ΜΠΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΙΑΤΡΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΑΓΓΕΛΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

Model Based Clustering on High Dimensional Data

ΑΘΗΝΑ 2012

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών και το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την……. από την εξεταστική επιτροπή:

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	***ΒΑΘΜΙΔΑ***	***ΥΠΟΓΡΑΦΗ***
π.χ.		
Δ. ΚΑΡΛΗΣ (Επιβλέπων)	**ΑΝ.ΚΑΘΗΓΗΤΗΣ**	**……**
Δ. ΠΑΝΑΓΙΩΤΑΚΟΣ	**ΑΝ. ΚΑΘΗΓΗΤΗΣ**	**…….**
Λ. ΜΕΛΙΓΚΟΤΣΙΔΟΥ	**ΛΕΚΤΟΡΑΣ**	**……..**

Ευχαριστίες

Ευχαριστώ την αρραβωνιαστικιά μου Φωτεινή Κοσμίδη για τη δύναμη και την αγάπη που μου δίνει καθημερινά στη ζωή μου.

Ευχαριστώ βαθιά και ειλικρινά τον Καθηγητή κύριο Καρλή Δημήτριο, που μου παρείχε τη δυνατότητα να πραγματοποιήσω την παρούσα διπλωματική υπό την καθοδήγησή του. Τον ευχαριστώ θερμά για την εμπιστοσύνη που έδειξε στο πρόσωπό μου αναθέτοντάς μου αυτό το θέμα της διπλωματικής χωρίς την ύπαρξη κάποιας πρωτύτερης συνεργασίας. Τον ευχαριστώ ιδιαίτερα για την άριστη συνεργασία, συνεννόηση και για την καθοδήγηση καθ' όλο το χρονικό διάστημα εκπόνησης της διπλωματικής μου. Επίσης τον ευχαριστώ θερμά για τις πολύτιμες συμβουλές του σχετικά με τη συνέχιση των σπουδών μου. Επίσης ευχαριστώ τα άλλα δύο μέλη της τριμελής επιτροπής τον Καθηγητή κύριο Δημοσθένη Παναγιωτάκο και τη Λέκτορα κυρία Λουκία Μελικοτσίδου για την παρακολούθηση της διπλωματικής και τα εποικοδομητικά σχόλια και διορθώσεις κατά την τελική διαμόρφωση του κειμένου.

Τέλος θα ήθελα να ευχαριστήσω ειλικρινά τον πρόεδρο του μεταπτυχιακού Καθηγητή κύριο Μπουρνέτα Απόστολο, τον Καθηγητή κύριο Δημοσθένη Παναγιωτάκο, τη Λέκτορα κυρία Λουκία Μελικοτσίδου και το Λέκτορα κύριο Παρασκευή Δημήτριο για τις πολύτιμες συμβουλές τους τόσο σε επιστημονική όσο και σε προσωπική βάση καθ' όλη τη διάρκεια των σπουδών μου στο μεταπτυχιακό αυτό πρόγραμμα.

Αθήνα 2012

Κωνσταντίνος Αγγελής

Κεφάλαιο 1: Εισαγωγή

1.1 Εισαγωγή

Μία από τις τεχνικές της Πολυμεταβλητής Ανάλυσης είναι η συσταδική ανάλυση (Cluster Analysis). Πρόκειται για μία μέθοδο που έχει ως σκοπό να κατατάξει σε ομάδες τις υπάρχουσες παρατηρήσεις, χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Με άλλα λόγια η ανάλυση κατά συστάδες εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών με σκοπό να δημιουργήσει ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους. Θέλουμε δηλαδή να δημιουργήσουμε ομάδες (clusters) από παρατηρήσεις για τις οποίες τα δεδομένα δείχνουν πως έχουν παρόμοια χαρακτηριστικά. Μία επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς, αλλά παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο. Μερικά παραδείγματα εφαρμογών της συσταδικής ανάλυσης είναι τα ακόλουθα:

- Οι βιολόγοι ενδιαφέρονται να κατατάξουν διαφορετικά είδη ζώων σε ομάδες με βάση κάποια χαρακτηριστικά τους.
- Στο margeting είναι ενδιαφέρον πως μπορούν να ομαδοποιηθούν οι πελάτες σύμφωνα με τα στοιχεία που υπάρχουν σχετικά με τις αγοραστικές τους συνήθειες και τα δημογραφικά χαρακτηριστικά τους. Κάτι τέτοιο είναι πολλαπλά χρήσιμο, κυρίως για διαφημιστικούς λόγους, για παράδειγμα κάποια προϊόντα απευθύνονται σε συγκεκριμένη αγοραστική ομάδα.
- Οι σχεδιαστές ηλεκτρονικών σελίδων ενδιαφέρονται να βρουν και να ομαδοποιήσουν τη συμπεριφορά των χρηστών του Internet ανάλογα με τον τρόπο με τον οποίο σερφάρουν ανάμεσα σε διαφορετικές σελίδες. Επομένως, η συμπεριφορά τους όπως καταγράφεται με τη διαδοχική εναλλαγή σελίδων προσφέρει δεδομένα με σκοπό την ομαδοποίηση των χρηστών.
- Οι γενετιστές ενδιαφέρονται να κατατάξουν ασθενείς σε διαφορετικές ομάδες ανάλογα με τη γενετική έκφραση κάποιων γονιδίων.

Υπάρχουν διαφορετικές προσεγγίσεις για το πώς μπορούμε να κατατάξουμε σε ομάδες τα δεδομένα μας. Κάποιες από αυτές είναι εμπειρικές και βασίζονται στην

έννοια της απόστασης, όπου απόσταση δύο παρατηρήσεων είναι ένα μέτρο που εκφράζει το βαθμό ετερογένειάς τους, το πόσο πολύ δηλαδή διαφέρουν. Σε αυτές τις μεθόδους οι παρατηρήσεις μέσα σε κάθε ομάδα που θέλουμε να είναι ομοιογενείς θα έχουν μικρή απόσταση, ενώ παρατηρήσεις μεταξύ των ομάδων, που είναι λιγότερο ομοιογενείς, θα έχουν μεγαλύτερη απόσταση. Φυσικά, υπάρχουν διάφορα μέτρα αποστάσεων που χρησιμοποιούνται στην πράξη. Τα μέτρα αυτά χωρίζονται σε ομάδες ανάλογα με το είδος των δεδομένων στα οποία μπορούν να εφαρμοστούν. Για συνεχή δεδομένα το πιο απλό και γνωστό μέτρο αποστάσεως είναι η ευκλείδεια απόσταση. Άλλα μέτρα είναι η City-block (Manhattan) distance, η απόσταση Minkowski, η απόσταση Chebychev κ.α. Ωστόσο, η επιλογή του κατάλληλου μέτρου απόστασης συνιστά ένα ερώτημα και έχει να κάνει κυρίως με τη μέθοδο που θα χρησιμοποιήσουμε αλλά και με τον τύπο των δεδομένων μας. Επίσης, είναι σημαντικό να γνωρίζουμε το σκοπό της ανάλυσης αλλά και κάποια επιμέρους χαρακτηριστικά. Συνεπώς, το πρόβλημα της επιλογής του κατάλληλου μέτρου απόστασης είναι αρκετά περίπλοκο. Παρ' όλα αυτά, οι μέθοδοι αποστάσεων έχουν ιδιαίτερη απήχηση σε εφαρμοσμένα προβλήματα, καθώς δε χρειάζονται υποθέσεις και λειτουργούν εύκολα στην πράξη. Παραδείγματα μεθόδων αποστάσεων είναι

- Ιεραρχική Ανάλυση: Εδώ ξεκινάμε με κάθε παρατήρηση να είναι από μόνη της μια ομάδα. Σε κάθε βήμα ενώνουμε τις 2 παρατηρήσεις που έχουν πιο μικρή απόσταση. Αν 2 παρατηρήσεις έχουν ενωθεί σε προηγούμενο βήμα ενώνουμε μια προϋπάρχουσα ομάδα με μια παρατήρηση μέχρι να φτιάξουμε μια ομάδα. Κοιτώντας τα αποτελέσματα (υπό μορφή κάποιου δέντρου/δενδρογράμματος) διαλέγουμε πόσες ομάδες τελικά προκύπτουν και σε ποια ομάδα ανήκει κάθε παρατήρηση. Ιεραρχικές μέθοδοι χρησιμοποιούνται ευρέως σε προβλήματα φυλογενετικής ανάλυσης.
- K-Means: Εδώ ο αριθμός των ομάδων θεωρείται γνωστός εκ των προτέρων. Με έναν επαναληπτικό αλγόριθμο μοιράζουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην κάθε παρατήρηση.

Αυτές οι μέθοδοι αποστάσεων δε στηρίζονται σε κανένα πιθανοθεωρητικό μοντέλο και στην πραγματικότητα προσφέρουν πολύ λίγα στοιχεία στατιστικής συμπερασματολογίας. Οι προσεγγίσεις τους είναι κυρίως μαθηματικές και σε κανένα σημείο δε λαμβάνεται υπ' όψιν η μεταβλητότητα που ίσως έχει σοβαρό ρόλο στα αποτελέσματα. Όμως, σε αρκετές εφαρμογές αυτή η έλλειψη κάποιου μοντέλου είναι επιθυμητή. Ουσιαστικά αφήνουμε τα δεδομένα να μιλήσουν χωρίς να τα προσαρμόζουμε σε κάποιο ιδεατό και πιθανότατα λανθασμένο μοντέλο. Από την

άλλη, αυτή η έλλειψη στατιστικού υποβάθρου μας εμποδίζει από το να κάνουμε στατιστική συμπερασματολογία. Προς αυτή την κατεύθυνση έχει αναπτυχθεί μια διαφορετική μεθοδολογία (model-based clustering) η οποία δεν έχει να κάνει με την έννοια της απόστασης, αλλά χρησιμοποιεί στατιστικά μοντέλα. Η μεθοδολογία αυτή έχει αφενός ένα σημαντικό θεωρητικό υπόβαθρο και αφετέρου προσφέρει μια σειρά από μεθοδολογικά εργαλεία για να μπορέσουμε να αξιολογήσουμε τα αποτελέσματα. Μέθοδοι ανάλυσης σε ομάδες με χρήση πιθανοθεωρητικών μοντέλων θα αναπτυχθούν εκτενώς στη συνέχεια.

Στο κεφάλαιο 2 γίνεται μια εισαγωγή στα μείγματα πολυμεταβλητών κανονικών κατανομών και πως αυτά χρησιμοποιούνται στο πλαίσιο του model-based clustering. Επίσης, παρουσιάζονται εκτενώς τα μοντέλα της GPCM (Gaussian Parsimonious Clustering Models) οικογένειας και ο αλγόριθμος EM που χρησιμοποιείται για την εκτίμηση των παραμέτρων τους. Επιπλέον, μελετώνται ζητήματα όπως το ποιο είναι το κατάλληλο μοντέλο, ποιες οι κατάλληλες αρχικές τιμές, πως θα εκτιμήσουμε το σωστό αριθμό ομάδων, πως θα αξιολογήσουμε τα αποτελέσματα της ομαδοποίησης κ.α.

Στο 3^{ο} κεφάλαιο παρουσιάζεται η χρήση των factor analyzers στο πλαίσιο του model-based clustering. Επίσης, παρουσιάζεται ο αλγόριθμος AECM που χρησιμοποιείται για την εκτίμηση των παραμέτρων καθώς και τρόποι ορισμού των αρχικών τιμών. Επιπλέον παρέχονται παραδείγματα και εφαρμογές.

Στο 4^{ο} κεφάλαιο παρουσιάζεται η PGMM (Parsimonious Gaussian Mixture Models) οικογένεια μοντέλων καθώς και η επέκταση αυτής, η EPGMM (Expanded Parsimonious Gaussian Mixture Models), που είναι κατάλληλες για εφαρμογή σε high dimensional δεδομένα. Επίσης παρέχονται παραδείγματα και συζήτηση γύρω από την επιλογή του κατάλληλου μοντέλου.

Στο 5^{ο} κεφάλαιο παρουσιάζεται η χρήση της πολυμεταβλητής t κατανομής στο πλαίσιο του model-based clustering καθώς και ο EM αλγόριθμος για την εκτίμηση των παραμέτρων. Επίσης, παρουσιάζεται η χρήση των factor analyzers για την περίπτωση της πολυμεταβλητής t-κατανομής καθώς και τα μοντέλα της MMtFA οικογένειας που προκύπτουν.

Τέλος, το 6^{ο} κεφάλαιο περιέχει μια εφαρμογή των μοντέλων της PGMM οικογένειας σε high dimensional δεδομένα από μια μελέτη έκφρασης γονιδίων.

Κεφάλαιο 2: Η τεχνική του model based clustering

2.1 Πεπερασμένα Μείγματα Κατανομών

Τα στατιστικά μοντέλα που θα αναπτυχθούν στη συνέχεια βασίζονται σε πεπερασμένες μίξεις κατανομών και ιδιαίτερα σε πεπερασμένες μίξεις κανονικών κατανομών. Γνωρίζουμε από το θεώρημα ολικής πιθανότητας ότι αν A_1, A_2,…,A_n είναι μία διαμέριση του δειγματικού χώρου S τότε για κάθε ενδεχόμενο E ισχύει $P(E)=\sum_{j=1}^{n}P\left(E\mid A_j\right)P\left(A_j\right)$. Το θεώρημα αυτό μπορεί να χρησιμοποιηθεί και στην περίπτωση που αντί για ενδεχόμενα, έχουμε τυχαίες μεταβλητές και συναρτήσεις πυκνότητας πιθανότητας. Έστω, ότι ο πληθυσμός μας αποτελείται από g υποπληθυσμούς, g ομάδες δηλαδή, και ότι γνωρίζουμε για κάθε ομάδα την κατανομή της. Ξέρουμε δηλαδή ότι

Ομάδα	**Κατανομή**
ομάδα 1	$f(x\mid\theta_1)$
ομάδα 2	$f(x\mid\theta_2)$
…	…
ομάδα g	$f(x\mid\theta_g)$

Διαλέγοντας ένα άτομο τυχαία από το γενικό πληθυσμό, χωρίς να γνωρίζουμε σε ποια ομάδα αυτός ανήκει, τότε η συνάρτηση πυκνότητας πιθανότητάς του θα είναι $f\left(x\mid\underset{\sim}{\theta}\right)=\sum_{j=1}^{g}\pi_j f_j\left(x\mid\theta_j\right)$ όπου $\pi_j>0$, με $\sum_{j=1}^{g}\pi_j=1$ και τα π_j δηλώνουν την πιθανότητα το άτομο αυτό να προέρχεται από την j ομάδα. Το διάνυσμα $\underset{\sim}{\theta}$, είναι το διάνυσμα των παραμέτρων, δηλαδή $\underset{\sim}{\theta}=\left(\pi_1,...,\pi_{g-1},\theta_1,...,\theta_g\right)$. Η παράμετρος θ_j της κάθε ομάδας δεν είναι απαραίτητο να είναι μία, αλλά μπορεί να είναι και διάνυσμα παραμέτρων. Για παράδειγμα αν η κατανομή της j ομάδας είναι η κανονική τότε θα

έχουμε ότι $\underset{\sim}{\theta}_j = (\mu_j, \sigma_j^2)$. Επίσης, το x δεν είναι απαραίτητο να είναι μία τυχαία μεταβλητή αλλά μπορεί να είναι διάνυσμα τυχαίων μεταβλητών, οδηγώντας έτσι σε πολυμεταβλητές κατανομές. Για παράδειγμα σε αυτή την περίπτωση η κατανομή της j ομάδας δε θα είναι η κανονική αλλά η πολυμεταβλητή κανονική. Η συνάρτηση πυκνότητας πιθανότητας $f\left(x \mid \underset{\sim}{\theta}\right)$ που μόλις ορίσαμε ονομάζεται πεπερασμένο μείγμα (finite mixture) g-ομάδων. Στην πράξη, και όταν χρησιμοποιούμε κάποια μέθοδο ομαδοποίησης (clustering), εμείς παρατηρούμε μόνο το x και όχι την ομάδα στην οποία ανήκει και επομένως οι παρατηρήσεις μας είναι ένα τυχαίο δείγμα από την $f\left(x \mid \underset{\sim}{\theta}\right)$ και όχι από τις επιμέρους $f_j\left(x \mid \theta_j\right)$. Αυτό βέβαια προϋποθέτει ότι η $f\left(x \mid \underset{\sim}{\theta}\right)$ είναι όντως αυτή η αληθινή συνάρτηση πυκνότητας πιθανότητας που περιγράφει τον πληθυσμό μας. Οι επιμέρους συναρτήσεις πυκνότητας πιθανότητας των διαφόρων κατανομών μπορεί να είναι διαφορετικές μεταξύ τους. Για παράδειγμα μπορεί η πρώτη να είναι κανονική κατανομή, η δεύτερη student κτλ. Όμως, συνήθως σε προβλήματα της cluster analysis και οι g συναρτήσεις πυκνότητας πιθανότητας λαμβάνονται ώστε να ανήκουν στην ίδια οικογένεια κατανομών. Η κατανομή που χρησιμοποιείται ευρέως είναι η κανονική, εξαιτίας των υπολογιστικών ευκολιών που παρουσιάζει.

Αν υποθέσουμε ότι $f\left(x \mid \mu, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\}$, $\mu, \sigma > 0$ τότε προκύπτει το μείγμα των κανονικών κατανομών με συνάρτηση πυκνότητας πιθανότητας $f\left(x \mid \underset{\sim}{\theta}\right) = \sum_{j=1}^{g} \pi_j \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_j^2}\left(x-\mu_j\right)^2\right\}$, όπου $\underset{\sim}{\theta} = \left(\pi_1, ..., \pi_{g-1}, \mu_1, ..., \mu_g, \sigma_1^2, ..., \sigma_g^2\right)$. Αν οι σ.π.π. των ομάδων είναι πολυμεταβλητές κανονικές κατανομές διαστάσεως p, δηλαδή $\varphi\left(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma\right) = \left(2\pi\right)^{-p/2} \left|\Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})^T\right\}$, όπου μ το διάνυσμα μέσων τιμών και Σ ο πίνακας διασποράς-συνδιασποράς τότε η σ.π.π. του μείγματος θα είναι $f\left(\underset{\sim}{x} \mid \underset{\sim}{\theta}\right) = \sum_{j=1}^{g} \pi_j \left(2\pi\right)^{-p/2} \left|\Sigma_j\right|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_j)\Sigma_j^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_j)^T\right\}$, με $\underset{\sim}{\theta} = \left(\pi_1, ..., \pi_{g-1}, \mu_1, ..., \mu_g, \Sigma_1, ..., \Sigma_g\right)$. Ωστόσο, στην ανάλυση σε ομάδες ανάλογα με τον τύπο των δεδομένων μας, θα μπορούσαμε να χρησιμοποιήσουμε διάφορες

κατανομές. Εδώ υποθέτοντας υποπληθυσμούς από την πολυμεταβλητή κανονική κατανομή, ουσιαστικά υποθέτουμε ότι τα δεδομένα μας είναι συνεχή. Άλλες επιλογές κατανομών θα ήταν πιο σωστές ανάλογα με το είδος των δεδομένων.

