 1, group = 1)}{\hat{h}(t / HER2 = 0, group = 1)} \; t \geq 0$$

Εάν λάβουμε υπόψη και το πως επηρεάζει ο βιοδείκτης τη συνάρτηση κινδύνου για κάθε είδος χημειοθεραπείας, τότε αναλύουμε το παρακάτω μοντέλο

παλινδρόμησης του Cox με όρο αλληλεπίδρασης μεταξύ του βιοδείκτη HER2 και του είδους χημειοθεραπείας group

$$h(t/Z) = h_0(t)\exp\left(b_1 \cdot HER2 + b_2 \cdot group + b_3 \cdot \left[HER2 \times group\right]\right) \quad t \geq 0$$

όπου $Z = \{HER2, group, HER2 \times group\}$ και εξετάζουμε την μηδενική υπόθεση $H_0 : b = (b_1, b_2, b_3)' = 0$, δηλαδή οι συμμεταβλητές του μοντέλου δεν επηρεάζουν τη συνάρτηση κινδύνου.

```
cox3=coxph(Surv(survival, death)~her2+group+her2*group, method="breslow")
summary(cox3)
```

```
n=289 (27 observations deleted due to missingness)

                coef exp(coef) se(coef)    z Pr(>|z|) exp(-coef) lower .95 upper .95
her2          0.1248    1.1329   0.7635  0.163  0.870    0.8827     0.2537     5.059
group        -0.3937    0.6745   0.7059 -0.558  0.577    1.4825     0.1691     2.691
her2:group    0.3575    1.4298   0.4667  0.766  0.444    0.6994     0.5728     3.569

Rsquare= 0.03   (max possible= 0.941 )
Likelihood ratio test  =  8.93  on 3 df,   p=0.03025
Wald test              =  9.62  on 3 df,   p=0.02207
Score (logrank) test   = 10.08  on 3 df,   p=0.01786
```

Οι ολικοί έλεγχοι απορρίπτουν την μηδενική υπόθεση, ενώ οι τοπικοί έλεγχοι αποδέχονται την μηδενική υπόθεση $H_0 : b_i = 0$, $i = 1,2,3$, δηλαδή ούτε ο βιοδείκτης, ούτε το είδος χημειοθεραπείας, αλλά ούτε και η αλληλεπίδρασή τους επηρεάζει την συνάρτηση κινδύνου. Ενώ στους ολικούς ελέγχους οι όροι αυτοί φαίνονται σημαντικοί, στους τοπικούς ελέγχους φαίνονται ασήμαντοι. Αυτό είναι ένδειξη ισχυρής συσχέτισης μεταξύ των όρων αυτών.

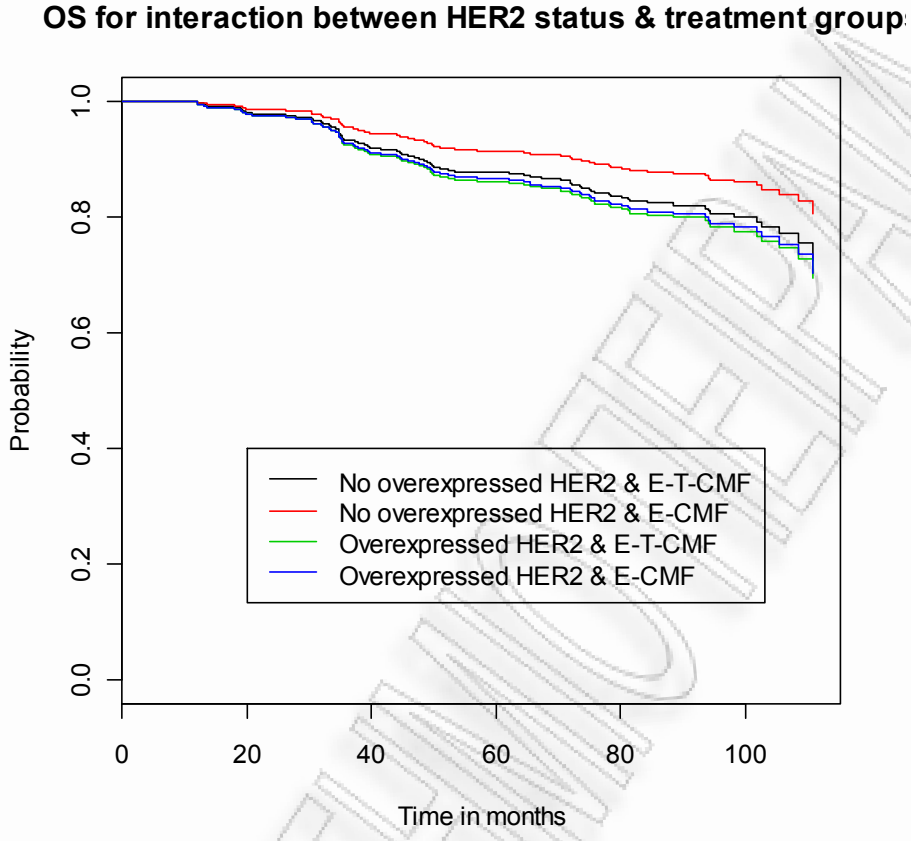
Η αλληλεπίδραση της χημειοθεραπείας με τον βιοδείκτη HER2 μπορεί να παρουσιαστεί γραφικά (Γράφημα Γ1.3) σύμφωνα με τις παρακάτω εντολές

```
cox3=coxph(Surv(survival, death)~her2+group+her2*group, method="breslow")

surv1=survfit(cox3, newdata=data.frame(her2=0, group=0), type="kaplan-meier",
conf. type="none")
surv2=survfit(cox3, newdata=data.frame(her2=0, group=1), type="kaplan-meier",
conf. type="none")
surv3=survfit(cox3, newdata=data.frame(her2=1, group=0), type="kaplan-meier",
conf. type="none")
surv4=survfit(cox3, newdata=data.frame(her2=1, group=1), type="kaplan-meier",
conf. type="none")

plot(surv1,xlab="Time in months",ylab="Probability", main="OS for interaction
between HER2 status & treatment groups",lty=1,col=1)
lines(surv2,col=2,lty=1)
lines(surv3,col=3,lty=1)
lines(surv4,col=4,lty=1)
legend(20, .4, c("No overexpressed HER2 & E-T-CMF", "No overexpressed HER2
& E-CMF", "Overexpressed HER2 & E-T-CMF", "Overexpressed HER2 &
E-CMF"), col=1:4, lty=1:1)
```

Ουσιαστικά εισάγουμε σε κοινό γράφημα 4 καμπύλες, δηλαδή όσοι είναι οι συνδυασμοί των επιπέδων της συμμεταβλητής her2 με τα επίπεδα της συμμεταβλητής group.

Το αποτέλεσμα των παραπάνω εντολών φαίνεται στο γράφημα Γ1.3

Γράφημα Γ1.3



Παρατηρούμε ότι για τις ασθενείς με υπερέκφραση του βιοδείκτη HER2 η πιθανότητα συνολικής επιβίωσης των ασθενών που πήραν E-T-CMF δεν διαφέρει από την πιθανότητα συνολικής επιβίωσης των ασθενών που πήραν E-CMF, σε αντίθεση με τις ασθενείς με μη-υπερέκφραση αυτού του βιοδείκτη, όπου οι ασθενείς που πήραν E-T-CMF έχουν λίγο μικρότερη πιθανότητα συνολικής επιβίωσης από τις ασθενείς που πήραν E-CMF. Οι σχετικοί λόγοι κινδύνων είναι αντίστοιχα

$$\widehat{HR}(t)=\frac{\hat{h}\left(t/HER2=0,group=1,HER2\times group=0\right)}{\hat{h}\left(t/HER2=0,group=0,HER2\times group=0\right)}=e^{\hat{b}_2}=0.6745 \quad t\geq 0 \text{ και}$$

$$\widehat{HR}(t)=\frac{\hat{h}\left(t/HER2=1,group=1,HER2\times group=1\right)}{\hat{h}\left(t/HER2=1,group=0,HER2\times group=0\right)}=e^{\hat{b}_2+\hat{b}_3}=0.96 \quad t\geq 0$$

Επιπλέον, στις ασθενείς που χορηγήθηκε E-T-CMF η πιθανότητα συνολικής επιβίωσης των ασθενών που έχουν υπερέκφραση του βιοδείκτη HER2 δεν διαφέρει από την πιθανότητα συνολικής επιβίωσης των ασθενών που έχουν μη-υπερέκφραση αυτού του βιοδείκτη, σε αντίθεση με τις ασθενείς στους οποίους χορηγήθηκε E-CMF, όπου οι ασθενείς με υπερέκφραση αυτού του βιοδείκτη έχουν λίγο μικρότερη

πιθανότητα συνολικής επιβίωσης από τις ασθενείς με μη-υπερέκφραση του βιοδείκτη. Οι σχετικοί λόγοι κινδύνων είναι αντίστοιχα

$$\widehat{HR}(t)=\frac{\hat{h}\left(t/HER2=1,group=0,HER2\times group=0\right)}{\hat{h}\left(t/HER2=0,group=0,HER2\times group=0\right)}=e^{\hat{b}_1}=1.1329\quad t\geq 0 \text{ και}$$

$$\widehat{HR}(t)=\frac{\hat{h}\left(t/HER2=1,group=1,HER2\times group=1\right)}{\hat{h}\left(t/HER2=0,group=1,HER2\times group=0\right)}=e^{\hat{b}_1+\hat{b}_3}=1.619\quad t\geq 0$$

Λαμβάνοντας υπόψη το γράφημα καταλαβαίνουμε ότι μόνο η χορήγηση E-CMF σε ασθενείς με μη-υπερέκφραση του βιοδείκτη HER2 φαίνεται να επηρεάζει την πιθανότητα επιβίωσής τους σε σχέση με τις ασθενείς που τους έχει χορηγηθεί E-T-CMF και έχουν μη-υπερέκφραση του βιοδείκτη HER2 ή τις ασθενείς που τους έχει χορηγηθεί E-CMF και έχουν υπερέκφραση του βιοδείκτη αυτού, αλλά η επίδραση αυτή δεν είναι (στατιστικά) σημαντική.

Μπορούμε να εξετάσουμε και άλλες συμμεταβλητές αν επηρεάζουν και πόσο την συνάρτηση κινδύνου. Μελετάμε όλους τους βιοδείκτες και τους κλινικοπαθολογικούς παράγοντες της έρευνας που θεωρούνται σημαντικοί σύμφωνα με την διαδικασία επιλογής μοντέλου, οπότε σχηματίζεται το παρακάτω μοντέλο παλινδρόμησης του Cox

$$h\left(t/Z\right)=h_0\left(t\right)\exp(b_1\cdot VEGF\_ich+b_2\cdot nodes13+b_3\cdot nodes49+b_4\cdot genes1+b_5\cdot ER+$$

$$+b_6\cdot PgR+b_7\cdot HER2+b_8\cdot genes2+b_9\cdot group+b_{10}\cdot\left[group\times PgR\right])$$

Να σημειώσουμε ότι η συμμεταβλητή *nodes* έχει τρία επίπεδα, οπότε πρέπει να θεωρήσουμε το πρώτο (συνήθως) επίπεδο ως επίπεδο αναφοράς (*reference category*) και να εισάγουμε τα υπόλοιπα δύο επίπεδα στο μοντέλο παλινδρόμησης του Cox. Γι' αυτό το λόγο κωδικοποιούμε τη συμμεταβλητή *nodes* ως εξής

Πίνακας Γ1.1

Επίπεδα της nodes	**nodes13**	**nodes49**
1 $\left(nodes<1\right)$	0	0
2 $\left(1\leq nodes\leq 3\right)$	1	0
3 $\left(4\leq nodes\leq 9\right)$	0	1

Παρατήρηση

Είναι προφανές ότι αν δεν είχαμε κωδικοποιήσει καταλλήλως τη συμμεταβλητή nodes τότε θα εξετάζαμε το παρακάτω μοντέλο

$$h(t/Z) = h_0(t)\exp(b_1 \cdot VEGF\_ich + b_2 \cdot nodes + b_3 \cdot genes1 + b_4 \cdot ER + b_5 \cdot PgR +$$
$$+b_6 \cdot HER2 + b_7 \cdot genes2 + b_8 \cdot group + b_9 \cdot [group \times PgR])$$

Η ερμηνεία του συντελεστή παλινδρόμησης για την nodes θα ήταν λανθασμένη, αφού με τον τρόπο που ορίσαμε τη συμμεταβλητή αυτή στο μοντέλο αφήνουμε να εννοηθεί ότι έχει μόλις δύο επίπεδα.

Για την μελέτη του πολυμεταβλητού μοντέλου παλινδρόμησης του Cox εφαρμόζουμε τις παρακάτω εντολές

```
cox4=coxph(Surv(survival,death)~vegf+nodes13+nodes49+genes1+erIhc+pgrIhc+
+her2+genes2+group+pgrIhc*group, method="breslow")
summary(cox4)
```

```
 n=235 (81 observations deleted due to missingness)
                 coef  exp(coef) se(coef)     z Pr(>|z|)
vegf           0.5048   1.6567   0.3551 1.422  0.15514
nodes13        1.2710   3.5643   0.4600 2.763  0.00572 **
nodes49        1.4628   4.3178   0.4521 3.235  0.00121 **
genes1        -0.5427   0.5812   0.1795 -3.024  0.00250 **
erIhc          0.1184   1.1257   0.3123 0.379  0.70467
pgrIhc         1.5292   4.6145   0.8476 1.804  0.07120 .
her2           0.8214   2.2737   0.2920 2.813  0.00490 **
genes2         0.6401   1.8967   0.2147 2.982  0.00287 **
group          0.6747   1.9633   0.3849 1.753  0.07965 .
pgrIhc:group  -1.0752   0.3412   0.5182 -2.075  0.03801 *

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

	exp(-coef)	lower .95	upper .95
vegf	0.6036	0.8260	3.3230
nodes13	0.2806	1.4470	8.7800
nodes49	0.2316	1.7800	10.4739
genes1	1.7207	0.4088	0.8262
erIhc	0.8884	0.6103	2.0761
pgrIhc	0.2167	0.8763	24.2986
her2	0.4398	1.2830	4.0296
genes2	0.5272	1.2453	2.8890
group	0.5093	0.9233	4.1748
pgrIhc:group	2.9307	0.1236	0.9422

Rsquare= 0.166 (max possible= 0.939)
Likelihood ratio test= 42.58 on 10 df, p=5.914e-06
Wald test = 37.63 on 10 df, p=4.4e-05
Score (logrank) test = 39.19 on 10 df, p=2.348e-05

Από τους ολικούς ελέγχους του μοντέλου καταλαβαίνουμε ότι τουλάχιστον ένας όρος του μοντέλου επηρεάζει την συνάρτηση κινδύνου. Δηλαδή, απορρίπτουμε την μηδενική υπόθεση $H_0 : b = (b_1, b_2, ..., b_{10})' = 0$ σε επίπεδο σημαντικότητας $a = 5\%$. Εξετάζουμε τους τοπικούς ελέγχους και παρατηρούμε ότι το πλήθος θετικών αδενωμάτων, ο βιοδείκτης HER2, το είδος χημειοθεραπείας, οι δύο πρώτες ομάδες γονιδίων και η αλληλεπίδραση μεταξύ χημειοθεραπείας και βιοδείκτη PgR-mRNA επηρεάζουν σημαντικά την συνάρτηση κινδύνου. Ο βιοδείκτης ER-mRNA δεν έχει καμία προγνωστική δύναμη $(p - value = 0.70467)$ στην συνάρτηση κινδύνου, ενώ ο παράγοντας VEGF δεν φαίνεται να επηρεάζει σημαντικά $(p - value = 0.15514)$ την συνάρτηση κινδύνου. Τέλος, ο βιοδείκτης PgR και η δεύτερη ομάδα γονιδίων οριακά δεν επηρεάζουν σημαντικά την συνάρτηση κινδύνου $(p - value = 0.07120$ και $p - value = 0.07965$ αντίστοιχα$)$

Χρησιμοποιούμε το λόγο σχετικών κινδύνων για να ερμηνεύσουμε τις επιδράσεις κάποιων σημαντικών συμμεταβλητών του μοντέλου στην συνάρτηση (συνολικού) κινδύνου

• Για το πλήθος θετικών αδενωμάτων παρατηρούμε ότι:

η ύπαρξη ενός μέχρι τριών θετικών αδενωμάτων αυξάνει περίπου 3.56 φορές το κίνδυνο θανάτου σε σχέση με την απουσία αδενωμάτων, εφόσον «κρατάμε» σταθερές τις υπόλοιπες συμμεταβλητές του μοντέλου, διότι

$$\widehat{HR}(t)=\frac{h\left(t/VEGF,nodes13=1,nodes49,...,group,group\times PgR\right)}{h\left(t/VEGF,nodes13=0,nodes49,...,group,group\times PgR\right)}=e^{\hat{b}_2}=3.5643$$

ενώ η ύπαρξη τεσσάρων μέχρι εννιά θετικών αδενωμάτων αυξάνει περίπου 4.32 φορές το κίνδυνο θανάτου σε σχέση με την απουσία αδενωμάτων, εφόσον «κρατάμε» σταθερές τις υπόλοιπες συμμεταβλητές του μοντέλου, διότι

$$\widehat{HR}(t)=\frac{h\left(t/VEGF,nodes13,nodes49=1,...,group,group\times PgR\right)}{h\left(t/VEGF,nodes13,nodes49=0,...,group,group\times PgR\right)}=e^{\hat{b}_3}=4.3178$$

και τέλος η ύπαρξη τεσσάρων μέχρι εννιά θετικών αδενωμάτων αυξάνει περίπου 1.21 φορές το κίνδυνο θανάτου σε σχέση με την ύπαρξη ενός μέχρι τριών θετικών αδενωμάτων, εφόσον «κρατάμε» σταθερές τις υπόλοιπες συμμεταβλητές του μοντέλου, διότι

$$\widehat{HR}(t)=\frac{h\left(t/VEGF,nodes13,nodes49=1,...,group,group\times PgR\right)}{h\left(t/VEGF,nodes13=1,nodes49,...,group,group\times PgR\right)}=e^{\hat{b}_3-\hat{b}_2}=1.2114$$

• Για την πρώτη ομάδα γονιδίων παρατηρούμε ότι:

όταν σε κάποια ασθενή η έκφραση των γονιδίων αυτής της ομάδας αυξάνει κατά ένα βαθμό, τότε ο κίνδυνος θανάτου μειώνεται σε σχέση με κάποια άλλη ασθενή όπου η έκφραση αυτών των γονιδίων έχει παραμείνει σταθερή, εφόσον «κρατάμε» σταθερές τις υπόλοιπες συμμεταβλητές του μοντέλου, διότι

$$\widehat{HR}(t)=\frac{h\left(t/VEGF,nodes13,nodes49,genes1=i+1,...,group\times PgR\right)}{h\left(t/VEGF,nodes13,nodes49,genes1=i,...,group\times PgR\right)}=e^{\hat{b}_4}=0.5812$$

Παρόμοια εργαζόμαστε για τη δεύτερη ομάδα γονιδίων.

• Για την αλληλεπίδραση μεταξύ χημειοθεραπείας και βιοδείκτη PgR παρατηρούμε ότι:

στις ασθενείς με θετικό βιοδείκτη PgR-mRNA η χορήγηση χημειοθεραπείας με πακλιταξέλη αυξάνει τον κίνδυνο θανάτου περίπου 1.57 φορές σε σχέση με την

χορήγηση χημειοθεραπείας χωρίς πακλιταξέλη, εφόσον «κρατάμε» σταθερές τις υπόλοιπες συμμεταβλητές του μοντέλου, διότι

$$\widehat{HR}(t)=\frac{h\left(t/VEGF,...,PgR=1,...,group=1,group\times PgR=1\right)}{h\left(t/VEGF,...,PgR=1,...,group=0,group\times PgR=0\right)}=e^{\hat{b}_8+\hat{b}_{10}}=1.5746$$

Στην επιλογή του βέλτιστου πολυμεταβλητού μοντέλου παλινδρόμησης του Cox θεωρούμε a priori ότι η συμμεταβλητή *group* (είδος χημειοθεραπείας) θα περιληφθεί οπωσδήποτε στο τελικό μοντέλο. Εφαρμόζουμε την μέθοδο του Collett και χρησιμοποιούμε Likelihood ratio έλεγχο για την επιλογή των συμμεταβλητών. Παρουσιάζουμε τις εντολές και τους πίνακες σύγκρισης των συμμεταβλητών σε κάθε βήμα της μεθόδου για την συνολική επιβίωση. Ανάλογα εργαζόμαστε και για την επιβίωση χωρίς ασθένεια.

ΒΗΜΑ 1: Συγκρίνουμε μάθε μονομεταβλητό μοντέλο με το μηδενικό μοντέλο σε επίπεδο σημαντικότητας $a=0.20$.

Πίνακας Γ1.2

Μοντέλο	**Μεταβλητές**	$-2\log\hat{L}$	**Y**	**df**	**p-value**	**AIC**	**Απόφαση**
0	καμία	907.593	-	-	-	910.593	-
1	VEGF_ich	710.784	196.810	1	0	713.784	✓
2	grade	901.561	6.032	1	0.014	904.561	✓
3	size	902.251	5.343	1	0.0208	905.251	✓
4	nodes	896.162	11.431	1	0.0007	899.162	✓
5	genes1	902.671	4.922	1	0.0265	905.671	✓
6	genes2	906.967	0.626	1	0.4287	909.967	×
7	genes3	906.668	0.925	1	0.3362	909.668	×
8	erIhc	882.660	24.933	1	5.9e-07	885.660	✓
9	pgrIhc	877.377	30.216	1	3.8e-08	880.377	✓
10	her2	807.499	100.095	1	0	810.499	✓

Το μοντέλο που προκύπτει στο βήμα αυτό είναι το

$$h\left(t/Z\right)=h_0\left(t\right)\exp(b_1\cdot VEGF\_ich+b_2\cdot gradeCat+b_3\cdot sizeCat+b_4\cdot nodes3+b_5\cdot genes1+$$

$$+b_6\cdot erIhc+b_7\cdot pgrIhc+b_8\cdot her2)\quad(1)$$

Για το υπολογισμό των ποσοτήτων $-2\log\hat{L}$ χρησιμοποιούμε την «ιδιότητα» $-2*object's\_name\$\log lik$ που προκύπτει από την εντολή **coxph()**. Για παράδειγμα για να υπολογίσουμε το $-2\log\hat{L}$ του «μηδενικού» μοντέλου εφαρμόζουμε τις παρακάτω εντολές