Επίσης, μπορούμε να θέσουμε διάφορους περιορισμούς στις κατανομές των επιμέρους ομάδων και ιδιαίτερα στους πίνακες διασποράς. Για παράδειγμα θα μπορούσαμε να θέσουμε $\Sigma_1 = \Sigma_2 = ... = \Sigma_g$, δηλαδή όλες οι ομάδες να έχουν τον ίδιο πίνακα διακύμανσης. Δεδομένου ότι ο πίνακας διακύμανσης ουσιαστικά καθορίζει το σχήμα της ομάδας υποθέτουμε ότι όλες οι ομάδες έχουν το ίδιο σχήμα. Επίσης σε αυτή την περίπτωση έχουμε τον μικρότερο αριθμό παραμέτρων προς εκτίμηση. Ωστόσο, μπορούμε να αφήσουμε τους πίνακες να διαφέρουν μεταξύ τους $\Sigma_1 \neq \Sigma_2 \neq ... \neq \Sigma_g$. Αυτή η περίπτωση εμφανίζει προβλήματα στην πράξη, αλλά από την άλλη επιτρέπει πολύ πιο ευέλικτα μοντέλα. Έτσι, μπορούμε να πάρουμε μια οικογένεια διαφορετικών κατανομών περισσότερο ή λιγότερο φειδωλές ως προς τον αριθμό των παραμέτρων. Τέλος, αυτό που πρέπει να επισημανθεί είναι ότι στον ορισμό της σ.π.π. του μείγματος κατανομών ο αριθμός των ομάδων g είναι γνωστός. Σε πολλές εφαρμογές, όπως και στην ομαδοποίηση δεδομένων, ο αριθμός των ομάδων δεν είναι γνωστός και πρέπει να εκτιμηθεί από τα δεδομένα. Περισσότερες πληροφορίες για μοντέλα μίξεων κατανομών και εφαρμογές τους δίνονται από τους McLachlan and Peel (2000a).

2.2 Αλγόριθμος ΕΜ

Από τον ορισμό των μοντέλων μίξεων κατανομών είναι σαφές ότι προϋποθέτουν την ύπαρξη ανομοιογενών πληθυσμών και συνεπώς είναι συνδεδεμένα με την ιδέα της ανάλυσης σε ομάδες. Υποθέτουμε ότι ο πληθυσμός αποτελείται από επιμέρους ομάδες που έχουν κάποια χαρακτηριστικά που εκφράζονται από τις παραμέτρους της κατανομής κάθε ομάδας, αλλά καθώς δε γνωρίζουμε σε ποια ομάδα ανήκει κάθε παρατήρηση, οι παρατηρήσεις μας προέρχονται από το μείγμα τους. Δηλαδή, στο model based clustering, θεωρούμε ότι τα δεδομένα μας προέρχονται από το μείγμα των κατανομών των επιμέρους ομάδων με σ.π.π. $f(x) = \sum_{j=1}^{g} \pi_j f_j(x \mid \theta)$, όπου f_j είναι η

σ.π.π. της κατανομής της j ομάδας και π_j είναι η πιθανότητα μία τυχαία παρατήρηση (πριν αρχίσουμε να παρατηρούμε δεδομένα) να ανήκει στην j ομάδα. Για την εφαρμογή στην περίπτωση της ομαδοποίησης δεδομένων το πρόβλημα είναι καθαρά στατιστικό: θέλουμε να εκτιμήσουμε τις άγνωστες παραμέτρους $\underset{\sim}{\theta} = \left(\pi_1,...,\pi_{g-1},\theta_1,...,\theta_g\right)$, όπου τα θ_j, όπως αναφέραμε και προηγουμένως, μπορεί να είναι διανύσματα (π.χ. μέση τιμή και διακύμανση για την περίπτωση μονοδιάστατων κανονικών κατανομών). Η κύρια μέθοδος που χρησιμοποιείται για να εκτιμήσουμε τις παραμέτρους μας είναι η μέθοδος μέγιστης πιθανοφάνειας. Η πιθανοφάνεια είναι $L = \prod_{i=1}^{n} f\left(x_i\right) = \prod_{i=1}^{n}\left[\sum_{j=1}^{g} \pi_j f\left(x_i \mid \theta_j\right)\right]$. Άρα, η λογαριθμική πιθανοφάνεια γίνεται $l\left(\theta \mid x\right) = \log\left\{\prod_{i=1}^{n}\left[\sum_{j=1}^{g} \pi_j f\left(x_i \mid \theta_j\right)\right]\right\} = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{g} \pi_j f\left(x_i \mid \theta_j\right)\right]$ που στην περίπτωση πολυμεταβλητών κανονικών κατανομών αντιστοιχεί σε

$$l\left(\theta \mid x\right) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{g} \pi_j \left(2\pi\right)^{-p/2} \left|\Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu})\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu})^T\right\}\right] =$$

$$\sum_{i=1}^{n} \log\left[\pi_1 \left(2\pi\right)^{-p/2} \left|\Sigma_1\right|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)\Sigma_1^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)^T\right\} + ...\right.$$

$$\left. ... + \pi_g \left(2\pi\right)^{-p/2} \left|\Sigma_g\right|^{-1/2} \exp\left\{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_g)\Sigma_g^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_g)^T\right\}\right].$$

Γενικά η εκτίμηση των παραμέτρων είναι αρκετά δύσκολη, καθώς η μεγιστοποίηση της log-likelihood δεν είναι καθόλου απλή και για αυτό το λόγο χρειαζόμαστε χρήση αριθμητικών μεθόδων και υπολογιστή. Συγκεκριμένα θα χρησιμοποιήσουμε τον αλγόριθμο EM.

Ο αλγόριθμος αυτός είναι ένας επαναληπτικός αλγόριθμος για εκτίμηση με τη μέθοδο μέγιστης πιθανοφάνειας. Τυπικά είναι μια αριθμητική μέθοδος μεγιστοποίησης που έχει όμως πολύ ενδιαφέρουσα στατιστική ερμηνεία και προσφέρει σημαντική βοήθεια στην απλοποίηση προβλημάτων εκτίμησης με τη μέθοδο μέγιστης πιθανοφάνειας. Ο αλγόριθμος είναι γνωστός στη στατιστική εδώ και πολλά χρόνια, αλλά από το 1977 (Dempster et al., 1977) έγινε γνωστός στη γενική του μορφή. Η εξάπλωση της χρήσης υπολογιστών βοήθησε ώστε ο αλγόριθμος να εφαρμοστεί σε μια μεγάλη ποικιλία εφαρμογών που μπορούν να ειδωθούν κάτω από το πρίσμα των "χαμένων δεδομένων" (missing data). Ο αλγόριθμος χρησιμοποιείται όταν έχουμε missing data ή όταν μπορούμε να εκφράσουμε το πρόβλημα σαν να

έχουμε missing data. Στην περίπτωση της cluster analysis, όπου είπαμε ότι χρησιμοποιούνται οι μίξεις κατανομών, ως missing data θεωρούμε τις ετικέτες που θα μας υποδείκνυαν από ποιον υποπληθυσμό προέρχεται κάθε παρατήρηση. Ο αλγόριθμος χρωστά το όνομά του στα δύο βήματα που τον απαρτίζουν. Το E-βήμα (Expectation step) και το M-βήμα (Maximization step). Γενικά η φιλοσοφία του αλγορίθμου είναι η εξής: στο E-step εκτιμάμε τα δεδομένα που λείπουν, χρησιμοποιώντας την πληροφορία που έχουμε μέχρι εκείνη τη στιγμή (δηλαδή τις παρατηρήσεις και τις τιμές των εκτιμητριών των παραμέτρων των κατανομών μέχρι τη στιγμή αυτή). Εκτιμάμε δηλαδή την πιθανότητα κάθε παρατήρηση να ανήκει στην κάθε ομάδα, δοθέντος των εκτιμήσεων των παραμέτρων που έχουμε μέχρι στιγμής. Στο M-step χρησιμοποιούμε τις εκτιμήσεις των πιθανοτήτων αυτών, για να μεγιστοποιήσουμε την πιθανοφάνεια και να υπολογίσουμε έτσι νέες εκτιμήσεις για τις παραμέτρους των κατανομών.

Αποδεικνύεται ότι ο αλγόριθμος μεγαλώνει την πιθανοφάνεια σε κάθε επανάληψη, επομένως συγκλίνει προς ένα μέγιστό της. Δεν είμαστε όμως ποτέ σίγουροι ότι αυτό το μέγιστο είναι το ολικό μέγιστο, επομένως θα πρέπει να προσέξουμε για να είμαστε σίγουροι ότι βρήκαμε το ολικό μέγιστο. Στην πράξη χρειάζεται να εκτελέσουμε τον αλγόριθμο περισσότερες από μία φορές, ξεκινώντας από διαφορετικές αρχικές τιμές, για να μεγιστοποιήσουμε την πιθανότητα να βρούμε το ολικό μέγιστο και όχι ένα τοπικό.

Το πρόβλημα με τον EM αλγόριθμο γενικά, είναι ότι πρέπει να βρούμε έναν τρόπο να ορίσουμε κατάλληλα τα missing data ώστε να διευκολυνόμαστε στη μεγιστοποίηση της πιθανοφάνειας. Γενικά, χρειάζεται αρκετή εμπειρία και φαντασία για να βρει κανείς την κατάλληλη δομή να συμπληρώσει τα παρατηρηθέντα δεδομένα με μη παρατηρηθέντα. Η τεχνική αυτή ονομάζεται data augmentation. Στην περίπτωση των πεπερασμένων μειγμάτων της πολυμεταβλητής κανονικής κατανομής, αυτό που εμείς παρατηρούμε είναι τα διανύσματα x_i, i=1,…,n που ακολουθούν ένα μείγμα πολυμεταβλητών κανονικών κατανομών όπως ορίστηκε παραπάνω. Έστω τυχαίες μεταβλητές Z_{ij}, i=1,…,n, j=1,…,g με τιμές Z_{ij}=1 αν η i παρατήρηση ανήκει στη j ομάδα και 0 αλλιώς. Αν γνωρίζαμε τα Z_{ij} τότε η εκτίμηση θα ήταν εύκολη, καθώς για κάθε παρατήρηση θα ξέραμε σε ποια ομάδα ανήκει και άρα το πρόβλημα θα ήταν απλά να εκτιμήσουμε τις παραμέτρους πολλών απλών πολυμεταβλητών κανονικών κατανομών. Συνεπώς, στην περίπτωσή μας τα Z_{ij} είναι τα missing data και

στο E-step εκτιμάμε τη δεσμευμένη αναμενόμενή τους τιμή, ενώ στο M-step χρησιμοποιούμε αυτές τις εκτιμήσεις, για να ανανεώσουμε τις παραμέτρους μας.

2.3 Βασική Θεωρία Αλγορίθμου ΕΜ

Έστω το παρακάτω μείγμα κατανομών

$$f\left(y_i \mid \Psi\right) = \sum_{j=1}^{g} \pi_j f_j\left(y_i \mid \theta_j\right) \tag{2.1}$$

όπου $y = \left(y_1^T, ..., y_n^T\right)^T$ τα δεδομένα που έχουμε παρατηρήσει και $\Psi = \left(\pi_1, ..., \pi_{g-1}, \xi\right)$ το διάνυσμα που περιέχει όλες τις άγνωστες παραμέτρους του μείγματος, με το διάνυσμα ξ να περιέχει τις άγνωστες παραμέτρους $\theta_1, ..., \theta_g$. Θεωρούμε ότι κάθε παρατήρηση y_i έχει προέλθει από μία συγκεκριμένη κατανομή του μείγματος, χωρίς όμως να γνωρίζουμε από ποια συγκεκριμένα. Έτσι, θεωρούμε τα διανύσματα $z_1, ..., z_n$ όπου κάθε z_i είναι ένα διάνυσμα διάστασης g, το οποίο μας υποδεικνύει αν η i παρατήρηση προέρχεται ή όχι από την j ομάδα, $z_{ij} = (z_i)_j = 1$ *ή* 0 αντίστοιχα. Αυτά τα z_{ij} όμως δεν τα παρατηρούμε και άρα δεν τα γνωρίζουμε. Έτσι, τα αντιμετωπίζουμε ως χαμένα δεδομένα (missing data). Κατά συνέπεια, το διάνυσμα $y = \left(y_1^T, ..., y_n^T\right)^T$ το θεωρούμε ως ημιτελές (incomplete), αφού δεν περιέχει τα z_{ij}. Αντίθετα το συμπληρωμένο (complete) διάνυσμα παρατηρήσεων είναι το $y_c = \left(y^T, z^T\right)^T$, όπου $z = \left(z_1^T, ..., z_n^T\right)^T$. Τότε η log-likelihood όλων των δεδομένων (complete-data log-likelihood) δίνεται από τον τύπο $\log L_c\left(\Psi\right) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij}\left\{\log \pi_j + \log f_j\left(y_i \mid \theta_j\right)\right\}$.

Ο ΕΜ αλγόριθμος εφαρμόζεται επαναληπτικά σε δύο βήματα (E & M steps). Στο Ε βήμα ο αλγόριθμος υπολογίζει τη δεσμευμένη αναμενόμενη τιμή της complete-data log-likelihood δοθέντος των παρατηρημένων δεδομένων y. Πάντα ξεκινάμε δίνοντας κάποιες αρχικές τιμές $\Psi^{(0)}$ στις παραμέτρους Ψ. Τότε, στην πρώτη επανάληψη του αλγορίθμου η δεσμευμένη αναμενόμενη τιμή της complete-data log-likelihood μπορεί να γραφεί ως $Q\left(\Psi \mid \Psi^{(0)}\right) = \mathrm{E}_{\Psi^{(0)}}\left\{\log L_c\left(\Psi\right) \mid y\right\}$. Ο δείκτης $\Psi^{(0)}$ του Ε δηλώνει ότι

αυτή η αναμενόμενη τιμή εξαρτάται από τις αρχικές τιμές του Ψ. Κατά συνέπεια στην $k+1$ επανάληψη υπολογίζεται η $Q\left(\Psi \mid \Psi^{(k)}\right)$, όπου $\Psi^{(k)}$ είναι η τιμή των Ψ μετά την k επανάληψη. Επίσης, καθώς η complete-data log-likelihood είναι γραμμική ως προς τα z_{ij}, στο Ε βήμα υπολογίζεται η δεσμευμένη αναμενόμενη τιμή των Z_{ij} δοθέντος των παρατηρήσεων $y=\left(y_1^T,...,y_n^T\right)^T$, όπου Z_{ij} είναι η τυχαία μεταβλητή που αντιστοιχεί στην παρατήρηση z_{ij} (την οποία επαναλαμβάνεται ότι δεν την γνωρίζουμε). Έτσι, στην $k+1$ επανάληψη έχουμε ότι $E_{\Psi^{(k)}}\left(Z_{ij} \mid y\right)=P_{\Psi^{(k)}}\left\{Z_{ij}=1 \mid y\right\}=\tau_j\left(y_i \mid \Psi^{(k)}\right)$ όπου αποδεικνύεται ότι

$$\tau_j\left(y_i \mid \Psi^{(k)}\right)=\frac{\pi_j^{(k)} f_j\left(y_i \mid \theta_j^{(k)}\right)}{f\left(y_i \mid \Psi^{(k)}\right)}=\frac{\pi_j^{(k)} f_j\left(y_i \mid \theta_j^{(k)}\right)}{\sum_{h=1}^{g} \pi_h^{(k)} f_h\left(y_i \mid \theta_h^{(k)}\right)}, \text{ i=1,…,n και j=1,…,g} \qquad (2.2)$$

Την ποσότητα $\tau_j\left(y_i \mid \Psi^{(k)}\right)$ θα την συμβολίζουμε από εδώ και στο εξής με $\tau_{ij}^{(k+1)}$ και είναι η εκ των υστέρων πιθανότητα η i παρατήρηση να ανήκει στην j ομάδα. Έτσι, η complete-data log-likelihood στην k+1 επανάληψη γίνεται $Q\left(\Psi \mid \Psi^{(k)}\right)=\sum_{j=1}^{g}\sum_{i=1}^{n} \tau_{ij}^{(k+1)}\left\{\log \pi_j+\log f_j\left(y_i \mid \theta_j\right)\right\}$.

Στο M step της $k+1$ επανάληψης μεγιστοποιούμε την $Q\left(\Psi \mid \Psi^{(k)}\right)$ ως προς Ψ, ώστε να πάρουμε μια νέα εκτίμηση $\Psi^{(k+1)}$ για το διάνυσμα των παραμέτρων Ψ. Οι εκτιμήσεις $\pi_j^{(k+1)}$ υπολογίζονται ανεξάρτητα από τις παραμέτρους $\xi^{(k+1)}$ του διανύσματος ξ, όπου $\xi=\left(\theta_1,...,\theta_g\right)$. Αν τα z_{ij} ήταν παρατηρήσιμα τότε η εκτιμήτρια μέγιστης πιθανοφάνειας των π_i θα δινόταν από $\hat{\pi}_j=\sum_{i=1}^{n}\frac{z_{ij}}{n}$ $(j=1,...,g)$. Καθώς όμως τα z_{ij} τα έχουμε εκτιμήσει από τα $\tau_{ij}^{(k+1)}$ στο Ε-βήμα, χρησιμοποιούμε αυτή τους την εκτίμηση για να υπολογίσουμε τα π_j. Άρα

$$\hat{\pi}_j=\sum_{i=1}^{n}\frac{\tau_{ij}^{(k+1)}}{n} \quad (j=1,...,g) \qquad (2.3)$$

Δηλαδή, η εκτίμηση του π_j στην k+1 επανάληψη είναι ο μέσος όρος των εκ των υστέρων πιθανοτήτων όλων των παρατηρήσεων να ανήκουν στην j ομάδα. Όσο αφορά την εκτίμηση των παραμέτρων ξ στο M βήμα της k+1 επανάληψης, αυτές δίνονται από την επίλυση της

$$\sum_{j=1}^{g}\sum_{i=1}^{n}\tau_{ij}^{(k+1)}\frac{\partial \log f_j\left(y_i \mid \theta_j\right)}{\partial \xi}=0 \qquad (2.4)$$

Ένα μεγάλο πλεονέκτημα του ΕΜ αλγορίθμου σε αυτό το σημείο είναι ότι η λύση αυτή συχνά είναι δυνατό να υπολογιστεί σε κλειστή μορφή, όπως για παράδειγμα στην περίπτωση μίξης πολυμεταβλητών κανονικών κατανομών.