```
cox11=coxph(Surv(survival,death)~1,method="breslow")
-2*cox11$loglik
```

Συνολικά εφαρμόζουμε 10 διαφορετικά μονομεταβλητά μοντέλα, υπολογίζουμε την ποσότητα $-2\log\hat{L}$ και συγκρίνουμε με την ποσότητα $-2\log\hat{L}$ του «μηδενικού» μοντέλου (δηλαδή υπολογίζουμε την ποσότητα Y). Για κάθε Likelihood ratio έλεγχο εξετάζουμε σε επίπεδο σημαντικότητας $a=20\%$ την μηδενική υπόθεση $H_0: b_i=0$, $i=1,2,...,10$, δηλαδή η αντίστοιχη συμμεταβλητή δεν βελτιώνει την επάρκεια του «μηδενικού» μοντέλου, οπότε την απορρίπτουμε.

Παρατήρηση

Η ποσότητα Y προκύπτει από την σχέση

$$Y=-2\log\frac{\hat{L}_0}{\hat{L}_i}=\left(-2\log\hat{L}_0\right)-\left(-2\log\hat{L}_i\right)=M_0-M_i$$

όπου $i=1,2,...,10$ και με M_i συμβολίζουμε το i μοντέλο συμμεταβλητών που εξετάζουμε. Είναι προφανές ότι μία μεγάλη θετική τιμή της ποσότητας Y δηλώνει «συμφωνία» μεταξύ του μοντέλου i και των δεδομένων και αυτό είναι ένδειξη ότι η προσθήκη της αντίστοιχης συμμεταβλητής βελτιώνει την επάρκεια του «μηδενικού» μοντέλου (*null model*). Να σημειώσουμε ότι η ποσότητα $\hat{L}$ είναι γινόμενο δεσμευμένων πιθανοτήτων, οπότε παίρνει τιμές μόνο μεταξύ του 0 και του 1, γι' αυτό η ποσότητα $-2\log\hat{L}$ είναι πάντα θετική και η τιμή της μειώνεται όσο προσθέτουμε συμμεταβλητές στο μοντέλο.

ΒΗΜΑ 2: Στην συνέχεια συγκρίνουμε το μοντέλο (1) με τα μοντέλα που προκύπτουν, όταν κάθε φορά αφαιρούμε μία συμμεταβλητή από το μοντέλο (1). Αυτή η σύγκριση γίνεται σε επίπεδο σημαντικότητας $a=0.10$ και σκοπός μας είναι

να ελέγξουμε αν οι συμμεταβλητές του μοντέλου (1) πρέπει να μείνουν τελικά (δηλαδή, αν το μοντέλο αυτό είναι επαρκές ή υπερεπαρκές με αυτές τις συμμεταβλητές, οπότε στην δεύτερη περίπτωση πρέπει να απορρίψουμε κάποια ή κάποιες).

Πίνακας Γ1.3

Μοντέλο	**Μεταβλητές**	$-2\log\hat{L}$	**Y**	**df**	**p-value**	**AIC**	**Από-φαση**
0	(1)	629.219	-	-	-	653.219	-
1	(1) -VEGF	747.967	118.748	1	0	768.967	✓
2	(1) -grade	629.534	0.315	1	0.5747	653.534	×
3	(1) -size	630.358	1.139	1	0.2858	651.358	×
4	(1) -nodes	639.316	10.097	1	0.0015	660.316	✓
5	(1) -genes1	632.170	2.951	1	0.0858	653.169	✓
6	(1) -erIhc	639.798	10.580	1	0.0011	660.798	✓
7	(1) -pgrIhc	642.966	13.747	1	0.0002	663.966	✓
8	(1) -her2	655.446	26.227	1	3.03e-07	676.446	✓

Σε αυτό το βήμα προκύπτει το παρακάτω μοντέλο

$$h(t/Z)=h_0(t)\exp(b_1\cdot VEGF\_ich+b_2\cdot nodes3+b_3\cdot genes1+b_4\cdot erIhc+b_5\cdot pgrIhc+$$
$$+b_6\cdot her2)\quad(2)$$

Για να υπολογίσουμε το $-2\log\hat{L}$ του μοντέλου (1)-VEGF ουσιαστικά βγάζουμε την συμμεταβλητή VEGF από το μοντέλο (1) και υπολογίζουμε τον λογάριθμο της μερικής πιθανοφάνειας του μοντέλου που προκύπτει

```
cox21=coxph(Surv(survival,death)~grade+size+nodes+genes1+erIhc+pgrIhc+her2,
method="breslow")
-2*cox21$loglik
```

Ανάλογα υπολογίζουμε την ποσότητα $-2\log\hat{L}$ για τα υπόλοιπα μοντέλα που περιέχονται στον πίνακα.

Παρατήρηση

Σε αντίθεση με το Βήμα 1, εδώ η ποσότητα Y ορίζεται διαφορετικά. Συγκεκριμένα προκύπτει από την σχέση

$$Y = -2\log\frac{\hat{L}_i}{\hat{L}_0} = \left(-2\log\hat{L}_i\right) - \left(-2\log\hat{L}_0\right) = M_i - M_0$$

όπου $i = 1,2,...,8$. Μάλιστα μέσω της Y εξετάζουμε σε επίπεδο σημαντικότητας $a = 10\%$ την μηδενική υπόθεση $H_0 : b_i = 0$, $i = 1,2,...,8$, δηλαδή η αντίστοιχη συμμεταβλητή είναι πλεονάζουσα για το μοντέλο (1) δοθέντων των άλλων συμμεταβλητών, οπότε την αφαιρούμε από το μοντέλο αυτό.

ΒΗΜΑ 3: Επαναλαμβάνουμε το Βήμα 2 για το μοντέλο (2), δηλαδή ελέγχουμε πάλι σε επίπεδο σημαντικότητας $a = 0.10$ αν το μοντέλο αυτό είναι επαρκές ή πλεονάζον σε κάποια ή κάποιες συμμεταβλητές.

Πίνακας Γ1.4

Μοντέλο	**Μεταβλητές**	$-2\log\hat{L}$	**Y**	**df**	**p-value**	**AIC**	**Από-φαση**
0	(2)	630.7681	-	-	-	648.7681	-
1	(2)-VEGF_	749.1538	118.3857	1	0	764.1538	✓
2	(2)-nodes	641.4679	10.6998	1	0.0011	655.4679	✓
3	(2)-genes1	634.7851	4.017	1	0.0450	649.7851	✓
4	(2)-erIhc	641.0803	10.3122	1	0.0013	656.0803	✓
5	(2)-pgrIhc	644.8040	14.0359	1	0.0002	659.8040	✓
6	(2)-her2	657.8797	27.1116	1	1.92e-07	672.8797	✓

Το μοντέλο (2) θεωρείται επαρκές με τις συμμεταβλητές που περιέχει, οπότε δεν υπάρχει λόγος να αφαιρέσουμε κάποια.

ΒΗΜΑ 4: Εξετάζουμε αν στο μοντέλο (2) μπορούν να εισαχθούν κάποιες από τις συμμεταβλητές που απορρίψαμε στο Βήμα 1, δηλαδή την genes2 και genes3 σε επίπεδο σημαντικότητας $a = 0.10$.

Πίνακας Γ1.5

Μοντέλο	**Μεταβλητές**	$-2\log\hat{L}$	**Y**	**df**	**p-value**	**AIC**	**Απόφαση**
0	(2)	630.768	-	-	-	648.768	-
1	(2)+genes2	621.709	9.059	1	0.0026	642.709	✓
2	(2)+genes3	630.768	0	1	1	651.768	×

Χρειάζεται να εισαχθεί η συμμεταβλητή genes2, οπότε προκύπτει το μοντέλο

$$h(t/Z) = h_0(t)\exp(b_1 \cdot VEGF\_ich + b_2 \cdot nodes3 + b_3 \cdot genes1 + b_4 \cdot erIhc + b_5 \cdot pgrIhc + \\ + b_6 \cdot her2 + b_7 \cdot genes2) \quad (3)$$

Παρατήρηση

Για τον υπολογισμό της ποσότητας Y βασιζόμαστε στην παρακάτω σχέση

$$Y = -2\log\frac{\hat{L}_0}{\hat{L}_i} = \left(-2\log\hat{L}_0\right) - \left(-2\log\hat{L}_i\right) = M_0 - M_i$$

όπου $i = 1,2$. Μέσω της Y εξετάζουμε σε επίπεδο σημαντικότητας $a = 10\%$ την μηδενική υπόθεση $H_0 : b_i = 0$, $i = 1,2$, δηλαδή η αντίστοιχη συμμεταβλητή είναι πλεονάζουσα για το μοντέλο (2) δοθέντων των άλλων συμμεταβλητών, οπότε δεν την εισάγουμε στο μοντέλο αυτό.

ΒΗΜΑ 5: Στο μοντέλο (3) εισάγουμε την προαποφασισμένη συμμεταβλητή, δηλαδή την θεραπεία και εξετάζουμε σε επίπεδο σημαντικότητας $a = 0.10$ αν μπορούν να εισαχθούν όροι αλληλεπίδρασης μεταξύ της θεραπείας και των συμμεταβλητών του μοντέλου (3). Με την εισαγωγή της χημειοθεραπείας προκύπτει το παρακάτω μοντέλο

$$h(t/Z) = h_0(t)\exp(b_1 \cdot VEGF\_ich + b_2 \cdot nodes3 + b_3 \cdot genes1 + b_4 \cdot erIhc + b_5 \cdot pgrIhc + \\ + b_6 \cdot her2 + b_7 \cdot genes2 + b_8 \cdot group) \quad (4)$$

Πίνακας Γ1.6

Μοντέλο	**Μεταβλητές**	$-2\log\hat{L}$	**Y**	**df**	**p-value**	**AIC**	**Από-φαση**
0	(3)	621.709	-	-	-	642.709	-
1	(3) +group	621.597	0.112	1	0.738	645.597	×
2	(4) +group*VEGF	621.590	0.007	1	0.935	648.590	×
3	(4) +group*nodes	621.304	0.292	1	0.588	648.304	×
4	(4) +group*genes1	621.363	0.234	1	0.629	648.363	×
5	(4) +group*genes2	619.250	2.346	1	0.126	646.250	×
6	(4) +group*erIhc	620.476	1.121	1	0.289	647.476	×
7	(4) +group*pgrIhc	617.909	3.687	1	0.055	644.909	✓
8	(4) +group*her2	621.552	0.044	1	0.834	648.552	×

Μόνο η αλληλεπίδραση μεταξύ του βιοδείκτη PgR-mRNA και της χημειοθεραπείας φαίνεται να βελτιώνει την επάρκεια του μοντέλου (4). Έτσι τελικά το βέλτιστο μοντέλο που θα μελετήσουμε είναι το εξής

$$h(t/Z) = h_0(t)\exp(b_1 \cdot VEGF\_ich + b_2 \cdot nodes3 + b_3 \cdot genes1 + b_4 \cdot erIhc + b_5 \cdot pgrIhc + \\ + b_6 \cdot her2 + b_7 \cdot genes2 + b_8 \cdot group + b_9 \cdot [group \times pgrIhc]) \quad (5)$$

Παρατήρηση

Για τον υπολογισμό της ποσότητας Y που αφορά την επάρκεια του μοντέλου (3) όταν εισάγουμε σε αυτό την χημειοθεραπεία χρησιμοποιήσαμε την σχέση

$$Y = -2\log\frac{\hat{L}_0}{\hat{L}_1} = \left(-2\log\hat{L}_0\right) - \left(-2\log\hat{L}_1\right) = M_0 - M_1 = 621.709\text{-}621.597 = 0.112$$

Επειδή $p-value = 0.738$ αποδεχόμαστε σε επίπεδο σημαντικότητας την μηδενική υπόθεση $H_0 : b_8 = 0$, δηλαδή η χημειοθεραπεία δεν βελτιώνει την επάρκεια του μοντέλου. Ωστόσο, έχουμε αποφασίσει a priori ότι πρέπει να εισαχθεί στο μοντέλο που θα προκύψει πριν την εξέταση για εισαγωγή αλληλεπιδράσεων, οπότε δεν μπορούμε να την απορρίψουμε. Μέσω της Y εξετάζουμε σε επίπεδο σημαντικότητας $a = 10\%$ την μηδενική υπόθεση $H_0 : b_i = 0$, $i = 2,...,8$, δηλαδή η αντίστοιχη συμμεταβλητή είναι πλεονάζουσα για το μοντέλο (4) δοθέντων των άλλων μεταβλητών, οπότε δεν την εισάγουμε στο μοντέλο αυτό.

Για τον υπολογισμό των Schoenfeld υπολοίπων εφαρμόζουμε την εντολή **residuals()** εισάγοντας το αντικείμενο της εντολής **coxph()**, όπως παρακάτω:

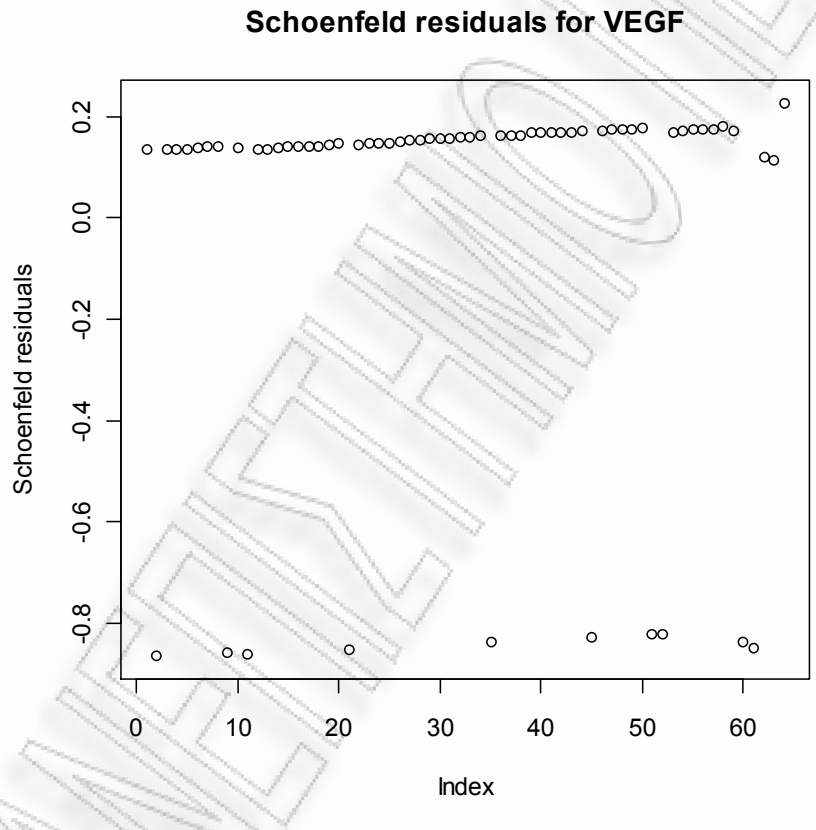
```
cox4=coxph(Surv(survival,death)~vegf+nodes13+nodes49+erIhc+pgrIhc+
+her2+genes1+genes2+group+pgrIhc*group, method="breslow")
schoen1=residuals(cox4, type="schoenfeld");schoen1
```

time	vegf	nodes13	nodes49	. . .	pgrIhc:group
12	0.1341307	-0.4297698	0.5078187	. . .	0.29488780
12.0655737704918	-0.8645663	-0.4339448	0.5127519	. . .	0.29775246
13.016393442623	0.1340576	-0.4346355	0.5135680	. . .	-0.70177360
. . .	. . .	. . .	. . .	. . .	. . .
105.540983606557	0.1214740	-0.3100634	-0.5808897	. . .	1.12812425
108.491803278689	0.1146968	0.6287671	-0.5506647	. . .	-0.67354228
110.852459016393	0.2281703	0.6030276	-0.4742277	. . .	1.29010059

Για κάθε μεταβλητή εμφανίζει τα Schoenfeld υπόλοιπα που αντιστοιχούν σε κάθε πλήρη χρόνο. Γνωρίζουμε ότι στους λογοκριμένους χρόνους τα Schoenfeld υπόλοιπα ισούται με μηδέν. Μπορούμε να παρουσιάσουμε γραφικά (Γράφημα Γ1.4) τα Schoenfeld υπόλοιπα της μεταβλητής που μας ενδιαφέρει έναντι του χρόνου ως εξής

```
plot(schoen1[,1])
```

Γράφημα Γ1.4



Ωστόσο, μεγαλύτερο ενδιαφέρον τόσο στη μελέτη όσο και στη γραφική παρουσίαση παρουσιάζουν τα scaled Schoenfeld υπόλοιπα, διότι μας δίνεται η δυνατότητα να εφαρμόσουμε στατιστικούς ελέγχους και να προσαρμόσουμε μία καμπύλης «εξομάλυνσης» στο γράφημα (*smoothing curve*). Για την εφαρμογή του ολικού ελέγχου για την επάρκεια του μοντέλου και των ατομικών ελέγχων για την εξέταση της υπόθεση του αναλογικού κινδύνου κάθε συμμεταβλητής χρησιμοποιούμε την εντολή **cox.zph()**, όπως παρακάτω:

```
test=cox.zph(cox4, transform="identity");test
```

	rho	chisq	p
vegf	0.000959	6.12e-05	0.9938
nodes13	-0.059890	2.35e-01	0.6281
nodes49	-0.173503	2.07e+00	0.1507
erIhc	0.193988	2.40e+00	0.1214
pgrIhc	*-0.249315*	*3.93e+00*	*0.0475*
her2	0.178116	2.34e+00	0.1260
genes1	0.043725	1.16e-01	0.7334
genes2	0.138693	1.19e+00	0.2744
group	-0.203522	2.75e+00	0.0970
pgrIhc:group	*0.288591*	*5.16e+00*	*0.0231*
GLOBAL	*NA*	*1.84e+01*	*0.0487*

Σε επίπεδο σημαντικότητας $a = 0.05$ απορρίπτουμε την μηδενική υπόθεση $H_0 : b_5(t) = b_5$ οριακά και την μηδενική υπόθεση $H_0 : b_{10}(t) = b_{10}$, δηλαδή ο βιοδείκτης PgR-mRNA οριακά δεν ικανοποιεί την υπόθεση του αναλογικού κινδύνου και επιπλέον η αλληλεπίδραση μεταξύ του βιοδείκτη PgR-mRNA και της χημειοθεραπείας δεν ικανοποιεί την υπόθεση του αναλογικού κινδύνου. Επίσης, σε επίπεδο σημαντικότητας $a = 0.05$ απορρίπτουμε οριακά την μηδενική υπόθεση $H_0 : b_1(t) = b_1, b_2(t) = b_2, ..., b_{10}(t) = b_{10}$ για τον ολικό έλεγχο, οπότε το μοντέλο είναι οριακά ανεπαρκής, αφού οι συμμεταβλητές δεν ικανοποιούν ταυτόχρονα την υπόθεση του αναλογικού κινδύνου.

Για να εμφανίσουμε τα scaled Schoenfeld υπολοίπα κάθε συμμεταβλητής του μοντέλου εργαζόμαστε όπως και στην περίπτωση των Schoenfeld υπολοίπων, δηλαδή

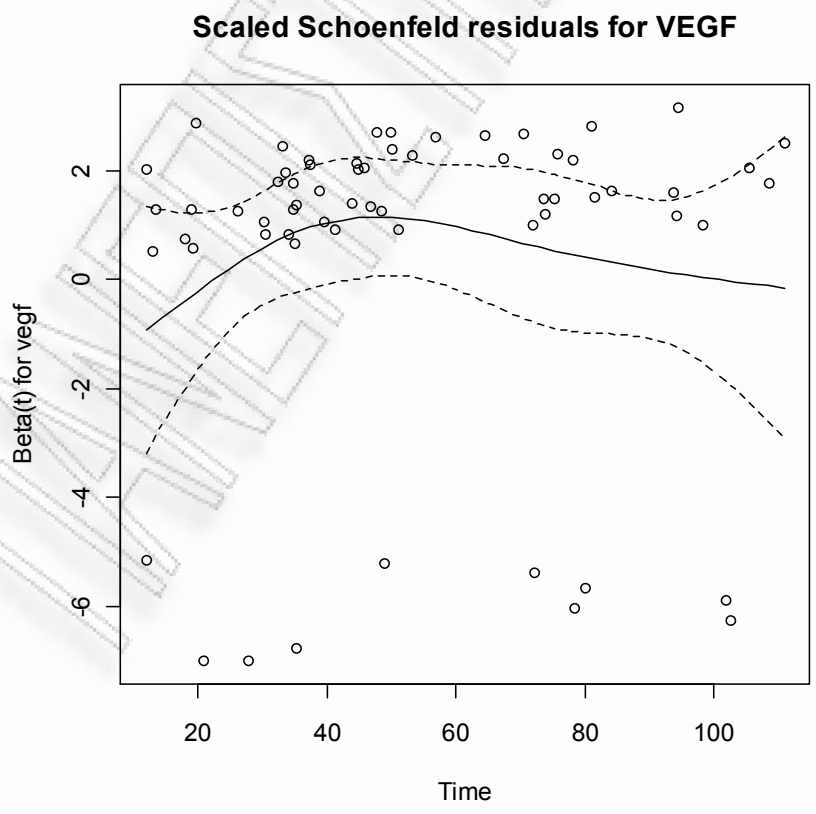
```
cox4=coxph(Surv(survival,death)~vegf+nodes13+nodes49+erIhc+pgrIhc+
+her2+genes1+genes2+group+pgrIhc*group, method="breslow")
scaled1=residuals(cox4, type="scaled");scaled1
```

time	vegf	nodes13	nodes49	. . .	pgrIhc:group
12	2.0217761	0.8947626	3.0931983	. . .	-4.610699
12.0655737704918	-5.1487577	0.3972150	4.2376252	. . .	-4.100900
13.016393442623	0.5277718	1.7645585	2.8173216	. . .	-5.045669
. . .	. . .	. . .	. . .	. . .	. . .
105.540983606557	2.0629492	-9.4667165	-9.2725760	. . .	4.219172
108.491803278689	1.7601612	3.3170757	0.5701094	. . .	3.606284
110.852459016393	2.5193978	3.6582001	1.3140792	. . .	0.907148

Για την γραφική (Γράφημα Γ1.5) απεικόνιση των scaled Schoenfeld υπολοίπων για κάθε συμμεταβλητή χρησιμοποιούμε την εντολή **plot()** ως εξής:

```
plot(test[1])
```

Γράφημα Γ1.5



Παρατηρούμε ότι η καμπύλης «εξομάλυνσης» έχει σχεδόν μηδενική τιμή και αυτό είναι αναμενόμενο αφού σύμφωνα με τον (τοπικό) έλεγχο της μηδενικής υπόθεση $H_0 : b_1(t) = b_1$ το $p-value$ είναι σχεδόν ίσο με τη μονάδα, οπότε συμπεραίνουμε ότι η πρωτεΐνη VEGF ικανοποιεί την υπόθεση του αναλογικού κινδύνου, δηλαδή σε κάθε ασθενή η τιμή της παραμένει σταθερή με την πάροδο του χρόνου (μεταβλητή ανεξάρτητη από τον χρόνο)

Στην συνέχεια μελετάμε τις διαφορές Δέλτα-Βήτα ή scaled score υπόλοιπα χρησιμοποιώντας την εντολή **residuals()**, όπως γνωρίζουμε ήδη

```
dbresid=residuals(cox4, type=”dfbeta”);dbresid
```

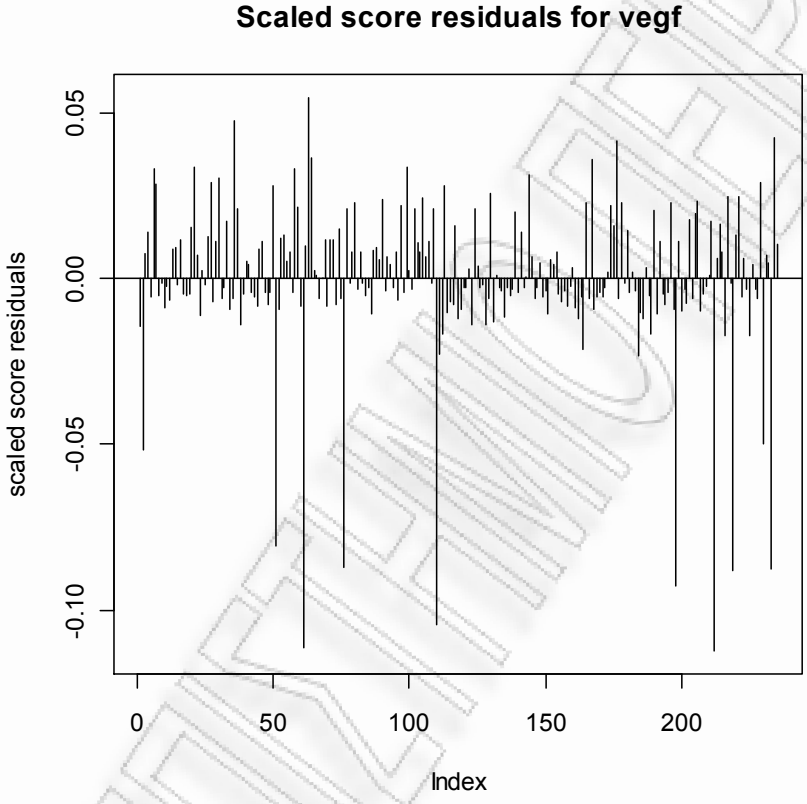
id	vegf	nodes13	nodes49	. . .	pgrIhc:group
1	-0.0143046934	-1.374072e-02	-3.278727e-02	. . .	-0.0626338165
2	-0.0514689130	7.036305e-03	2.758396e-02	. . .	-0.0184425987
3	0.0074818202	-2.871107e-03	-1.188780e-03	. . .	0.0056325939
. . .	. . .	. . .	. . .	. . .	. . .
313	-0.0875592963	-2.226831e-03	2.579048e-02	. . .	0.0274723207
314	0.0422632400	5.369133e-02	5.793806e-02	. . .	0.0133295708
315	0.0104664207	5.316115e-03	2.044898e-02	. . .	-0.0110892178

Για κάθε μεταβλητή υπολογίζονται τα scaled score υπόλοιπα, όπως συμβαίνει και στην περίπτωση των Schoenfeld και scaled Schoefeld υπολοίπων. Ωστόσο, στην περίπτωση των scaled score υπολοίπων λαμβάνουμε υπόψη και τους λογοκριμένους χρόνους.

Για την γραφική αναπαράσταση (Γράφημα Γ1.6) των scaled score υπολοίπων έναντι των ασθενών εργαζόμαστε πάλι όπως στα Schoenfeld υπόλοιπα και επιπλέον εισάγουμε την παράμετρο **type="h"**, στην εντολή **plot()** έτσι ώστε να εμφανίζονται τα σημεία με την μορφή ευθείων γραμμών παράλληλων προς τον άξονα y. Συγκεκριμένα εφαρμόζουμε τις παρακάτω εντολές:

```
plot(deltabeta1[,1],type="h",ylab="scaled score residuals",main="Scaled score
residuals for vegf")
abline(h=0)
```

Γράφημα Γ1.6



Παρατηρούμε ότι σχεδόν όλοι οι ασθενείς ασκούν μικρή επιρροή στο υπό μελέτη μοντέλο, αφού τα αντίστοιχα scaled score υπόλοιπα κυμαίνονται στο διάστημα $(-0.025, 0.05)$. Περίπου 10 ασθενείς έχουν scaled score υπόλοιπα μικρότερα του -0.05, εκ των οποίων η $272^{η}$ ασθενής φαίνεται να επηρεάζει συγκριτικά περισσότερο το μοντέλο, αφού έχει scaled score υπόλοιπο ίσο με -0.112145. Ωστόσο η τιμή αυτών των scaled score υπολοίπων δεν θεωρείται μεγάλη έτσι ώστε να συμπεράνουμε ότι οι αντίστοιχες ασθενείς επηρεάζουν σημαντικά το μοντέλο.

Παράρτημα Γ2

Παρουσίαση εντολών του προγράμματος R στην Ανάλυση επιβίωσης

Cross tabulation with test for factor independence (CrossTable {gmodels})

Κατασκευάζει πίνακες συνάφειας που περιέχουν απόλυτες και αναμενόμενες τιμές, ποσοστά (ανά γραμμή, ανά στήλη, ανά σύνολο) και ελέγχο ανεξαρτησίας των παραγόντων (Fisher, Chi-square, McNemar).

CrossTable(x, y, digits, expected, prop.r, prop.c, prop.t, prop.chisq, fisher, mcnemar, resid, sresid, asresid, missing.include)

x: ο παράγοντας που θα αποτελέσει τις γραμμές του πίνακα.

y: ο παράγοντας που θα αποτελέσει τις στήλες του πίνακα.

digits: το πλήθος των δεκαδικών ψηφίων στα ποσοστά μόνο.

expected: εάν δηλώσουμε **TRUE**, τότε σε κάθε κελί θα εμφανίζεται η αναμενόμενη τιμή σύμφωνα με το chi-square στατιστικό.

prop.r: εάν δηλώσουμε **TRUE**, τότε σε κάθε κελί θα εμφανίζονται τα ποσοστά ανά γραμμή.

prop.c: εάν δηλώσουμε **TRUE**, τότε σε κάθε κελί θα εμφανίζονται τα ποσοστά ανά στήλη.

prop.t: εάν δηλώσουμε **TRUE**, τότε σε κάθε κελί θα εμφανίζονται τα ποσοστά ανά σύνολο.

prop.chisq: εάν δηλώσουμε **TRUE**, τότε σε κάθε κελί θα εμφανίζονται τα αναμενόμενα ποσοστά σύμφωνα με το chi-square στατιστικό.

fisher: εάν δηλώσουμε **TRUE**, τότε εμφανίζει τη στατιστική συνάρτηση, τους βαθμούς ελευθερίας και το p-value του fisher test.

chisq: παρόμοια με το **fisher.**

mcnemar: παρόμοια με το **fisher.**

resid: εάν δηλώσουμε **TRUE**, τότε εμφανίζονται τα υπόλοιπα του Pearson.

sresid: εάν δηλώσουμε **TRUE**, τότε εμφανίζονται τα standardized υπόλοιπα.

asresid: εάν δηλώσουμε **TRUE**, τότε εμφανίζονται τα adjusted standardized υπόλοιπα.

missing.include: εάν δηλώσουμε **TRUE**, τότε παραλείπει οποιοδήποτε ασυνήθιστο χαρακτήρα ή «κενή» παρατήρηση.

Compute a survival curve for censored data (survfit.formula {survival})

Κατασκευάζει τον πίνακα επιβίωσης και χρησιμοποιείται στην κατασκευή γραφημάτων επιβίωσης

survfit(Surv(time, status) ~ x, subset, type, error, conf. type, conf. int, se. fit)

time: πρόκειται για το διάνυσμα με τους χρόνους ζωής (πλήρεις και λογοκριμένους).

status: η δίτιμη μεταβλητή, όπου η μονάδα αντιστοιχεί σε πλήρη χρόνο και το μηδέν σε λογοκριμένο χρόνο.

x: μία κατηγορική μεταβλητή. Στον πίνακα θα εμφανίζονται πληροφορίες για τους διαφορετικούς διατεταγμένους πλήρεις χρόνους, την πιθανότητα επιβίωσης και διάφορες άλλες ποσότητες για κάθε επίπεδο της μεταβλητής αυτής. Αν στη θέση του **x** ορίσουμε την μονάδα, τότε μελετάμε τους χρόνους ζωής χωρίς να λάβουμε υπόψη κάποια μεταβλητή. Αν στη θέση του **x** ορίσουμε την έκφραση **x+strata(y)**, όπου **y** μία κατηγορική μεταβλητή, τότε μελετάμε τους χρόνους ζωής για κάθε επίπεδο της μεταβλητής **x** και κάθε στρώμα της μεταβλητής (στρωματοποίησης) **y**.

subset: δηλώνουμε ποιο επίπεδο της μεταβλητής θέλουμε να μελετήσουμε. Χρησιμοποιείται και όταν θέλουμε να μελετήσουμε την επιβίωση στα επίπεδα μιας μεταβλητής, αλλά σε συγκεκριμένο στρώμα της μεταβλητής στρωματοποίησης. Στην πρώτη περίπτωση δηλώνουμε **subset = x = = επίπεδο_ενδιαφέροντος**, ενώ στη δεύτερη περίπτωση ορίζουμε **subset = y = = στρώμα_ενδιαφέροντος**.

type: αν μας ενδιαφέρει η Kaplan-Meier εκτίμηση της συνάρτησης επιβίωσης, τότε δηλώνουμε **"kaplan-meier"**, ενώ αν μας ενδιαφέρει η Fleming-Harrington εκτίμηση της συνάρτησης επιβίωσης, τότε δηλώνουμε **"fleming-harrington**".

error: καθορίζουμε την εκτίμηση της διακύμανσης της συνάρτησης επιβίωσης. Για την εκτίμηση της διακύμανσης της Fleming-Harrington και της Kaplan-Meier

συνάρτησης επιβίωσης ορίζουμε συνήθως **"greenwood"**, ενώ για την εκτίμηση της διακύμανσης της Fleming-Harrington αθροιστικής συνάρτησης κινδύνου συνήθως ορίζουμε **"tsiatis"**.

conf. type: αν δεν θέλουμε να υπολογίσουμε το διάστημα εμπιστοσύνης για την συνάρτηση επιβίωσης, τότε ορίζουμε **"plain"**, **"log"** ή **"log-log".** Να σημειώσουμε ότι σε κάποια σετ δεδομένων είναι δυνατόν το διάστημα εμπιστοσύνης τύπου plain να περιλαμβάνει τιμές που βρίσκονται εκτός του διαστήματος $[0,1]$, οπότε καλό είναι να αποφεύγεται σ' αυτές τις περιπτώσεις. Το διάστημα εμπιστοσύνης τύπου log βασίζεται στην αθροιστική συνάρτηση κινδύνου ή στον λογάριθμο της αθροιστικής συνάρτησης επιβίωσης, ενώ το διάστημα εμπιστοσύνης τύπου log-log βασίζεται στον λογάριθμο της αθροιστικής συνάρτησης κινδύνου ή στον λογάριθμο του αρνητικού λογαρίθμου της αθροιστικής συνάρτησης επιβίωσης.

conf. int: το επίπεδο εμπιστοσύνης για το δίπλευρο (*two-sided*) διάστημα εμπιστοσύνης για τη συνάρτηση επιβίωσης. Ως προεπιλογή είναι το 0.95.

se. fit: με την επιλογή **TRUE** (όπου είναι η προεπιλογή) υπολογίζει τα τυπικά σφάλματα της συνάρτησης επιβίωσης.

Κάποια από τα αντικείμενα της survfit( )

$time: διατεταγμένοι διαφορετικοί πλήρεις χρόνοι ζωής.

$n. risk: πλήθος ατόμων σε κίνδυνο σε κάθε (πλήρη) χρονική στιγμή.

$n. event: πλήθος αποτυχιών σε κάθε χρονική στιγμή.

$surv: η συνάρτηση επιβίωσης που αντιστοιχεί σε κάθε χρονική στιγμή. Σύμφωνα με την θεωρία υπολογίζεται για κάθε διάστημα $\left(t_j, t_{j+1}\right]$.

$std. error: το τυπικό σφάλμα της συνάρτησης επιβίωσης.

$lower: το κάτω άκρο του $(1-a)\%$ διαστήματος εμπιστοσύνης για τη συνάρτηση επιβίωσης για κάποιο συγκεκριμένο τύπο (plain, log, log-log) που έχουμε ορίσει.

$upper: το κάτω άκρο του $(1-a)\%$ διαστήματος εμπιστοσύνης για τη συνάρτηση επιβίωσης.

$n. censor: το πλήθος λογοκριμένων δεδομένων που συμβαίνουν σε κάθε διάστημα $\left[t_j, t_{j+1}\right)$ σύμφωνα με τη θεωρία.

$n: στην περίπτωση που δεν εξετάζεται κάποιος παράγοντας, αυτό το αντικείμενο μας πληροφορεί για το πλήθος των χρόνων ζωής του σετ δεδομένων. Σε περίπτωση που εξετάζεται κάποιος παράγοντας, τότε για κάθε επίπεδό του δίνεται το πλήθος των χρόνων ζωής που περιέχει. Τέλος, αν λαμβάνεται υπόψη και κάποιος παράγοντας στρωματοποίησης, τότε για κάθε συνδυασμό επιπέδου παράγοντα και στρώματος του παράγοντα στρωματοποίησης δίνει το πλήθος των χρόνων ζωής που αντιστοιχούν στον συνδυασμό αυτό.

Παρατήρηση

Εισάγοντας το όνομα της εντολής **survfit( )** στην εντολή **plot( )**, τότε παίρνουμε το γράφημα της συνάρτησης επιβίωσης για τις παραμέτρους που έχουμε ορίσει στην εντολή **survfit( )** είτε με είτε χωρίς συμμεταβλητή, είτε με είτε χωρίς παράγοντα στρωματοποίησης και τέλος είτε με είτε χωρίς διάστημα εμπιστοσύνης.

Test survival curve differences (survdiff {survival})

Με αυτή την εντολή συγκρίνουμε δύο ή περισσότερες καμπύλες επιβίωσης χρησιμοποιώντας τον logrank έλεγχο ή τον Gehan-Wilcoxon έλεγχο, όπου είναι η τροποποίηση του Peto-Peto ελέγχου. Και οι δύο αυτοί έλεγχοι ανήκουν στην $G-rho$ οικογένεια ελέγχων.

survdiff(Surv(time, status) ~ x, subset, rho)

Στην περίπτωση που θέλουμε να συγκρίνουμε τα επίπεδα ενός παράγοντα **x**, τότε παραλείπουμε την παράμετρο **subset**. Για ολικό στρωματοποιημένο έλεγχο χρησιμοποιούμε την «έκφραση» **x + strata(y)**. Για τοπικό στρωματοποιημένο έλεγχο χρησιμοποιούμε την παράμετρο **subset** ορίζοντας το στρώμα που μας ενδιαφέρει να μελετήσουμε.

rho: αν μας ενδιαφέρει τον logrank έλεγχο ορίζουμε την παράμετρο αυτή ίση με μηδέν, αλλιώς την ορίζουμε ίση με μονάδα αν μας ενδιαφέρει ο Gehan-Wilcoxon έλεγχος.

Παρατήρηση

Ο Gehan-Wilcoxon έλεγχος είναι γενίκευση του Kruskal-Wallis ελέγχου για λογοκριμένα δεδομένα. Ως εναλλακτική δοκιμάζουμε την συνάρτηση survdiff( ) με παράμετρο rho=1. Αυτή η μέθοδος χρησιμοποιεί για σταθμίσεις την Kaplan-Meier εκτίμηση της συνάρτησης επιβίωσης, δηλαδή $w_j = \hat{S}\left(t_j\right)$, όπου αποτελεί την Fleming-Harrington εκδοχή του Kruskal-Wallis ελέγχου για λογοκριμένα δεδομένα. Λόγω αυτής της στάθμισης αυτός ο έλεγχος δίνει περισσότερο βάρος στο αριστερό άκρο των πλήρων χρόνων ζωής (όπως και ο Gehan έλεγχος), όπου υπάρχουν περισσότερα άτομα σε κίνδυνο για να μελετήσουμε, άρα υπάρχει περισσότερη πληροφορία γύρω από τις συναρτήσεις επιβίωσης των ομάδων που θέλουμε να συγκρίνουμε.

Αντικείμενα της εντολής survdiff( )

$n: το πλήθος παρατηρήσεων σε κάθε επίπεδο του υπό εξέταση παράγοντα.

$obs: το σταθμισμένο παρατηρούμενο πλήθος αποτυχιών σε κάθε επίπεδο. Εάν έχουμε στρωματοποιημένο έλεγχο, τότε εμφανίζει έναν πίνακα, όπου οι στήλες είναι τα στρώματα και οι γραμμές είναι οι ομάδες που συγκρίνουμε.

$exp: το σταθμισμένο αναμενόμενο πλήθος αποτυχιών σε κάθε επίπεδο. Εάν έχουμε στρωματοποιημένο έλεγχο, τότε εμφανίζει έναν πίνακα, όπου οι στήλες είναι τα στρώματα και οι γραμμές είναι οι ομάδες που συγκρίνουμε.

$chisq: δίνει την τιμή της στατιστικής συνάρτησης Q για τον έλεγχο.

$var: δίνει ολόκληρο τον πίνακα διακυμάνσεων-συνδιακυμάνσεων $\Sigma = (\sigma_{is})_{k\times k}$, όπου k είναι το πλήθος επιπέδων του παράγοντα.