Τα Ε και Μ βήματα επαναλαμβάνονται διαδοχικά μέχρι τη σύγκλιση, η οποία μπορεί να καθοριστεί χρησιμοποιώντας μία κατάλληλη συνθήκη τερματισμού, όπως για παράδειγμα η διαφορά $\left|L\left(\Psi^{(k+1)}\right)-L\left(\Psi^{(k)}\right)\right|<\varepsilon$, για κάποια μικρή θετική ποσότητα ε. Οι Dempster et al. (1977) έδειξαν ότι η πιθανοφάνεια L(Ψ) δε μειώνεται μετά από μία επανάληψη του αλγορίθμου, δηλαδή ισχύει $L\left(\Psi^{(k+1)}\right)\geq L\left(\Psi^{(k)}\right)$ για k=0,1,2,…Περισσότερες λεπτομέρειες σχετικά μπορούν να βρεθούν και στο Ng et al. (2004). Ωστόσο, δεν υπάρχει καμία εγγύηση ότι η σύγκλιση η οποία επιτυγχάνεται είναι στο ολικό μέγιστο και όχι σε ένα τοπικό μέγιστο. Για συναρτήσεις πιθανοφάνειας με πολλά τοπικά μέγιστα η σύγκλιση στο ολικό μέγιστο εξαρτάται από τις αρχικές τιμές $\Psi^{(0)}$.

Μία σημαντική παρατήρηση είναι ότι τα $\tau_{ij}^{(k+1)}$ περιέχουν την εκ των υστέρων πιθανότητα η i παρατήρηση να ανήκει στην j ομάδα. Όταν ο αλγόριθμος έχει συγκλίνει, αυτές οι πιθανότητες είναι διαθέσιμες και μπορούν να χρησιμοποιηθούν για να κατατάξουν τις παρατηρήσεις σε ομάδες. Συνήθως, κατατάσσουμε κάθε παρατήρηση στην ομάδα με τη μεγαλύτερη πιθανότητα. Ένα από τα πλεονεκτήματα της μεθόδου (model based clustering) είναι πως έχουμε πιθανότητες για κάθε ομάδα, δηλαδή κάθε παρατήρηση μπορεί να ανήκει σε περισσότερες από μία ομάδες με κάποια πιθανότητα. Γι' αυτό και οι ομάδες μπορούν να επικαλύπτονται, κάτι που δεν είναι δυνατό στην ομαδοποίηση μέσω μεθόδων αποστάσεων, όπως ο K-means.

2.4 Μίξεις Πολυμεταβλητών Κανονικών Κατανομών στον ΕΜ Αλγόριθμο

Παραπάνω περιγράψαμε τη γενική μορφή του αλγορίθμου στην περίπτωση που το μείγμα αποτελείται από διαφόρων ειδών κατανομές. Τώρα θα δούμε το αποτέλεσμα

του αλγορίθμου στην περίπτωση που όλες οι κατανομές είναι πολυμεταβλητές κανονικές. Πρώτα θα θεωρήσουμε την ετεροσκεδαστική περίπτωση όπου δηλαδή όλοι οι πίνακες διασποράς-συνδιασποράς των ομάδων επιτρέπεται να είναι διαφορετικοί μεταξύ τους.

Έστω δηλαδή ότι έχουμε το μείγμα

$$f\left(y_i \mid \Psi\right) = \sum_{j=1}^{g} \pi_j \varphi\left(y_i \mid \mu_j, \Sigma_j\right) \tag{2.5}$$

όπου $\Psi = \left(\pi_1, ..., \pi_{g-1}, \xi\right)$ και το ξ περιέχει τους μέσους μ_j (διάστασης p) και τους πίνακες διασποράς-συνδιασποράς Σ_j (διάστασης pxp) των ομάδων για j=1,…,g. Στο Ε-βήμα, από την (2.2) έχουμε ότι η εκτίμηση των z_{ij}, που μας δείχνουν από ποια ομάδα προέρχεται κάθε παρατήρηση, στην k+1 επανάληψη είναι

$$\tau_{ij}^{(k+1)} = \frac{\pi_j^{(k)} \varphi_j\left(y_i \mid \mu_j^{(k)}, \Sigma_j^{(k)}\right)}{\sum_{h=1}^{g} \pi_h^{(k)} \varphi_h\left(y_i \mid \mu_h^{(k)}, \Sigma_h^{(k)}\right)} =$$

$$= \frac{\pi_j^{(k)} \left(2\pi\right)^{-p/2} \left|\Sigma_j\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(y_i - \mu_j\right)\Sigma_j^{-1}\left(y_i - \mu_j\right)^T\right\}}{\sum_{h=1}^{g} \pi_h^{(k)} \left(2\pi\right)^{-p/2} \left|\Sigma_h\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(y_i - \mu_h\right)\Sigma_h^{-1}\left(y_i - \mu_h\right)^T\right\}}, i = 1, ..., n \;\&\; j = 1, ..., g \tag{2.6}$$

Στο Μ βήμα, στην k+1 επανάληψη, η εκτίμηση των π_j είναι η ίδια όπως τη βρήκαμε στην (2.3), δηλαδή $\hat{\pi}_j = \sum_{i=1}^{n} \frac{\tau_{ij}^{(k+1)}}{n} \quad (j = 1, ..., g)$. Οι εκτιμήσεις των μέσων τιμών μ_j και πινάκων διακύμανσης-συνδιακύμανσης Σ_j υπάρχουν σε κλειστή μορφή καθώς η (2.4) επιλύεται αναλυτικά. Έτσι έχουμε:

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k+1)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k+1)}} \quad \text{και} \tag{2.7}$$

$$\Sigma_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k+1)} \left(y_i - \mu_j^{(k+1)}\right)\left(y_i - \mu_j^{(k+1)}\right)^T}{\sum_{i=1}^{n} \tau_{ij}^{(k+1)}} \quad \text{για i=1,…,n και j=1,…,g.} \tag{2.8}$$

Ωστόσο, για διευκόλυνση των υπολογισμών και κυρίως για να μειώσουμε το χρόνο εκτέλεσης του αλγορίθμου στον υπολογιστή αλλά και για να κερδίσουμε χώρο στη μνήμη είναι καλύτερα να γράψουμε τη (2.8) μέσω των ποσοτήτων T_{j1}, T_{j2}, T_{j3}

όπου: $T_{j1}^{(k+1)} = \sum_{i=1}^{n} \tau_{ij}^{(k+1)}$, $T_{j2}^{(k+1)} = \sum_{i=1}^{n} \tau_{ij}^{(k+1)} y_i$ και $T_{j3}^{(k+1)} = \sum_{i=1}^{n} \tau_{ij}^{(k+1)} y_i y_i^T$. Άρα η (2.8) γίνεται

$$\Sigma_j^{(k+1)} = \frac{T_{j3}^{(k+1)} - T_{j1}^{(k+1)^{-1}} T_{j2}^{(k+1)} T_{j2}^{(k+1)^T}}{T_{j1}^{(k+1)}} \quad (j = 1, ..., g) \tag{2.9}$$

Στην πράξη, και όπως θα δούμε παρακάτω, συχνά τίθενται διάφοροι περιορισμοί, κυρίως στους πίνακες διακυμάνσεων. Για παράδειγμα μπορεί να ζητάμε να ισχύει $\Sigma_j=\Sigma$ (j=1,…,g), δηλαδή όλοι οι πίνακες διασποράς-συνδιασποράς των ομάδων να ισούνται μεταξύ τους (ομοσκεδαστικότητα). Σε αυτή την περίπτωση οι πίνακες Σ_j, όπου βέβαια τώρα έχουμε μόνο έναν πίνακα Σ αφού όλοι ισούνται μεταξύ τους, δίνεται από

$$\Sigma^{(k+1)} = \frac{\sum_{j=1}^{g} T_{j1}^{(k+1)} \Sigma_j^{(k+1)}}{n} \tag{2.10}$$

όπου ο $\Sigma_j^{(k+1)}$ δίνεται από την σχέση (2.9). Οι εκτιμήσεις των π_j και μ_j είναι ίδιες όπως στην ετεροσκεδαστική περίπτωση. Μπορούμε να θέσουμε επιπλέον περιορισμούς στους πίνακες διασποράς-συνδιασποράς όπως για παράδειγμα να απαιτήσουμε ο κοινός πίνακας Σ να είναι διαγώνιος με ίσα διαγώνια στοιχεία σ^2, δηλαδή $\Sigma = \sigma^2 \mathrm{I}_p$ McLachlan and Peel (2000a).

2.5 Αρχικές τιμές

Όπως αναφέρθηκε και παραπάνω ο αλγόριθμος για να τρέξει απαιτεί να του δώσουμε αρχικές τιμές $\Psi^{(0)}$. Καλές αρχικές τιμές μας εξασφαλίζουν γρήγορη σύγκλιση μέσα σε λίγες επαναλήψεις. Αντίθετα, αν οι αρχικές τιμές δεν είναι καλές ο αλγόριθμος θα αργήσει πολύ να συγκλίνει. Επίσης σε εφαρμογές όπου η συνάρτηση πιθανοφάνειας έχει πολλές ρίζες, που αντιστοιχούν σε πολλά τοπικά μέγιστα, ο αλγόριθμος θα πρέπει να εφαρμόζεται για διαφορετικές αρχικές τιμές ώστε να εντοπίσουμε το ολικό μέγιστο.

Είδαμε ότι στο Ε-βήμα εκτιμούμε τις posterior πιθανότητες τ_{ij}, τις πιθανότητες δηλαδή η i παρατήρηση να ανήκει στην j ομάδα έχοντας παρατηρήσει δεδομένα. Με βάση τη (2.2) βλέπουμε ότι αυτές οι τ_{ij}, εξαρτώνται από τις πιθανότητες π_j, την

πιθανότητα δηλαδή μία παρατήρηση να ανήκει στην j ομάδα πριν παρατηρήσουμε δεδομένα, και τις κατανομές των διαφόρων ομάδων. Δίνουμε λοιπόν αρχικές τιμές στις πιθανότητες π_j καθώς επίσης και στις μέσες τιμές μ_j και στους πίνακες διασποράς-συνδιασποράς των ομάδων. Αυτό δηλαδή που γίνεται επί της ουσίας είναι μία διαμέριση των παρατηρήσεών μας στις επιμέρους ομάδες.

Εδώ όμως προκύπτει το ερώτημα με ποιον τρόπο μπορεί να γίνει αυτή η αρχική διαμέριση. Ένα παράδειγμα αρχικής διαμέρισης των παρατηρήσεων για την περίπτωση g=2 ομάδων με ίδιο πίνακα διασποράς-συνδιασποράς, θα μπορούσε να γίνει παραστώντας γραφικά τα δεδομένα για ανά δύο μεταβλητές από τις p και τραβώντας μια γραμμή που να χωρίζει τα διμεταβλητά δεδομένα σε δύο ομάδες έτσι ώστε οι δύο ομάδες να είναι όσο γίνεται πιο κοντά στην κανονικότητα. Ωστόσο αυτό είναι δύσκολο όταν έχουμε δεδομένα πολλών διαστάσεων. Σε αυτή την περίπτωση, αρχική διαμέριση θα μπορούσε να γίνει με τη χρήση κάποιας μεθόδου αποστάσεων, όπως ο K-means αλγόριθμος ή κάποια μέθοδο ιεραρχικής ανάλυσης αν το πλήθος των δεδομένων δεν είναι πολύ μεγάλο (γιατί τότε οι ιεραρχικοί αλγόριθμοι είναι αργοί), (McLachlan and Peel 2000a).

Άλλος τρόπος για να πάρουμε μια αρχική διαμέριση των παρατηρήσεών μας είναι να κατατάξουμε τις παρατηρήσεις στις ομάδες τυχαία. Ωστόσο, με αυτό τον τρόπο όταν οι παρατηρήσεις μας είναι πολλές οι αρχικές τιμές των π_j που θα πάρουμε θα είναι παρόμοιες. Ένας τρόπος για να μειώσουμε αυτήν την επίδραση των ίσων π_j είναι να πάρουμε ένα μικρό τυχαίο υποδείγμα από τις παρατηρήσεις μας και να διαμερίσουμε τυχαία τις παρατηρήσεις του στις g ομάδες. Τότε το πρώτο E & M βήμα θα εκτελεστεί με βάση αυτό το υποδείγμα. Το υποδείγμα όμως θα πρέπει να είναι αρκετά μεγάλο ώστε να μπορούν να εκχωρηθούν τουλάχιστο p+1 παρατηρήσεις σε κάθε ομάδα για να μην έχουμε υπολογιστικά προβλήματα (McLachlan and Peel 2000a, section 2.12).

Ένας εναλλακτικός τρόπος τυχαίας αρχικοποίησης είναι να πάρουμε αρχικές τιμές για τους μέσους των ομάδων μέσω προσομοίωσης. Συγκεκριμένα παίρνουμε τα $\mu_j^{(0)}$ έτσι ώστε $\mu_1^{(0)},...,\mu_g^{(0)} \overset{i.i.d.}{\sim} N(\bar{y},S)$ όπου $\bar{y}$ είναι ο δειγματικός μέσος και S είναι ο δειγματικός πίνακας διασποράς-συνδιασποράς όλων των παρατηρημένων δεδομένων. Με αυτό τον τρόπο υπάρχει μεγαλύτερη μεταβλητότητα μεταξύ των αρχικών τιμών των $\mu_j^{(0)}$ απ' ότι να διαμερίζαμε τυχαία τις παρατηρήσεις σε ομάδες και είναι και

υπολογιστικά λιγότερο απαιτητικό. Οι πίνακες διασποράς-συνδιαποράς και οι πιθανότητες π_j μπορούν να οριστούν ως εξής: $\Sigma_j^{(0)} = S \ \ και \ \ \pi_j^{(0)} = 1/g \quad (j = 1,...,g)$ Ng S.K. et al. (2004). Σημαντικό σημείο για το καλό fit στα δεδομένα μοντέλων μίξεων κατανομών είναι η ακριβής εκτίμηση των πιθανοτήτων π_j. Ένας τρόπος εκτίμησης των π_j για την περίπτωση μονοδιάστατων κανονικών κατανομών προτείνεται από τον Fowlkes (1979).

2.6 Gaussian Parsimonious Clustering Models (GPCM)

Οι πίνακες διασποράς-συνδιασποράς των επιμέρους ομάδων είναι αυτοί που καθορίζουν τα γεωμετρικά χαρακτηριστικά κάθε ομάδας, όπως το σχήμα το μέγεθος και τον προσανατολισμό της. Στην παράγραφο 2.4 είδαμε την περίπτωση όπου οι πίνακες διασποράς-συνδιασποράς των επιμέρους ομάδων ήταν διαφορετικοί μεταξύ τους (ετεροσκεδαστικότητα). Ωστόσο, είπαμε ότι μπορεί να τεθούν διάφοροι περιορισμοί στους πίνακες Σ_j, και είδαμε την περίπτωση όπου $\Sigma_j=\Sigma$ καθώς επίσης και $\Sigma_j = \sigma^2 \mathrm{I}_p$ για την περίπτωση των πολυμεταβλητών κανονικών κατανομών. Για την ακρίβεια υπάρχει η δυνατότητα να θέσουμε διάφορους περιορισμούς στα στοιχεία των πινάκων Σ_j, και άρα να πάρουμε μια οικογένεια διαφορετικών μοντέλων για τον ίδιο αριθμό ομάδων. Μια οικογένεια τέτοιων μοντέλων υπάρχει στο πακέτο mclust της R (Fraley and Raftery 2006, September) και μια περιγραφή τους παρέχεται από τους Fraley and Raftery (2007) καθώς επίσης και από τον McNicholas (2011). Ένα επιπλέον πλεονέκτημα που έχουμε θέτοντας περιορισμούς, είναι η μείωση των παραμέτρων προς εκτίμηση. Αν δε θέσουμε κανένα περιορισμό στους πίνακες Σ_j, τότε ο αριθμός των παραμέτρων προς εκτίμηση για κάθε πίνακα Σ_j είναι $\frac{1}{2}p(p+1)$. Δηλαδή έχουμε να εκτιμήσουμε αρκετές παραμέτρους για κάθε ομάδα. Επιβάλλοντας περιορισμούς μειώνουμε τον αριθμό παραμέτρων που πρέπει να εκτιμήσουμε, αλλά όμως ελαττώνουμε την ευελιξία του μοντέλου μας.

Ένα γενικό πλαίσιο για να θέτουμε διάφορους περιορισμούς στους πίνακες Σ_j παρέχεται μέσω της φασματικής ανάλυσης ενός πίνακα και είναι το εξής:

$$\Sigma_j = \lambda_j D_j A_j D_j^T \tag{2.11}$$

(Banfield and Raftery 1993), όπου D_j είναι ο ορθογώνιος πίνακας των ιδιοδιανυσμάτων, A_j είναι ο διαγώνιος πίνακας του οποίου τα στοιχεία είναι ανάλογα προς τις ιδιοτιμές και λ_j είναι μία σταθερά κανονικοποίησης. Τους παράγοντες D_j, A_j, λ_j τους χειριζόμαστε σα να είναι ανεξάρτητοι και μπορεί να είναι είτε ίδιοι μεταξύ των ομάδων είτε διαφορετικοί. Κάθε ένας από αυτούς τους παράγοντες προσδίδει και ένα διαφορετικό γεωμετρικό χαρακτηριστικό σε κάθε ομάδα και άρα όταν θέτουμε περιορισμούς, οι ομάδες μοιράζονται ορισμένες κοινές γεωμετρικές ιδιότητες. Για παράδειγμα ο πίνακας D_j καθορίζει τον προσανατολισμό της ομάδας, ο A_j το σχήμα και τα λ_j το μέγεθος τα οποία είναι ανάλογα με $\lambda_j^d \det(A_j)$.

Τα διαφορετικά μοντέλα που προκύπτουν για διάφορους περιορισμούς των D_j, A_j, λ_j και είναι διαθέσιμα στο πακέτο mclust της R συνοψίζονται στον παρακάτω πίνακα. Αυτή η οικογένεια μοντέλων θα αναφέρεται από εδώ και στο εξής ως MCLUST (Fraley and Raftery 2007). Η πρώτη στήλη μας υποδεικνύει το όνομα του μοντέλου. Το γράμμα E (equal) συμβολίζει την ισότητα μεταξύ των ομάδων, το V (varying) τη διαφορετικότητα και το I (identity) το μοναδιαίο πίνακα. Για παράδειγμα το όνομα EVI του μοντέλου, μας υποδεικνύει ότι πρόκειται για ένα μοντέλο στο οποίο το μέγεθος των ομάδων (λ_j) είναι το ίδιο (E), το σχήμα των ομάδων (A_j) μπορεί να διαφέρει (V) και ο προσανατολισμός (D_j) είναι ο ίδιος και καθορίζεται από το μοναδιαίο πίνακα (I). Δηλαδή το πρώτο γράμμα αναφέρεται στα λ_j, το δεύτερο στους πίνακες A_j και το τρίτο στους πίνακες D_j.