$strata: δίνει το πλήθος παρατηρήσεων που περιέχει κάθε στρώμα. Για να δούμε το πλήθος των παρατηρήσεων που περιέχονται σ' ένα από αυτά τα στρώματα θα πρέπει στην εντολή **survdiff( )** να έχουμε ορίσει τις παραμέτρους **x + strata(y), subset = y = = στρώμα_της_y**. Αν δεν περιλάβουμε την παράμετρο **strata( )**, τότε το αντικείμενο **$strata** δίνει το αποτέλεσμα **NULL**. Και οι δύο αυτές περιπτώσεις που περιγράφηκαν συμβαίνουν στους τοπικούς στρωματοποιημένους ελέγχους.

Fit proportional hazard regression model (coxph {survival})

Προσαρμόζει το μοντέλο παλινδρόμησης του Cox. Εξετάζουμε σταθερές (*fixed* ή *time-independent)* μεταβλητές, μεταβλητές που εξαρτώνται από τον χρόνο (*time dependent*) και το στρωματοποιημένο μοντέλο του Cox.

coxph(Surv(start, stop, event) ~ variables, subset,
method=c("efron" , "breslow","exact"))

(start, stop): ουσιαστικά πρόκειται για τα προκαθορισμένα χρονικά διαστήματα εξέτασης κάθε ασθενή. Σε κάθε τέτοιο χρονικό διάστημα ορίζεται η τιμή της μεταβλητής περί λογοκρισίας, **event** . Είναι προφανές ότι οι ποσότητες **start** και **stop** χρησιμοποιούνται όταν μελετάμε μεταβλητές που εξαρτώνται από τον χρόνο. Στην περίπτωση των σταθερών μεταβλητών γνωρίζουμε μόνο την **start**.

event: είναι η μεταβλητή λογοκρισίας, οπότε παίρνει την τιμή 0 ή 1, εάν ο αντίστοιχος χρόνος είναι λογοκριμένος ή πλήρης αντίστοιχα.

variables: κατηγορικές μεταβλητές, συνεχείς μεταβλητές και αλληλεπιδράσεις μεταβλητών, όπου είναι εφικτό (δηλαδή κατηγορική με κατηγορική και κατηγορική με συνεχή).

method: στην περίπτωση απουσίας δεσμών και οι τρεις μέθοδοι δίνουν ίδια αποτελέσματα. Στην περίπτωση αρκετών δεσμών προτιμάται η μέθοδος **Efron** ως η πιο αξιόπιστη. Στην περίπτωση που υποθέτουμε ότι τα δεδομένα μας προέρχονται από κάποια διακριτή κατανομή του χρόνου ζωής, τότε προτιμάται η μέθοδος **exact**. Ως προεπιλογή είναι η μέθοδος **Breslow**.

Παρατηρήσεις

1. Όταν εφαρμόζουμε το στρωματοποιημένο μοντέλο του Cox (*Cox's stratified model, SC model*), τότε ορίζουμε την «έκφραση» **variables + strata(y)**. Εάν θέλουμε να προσαρμόσουμε το SC μοντέλο για συγκεκριμένο επίπεδο της μεταβλητής στρωματοποίησης **y**, τότε ορίζουμε και την παράμετρο **subset = y = =στρώμα_της_y** καταλλήλως. Για παράδειγμα

   coxph(Surv(time, status) ~ her2+erihc+strata(group),
   subset = group = = 1, method="breslow")

2. Για να ελέγξουμε την υπόθεση μη-αλληλεπίδρασης αρκεί να κατασκευάσουμε δύο SC μοντέλα, όπου το ένα μοντέλο δεν θα περιέχει αλληλεπίδραση μεταξύ του παράγοντα στρωματοποίησης και των παραγόντων του μοντέλου (υπόθεση μη αλληλεπίδρασης), ενώ το άλλο μοντέλο θα περιέχει αλληλεπίδραση (υπόθεση αλληλεπίδρασης). Έπειτα συγκρίνουμε τα δύο αυτά μοντέλα χρησιμοποιώντας Likelihood ratio έλεγχο για να ελέγξουμε την υπόθεση $H_0 : a = 0$, όπου a είναι το διάνυσμα των παραμέτρων των αλληλεπιδράσεων.

3. Είναι δυνατόν σε κάποια σετ δεδομένων ο εκτιμητής μεγίστης πιθανοφάνειας ενός συντελεστή παλινδρόμησης να είναι άπειρος, για παράδειγμα στην περίπτωση μίας διχοτόμου μεταβλητής όπου μία από τις δύο ομάδες της δεν έχει αποτυχίες (*events*), δηλαδή έχει μόνο λογοκριμένα γεγονότα. Όταν συμβαίνει αυτό, τότε ο αντίστοιχος συντελεστής (*coefficient*) παλινδρόμησης θα πάρει μεγάλη τιμή και στην εντολή **coxph( )** θα συμβεί ένα από τα παρακάτω:

- Είτε ο παρατηρούμενος πίνακας πληροφορίας του Fisher θα γίνει ιδιόμορφος (*singular*), δηλαδή δεν θα αντιστρέφεται, οπότε δεν μπορεί να υπολογιστεί ο πίνακας διακυμάνσεων-συνδιακυμάνσεων του διανύσματος παραμέτρων b, αλλά ούτε μπορεί να υπολογιστεί η στατιστική συνάρτηση του score ελέγχου.
- Είτε ο μέγιστος αριθμός επαναλήψεων του αλγόριθμου είναι υπερβολικά μεγάλος.

Συνήθως συμβαίνει η πρώτη περίπτωση. Συνέπεια αυτής της «αδυναμίας» είναι ότι ο τοπικός Wald έλεγχος και ό,τι αφορά τη συγκεκριμένη μεταβλητή δεν είναι αξιόπιστα. Μόνο οι ολικοί έλεγχοι Wald και Likelihood ratio είναι αξιόπιστοι, διότι δεν εξαρτώνται από τον αντίστροφο του παρατηρούμενου πίνακα πληροφορίας του Fisher.

4. Παράδειγμα δεδομένων όπου η μεταβλητή LBRT εξαρτάται από τον χρόνο

Patient	**Start**	**Stop**	**Status**	**AGE**	**LBRT**
. . .	. . .	. . .	. . .	. . .	. . .
5	0	87	0	73	2.8
5	87	192	0	73	2.6
5	192	341	0	73	2.9
5	341	450	1	73	3.4
6	0	94	0	64	2.4
6	94	197	1	64	2.3
6	197	384	0	64	2.8
. . .	. . .	. . .	. . .	. . .	. . .

Κάποια από τα αντικείμενα της εντολής coxph( )

$loglik: δίνει την ποσότητα $-2\log\hat{L}(0)$ για το «μηδενικό» μοντέλο (*null model*) του Cox, όπου δεν περιέχει κάποια μεταβλητή, και την ποσότητα $-2\log\hat{L}(\hat{b})$ για το μοντέλο του Cox, όπου περιέχει τουλάχιστον μία μεταβλητή. Δηλαδή, το αντικείμενο **$loglik** είναι ένα διάνυσμα μήκους ίσο με 2.

$coefficient: δίνει τους εκτιμητές μεγίστης πιθανοφάνειας των συντελεστών παλινδρόμησης του μοντέλου. Πρόκειται για ένα διάστημα μήκους όσο το πλήθος των μεταβλητών του μοντέλου.

$var: ο πίνακας διακυμάνσεων-συνδιακυμάνσεων του διανύσματος $\hat{b}$, όπου ουσιαστικά είναι ο αντίστροφος του παρατηρούμενου πίνακα πληροφορίας του Fisher (*Fisher's observed information matrix*).

$score: η τιμή της στατιστικής συνάρτησης του score ελέγχου. Χρησιμοποιούνται οι αρχικές τιμές του διανύσματος των συντελεστών, b_0 και οι τελικές (εκτιμημένες) τιμές του διανύσματος των συντελεστών, $\hat{b}$.

$wald. test: η τιμή της στατιστικής συνάρτησης του ολικού Wald ελέγχου που εξετάζει αν οι τελικές τιμές των συντελεστών παλινδρόμησης διαφέρουν από τις αρχικές τιμές τους, δηλαδή $\hat{b} - b_0$.

$iter: το πλήθος των επαναλήψεων μέχρι να συγκλίνει ο αλγόριθμος.

$residuals: δίνει τα martingales residuals, ένα για κάθε παρατήρηση όλου του σετ δεδομένων.

$means: για τις δίτιμες κατηγορικές μεταβλητές δίνει το μέσο πλήθος παρατηρήσεων που δεν ανήκουν στο επίπεδο αναφοράς (*reference category*), ενώ για τις συνεχείς μεταβλητές δίνει τη μέση τιμή τους. Οι καμπύλες επιβίωσης βασίζονται σε αυτή την τιμή, δηλαδή υπολογίζεται η ποσότητα $\hat{S}\left(t/\overline{Z}\right)$, όπου $\overline{Z} = \left(\overline{Z}_1, \overline{Z}_2, ..., \overline{Z}_p\right)$ είναι το διάνυσμα με τη μέση τιμή των μεταβλητών του μοντέλου.

Compute a survival curve from Cox model (survfit.coxph {survival})

Υπολογίζει την συνάρτηση επιβίωσης για το μοντέλο αναλογικού κινδύνου του Cox.

survfit(formula, newdata, conf. int, se. fit, conf. type = c("none","plain", "log", "log-log"), type = c("kaplan-meier", "aalen"))

formula: το αντικείμενο της εντολής **coxph( )**.

newdata: αν θέλουμε να υπολογίσουμε την συνάρτηση επιβίωσης, έστω για το πρώτο επίπεδο της μεταβλητής x και το δεύτερο επίπεδο της μεταβλητής y που περιέχονται στο μοντέλο παλινδρόμησης του Cox, τότε ορίζουμε την παράμετρο αυτή ως εξής **newdata = data. frame(x = 1, y = 2)**.

Παρατήρηση

Για να υπολογίσουμε την αναφορική συνάρτηση επιβίωσης, αρκεί αρχικά πρέπει να ορίσουμε την εντολή **coxph( )** χωρίς μεταβλητή, δηλαδή **coxph(Surv(time, status) ~ 1, method=".")** και να τοποθετήσουμε το αντικείμενο της εντολής αυτής στην παράμετρο **formula**.

Calculate residuals for a Cox fit (residuals. cox {survival})

Μέσω της ανάλυσης των υπολοίπων εξετάζουμε γραφικά και με στατιστικούς ελέγχους, αν οι μεταβλητές του μοντέλου παλινδρόμησης του Cox ικανοποιούν την υπόθεση αναλογικού κινδύνου, οπότε θα είναι σταθερές στον χρόνο.

residuals(object, type = c("martingale","deviance","score", "schoenfeld", "dfbeta","scaledsch"))

object: εισάγουμε το αντικείμενο της εντολής **coxph( )**.

type: καθορίζουμε ποιο είδος υπολοίπου θέλουμε να μελετήσουμε.

Παρατηρήσεις

1. Για τα **martingale** και **deviance** υπόλοιπα δίνεται ένα υπόλοιπο για κάθε παρατήρηση όλου του σετ δεδομένων. Εισάγοντας το αντικείμενο της εντολής **residuals( )** στην εντολή **plot( )** παίρνουμε την γραφική παράσταση των υπολοίπων αυτών, όπου ο άξονας x περιέχει τη θέση των παρατηρήσεων στο σύνολο των δεδομένων. Κάθε παρατήρηση έχει το ίδιο υπόλοιπο για κάθε μεταβλητή. Αυτό γνωρίζουμε ότι δεν ισχύει στα **schoenfeld**, **scaled schoenfeld**, **dfbeta** και **score** υπόλοιπα.

2. Για τα **schoenfeld**, **scaled schoenfeld**, **dfbeta** και **score** υπόλοιπα κατασκευάζεται ένας πίνακας με τόσες στήλες όσες είναι οι μεταβλητές του μοντέλου του Cox και τόσες γραμμές όσοι είναι οι διαφορετικοί πλήρεις χρόνοι ζωής και παρουσιάζονται στον πίνακα αυτό διατεταγμένοι κατ' αύξουσα τάξη. Για τον ολικό έλεγχο της επάρκειας του μοντέλου αναλογικού κινδύνου, δηλαδή της υπόθεσης $H_0 : b_1(t) = b_1, b_2(t) = b_2, ..., b_p(t) = b_p$, όπου λαμβάνουμε υπόψη όλες τις μεταβλητές του υπό εξέταση μοντέλου του Cox, αλλά και για τον ατομικό έλεγχο της υπόθεσης αναλογικού κινδύνου για κάθε μεταβλητή χωριστά (ή ισοδύναμα της υπόθεσης της μηδενικής κλίσης της *smoothing curve*), δηλαδή της υπόθεσης $H_0 : b_g(t) = b_g$ για την g μεταβλητή $(1 \le g \le p)$ χρησιμοποιούμε την παρακάτω εντολή:

cox. zph(object, transform = c("log", "identity"))

object: εισάγουμε το αντικείμενο της εντολής **coxph( )**.

transform: εάν θέλουμε να μελετήσουμε τον λογάριθμο των πλήρων χρόνων ζωής, τότε δηλώνουμε **"log"**, ενώ αν μας ενδιαφέρουν οι πλήρεις χρόνοι ζωής όπως είναι, τότε δηλώνουμε **"identity"**.

Αντικείμενα της εντολής cox. zph( )

$table: δίνεται ένας πίνακας όπου κάθε γραμμή αντιστοιχεί στον τοπικό έλεγχο για κάθε μεταβλητή του Cox μοντέλου, ενώ η τελευταία γραμμή αφορά τον ολικό (*global*) έλεγχο. Η πρώτη στήλη του πίνακα είναι ο συντελεστής συσχέτισης του Pearson μεταξύ των πλήρων χρόνων ζωής (όπως ορίστηκαν σύμφωνα με την παράμετρο **type**) και των scaled Schoenfeld υπολοίπων. Φυσικά, για τον ολικό έλεγχο δεν υφίσταται κάποιος συντελεστής συσχέτισης γι΄ αυτό εμφανίζεται η ένδειξη **NA**. Η δεύτερη στήλη δίνει την τιμή της στατιστικής συνάρτησης για κάθε τοπικό έλεγχο αλλά και για τον ολικό έλεγχο. Η τρίτη στήλη δίνει τα p-values. Τόσο οι τοπικοί έλεγχοι όσο και ο ολικός έλεγχος ακολουθούν ασυμπτωτικά την χ^2 κατανομή με 1 και p βαθμούς ελευθερίας αντίστοιχα, όπου p είναι το πλήθος των μεταβλητών του μοντέλου του Cox.