Πίνακας 2.1. Μοντέλα της MCLUST οικογένειας.

Model	**Volume** λ_j	**Shape** A_j	**Orientation** D_j	Σ_j	**Free Covariance Parameters**	**Distribution**
EII	Equal	Spherical	-	$\lambda \mathrm{I}$	1	Spherical
VII	Variable	Spherical	-	$\lambda_j \mathrm{I}$	g	Spherical
EEI	Equal	Equal	Axis-Alligned	λA	p	Diagonal
VEI	Variable	Equal	Axis-Alligned	$\lambda_j A$	$\mathrm{p}+g-1$	Diagonal
EVI	Equal	Variable	Axis-Alligned	λA_j	$\mathrm{pg}-g+1$	Diagonal
VVI	Variable	Variable	Axis-Alligned	$\lambda_j A_j$	pg	Diagonal
EEE	Equal	Equal	Equal	λDAD^T	$\mathrm{p}(\mathrm{p}+1)/2$	Ellipsoidal
EEV	Equal	Equal	Variable	$\lambda D_j A D_j^T$	$\mathrm{gp}(\mathrm{p}+1)/2-(g-1)\mathrm{p}$	Ellipsoidal
VEV	Variable	Equal	Variable	$\lambda_j D_j A D_j^T$	$\mathrm{gp}(\mathrm{p}+1)/2-(g-1)(\mathrm{p}-1)$	Ellipsoidal
VVV	Variable	Variable	Variable	$\lambda_j D_j A_j D_j^T$	$\mathrm{gp}(\mathrm{p}+1)/2$	Ellipsoidal

Η στήλη Covariance parameters περιέχει το συνολικό αριθμό παραμέτρων που πρέπει να εκτιμήσουμε για τους πίνακες διασποράς-συνδιασποράς, όπου g ο αριθμός των ομάδων και p ο αριθμός των μεταβλητών. Οι πίνακες διασποράς-συνδιασποράς στο EVI μοντέλο έχουν τη μορφή $\Sigma_j = \lambda \mathrm{I}_p \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_p \end{bmatrix} \mathrm{I}_p$. Δηλαδή για κάθε ομάδα έχουμε να εκτιμήσουμε p-1 παραμέτρους για τον πίνακα A_j. Επειδή όμως έχουμε g ομάδες, συνολικά για όλους τους A_j πίνακες πρέπει να εκτιμήσουμε (p-1)g παραμέτρους. Αν βάλουμε και άλλη μία παράμετρο, την λ, που πρέπει να βρούμε, συνολικά θα πρέπει να εκτιμήσουμε $(p-1)g+1 = pg-g+1$ παραμέτρους, όπως άλλωστε υποδεικνύεται και στην τελευταία στήλη του πίνακα 1.

Άλλο παράδειγμα μοντέλου είναι το VEI. Σε αυτό το μοντέλο τα λ_j διαφέρουν μεταξύ των ομάδων, οι πίνακες A_j ισούνται μεταξύ τους και οι πίνακες D_j είναι μοναδιαίοι. Δηλαδή οι πίνακες Σ_j σε αυτό το μοντέλο είναι της μορφής $\Sigma_j = \lambda_j \mathrm{I}_p \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_p \end{bmatrix} \mathrm{I}_p$, όπου ο πίνακας $\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_p \end{bmatrix}$ είναι ίδιος για όλες τις ομάδες. Εδώ, πρέπει να εκτιμήσουμε τις g παραμέτρους λ_j και άλλες p-1 παραμέτρους για τον πίνακα A_j=A. Δηλαδή συνολικά πρέπει να εκτιμήσουμε g+p-1 παραμέτρους για τους πίνακες διασποράς-συνδιασποράς των ομάδων. Σε κάθε περίπτωση για να βρούμε το συνολικό αριθμό παραμέτρων προς εκτίμηση αρκεί στις παραμέτρους της τελευταίας στήλης να προσθέσουμε gp παραμέτρους για την εκτίμηση των μέσων μ_j και g-1 παραμέτρους για τις πιθανότητες π_j.

Με παρόμοιο τρόπο αναλύονται και τα υπόλοιπα μοντέλα. Το πιο φειδωλό από αυτά είναι το πρώτο (EII) το οποίο υποθέτει ίσους πίνακες Σ_j για όλες τις ομάδες, διαγώνιους με ίσα διαγώνια στοιχεία. Το EII μοντέλο είναι το ίδιο με την περίπτωση $\Sigma_j = \Sigma = \sigma^2 \mathrm{I}_p$ που αναφέρθηκε στην 2.4. Αντίθετα, το πιο "πλούσιο" μοντέλο είναι το VVV όπου όλοι οι πίνακες και τα στοιχεία τους επιτρέπεται να είναι διαφορετικά. Αυτό το μοντέλο έχει το πλεονέκτημα ότι είναι το πιο γενικό μοντέλο και άρα το πιο ευέλικτο και κατά συνέπεια μπορεί να εφαρμοστεί σε κάθε περίπτωση, αλλά έχει το

μειονέκτημα ότι περιέχει το μέγιστο αριθμό παραμέτρων που πρέπει να εκτιμηθούν και άρα απαιτεί περισσότερα δεδομένα σε κάθε ομάδα.

Αντιλαμβάνεται κανείς ότι μέσω της (2.11) δημιουργείται μια οικογένεια μοντέλων, όπου επιτρέποντας σε μερικές και όχι σε όλες τις παραμέτρους να διαφοροποιούνται, έχουμε μεγαλύτερη δυνατότητα στο να μοντελοποιήσουμε καλύτερα δεδομένα που έχουν ποικίλα χαρακτηριστικά. Επίσης, έχουμε μοντέλα περισσότερο ή λιγότερο φειδωλά ως προς τον αριθμό παραμέτρων προς εκτίμηση.

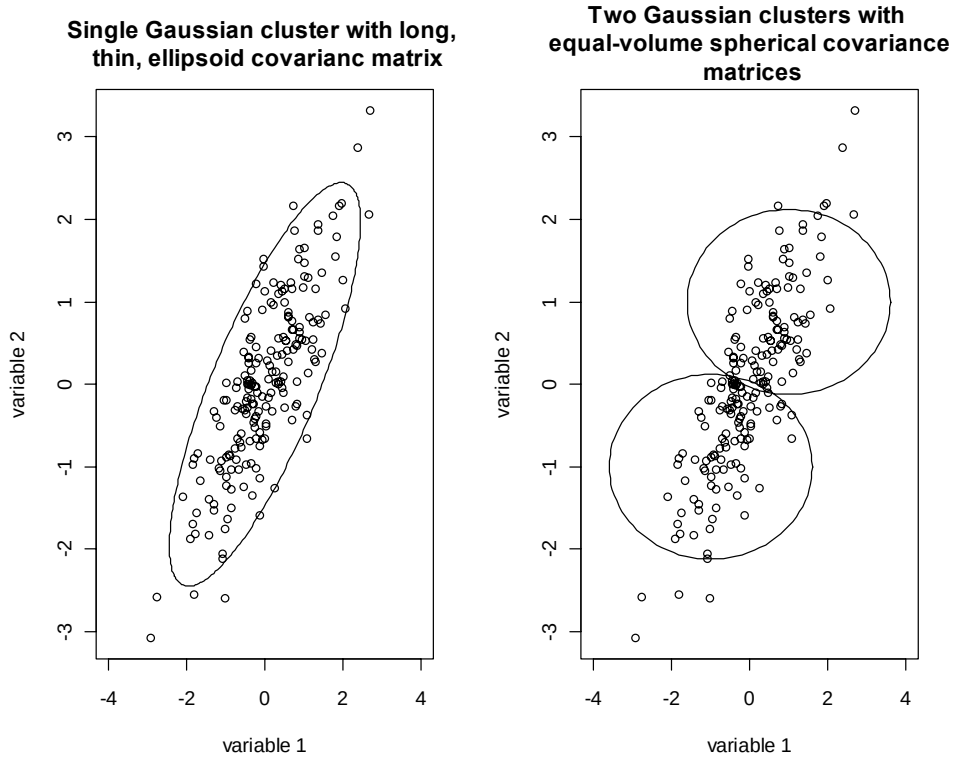
2.7 Επιλογή μοντέλου (Model selection)

Από όσα έχουμε δει μέχρι στιγμής ο ΕΜ αλγόριθμος ταξινομεί τις παρατηρήσεις μας σε ομάδες, έτσι ώστε οι περισσότερο όμοιες παρατηρήσεις να ανήκουν σε ίδιες ομάδες. Ωστόσο, ο ίδιος ο αλγόριθμος δεν έχει τη δυνατότητα να εκτιμήσει ποιος είναι ο κατάλληλος αριθμός ομάδων στις οποίες πρέπει να κατατάξουμε τις παρατηρήσεις. Εμείς είμαστε αυτοί που του ορίζουμε εξ' αρχής σε πόσες ομάδες να κατατάξει τις παρατηρήσεις και στη συνέχεια ο ΕΜ εκτιμά τις παραμέτρους και κάνει την ομαδοποίηση των παρατηρήσεων. Άλλη παράμετρος που πρέπει να καθορίσουμε εκ των προτέρων, πριν τρέξουμε τον αλγόριθμό μας, είναι να καθορίσουμε ποιον περιορισμό να θέσουμε στους πίνακες Σ_j, να αποφασίσουμε δηλαδή ποιο μοντέλο (σαν αυτά που περιγράφηκαν στην 2.6) είναι κατάλληλο να χρησιμοποιήσουμε. Κάθε συνδυασμός μοντέλου του πίνακα 2.1 και αριθμού ομάδων, συνιστά ένα ξεχωριστό μοντέλο. Άρα, το πρόβλημα επιλογής του κατάλληλου περιορισμού των πινάκων Σ_j και του αριθμού ομάδων συνιστά ένα πρόβλημα επιλογής μοντέλου. Αυτό είναι σημαντικό γιατί υπάρχει μια σχέση (tradeoff) μεταξύ μοντέλου και του αριθμού ομάδων. Για να γίνει πιο κατανοητό, αν κάποιος χρησιμοποιήσει ένα περίπλοκο μοντέλο (π.χ. VVV) ίσως ένας μικρός αριθμός ομάδων να είναι αρκετός ώστε να επιτευχθεί ικανοποιητικό fit στα δεδομένα, αλλά αν χρησιμοποιήσει ένα φειδωλό μοντέλο (π.χ. ΕΙΙ) τότε ίσως να χρειάζεται μεγαλύτερος αριθμός ομάδων.

Για παράδειγμα, ας θεωρήσουμε τα δεδομένα του παρακάτω σχήματος. Στην πρώτη περίπτωση χρησιμοποιούμε μόνο μία ομάδα, της οποίας ο πίνακας διακύμανσης-συνδιακύμανσης αντιστοιχεί σε μια έλλειψη με μεγάλο τον ένα της

άξονα. Ας φανταστούμε ότι η περίπτωση αυτή αντιστοιχεί σε ένα "πλούσιο" μοντέλο με μικρό αριθμό ομάδων. Αν όμως αντί για έλλειψη περιορίζαμε τον πίνακα διακύμανσης-συνδιακύμανσης να είναι σφαιρικός, τότε θα χρειαζόμασταν 2 ομάδες για να πετύχουμε καλύτερο fit στα δεδομένα. Η περίπτωση αυτή αντιστοιχεί σε πιο φειδωλό μοντέλο αλλά με περισσότερες ομάδες.

Γράφημα 2.1. Διαφορετικοί τρόποι ομαδοποίησης των ίδιων δεδομένων σε 1 και 2 ομάδες.



Η επιλογή του κατάλληλου μοντέλου και του αριθμού ομάδων μπορεί να γίνει με βάση το Bayesian Information Criterion (BIC). Το BIC ισούται με $BIC = 2l\left(y, \hat{\theta}\right) - m\log(n)$, όπου m είναι ο αριθμός των παραμέτρων προς εκτίμηση, n ο αριθμός των παρατηρήσεων, και $l\left(y, \hat{\theta}\right)$ η εκτίμηση μέγιστης πιθανοφάνειας. Ο δεύτερος όρος θέτει μια ποινή στην πιθανοφάνεια ανάλογα με τις παραμέτρους του μοντέλου προς εκτίμηση, διότι η πιθανοφάνεια μπορεί να αυξηθεί απλά προσθέτοντας περισσότερες παραμέτρους προς εκτίμηση. Όσες περισσότερες παραμέτρους έχουμε να εκτιμήσουμε τόσο μεγαλύτερη η ποινή. Μάλιστα η ποινή που θέτει το BIC είναι μεγαλύτερη από την αντίστοιχη που χρησιμοποιεί το AIC κριτήριο $\left(AIC = 2l\left(y, \hat{\theta}\right) - 2m\right)$ καθώς για $n \geq 8$ ισχύει $m\log(n) > 2m$. Όσο μεγαλύτερη είναι

η τιμή BIC τόσο καλύτερο είναι το μοντέλο. Κατά συνέπεια, μπορούμε να χρησιμοποιήσουμε αυτό το κριτήριο για να συγκρίνουμε μεταξύ τους μοντέλα με διαφορετικούς περιορισμούς στις παραμέτρους των πινάκων διασποράς-συνδιασποράς αλλά και με διαφορετικό αριθμό ομάδων. Το BIC παρέχει επιπλέον το πλεονέκτημα της σύγκρισης δύο μη nested μοντέλων. Συνήθως διαφορές στην τιμή του BIC μεγαλύτερες των 10 μονάδων συνιστούν ισχυρή ένδειξη υπέρ του ενός μοντέλου.

Όμως, τα μοντέλα μίξεων κατανομών δεν ικανοποιούν πλήρως τις συνθήκες κανονικότητας και γενικά τις υποθέσεις πάνω στις οποίες βασίζεται η εξαγωγή του BIC. Ωστόσο, διάφορα αποτελέσματα μελετών υποδεικνύουν την καταλληλότητα και την καλή επίδοση του BIC στο πλαίσιο του model-based clustering (Fraley and Raftery 1998, 2002).

Φυσικά υπάρχουν και άλλα κριτήρια που έχουν προταθεί για την επιλογή του καλύτερου μοντέλου. Ένα από αυτά είναι το Normalized Entropy Criterion (NEC) (Biernacki and Govaert 1999), το οποίο είναι ένα κριτήριο ταξινόμησης και μας δείχνει πόσο καλή είναι η ομαδοποίηση. Το κριτήριο αυτό δίνεται από τον τύπο $NEC_g = \frac{I_g}{l_g - l_1}, \quad για\ g > 1$, όπου $I_g = -\sum_{j=1}^{g}\sum_{i=1}^{n} \tau_{ij} \log \tau_{ij}$, και όπου g είναι ο αριθμός των ομάδων. Επίσης χρησιμοποιούμε την συνθήκη $0 \log 0 = 0$ και τα τ_{ij} όπως έχουμε δει είναι διαθέσιμα μετά το τέλος του EM αλγορίθμου και μπορούμε να τα χρησιμοποιήσουμε στο NEC_g. Ωστόσο, η ποσότητα I_g παίρνει την τιμή 0 στην περίπτωση τέλειας ομαδοποίησης. Αυτή αντιστοιχεί στην περίπτωση που κάθε παρατήρηση ανήκει με πιθανότητα 1 σε μια ομάδα και άρα τα τ_{ij} για κάθε παρατήρηση αποτελούνται από μία μονάδα και g-1 μηδενικά. Στην περίπτωση που η ομαδοποίηση είναι η χειρότερη δυνατή, θα ισχύει ότι $\tau_{ij} = 1/g \quad i = 1,...,n, \quad j = 1,...,g$, και επομένως $I_g = -n \log\left(\frac{1}{g}\right)$. Γι' αυτό πολλές φορές χρησιμοποιούμε ως κριτήριο την ποσότητα $J(g,n) = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n} \tau_{ij} \ln \tau_{ij}}{n \log(1/g)}$. Τιμές της $J(g,n)$ κοντά στο 1 υποδεικνύουν πολύ καλή ομαδοποίηση.

Άλλα παρόμοια κριτήρια με το BIC και το NEC, είναι τα AIC, AIC3, AWE (για το οποίο όμως έχει βρεθεί ότι δεν είναι τόσο καλό όσο το BIC), ICOM, ICL

(Integrated Completed Likelihood) και άλλα. Οι επιδόσεις μερικών από αυτών στο model based clustering συγκρίνονται στην εργασία των Biernacki and Govaert (1999). Το ICL μάλιστα εμφανίζεται να είναι πιο ανθεκτικό από το BIC σε τυχόν αποκλίσεις των υποθέσεων που αφορούν τα μείγματα κατανομών και δοκιμές σε προσομοιωμένα και αληθινά δεδομένα έχουν δείξει ότι δίνει συγκρίσιμα αποτελέσματα με το BIC (Biernacki et al. 2000). Το ICL (ή καλύτερα η προσέγγιση του ICL, *approximate ICL*) δίνεται από τον τύπο $ICL \approx BIC + \sum_{i=1}^{n}\sum_{j=1}^{g} MAP\{\tau_{ij}\}\log\tau_{ij}$, όπου $MAP\{\tau_{ij}\}$ είναι η ταξινόμηση που προκύπτει μετά το τέλος του αλγορίθμου ομαδοποίησης βάσει των εκτιμώμενων εκ των υστέρων πιθανοτήτων τ_{ij} και δίνεται από $MAP\{\tau_{ij}\} = \begin{cases} 1, & \alpha\nu\ to\ \max_j\{\tau_{ij}\}\ \pi\rho o\kappa\acute{\upsilon}\pi\tau\epsilon\iota\ \sigma\tau\eta\nu\ j\ o\mu\acute{\alpha}\delta\alpha \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$. Το διπλό άθροισμα αναφέρεται στη βιβλιογραφία ως *estimated mean entropy* και είναι ένα μέτρο που μας δείχνει την αβεβαιότητα για την ταξινόμηση της i παρατήρησης στην j ομάδα. Κατά συνέπεια το ICL θα μας υποδεικνύει λιγότερο συχνά μεγάλο αριθμό ομάδων σε σχέση με το BIC, καθώς είναι πιο πιθανό για μια παρατήρηση να ανήκει σε μια μεγαλύτερη ομάδα απ' ότι σε μια μικρότερη (δεδομένου του ότι όταν έχουμε πολλές ομάδες αυτές θα είναι και μικρότερου μεγέθους για σταθερό αριθμό δεδομένων).

Ανάμεσα σε όλα αυτά τα κριτήρια, αυτό που συνήθως χρησιμοποιείται στην πράξη είναι το BIC, εξαιτίας της υπολογιστικής ευκολίας του και της καλής επίδοσής του. Σε περιπτώσεις που τα αποτελέσματα βάση αυτών των κριτηρίων δεν είναι ξεκάθαρα, η καταλληλότητα της ομαδοποίησης μπορεί να ποσοτικοποιηθεί και να αξιολογηθεί χρησιμοποιώντας τον Rand Index που αναφέρεται παρακάτω αρκεί να υπάρχει κάποια εκ των προτέρων γνώση της ομαδοποίησης.

2.8 Rand Index & Adjusted Rand Index

Το αποτέλεσμα του EM αλγορίθμου στο model-based clustering είναι μια διαμέριση των παρατηρήσεών μας σε διάφορες ομάδες. Έτσι, το να συγκρίνουμε την ομαδοποίηση που προκύπτει από τον EM με κάποια γνωστή εξωτερική ομαδοποίηση

είναι σα να συγκρίνουμε δύο διαφορετικές ομαδοποιήσεις μεταξύ τους. Ένα μέτρο συμφωνίας μεταξύ δύο ομαδοποιήσεων είναι ο adjusted Rand Index. Ο δείκτης αυτός μπορεί να χρησιμοποιηθεί ακόμα και στην περίπτωση διαφορετικού αριθμού ομάδων.

Ο Rand Index είναι ο λόγος των συμφωνιών (agreements) προς όλα τα δυνατά ζεύγη παρατηρήσεων μεταξύ των δύο διαμερίσεων. Δηλαδή είναι ο αριθμός των ζευγών παρατηρήσεων που ανήκουν είτε στις ίδιες ομάδες και στις δύο διαμερίσεις, είτε σε διαφορετικές ομάδες και στις δύο διαμερίσεις, προς τον ολικό αριθμό ζευγών παρατηρήσεων. Για να γίνει πιο κατανοητό ας θεωρήσουμε δύο διαφορετικές διαμερίσεις U και V όπως παρακάτω.

Πίνακας 2.2. Δύο τυχαίες διαμερίσεις U και V.

		Method V				
		Cluster 1	Cluster 2	…	Cluster c_2	
Method U	Cluster 1	n_{11}	n_{12}	…	n_{1c_2}	$n_{1\cdot}$
	Cluster 2	n_{21}	n_{22}		n_{2c_2}	$n_{2\cdot}$
	…	⋮	⋮	⋱	⋮	⋮
	Cluster c_1	$n_{c_1 1}$	$n_{c_1 2}$	…	$n_{c_1 c_2}$	$n_{c_1\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c_2}$	$n_{\cdot\cdot}$

Έστω α ο αριθμός ζευγών παρατηρήσεων που ανήκουν στην ίδια ομάδα στη U διαμέριση και στην ίδια στη V διαμέριση, b ο αριθμός των ζευγών παρατηρήσεων που ανήκουν στην ίδια ομάδα στη U αλλά όχι στην ίδια ομάδα στη V, c ο αριθμός των ζευγών που ανήκουν στην ίδια ομάδα στη V αλλά όχι στην ίδια ομάδα στη U και d ο αριθμός των ζευγών παρατηρήσεων που ανήκουν σε διαφορετικές ομάδες στη U και σε διαφορετικές στη V. Τότε οι ποσότητες α και d μπορούν να ερμηνευθούν ως "συμφωνίες" (agreements ή concordances) και οι b, c ως "διαφωνίες" (disagreements ή discordances). Έτσι, ο Rand Index ισούται με

$$\text{Rand Index} = \frac{\alpha + d}{\alpha + b + c + d} = \frac{\binom{n}{2} - b - c}{\binom{n}{2}},$$ με τα b και c να ισούνται με

$b = \sum_{i=1}^{c_1} \binom{n_{i.}}{2} - \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \binom{n_{ij}}{2}$ και $c = \sum_{j=1}^{c_2} \binom{n_{.j}}{2} - \sum_{i=1}^{c_1}\sum_{j=1}^{c_2} \binom{n_{ij}}{2}$. Διαφορετικά ο Rand Index μπορεί να γραφεί ως $Rand\ \ Index = \dfrac{\binom{n}{2} + 2\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{c_1}\binom{n_{i.}}{2} + \sum_{j=1}^{c_2}\binom{n_{.j}}{2}\right]}{\binom{n}{2}}$.

Ο Rand Index κυμαίνεται από 0 έως 1. Όσο μεγαλύτερη είναι η τιμή του δείκτη τόσο μεγαλύτερη είναι η συμφωνία των δύο ομαδοποιήσεων. Τιμή 1 αντιστοιχεί σε πλήρη συμφωνία των δύο διαμερίσεων. Το πρόβλημα όμως με αυτό το δείκτη είναι ότι δε λαμβάνει υπ' όψιν του την τυχαιότητα. Δηλαδή ο δείκτης δεν παίρνει μια σταθερή τιμή (ας πούμε 0) για δύο τυχαίες διαμερίσεις. Σε αυτή την περίπτωση των τυχαίων διαμερίσεων ο Rand Index δεν είναι απαραίτητα μηδέν όπως θα αναμενόταν, γιατί περιμένουμε μόνο και μόνο λόγω τύχης να έχουμε κάποιες "συμφωνίες" ή/και "διαφωνίες". Ο adjusted Rand Index διορθώνει για αυτό ακριβώς το πρόβλημα (Hubert and Arabie 1985, Yeung and Ruzzo 2001). Έτσι, η αναμενόμενη τιμή του adjusted Rand Index για την περίπτωση δύο τυχαίων διαμερίσεων είναι μηδέν. Ο adjusted Rand Index δίνεται από τον τύπο $Adjusted\ \ Rand\ \ Index = \dfrac{\binom{n}{2}\sum_{i=1}^{c_1}\sum_{j=1}^{c_2}\binom{n_{ij}}{2} - \sum_{i=1}^{c_1}\binom{n_{i.}}{2}\sum_{j=1}^{c_2}\binom{n_{.j}}{2}}{\frac{1}{2}\binom{n}{2}\left[\sum_{i=1}^{c_1}\binom{n_{i.}}{2} + \sum_{j=1}^{c_2}\binom{n_{.j}}{2}\right] - \sum_{i=1}^{c_1}\binom{n_{i.}}{2}\sum_{j=1}^{c_2}\binom{n_{.j}}{2}}$ και προέρχεται από τον τύπο $\text{Adjusted Rand Index} = \dfrac{\text{Rand Index - Expected Rand Index}}{\text{Max Rand Index - Expected Rand Index}}$. Ο δείκτης αυτός παίρνει τιμές από -1 έως 1, με τις αρνητικές τιμές να υποδεικνύουν πολύ κακή συμφωνία και χρησιμοποιείται ευρέως για να συγκρίνουμε ομαδοποιήσεις μεταξύ τους.

2.9 Παραδείγματα

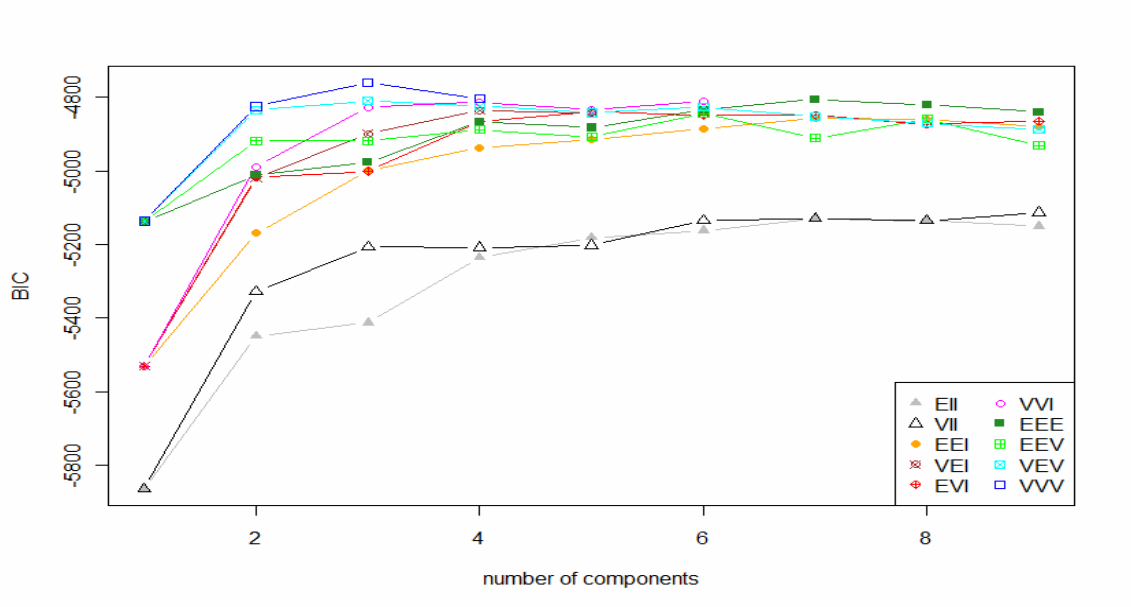
2.9.1 Παράδειγμα 1

Θα εφαρμόσουμε την τεχνική του model-based clustering μέσω του EM αλγορίθμου για την οικογένεια μοντέλων MCLUST, για τα δεδομένα *diabetes*, τα οποία υπάρχουν στο πακέτο mclust της R. Τα δεδομένα αυτά αφορούν 145 ασθενείς οι οποίοι εξετάστηκαν κλινικά και χωρίστηκαν σε 3 ομάδες: normal, chemically diabetic και overtly diabetic. Για τους ασθενείς αυτούς έχουμε δεδομένα για τις ακόλουθες 3 μεταβλητές:

Variable	**Description**
glucose	plasma glucose response to oral glucose
insulin	plasma insulin response to oral glucose
sspg	steady-state plasma glucose (measures insulin resistance)

Θέλουμε να ομαδοποιήσουμε αυτούς τους ασθενείς. Για να πετύχουμε την καλύτερη ομαδοποίηση εφαρμόζουμε τον EM αλγόριθμο και για τα 10 μοντέλα της οικογένειας MCLUST για διάφορους αριθμούς ομάδων. Η σύγκριση των μοντέλων γίνεται με βάση το κριτήριο BIC που περιγράφηκε προηγουμένως. Τα αποτελέσματα φαίνονται στο ακόλουθο γράφημα.

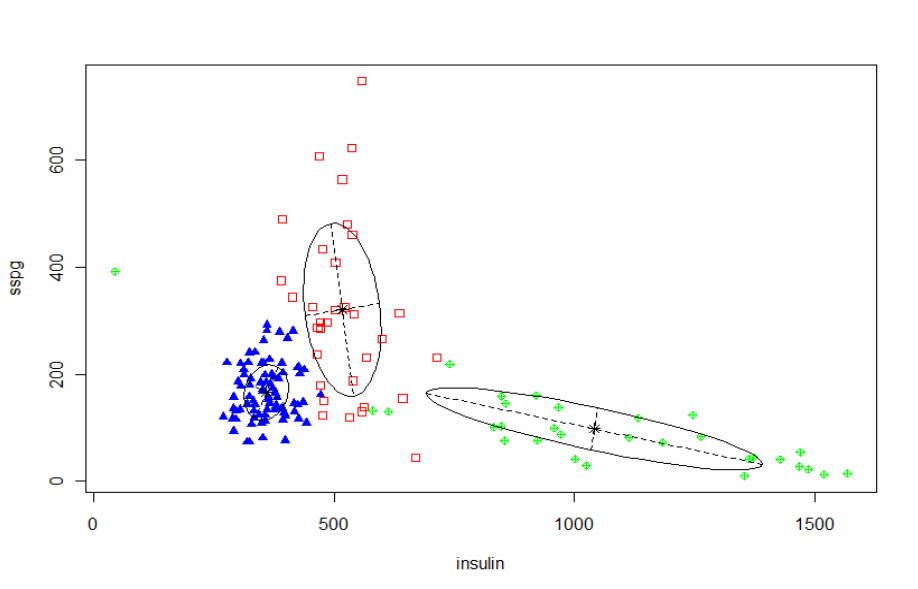
Γράφημα 2.2: BIC κριτήριο για τα 10 μοντέλα της MCLUST μέχρι 9 ομάδες για τα δεδομένα diabetes της βιβλιοθήκης mclust της R.



Από το γράφημα 1 γίνεται αντιληπτό ότι το καλύτερο μοντέλο με βάση το BIC κριτήριο είναι το VVV για την περίπτωση των 3 ομάδων, καθώς εκεί η τιμή BIC είναι η μέγιστη. Και οι 3 ομάδες που δημιουργούνται εδώ έχουν διαφορετικό πίνακα διακύμανσης. Σημειώνεται ότι οι 3 διαφορετικές ομάδες που προτείνονται εδώ από το BIC κριτήριο ως καλύτερη ομαδοποίηση, συμπίπτουν με τον αριθμό ομάδων στις οποίες είχαν χωριστεί οι ασθενείς κλινικά.

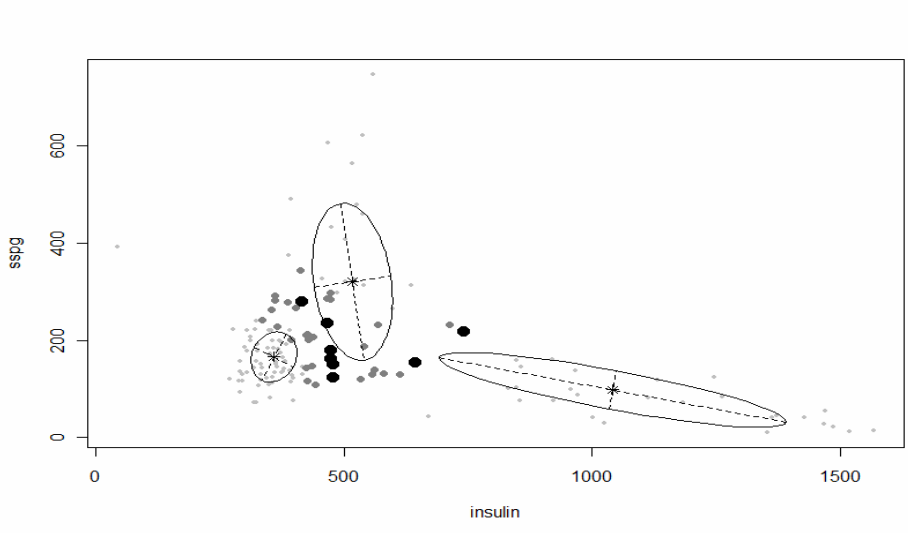
Στο παρακάτω γράφημα φαίνεται μια προβολή των δεδομένων πάνω στις μεταβλητές insulin και sspg και η ομαδοποίησή τους με βάση το καλύτερο VVV μοντέλο. Διαφορετικά σύμβολα υποδεικνύουν παρατηρήσεις διαφορετικών ομάδων. Με αστεράκι συμβολίζονται οι μέσοι των ομάδων και οι ελλείψεις αντιστοιχούν στους πίνακες διασποράς-συνδιασποράς.

Γράφημα 2.3: Προβολή των diabetes δεδομένων πάνω στις insulin & sspg μεταβλητές και ομαδοποίησής τους με βάση το VVV μοντέλο.



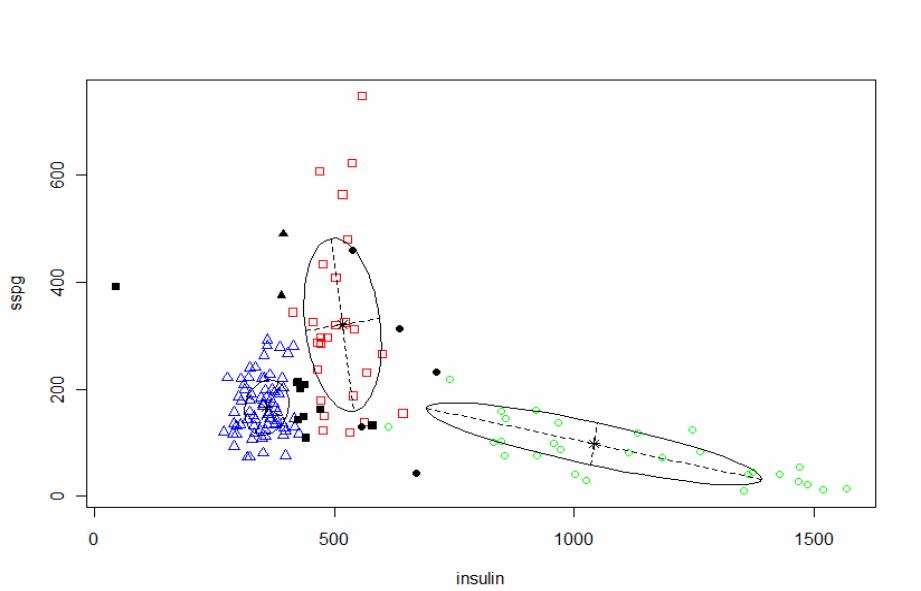
Στο ακόλουθο γράφημα παριστάνονται σημεία με μεγάλη αβεβαιότητα ως προς την ομαδοποίησή τους (έντονο μαύρο χρώμα). Η αβεβαιότητα ταξινόμησης μιας παρατήρησης δίνεται ως $1-\max_j \tau_{ij}^*$, όπου τ_{ij}^* είναι η τιμή της πιθανότητας τ_{ij} στο σημείο μεγίστου της $f\left(y \mid \Psi\right)=\sum_{j=1}^{g}\pi_j f_j\left(y \mid \theta_j\right)$ (Fraley and Raftery 2002).

Γράφημα 2.4: Ομαδοποίηση των diabetes δεδομένων με βάση το VVV μοντέλο και σημεία με μεγάλη αβεβαιότητα ως προς την ομαδοποίηση.



Επειδή όμως για την περίπτωση των 3 ομάδων, που προέκυψε εδώ ως ο κατάλληλος αριθμός ομάδων για ομαδοποίηση, υπάρχει ήδη μια γνωστή ομαδοποίηση, η κλινική: normal, chemically diabetic και overtly diabetic, μπορούμε να βρούμε ποια και πόσα σημεία έχουν καταταχτεί εσφαλμένα με βάση αυτή την αρχική κλινική ομαδοποίηση. Έτσι, το ακόλουθο γράφημα παρουσιάζει τις παρατηρήσεις που έχουν καταταχτεί λάθος (έντονο μαύρο χρώμα).