$x: οι πλήρεις χρόνοι ζωής σύμφωνα με την παράμετρο **transform**. Τοποθετούνται στον άξονα x του γραφήματος για τον γραφικό έλεγχο της υπόθεσης του αναλογικού κινδύνου χωριστά για κάθε μεταβλητή του μοντέλου του Cox.

$y: τα scaled Schoenfeld υπόλοιπα ως πίνακας με τόσες στήλες όσες οι μεταβλητές του μοντέλου του Cox που μελετάμε.

$var: στην ουσία είναι ο πίνακας διακυμάνσεων-συνδιακυμάνσεων των συντελεστών παλινδρόμησης του μοντέλου του Cox.

Για την γραφική αναπαράσταση των scaled Schoenfeld υπολοίπων για κάθε μεταβλητή του μοντέλου του Cox που μελετάμε τοποθετούμε το αντικείμενο της εντολής **cox. zph( )** στην εντολή **plot( )** μαζί με την κατάλληλη ένδειξη για την μεταβλητή που μας ενδιαφέρει. Για παράδειγμα, έστω το μοντέλο του Cox

cox=coxph(Surv(time, status) ~ group + her2, method="breslow")

Έπειτα εφαρμόζω την εντολή cox. zph( ) με τους πλήρεις χρόνους ως είναι:

phfit=cox. zph(cox, transform = "identity")

Λαμβάνοντας υπόψη τη σειρά που βρίσκεται κάθε μεταβλητή στο μοντέλο του Cox καθορίζουμε τον γραφικό έλεγχο για το αν ικανοποιεί την υπόθεση του αναλογικού κινδύνου ως εξής:

plot(phfit[1]) για την group και **plot(phfit[2])** για την her2

Σε κάθε γράφημα ο άξονας y περιέχει τα scaled Schoenfeld υπόλοιπα της μεταβλητής που μας ενδιαφέρει, ο άξονας x περιέχει τους πλήρεις χρόνους ζωής όπως έχουν οριστεί με την παράμετρο **transform** και επιπλέον εμφανίζεται και μία *smoothing curve*.

3. Για την γραφική απεικόνιση των Schoenfeld υπολοίπων καθορίζουμε ποια στήλη του πίνακα (δηλαδή του αντικειμένου **$table** της εντολής **residuals( )**) θέλουμε να εμφανίζεται στον άξονα y του γραφήματος. Για παράδειγμα, λαμβάνοντας υπόψη το μοντέλο του Cox της Παρατήρησης 2, τα Schoenfeld υπόλοιπα για κάθε μεταβλητή αυτού του μοντέλου δίνονται από την εντολή:

resid = residuals(cox, type= "schoenfeld")

οπότε η γραφική παράσταση των Schoenfeld υπολοίπων για την group δίνονται από την εντολή

plot(resid[,1])

ενώ για τη her2 δίνονται από την εντολή

plot(resid[,2])

Να σημειώσουμε ότι στα γραφήματα των Schoenfeld υπολοίπων δεν εμφανίζεται η *smoothing curve*. Παρόμοια εμφανίζουμε γραφικά και τα score υπόλοιπα. Σε αντίθεση με τα Schoenfeld υπόλοιπα στην γραφική αναπαράσταση των score υπολοίπων ο άξονας x περιέχει τη θέση κάθε παρατήρησης στο σετ των δεδομένων.

4. Για την γραφική απεικόνιση των **dfbeta** υπολοίπων για κάθε μεταβλητή ακολουθούμε παρόμοια διαδικασία με τα Schoenfeld και score υπόλοιπα, μόνο που στην εντολή **plot( )** εισάγουμε και την παράμετρο **type= "h"**, έτσι ώστε τα σημεία του γραφήματος να εμφανίζονται ως κάθετες γραμμές που ξεκινούν από μία ευθεία που είναι παράλληλη με τον άξονα x και αρχίζει από το σημείο $(0,0)$

του γραφήματος. η ευθεία αυτή εμφανίζεται σύμφωνα με την εντολή **abline(h=0)**. Σε κάθε παρατήρηση αντιστοιχεί ένα Δέλτα-Βήτα υπόλοιπο για την μεταβλητή που μας ενδιαφέρει.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

1. Αντζουλάκος Δημήτριος (2009). Ανάλυση Επιβίωσης (Β' Έκδοση). Σημειώσεις Παραδόσεων. Πρόγραμμα Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς.

2. Κούτρας Μάρκος (2005). Εφαρμοσμένη Πολυμεταβλητή Ανάλυση: Ανάλυση κατά συστάδες. Πρόγραμμα Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς.

3. Τσακανίκας Παναγιώτης (2005). Αναγνώριση γονιδιακών εκφράσεων νεοπλασιών με microarrays. Διπλωματική Εργασία. Τμήμα Φυσικής, Σχολή Θετικών Επιστημών, Πανεπιστήμιο Πατρών.

Ξένη

1. Abe, T., Kanaya, S., Kinouci, M., Matsuda, H., Kudo, Y., Mori, H., Carlos, D.C., Ikemura, T. (1999). Gene classification method based on batch-learning SOM. Genome Informatics, 10: 314-315.

2. Bastien, P., Vinzi, E., Tenenhaus, M. (2005). PLS generalised linear regression. Computational Statistics and Data Analysis, 48, 17-46.

3. Berrar, D.P., Dubitzky, W., Granzow, M., eds. (2003) A Practical Approach to Microarray Data Analysis. Kluwer: Norwell, MA, Chapter 5, 91-109.

4. Brock, G., Datta, Susmita, Datta, Somnath, Pihur, V. (2008). cl Valid: An R Package for Cluster Validation. Journal of Statistical Software, Volume 25, Issue 4.

5. Causton, H.C., Quackenbush, J., Brazma, A. (2003). Microarray Gene Expression Data Analysis: A Beginner's Guide. Blackwell Publishing Company.

6. Chen, G., Banerjee, N., Jaradat, S.A., Tanaka, T.K., Ko, M.S.H., Zhankg, M.Q. (2002). Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data. Statistica Sinica, 12: 241-262.

7. Chen, Y., Dougherty, E.R., Bittner, M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical Optics, 2 (4), 364-374.

8. Colosimo, E.A., Ferreira, F.F., Oliveira, M.D., Sousa, C.B. (2002). Empirical comparisons between nelson-Aalen and Kaplan-Meier survival function estimators. J. Statist. Comput. Simul., Vol. 72 (4), 299-308.

9. Daxin, J., Chung, T., Aidong, Z. (2004). Cluster Analysis for Gene Expression Data: A Survey. Department of Computing Science & Engineering, State of University of New York, USA.

10. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci., Vol. 95, 14863-14868, Genetics, USA.

11. Erfaneh, N., Yonghong, P. (2009). Microarray Gene Expression Data Mining: Clustering Analysis Review. Department of Computing, University of Bradford.

12. Handl, J., Knowles, J., Kell, D.B. (2005). Computational cluster validation in post-genomic data analysis. Bioinformatics, Vol. 21, no. 15, 3201-3213.

13. Hautaniemi, S., Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A., Wolf, M., Rutz, J., Mousses, S., Kallioniemi, O.P. (2003). Analysis and Visualization of Gene Expression microarray Data in Human Cancer Using Self-Organizing Maps. Machine Learning, 52, 45-66.

14. Hess, K.R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Statistics in Medicine, 14, 1707-1723.

15. Hollmen Jaakko (1996). Process Modeling Using the Self-Organizing Map. Department of Computer Science. Helsinki of University of technology, Finland.

16. Holmes, S. (2009). Clustering (Courses). Statistics Dept., Sequoia Hall, Stanford University, California.

17. Hourani Mouath and El Emary Ibrahiem M. M. (2009). Microarray Missing Values Imputation Methods: Critical Analysis Review. ComSIS Vol. 6, No. 2, 165-190.

18. Miller, J. E. (2005). Writing about Hazard Analysis. www.policy.rutgers.edu/faculty/miller/HazardsEcon.pdf

19. Johnson, R.A., Wichern, D.W. (1998). Applied Multivariate Statistical Analysis (Fourth Edition). Prentice Hall.

20. McLachlan, G.J., Do, K.A., Ambroise, C. (2004). Analyzing Microarray Gene Expression Data. Wiley Series in Probability and Statistics.

21. Patrik D'haeseleer (2005). How does gene expression clustering work? Nature Biotechnology, Volume 23, Number 12, 1499-1501.

22. Rachel MacKay Altman (spring 2009). Lecturer 33: Empirical Survivor Function. Teaching in Stat 402/602.

23. Richard Kay (2006). Cox's Proportional Hazards model. Encyclopedia of Statistical Sciences, John Wiley & sons.

24. Rodriguez, G. (2005). Non- parametric estimation in survival models. www.data.princeton.edu/pop509a/

25. Saed Sayad (2010). Clustering: Self Organizing Map (SOM). Data Mining in Engineering, Business and Medicine, University of Toronto, Canada.

26. Sawyer, S. (2005). Nonparametric Survival Analysis: Cox-Mantel tests and Permutation tests.

27. Sharan, R, Elkon, R., Shamir, R. (2002). Cluster Analysis and its Applications to Gene Expression Data. In Ernst Schering workshop on Bioinformatics and Genome Analysis.

28. Shoemaker, J.S., Lin, S.M. (2005). Methods of Microarray Data Analysis. Chapter 2, Springer Science, Boston. www.springerlink.com/index/w2173757q3q35121.pdf

29. Theodoridis, S., Koutroubas, K. (2006). Pattern Recognition (Third Edition). Academic Press, pp. 635. Orlando, FL, USA.

30. Wieringen, W. van W., Kun, D., Hampel, R., Boulesteix, A.L. (2009). Survival prediction using gene expression data: A review and comparison. Computational Statistics and Data Analysis, 53, 1590-1603.

31. Wilkinson, L., Friendly, M. (2009). The History of Cluster Heat Map. The American Statistitian, Volume 63 (2), 179-184.

32. Winnett, A., Sasieni, P. (2001). A note on scaled Schoenfeld residuals for the proportional hazards model. Biometrika, 88, 2, 565-571.

33. Witten, D.M., Tibshirani, R. (2007). A comparison of fold-change and the t-statistic for microarray data analysis. Technical Report. Stanford University, California.

34. Yeung, K.Y., Haynor, D.R., Ruzzo, W.L. (2001). Validating Clustering for Gene Expression Data. Bioinformatics, Vol. 17, no. 4, 309-318.

35. Yi Li (2010). Model Selection in Survival Analysis (Lecture notes for the Survival Analysis Course at Bocconi University, Italy). Department of Biostatistics, Havard SPH.

Ιστοσελίδες

1. Ai-junkie: www.ai-junkie.com/ann/som/som1.html

2. Nagpaul, P.S. (1999). Guide to Advanced Data Analysis using IDAMS Software. Submitted to Division of Information and Informaties, UNESCO. New Delhi, India. www.unesco.org/webworld/idams/advguide/TOC.html

3. Teknomo, K. (2009). Hierarchical Clustering Tutorial www.people.revoledu.com/kardi/tutorial/clustering/index.html

4. Tibshirani, R. (2004). Super PC for R: Tutorial. www.stat.stanford.edu/~tibs/superpc/tutorial.html