Γράφημα 2.5: Σημεία εσφαλμένης ομαδοποίησης (έντονο μαύρο χρώμα) των diabetes δεδομένων με βάση το VVV μοντέλο.



Αυτό που παρατηρούμε από τα γραφήματα 3 και 4 είναι ότι κάποιες από τις παρατηρήσεις με μεγάλη αβεβαιότητα έχουν καταταχθεί και λάθος. Αυτά τα σημεία αντιστοιχούν σε outliers των κατανομών των ομάδων. Περισσότερα για το παράδειγμα αυτό παρέχονται από τους Fraley and Raftery (2007).

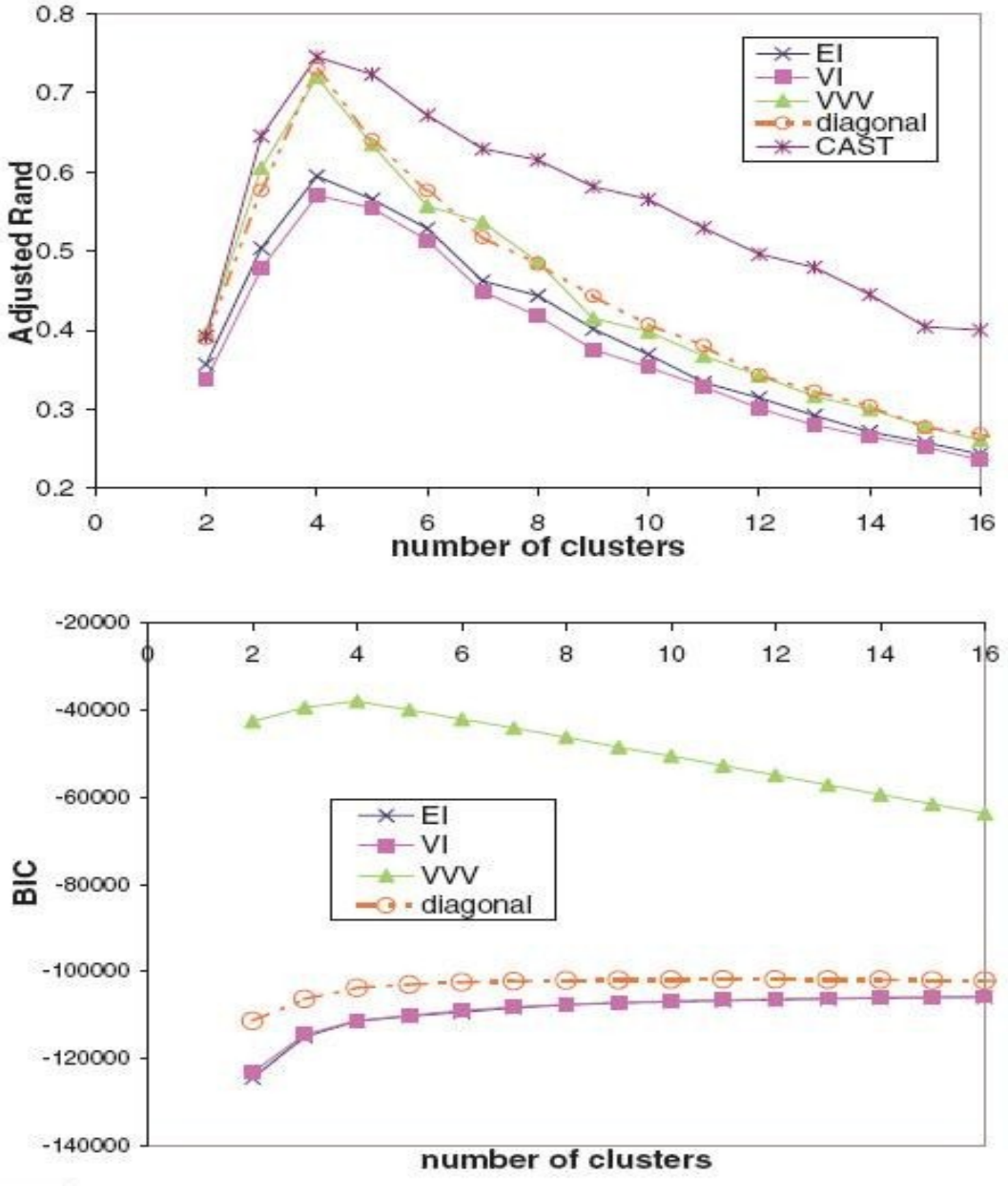
2.9.2 Παράδειγμα 2

Στο παράδειγμα αυτό χρησιμοποιούμε ένα μέρος των Ovary data. Τα δεδομένα αυτά αφορούν 235 αλληλουχίες DNA (235 παρατηρήσεις) από 24 ιστούς γυναικών (24 μεταβλητές) όπου κάποιοι ιστοί προέρχονται από υγιείς γυναίκες και κάποιοι από γυναίκες με καρκίνο ωοθηκών διαφόρων σταδίων. Βρέθηκε ότι αυτές οι 235 αλληλουχίες προέρχονται από 4 διαφορετικά γονίδια, δηλαδή σχηματίζουν 4 ομάδες, όπου η $1^{η}$ περιέχει 58 αλληλουχίες, η $2^{η}$ 88, η $3^{η}$ 57 και η $4^{η}$ 32. Αυτή η ομαδοποίηση

(την οποία θα αναφέρουμε ως "πραγματική") θα χρησιμοποιηθεί ως βάση για να συγκρίνουμε τις διαφορετικές ομαδοποιήσεις που θα προκύψουν από τα διαφορετικά μοντέλα. Ωστόσο, στο παράδειγμα που θα ακολουθήσει δε θα χρησιμοποιήσουμε τα αυθεντικά δεδομένα, αλλά αρχικά θα προσομοιώσουμε δεδομένα από 4 πολυμεταβλητές κανονικές κατανομές με μέσους το δειγματικό μέσο κάθε μίας από τις 4 ομάδες της πραγματικής ομαδοποίησης, και πίνακες διασποράς-συνδιασποράς τους δειγματικούς. Συνολικά μέσω προσομοίωσης παίρνουμε 2350 παρατηρήσεις και επαναλαμβάνουμε την προσομοίωση 10 φορές παίρνοντας έτσι 10 διαφορετικά σετ δεδομένων. Το μέγεθος κάθε ομάδας στα προσομοιωμένα δεδομένα είναι 10πλάσιο από το αντίστοιχο των αυθεντικών δεδομένων, δηλαδή οι ομάδες αποτελούνται από 580, 880, 570 και 320 αλληλουχίες η κάθε μία. Αυτά τα 10 προσομοιωμένα σετ περιέχουν δεδομένα που προέρχονται από πολυμεταβλητές κανονικές κατανομές για κάθε ομάδα. Οπότε η πολυμεταβλητή κανονικότητα δεν είναι μια υπόθεση εδώ, είναι μια πραγματικότητα.

Στο προηγούμενο παράδειγμα είχαμε απλά δει την εφαρμογή του model-based clustering για τα μοντέλα της οικογένειας MCLUST. Εδώ θα δούμε και μια σύγκριση μοντέλων αυτής της οικογένειας με δύο άλλα μοντέλα, το diagonal μοντέλο και το CAST καθώς επίσης και τη χρήση του adjusted Rand Index για σύγκριση. Στο diagonal μοντέλο οι πίνακες διασποράς-συνδιασποράς κάθε ομάδας ισούνται με $\Sigma_j = \lambda_j \mathrm{B}_j$ όπου $\mathrm{B_j}$ είναι ένας διαγώνιος πίνακας με $\left|B_j\right| = 1$. Το CAST είναι ένας ευρετικός αλγόριθμος ομαδοποίησης ο οποίος συγκρινόμενος με άλλες ιεραρχικές μεθόδους ομαδοποίησης έχει βρεθεί ότι είναι καλύτερος (Yeung et al. 2001a). Στο παρακάτω γράφημα φαίνονται τα αποτελέσματα.

Γράφημα 2.6: Μέσος Adjusted Rand Index και μέσο BIC κριτήριο για την ομαδοποίηση των 10 προσομοιωμένων datasets των Ovary data.



Το πρώτο γράφημα απεικονίζει τη μέση τιμή του adjusted Rand Index για τα 10 προσομοιωμένα σετ δεδομένων για επιλογή διαφορετικού αριθμού ομάδων. Κάθε ομαδοποίηση που προκύπτει από διαφορετικό μοντέλο και αριθμό ομάδων συγκρίνεται με την πραγματική, και προκύπτει ο αντίστοιχος adjusted Rand Index. Επειδή όμως έχουμε 10 διαφορετικά σετ δεδομένων, άρα και 10 διαφορετικούς δείκτες για το ίδιο μοντέλο και αριθμό ομάδων, παίρνουμε το μέσο όρο αυτών και

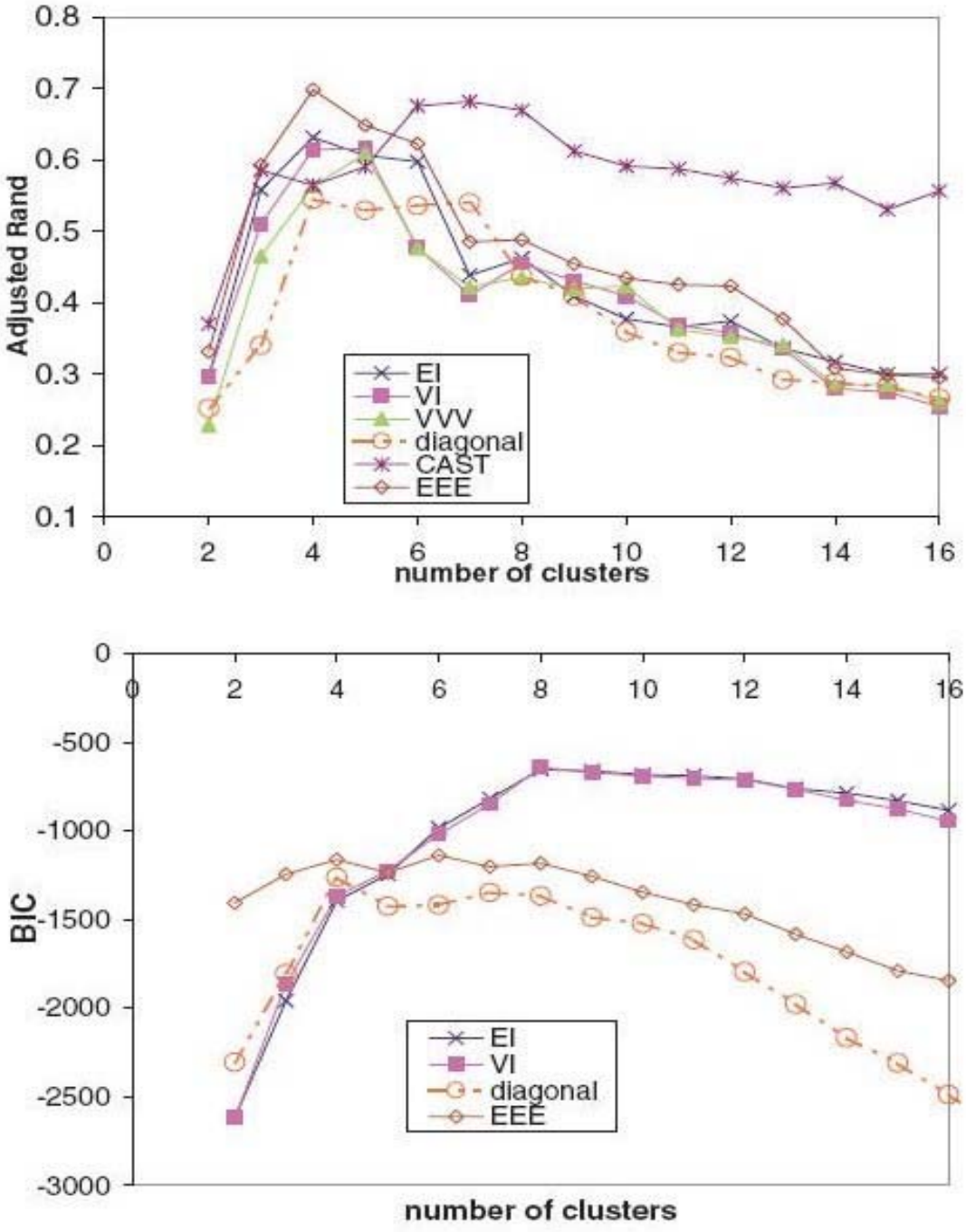
αυτός ο μέσος όρος είναι που απεικονίζεται στο γράφημα. Το ίδιο γίνεται και για το BIC κριτήριο το οποίο απεικονίζεται στο δεύτερο γράφημα.

Η μέση τιμή του adjusted Rand Index γίνεται μέγιστη για 4 ομάδες (όσες και στην πραγματική ομαδοποίηση) με το VVV, το diagonal μοντέλο και το CAST να έχουν περίπου παρόμοιες μέσες τιμές. Τα μοντέλα VI και EI είναι τα ίδια με τα VII και EII που είδαμε στην 2.6. Όσο αφορά το BIC κριτήριο το καλύτερο μοντέλο, μεταξύ των τεσσάρων είναι το VVV, πάλι για 4 ομάδες. Επίσης, το diagonal μοντέλο παράγει τιμές BIC που είναι υψηλότερες από τα VII και EII μοντέλα, κάτι που είναι σε συμφωνία και με τα αποτελέσματα του adjusted Rand Index. Κατά συνέπεια η ανάλυση που βασίζεται στο BIC κριτήριο διαλέγει το σωστό μοντέλο και το σωστό αριθμό ομάδων σε αυτά τα προσομοιωμένα δεδομένα που ενισχύει το BIC ως κατάλληλο κριτήριο για την επιλογή του σωστού μοντέλου.

Όλα τα παραπάνω μοντέλα του model-based clustering που έχουμε δει μέχρι στιγμής υποθέτουν ότι τα δεδομένα μας ακολουθούν πολυμεταβλητή κανονική κατανομή, πράγμα που σπανίως ικανοποιείται στην πράξη. Ωστόσο, τα μοντέλα είναι αρκετά ανθεκτικά σε αποκλίσεις από την κανονικότητα αλλά δουλεύουν καλύτερα όταν τα δεδομένα είναι κανονικά ή μπορούν να μετασχηματιστούν όσο πιο κοντά γίνεται στην κανονικότητα. Επειδή είναι δύσκολο να ελέγξουμε για πολυμεταβλητή κανονικότητα αλλά και επειδή σπάνια αυτή ικανοποιείται, μια συνήθης πρακτική είναι να ελέγχουμε για κάθε μεταβλητή αν τα δεδομένα της ακολουθούν μονοδιάστατη κανονική κατανομή και αν όχι, να μετασχηματίζουμε τα δεδομένα κατάλληλα ώστε να υπάρχει η ελάχιστη δυνατή απόκλιση από την κανονικότητα.

Στο παράδειγμα των Ovary δεδομένων βρέθηκε ότι ο κατάλληλος μετασχηματισμός για τα **original** δεδομένα είναι η τετραγωνική ρίζα. Έτσι μετασχηματίστηκαν όλες οι μεταβλητές με βάση την τετραγωνική ρίζα και ξαναέτρεξε η ομαδοποίηση. Τα αποτελέσματα φαίνονται παρακάτω:

Γράφημα 2.7: Adjusted Rand Index και BIC κριτήριο για την ομαδοποίηση των Ovary data μετά από μετασχηματισμό τετραγωνικής ρίζας για 6 μοντέλα.



Το πρώτο γράφημα περιέχει τον adjusted Rand Index όπως αυτός προκύπτει συγκρίνοντας το αποτέλεσμα κάθε μοντέλου με την πραγματική ομαδοποίηση για διάφορους αριθμούς ομάδων και το δεύτερο περιέχει το BIC κριτήριο για τα αντίστοιχα μοντέλα. Από το πρώτο γράφημα φαίνεται ότι τα μοντέλα EII,VII και EEE δίνουν καλύτερη ομαδοποίηση σε σχέση με τα diagonal, CAST και VVV για την περίπτωση των 4 ομάδων (που είναι και οι πραγματικές ομάδες). Ωστόσο, ο ρυθμός

με τον οποίο μειώνεται ο adjusted Rand Index για το CAST είναι μικρότερος από τα υπόλοιπα μοντέλα οδηγώντας έτσι το μοντέλο αυτό να είναι προτιμότερο όταν οι ομάδες είναι περισσότερες. Όσο αφορά τα αποτελέσματα του BIC κριτηρίου, βλέπουμε ότι το EEE μοντέλο που ήταν το καλύτερο με βάση τον adjusted Rand Index στις 4 ομάδες, παρουσιάζει το ένα από τα δύο μέγιστά του στις 4 ομάδες (που είναι και οι πραγματικές) και το diagonal μοντέλο παρουσιάζει και αυτό το ολικό μέγιστό του στις 4 ομάδες. Με βάση το BIC το EEE είναι καλύτερο από όλα τα άλλα στις 4 ομάδες και αυτό είναι σύμφωνο και με την Rand Index ανάλυση. Τα άλλα δύο μοντέλα, EII και VII, παρουσιάζουν μέγιστο στις 8 ομάδες. Ωστόσο, η λύση των 8 ομάδων ίσως έχει κάποια ερμηνεία καθώς διαφέρει από την πραγματική λύση στο ότι χωρίζει τις μεγάλες ομάδες σε δύο ή τρεις επιμέρους (κάτι το οποίο μπορεί να αντανακλά διαφορές στις αλληλουχίες DNA, για παράδειγμα). Οπότε, η λύση των 8 ομάδων που προτείνεται από το BIC δεν είναι παράλογη.

Άρα, γενικά θα λέγαμε ότι η model-based προσέγγιση δίνει ελαφρώς καλύτερα αποτελέσματα από την CAST (βάσει του Rand Index) για αυτά τα πραγματικά δεδομένα και η ανάλυση μέσω του BIC έδωσε επίσης μια λογική υπόδειξη για τον αριθμό των ομάδων. Περισσότερες πληροφορίες σχετικά με αυτό το παράδειγμα υπάρχουν στο Yeung et al. (2001b).

2.10 Παρατηρήσεις

Στο σημείο αυτό, μετά και τα προηγούμενα παραδείγματα αλλά κατόπιν και όλων όσων έχουν αναφερθεί προηγουμένως, αξίζει να επισημάνουμε το μεγάλο πλεονέκτημα του model-based clustering έναντι των κλασικών μεθόδων αποστάσεων στην ομαδοποίηση δεδομένων. Η τεχνική του model-based clustering βασίζεται σε πιθανοθεωρητικά μοντέλα και αυτό είναι ένα σημαντικό πλεονέκτημα έναντι των μεθόδων αποστάσεων διότι μπορούν να υποστηρίξουν στατιστική συμπερασματολογία. Υπό την υπόθεση ότι τα δεδομένα μας ακολουθούν πολυμεταβλητή κανονική κατανομή μέσα σε κάθε ομάδα (cluster), ή οποιαδήποτε άλλη κατανομή, το πρόβλημα καθορισμού του σωστού αριθμού ομάδων αλλά και του κατάλληλου αλγόριθμου ομαδοποίησης (το μοντέλο δηλαδή) ανάγεται σε ένα στατιστικό πρόβλημα επιλογής μοντέλου και είδαμε τρόπους για να το επιτύχουμε

αυτό. Κατά συνέπεια μπορούμε να προβούμε σε στατιστική συμπερασματολογία. Αντιθέτως, στις μεθόδους αποστάσεων, η απουσία οποιουδήποτε στατιστικού μοντέλου καθιστά δύσκολο τον προσδιορισμό του κατάλληλου αλγορίθμου και του σωστού αριθμού ομάδων. Έτσι, το model-based clustering υπερέχει στον τομέα της στατιστικής συμπερασματολογίας.

Όσο αφορά τον ΕΜ αλγόριθμο, αυτός έχει μερικές ελκυστικές ιδιότητες κάποιες από τις οποίες είναι οι εξής:

- Έχει μονότονη σύγκλιση καθώς σε κάθε επανάληψη η πιθανοφάνεια μεγαλώνει. Αυτό μας επιτρέπει να αποφασίσουμε πότε θα σταματήσουμε τις επαναλήψεις θέτοντας κάποιο κριτήριο τερματισμού.
- Για καλές αρχικές τιμές ο αλγόριθμος συγκλίνει σε ολικό μέγιστο της πιθανοφάνειας.
- Μπορεί εύκολα να υλοποιηθεί, αναλυτικά και υπολογιστικά. Ιδιαίτερα είναι εύκολο να προγραμματιστεί και απαιτεί λίγο χώρο στη μνήμη του υπολογιστή. Ελέγχοντας τη μονότονη σύγκλιση της πιθανοφάνειας με το πέρασμα των επαναλήψεων μπορούμε να ελέγξουμε πότε έχει συγκλίνει ο αλγόριθμος.
- Μπορεί να χρησιμοποιηθεί για να πάρουμε εκτιμήσεις και άλλων ποσοτήτων που τις θεωρούμε ως missing data (όπως οι πιθανότητες τ_{ij}).
- Δίνει εκτιμήσεις μέσα στα επιτρεπτά όρια (π.χ. δε δίνει αρνητική διακύμανση), αν οι αρχικές τιμές είναι μέσα στα επιτρεπτά όρια. Αυτό δεν είναι σίγουρο όταν χρησιμοποιούμε άλλες αριθμητικές μεθόδους μεγιστοποίησης.

Ωστόσο υπάρχουν και μειονεκτήματα όπως:

- Το αποτέλεσμα εξαρτάται από τις αρχικές τιμές και άρα χρειαζόμαστε καλές αρχικές τιμές. Στον ΕΜ οι καλές αρχικές τιμές ισοδυναμούν με γρήγορη σύγκλιση μέσα σε λίγες επαναλήψεις, ενώ σε άλλες μεθόδους αν δεν έχουμε καλές αρχικές τιμές αυτό μπορεί να σημαίνει και ότι ο αλγόριθμος δε συγκλίνει ποτέ.
- Μπορεί να καταλήξουμε σε μία λύση που να αντιστοιχεί σε τοπικό και όχι σε ολικό μέγιστο. Επομένως, χρειάζεται να ξεκινήσουμε από διαφορετικές αρχικές τιμές για να είμαστε σίγουροι πως θα βρούμε τη λύση που αντιστοιχεί στο ολικό μέγιστο της πιθανοφάνειας.

- Ο αλγόριθμος δεν παρέχει αυτόματα εκτίμηση των διασπορών & συνδιασπορών των εκτιμημένων παραμέτρων. Ωστόσο το μειονέκτημα αυτό μπορεί να ξεπεραστεί χρησιμοποιώντας κατάλληλη μεθοδολογία.
- Μερικές φορές η σύγκλιση είναι πολύ αργή.
- Σε μερικές περιπτώσεις τα Ε ή Μ βήματα μπορεί να είναι αναλυτικά δύσκολο να υπολογιστούν.

2.11 Παραλλαγές του ΕΜ

Υπάρχουν διάφορες παραλλαγές του ΕΜ αλγορίθμου. Ο ECM (expectation conditional maximization) είναι μία από αυτές. Ένας από τους κύριους λόγους που ο ΕΜ αλγόριθμος είναι δημοφιλής είναι επειδή στο Μ-βήμα του περιέχει την επίλυση της complete-data log-likelihood η οποία είναι υπολογιστικά εύκολη. Αλλά αν η επίλυση της complete-data log-likelihood είναι δύσκολη, ανάλογα με τις κατανομές που χρησιμοποιούμε, τότε ο ΕΜ δε φαίνεται και τόσο ελκυστικός. Ωστόσο, σε πολλές περιπτώσεις η επίλυση της complete-data log-likelihood μπορεί να γίνει σχετικά εύκολα αν η μεγιστοποίηση υπολογιστεί δοθέντος κάποιων από τις παραμέτρους (ή δοθέντος κάποιων συναρτήσεων των παραμέτρων). Προς αυτή την κατεύθυνση οι Meng and Rubin (1993) δημιούργησαν τον ECM αλγόριθμο. Ο αλγόριθμος αυτός εκμεταλλεύεται την ευκολία του υπολογισμού της δεσμευμένης complete-data log-likelihood αντικαθιστώντας το περίπλοκο Μ-βήμα του ΕΜ αλγορίθμου με άλλα, υπολογιστικώς πιο απλά CM-βήματα. Το CM βήμα μπορεί να υπάρχει σε κλειστή μορφή ή μπορεί να χρειάζεται επαναληπτική διαδικασία για να υπολογιστεί, αλλά επειδή ο παραμετρικός χώρος πάνω στον οποίο πραγματοποιείται έχει μικρότερες διαστάσεις, συχνά είναι πιο απλό και πιο γρήγορο από το αντίστοιχο Μ-βήμα του ΕΜ αλγορίθμου. Κατά συνέπεια ο ECM συγκλίνει πιο αργά από τον ΕΜ ως προς τον αριθμό των επαναλήψεων που απαιτούνται για τη σύγκλιση αλλά μπορεί να είναι πιο γρήγορος χρονικά. Και το πιο σημαντικό, ο ECM διατηρεί τις ιδιότητες του ΕΜ όπως τη μονότονη σύγκλιση. Περισσότερες πληροφορίες όπως και αναλυτική περιγραφή του ECM παρέχεται από τους McLachlan and Krishnan (2008).

Άλλη παραλλαγή του EM είναι ο CEM (conditional expectation maximization). Η διαφορά του CEM είναι ότι στο Ε-βήμα και αφού υπολογιστούν τα τ_{ij} για κάθε παρατήρηση βρίσκουμε την ομάδα που έχει τη μεγαλύτερη πιθανότητα να ανήκει και θέτουμε $\tau_{ij}^{*}=1$ για αυτή την ομάδα και $\tau_{ij}^{*}=0$ για όλες τις υπόλοιπες. Δηλαδή κατατάσσουμε την κάθε παρατήρηση σε μία μόνο ομάδα. Στη συνέχεια στο Μ-βήμα χρησιμοποιούμε τα τ_{ij}^{*} στη θέση των τ_{ij}. Ουσιαστικά ο αλγόριθμος αυτός μοιάζει πολύ με τον αλγόριθμο K-means. Αποδεικνύεται μάλιστα ότι ο CEM ταυτίζεται με τον K-means, όταν υποθέσουμε κοινό πίνακα διακύμανσης για όλες τις ομάδες ο οποίος να είναι και διαγώνιος. Ουσιαστικά ο K-means φτιάχνει στρογγυλές ομάδες (επειδή ο πίνακας διακύμανσης είναι διαγώνιος) και οι ομάδες δεν τέμνονται (γιατί κάθε παρατήρηση ανήκει αναγκαστικά σε μία ομάδα) κάτι το οποίο αντιστοιχεί σε περιορισμούς όχι ρεαλιστικούς στην πράξη. Αντίθετα, με τον EM μπορούμε να πάρουμε ομάδες που τέμνονται και έχουν διαφορετικό σχήμα, όχι αναγκαστικά σφαιρικό, όπως είδαμε στο παράδειγμα 1 (γράφημα 2.3).

Υπάρχουν επίσης διάφορες άλλες παραλλαγές του EM αλγορίθμου, όπως οι ECME (expectation conditional maximization either), SAGE (space-alternating generalized EM), PX-EM (parameter-expanded EM), AECM (alternating expectation conditional maximization) κτλ. οι οποίοι δεν αναπτύσσονται εδώ. Παρακάτω θα εξετάσουμε αναλυτικά μόνο την περίπτωση του AECM στη χρήση των factor analyzers.

Κεφάλαιο 3: Το πρόβλημα των high-dimensional data

3.1 High-Dimensional Data

Είδαμε στο προηγούμενο κεφάλαιο την εφαρμογή του ΕΜ αλγορίθμου στο model-based clustering, παρουσιάσαμε κάποια από τα πλεονεκτήματά του και είδαμε και κάποια παραδείγματα εφαρμογής του στην ομαδοποίηση δεδομένων. Παρά την πολύ καλή επίδοσή του σε μια πληθώρα προβλημάτων της cluster analysis η εφαρμογή του σε μερικά είδη δεδομένων παρουσιάζει κάποιες δυσκολίες. Μια τέτοια περίπτωση είναι αυτή των high-dimensional δεδομένων. High-dimensional δεδομένα είναι αυτά όπου έχουμε πολύ μεγάλο αριθμό μεταβλητών (p). Ο αριθμός των μεταβλητών μπορεί να είναι οποιοσδήποτε, από μερικές δεκάδες μέχρι και χιλιάδες. Φυσικά όταν έχουμε ένα τόσο μεγάλο αριθμό μεταβλητών είναι δύσκολο να συγκεντρώσουμε περισσότερες παρατηρήσεις και έτσι καταλήγουμε να έχουμε διαθέσιμες λιγότερες παρατηρήσεις (n) από μεταβλητές. Βρισκόμαστε δηλαδή στην περίπτωση όπου $n \ll p$. Τέτοιου είδους δεδομένα συναντάμε συνήθως στη γενετική, σε gene-expression πειράματα.

Ένα πρόβλημα του ΕΜ αλγορίθμου όταν χρησιμοποιούμε πολύ ευέλικτα μοντέλα, όπως το VVV, είναι ο μεγάλος αριθμός παραμέτρων που πρέπει να εκτιμήσουμε. Το πρόβλημα αυτό γιγαντώνεται στην περίπτωση των high-dimensional δεδομένων, καθώς εκεί ο αριθμός των παραμέτρων που θα πρέπει να εκτιμήσουμε είναι πραγματικά τεράστιος ειδικά αν χρησιμοποιούμε και πολλές ομάδες. Επιπλέον, η εφαρμογή του ΕΜ αλγορίθμου σε ελλειπτικά μοντέλα (EEE, EEV, VEV, VVV βλ. πίνακα 1.1) για την περίπτωση των high-dimensional δεδομένων καθίσταται εντελώς αδύνατη καθώς προκύπτουν αριθμητικά προβλήματα (singularities) στον υπολογισμό της log-likelihood. Ωστόσο, είναι πιθανό να μπορεί να εφαρμοστεί για σφαιρικά (EII, VII) και διαγώνια (EEI, VEI, EVI, VVI) μοντέλα. Όμως τα μοντέλα αυτά, παρ' ότι απαιτούν λιγότερες παραμέτρους προς εκτίμηση, δεν παρέχουν την ευελιξία των ελλειπτικών μοντέλων και άρα δεν είναι πάντοτε κατάλληλα.

Όταν λοιπόν τα δεδομένα μας είναι πολλών διαστάσεων, χωρίς απαραίτητα οι παρατηρήσεις να είναι λιγότερες από τις μεταβλητές, είναι φανερό ότι τα δεδομένα μας απαιτούν κάποια τροποποίηση, ώστε να μειώσουμε τον αριθμό των διαστάσεών

τους. Μερικές φορές συσχετίσεις ή άλλες σχέσεις μεταξύ των μεταβλητών είναι προφανείς, οπότε μπορούμε εύκολα να διαλέξουμε μερικές μεταβλητές από όλο το πλήθος των μεταβλητών και να προχωρήσουμε στην ομαδοποίηση χρησιμοποιώντας μόνο αυτές. Επίσης, υπάρχουν τεχνικές της πολυμεταβλητής ανάλυσης για τη μείωση των διαστάσεων όπως η μέθοδος κύριων συνιστωσών (Principal Components Analysis) και η παραγοντική ανάλυση (Factor Analysis).

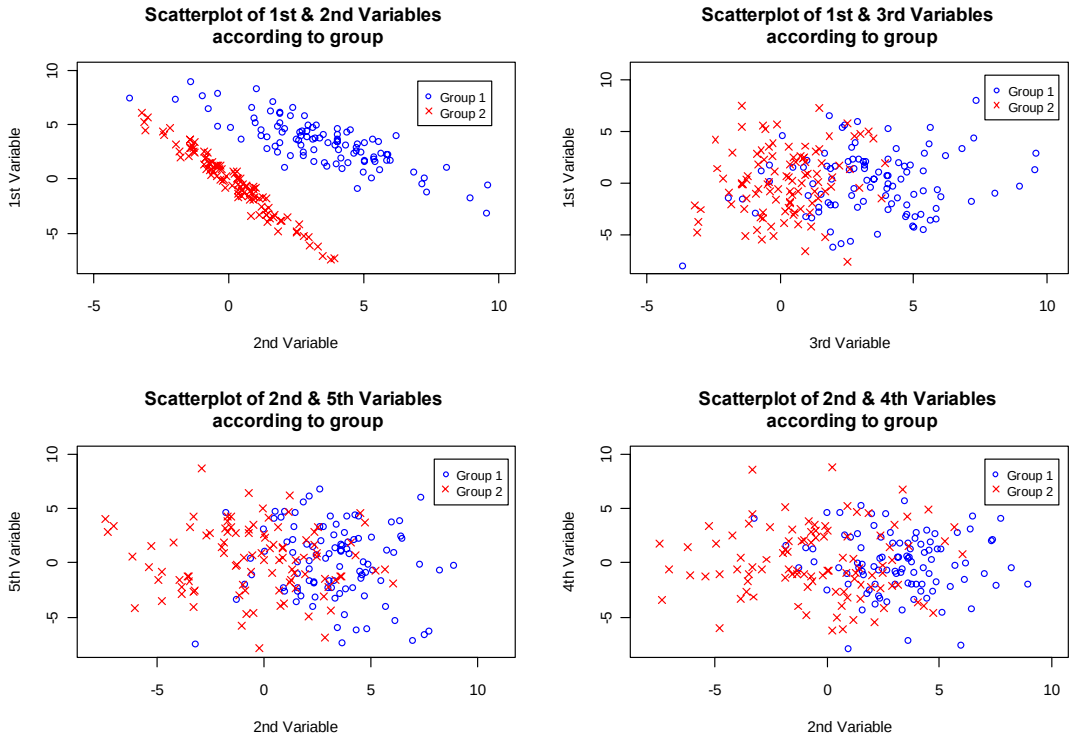
Με βάση τη μέθοδο κύριων συνιστωσών, μετασχηματίζουμε όλα τα αρχικά δεδομένα μας σε κύριες συνιστώσες και προχωρούμε στην ομαδοποίηση με βάση τις κύριες συνιστώσες ως νέες μεταβλητές (βλ. Καρλής 2005 για μέθοδο κύριων συνιστωσών). Οι κύριες συνιστώσες που υπολογίζουμε προφανώς θα είναι πολύ μικρότερες σε αριθμό από τις μεταβλητές, διότι σε διαφορετική περίπτωση δεν έχει νόημα η όλη μετατροπή. Όμως έχει βρεθεί ότι σε μερικές περιπτώσεις αυτός ο μετασχηματισμός των δεδομένων μπορεί να αποκρύψει παρά να αναδείξει διαφορετικές ομάδες δεδομένων. Εάν υπάρχουν λίγες διαφορετικές ομάδες και είναι καλά διαχωρισμένες μεταξύ τους και η διασπορά μεταξύ των ομάδων είναι πολύ μεγαλύτερη των διασπορών μέσα στις ομάδες, τότε αναμένουμε οι πρώτες κύριες συνιστώσες να αναδεικνύουν τη δομή των δεδομένων και να αποκαλύπτουν τις ομάδες. Ωστόσο, εάν αυτές οι προϋποθέσεις δεν ισχύουν η μέθοδος κύριων συνιστωσών αποτυγχάνει να διαχωρίσει τις ομάδες, ανεξάρτητα του αν θα χρησιμοποιήσουμε τον πίνακα διασποράς ή τον πίνακα συσχετίσεων για την ανάλυση κύριων συνιστωσών. Το ακόλουθο παράδειγμα βοηθά στην κατανόηση του συγκεκριμένου προβλήματος.

Έστω ότι προσομοιώνουμε 100 παρατηρήσεις από την 5-μεταβλητή κανονική κατανομή για κάθε μία από δύο ομάδες παρατηρήσεων. Συνολικά δηλαδή έχουμε 200 παρατηρήσεις, 100 για κάθε ομάδα. Η πρώτη ομάδα έχει μέσο και πίνακα διασποράς $\mu_1 = (5,0,0,0,0)^T$ *και* $\Sigma_1 = diag(1,10,10,10,10)$ αντίστοιχα και η δεύτερη $\mu_2 = (0,0,0,0,0)^T$ *και* $\Sigma_2 = \begin{pmatrix} 1 & 3 & 0 & 0 & 0 \\ 3 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 10 \end{pmatrix}$ αντίστοιχα. Δηλαδή για την 1η ομάδα ισχύει $y_i \sim N_5(\mu_1, \Sigma_1)$ και για τη δεύτερη $y_i \sim N_5(\mu_2, \Sigma_2)$. Στη συνέχεια πολλαπλασιάζουμε κάθε μία από τις 200 παρατηρήσεις με τον ορθογώνιο πίνακα

$$H = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) & 0 & 0 & 0 \\ \sin(\pi/4) & \cos(\pi/4) & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \cos(\pi/4) & -\sin(\pi/4) \\ 0 & 0 & 0 & \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}.$$

Εκτελούμε δηλαδή ένα ορθογώνιο μετασχηματισμό στα δεδομένα. Έτσι, υπό τα μετασχηματισμένα δεδομένα οι δύο ομάδες είναι σχετικά καλά διαχωρισμένες μεταξύ τους. Στο ακόλουθο σχήμα φαίνονται οι ομάδες προβάλλοντας τα δεδομένα πάνω σε κάποιες από τις πέντε μεταβλητές.

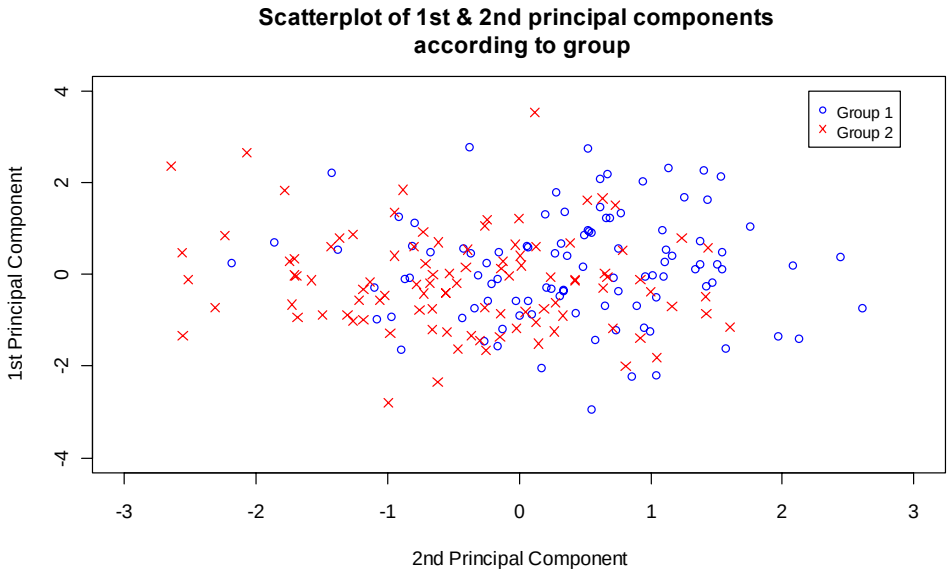
Γράφημα 3.1. Scatterplot για διάφορες ανά δύο μεταβλητές των μετασχηματισμένων δεδομένων ανάλογα με την ομάδα.



Βλέπουμε από τα παραπάνω γραφήματα ότι οι ομάδες διαχωρίζονται μεταξύ τους καλά, μόνο στις 2 πρώτες περιπτώσεις (γραφήματα της $1^{ης}$ γραμμής), για τις μεταβλητές 1, 2 και 3. Επίσης αντιλαμβανόμαστε ότι οι μέσοι των ομάδων διαφέρουν σημαντικά μόνο κατά τις 2 πρώτες μεταβλητές και κατά συνέπεια μόνο οι 2 πρώτες μεταβλητές παρέχουν σημαντική πληροφορία για το διαχωρισμό των ομάδων. Αυτό έχει σαν συνέπεια η μέθοδος κύριων συνιστωσών να αποτυγχάνει στον καλό

διαχωρισμό των ομάδων. Από το ακόλουθο γράφημα το οποίο παριστάνει τις δύο πρώτες κύριες συνιστώσες (με βάση τον πίνακα συσχετίσεων) δε φαίνεται να υπάρχει καμία ένδειξη ότι τα δεδομένα μας προέρχονται από 2 διαφορετικές ομάδες.

Γράφημα 3.2. Scatterplot των 2 πρώτων κύριων συνιστωσών των μετασχηματισμένων δεδομένων ανάλογα με την ομάδα.



Πράγματι αν εφαρμόζαμε τον EM αλγόριθμο για την ομαδοποίηση των παραπάνω δεδομένων σε δύο ομάδες, χρησιμοποιώντας τις 2 πρώτες κύριες συνιστώσες θα καταλήγαμε να κατατάξουμε λάθος το 41% των παρατηρήσεων. Αντίθετα αν χρησιμοποιούσαμε την παραγοντική ανάλυση μέσα σε κάθε ομάδα για τη μείωση των διαστάσεων και εφαρμόζαμε τον αντίστοιχο αλγόριθμο (AECM) δε θα κατατάσσαμε καμία παρατήρηση λάθος.

Ένα επιπλέον μειονέκτημα της χρησιμοποίησης της PCA ανάλυσης στα αρχικά high-dimensional δεδομένα είναι, ότι ακόμα και να πετύχουμε μια πολύ καλή ομαδοποίηση, στο τέλος δε θα ξέρουμε ποιες ακριβώς από τις μεταβλητές είναι αυτές που διαφοροποιούνται περισσότερο μεταξύ τους και άρα συμβάλλουν περισσότερο στο διαχωρισμό των ομάδων. Και αυτό γιατί η όλη ανάλυση πλέον δε βασίζεται στις αρχικές μεταβλητές αλλά στις κύριες συνιστώσες. Οπότε, έτσι χάνουμε τη δυνατότητα εύκολης ερμηνείας των αποτελεσμάτων.

Το πρόβλημα αυτής της προσέγγισης είναι ότι η PCA επιβάλει ένα γραμμικό μετασχηματισμό καθολικά, σε όλα δηλαδή τα δεδομένα και αυτό συνιστά ένα μεγάλο περιορισμό, ο οποίος οδηγεί τη μέθοδο συχνά σε αποτυχία. Εντούτοις, η PCA μπορεί να χρησιμοποιηθεί επιτυχώς στην cluster analysis εάν εφαρμοστεί τοπικά μέσα σε

κάθε ομάδα. Αν επιπλέον εφαρμόσουμε ένα πεπερασμένο μείγμα κατανομών για να περιγράψουμε τα νέα δεδομένα οδηγούμαστε σε ένα πιθανοθεωρητικό μοντέλο όπου όλες οι παράμετροι μπορούν να εκτιμηθούν από μία και μόνη συνάρτηση πιθανοφάνειας. Η προσέγγιση αυτή οδηγεί στην ανάπτυξη του μοντέλου PPCA (probabilistic principal component analysis). Έτσι αποκτούμε όχι μόνο έναν αλγόριθμο επίλυσης αλλά και μια συνάρτηση πυκνότητας πιθανότητας του μοντέλου. Με τον τρόπο αυτό πετυχαίνουμε μία αρχική μείωση στον αριθμό παραμέτρων προς εκτίμηση. Επίσης μπορούμε να πετύχουμε περαιτέρω μείωση των παραμέτρων αν επιβάλουμε κάποιους περιορισμούς στους πίνακες διασποράς των πολυμεταβλητών κανονικών κατανομών των ομάδων, όπως και στην περίπτωση της παραγράφου 2.6. Οι δύο περιορισμοί που εφαρμόζονται εδώ αναγκάζουν τον κάθε πίνακα διασποράς να είναι διαγώνιος με ίσα ή διαφορετικά στοιχεία. Στο PPCA μοντέλο ο πίνακας διασποράς κάθε ομάδας είναι της μορφής $C = \sigma^2 I + WW^T$, όπου W είναι ο πίνακας που περιέχει ως γραμμές τα ιδιοδιανύσματα του δειγματικού πίνακα διακύμανσης των δεδομένων κάθε ομάδας. Αναλυτική περιγραφή του PPCA μοντέλου, καθώς επίσης παραδείγματα και εφαρμογές μπορούν να βρεθούν στην εργασία των Tipping and Bishop (1999).

Φυσικά υπάρχουν και άλλοι τρόποι προσέγγισης του προβλήματος ομαδοποίησης high-dimensional δεδομένων. Ένας από αυτούς είναι ο αλγόριθμος HDDC (High-Dimensional Data Clustering) ο οποίος εφαρμόζεται κυρίως για να εντοπίζει τη θέση διαφόρων αντικειμένων μέσα σε εικόνες χρησιμοποιώντας ένα πιθανοθεωρητικό πλαίσιο (Bouveyron et al. 2007). Άλλη προσέγγιση είναι η μέθοδος EMMIX-GENE, η οποία χρησιμοποιείται στην ομαδοποίηση microarray expression data όπου ο αρχικός μεγάλος αριθμός των γονιδίων (μεταβλητές) μειώνεται με τη χρήση του λόγου πιθανοφανειών πριν στη συνέχεια εφαρμοστεί κάποιος αλγόριθμος ομαδοποίησης, McLachlan *et al.* (2002).

3.2 Single-factor Analysis Model

Είδαμε προηγουμένως ότι η εφαρμογή της ανάλυσης κύριων συνιστωσών πάνω σε όλα τα δεδομένα απέτυχε να συμβάλει ικανοποιητικά στην ομαδοποίηση των δεδομένων. Αντίθετα η εφαρμογή της μεθόδου μέσα σε κάθε ομάδα οδήγησε στην ανάπτυξη του PPCA μοντέλου το οποίο εφαρμόζεται στη συσταδική ανάλυση. Αντί να εφαρμόσουμε ανάλυση κύριων συνιστωσών μέσα σε κάθε ομάδα ώστε να μειώσουμε τον αριθμό των παραμέτρων προς εκτίμηση, μπορούμε εναλλακτικά να εφαρμόσουμε παραγοντική ανάλυση οδηγώντας έτσι σε ένα μοντέλο ενός μείγματος από factor analyzers (mixtures of factor analyzers model). Αυτό το μοντέλο παρουσιάστηκε για πρώτη φορά από τους Ghahramani and Hinton (1997) οι οποίοι υπέθεσαν για τον πίνακα διασποράς κάθε ομάδας ότι είναι της μορφής $\Sigma_j = B_j B_j^T + D, \ (j = 1,...,g)$, όπου B_j είναι ο πίνακας επιβαρύνσεων και D ο πίνακας ιδιαιτεροτήτων, που προκύπτουν από την εφαρμογή της παραγοντικής ανάλυσης. Αργότερα οι McLachlan and Peel (2000b) επέκτειναν αυτό το μοντέλο υποθέτοντας $\Sigma_j = B_j B_j^T + D_j, \ (j = 1,...,g)$, δηλαδή διαφορετικό πίνακα ιδιαιτεροτήτων για κάθε ομάδα. Επίσης περισσότερες πληροφορίες σχετικά μπορούν να βρεθούν και στα McLachlan and Peel (2000a, chapter 8), McLachlan et al. (2002). Τέλος, το PPCA μοντέλο των Tipping and Bishop (1999) που αναφέρθηκε παραπάνω υποθέτει ότι $\Sigma_j = B_j B_j^T + d_j I_p, \ (j = 1,...,g)$.

Η παραγοντική ανάλυση χρησιμοποιείται ιδιαίτερα για να διερευνούμε συσχετίσεις μεταξύ των μεταβλητών σε πολυμεταβλητά δεδομένα, αλλά επίσης μπορεί να χρησιμοποιηθεί και για μείωση διαστάσεων. Το μοντέλο παραγοντικής ανάλυσης είναι το $Y_i = \mu + BU_i + e_i \ \ (i = 1,...,n)$, όπου Y_i είναι η i παρατήρηση (διάνυσμα p-διαστάσεων), U_i είναι ένα διάνυσμα q-διαστάσεων (q<p) που περιέχει τους παράγοντες, B είναι ένας $p \times q$ πίνακας όπου τα στοιχεία του είναι οι επιβαρύνσεις (loadings) των παραγόντων πάνω στις μεταβλητές, μ είναι το διάνυσμα των μέσων και e_i είναι το σφάλμα, το μέρος δηλαδή της μεταβλητής το οποίο δεν μπορεί να εξηγηθεί από τους παράγοντες. Για τους παράγοντες θεωρούμε ότι ισχύει $U_i \overset{i.i.d.}{\sim} N(0, I_q)$ και για τα σφάλματα $e_i \overset{i.i.d.}{\sim} N(0, D)$, όπου D είναι ένας διαγώνιος πίνακας $D = diag(\sigma_1^2,...,\sigma_p^2)$. Επίσης υποθέτουμε ότι οι παράγοντες είναι

ανεξάρτητοι των σφαλμάτων, δηλαδή ισχύει $\mathrm{cov}(U_k, e_i) = 0, \ \forall i \neq k$. Επειδή ακριβώς υποθέτουμε ότι οι παράγοντες είναι ανεξάρτητοι μεταξύ τους, το μοντέλο αυτό της παραγοντικής ανάλυσης λέγεται ορθογώνιο. Με βάση τις παραπάνω υποθέσεις έχουμε ότι $Y_i \,|\, U_i \sim N_p(\mu + BU_i, D)$. Άρα οι παρατηρήσεις μας θα κατανέμονται κανονικά με μέσο μ και διασπορά

$$\Sigma = \mathrm{BB}^{\mathrm{T}} + D \tag{3.1}$$

δηλαδή θα ισχύει $Y_i \sim N(\mu, BB^T + D)$. Αυτό ισχύει επειδή:

$\Sigma = Cov(Y) = Cov(\mu + \mathrm{B}U + e) = BCov(U)B^T + Cov(e) = BIB^T + D = BB^T + D$.

Φυσικά μπορούμε να κεντροποιήσουμε τις παρατηρήσεις μας, οπότε τότε καταλήγουμε ότι $Y_i \sim N(0, BB^T + D)$. Αυτό που στην ουσία έχουμε υποθέσει εδώ και από το οποίο προκύπτουν τα υπόλοιπα είναι ότι το από κοινού διάνυσμα των Y_i, U_i ακολουθεί πολυμεταβλητή κανονική κατανομή, δηλαδή $\begin{pmatrix} Y_i \\ U_i \end{pmatrix} \sim N_{p+q}(\mu^*, \Sigma)$,

όπου $\mu^* = (\mu, 0)$ και $\Sigma = \begin{bmatrix} BB^T + D & B \\ B^T & I_q \end{bmatrix}$. Περισσότερες λεπτομέρειες για την παραγοντική ανάλυση μπορούν να βρεθούν στο Καρλής (2005).

Αν επιλέξουμε το q να είναι πολύ μικρότερο από το p, η αναπαράσταση του πίνακα διασποράς Σ μέσω της (3.1) επιβάλει κάποιους περιορισμούς στα στοιχεία του πίνακα Σ και κατά συνέπεια μειώνεται έτσι ο αριθμός των παραμέτρων προς εκτίμηση. Ωστόσο, πρέπει να σημειωθεί ότι στην περίπτωση που $q > 1$, υπάρχουν άπειρες επιλογές για τον πίνακα B, αφού η (3.1) θα ικανοποιείται αν ο B αντικατασταθεί από τον πίνακα BC, όπου C ένας οποιοσδήποτε ορθογώνιος πίνακας τάξης q, και αυτό γιατί $(BC)(BC)^T + D = BCC^T B^T + D = BB^T + D$ (επειδή C ορθογώνιος ισχύει $C^T = C^{-1}$). Ένας αυθαίρετος τρόπος για να ορίσουμε μοναδικά τον B είναι να διαλέξουμε τον ορθογώνιο πίνακα C έτσι ώστε ο πίνακας $B^T D^{-1} B$ να είναι διαγώνιος και τα στοιχεία του να είναι σε φθίνουσα σειρά. Υποθέτοντας ότι οι ιδιοτιμές του πίνακα BB^T είναι θετικές και διαφορετικές μεταξύ τους, η παραπάνω συνθήκη για τον πίνακα $B^T D^{-1} B$ επιβάλει $\frac{1}{2}q(q-1)$ περιορισμούς στις παραμέτρους προς εκτίμηση του πίνακα $\Sigma = \mathrm{BB}^{\mathrm{T}} + D$. Έτσι ο αριθμός των παραμέτρων προς

εκτίμηση ισούται πλέον με $pq + p - \frac{1}{2}q(q-1)$, γιατί χρειαζόμαστε pq παραμέτρους για να προσδιορίσουμε τον πίνακα B, p παραμέτρους για να προσδιορίσουμε τον D, μείον $\frac{1}{2}q(q-1)$ περιορισμούς που έχουμε επιβάλει στις παραμέτρους.

Άρα, για να επιστρέψουμε στο πρόβλημα της ομαδοποίησης δεδομένων μέσω πεπερασμένου μείγματος πολυμεταβλητών κανονικών κατανομών, ενώ προηγουμένως είχαμε να εκτιμήσουμε $\frac{1}{2}p(p+1)$ παραμέτρους για τον πίνακα διασποράς κάθε ομάδας, τώρα μέσω της εφαρμογής της παραγοντικής ανάλυσης έχουμε να εκτιμήσουμε $pq + p - \frac{1}{2}q(q-1)$ παραμέτρους. Έτσι, αν επιλέξουμε το q να είναι αρκετά μικρότερο από το p, ώστε η διαφορά $C = \frac{1}{2}p(p+1) - \left\{ pq + p - \frac{1}{2}q(q-1) \right\} = \frac{1}{2}\left\{ (p-q)^2 - (p+q) \right\}$ να είναι θετική, τότε μειώνουμε τον αριθμό των παραμέτρων που πρέπει να εκτιμήσουμε.

Το πρόβλημα πλέον βρίσκεται στο να εκτιμήσουμε τους πίνακες B και D. Ας πάρουμε την περίπτωση όπου έχουμε μία μόνο ομάδα και θέλουμε να εφαρμόσουμε παραγοντική ανάλυση. Τότε, η εκτίμηση των πινάκων B και D γίνεται με χρήση του EM αλγορίθμου επαναληπτικά, καθώς δεν υπάρχει λύση σε κλειστή μορφή για τις εκτιμήσεις μέγιστης πιθανοφάνειας των B και D. Αυτό που πρέπει να σημειωθεί είναι ότι για τον υπολογισμό της log-likelihood $\log L(\Psi) = -\frac{1}{2}n\left\{ \log\left|BB^T + D\right| + \sum_{i=1}^{n}\left(y_i - \mu\right)^{\mathrm{T}}\left(BB^T + D\right)^{-1}\left(y_i - \mu\right) \right\}$ (η log-likelihood εδώ είναι χωρίς το σταθερό όρο) δε χρειάζεται να υπολογίζουμε άμεσα τον αντίστροφο του $p \times p$ πίνακα $\Sigma = \mathrm{BB}^{\mathrm{T}} + D$, ο οποίος μάλιστα μπορεί να μην υπολογίζεται για περιπτώσεις όπου $n \ll p$ εξαιτίας αριθμητικών προβλημάτων (π.χ. μηδενική ορίζουσα), αλλά μπορούμε να τον υπολογίσουμε έμμεσα μέσω της ταυτότητας του Woodbury

$$\left(\mathrm{BB}^{\mathrm{T}} + D\right)^{-1} = D^{-1} - D^{-1}B\left(I_q + B^T D^{-1}B\right)^{-1} B^T D^{-1} \qquad (3.2)$$

Με αυτό τον τρόπο αντί να αντιστρέφουμε $p \times p$ πίνακες που είναι υπολογιστικά δύσκολο και χρονοβόρο, αντιστρέφουμε $q \times q$ πίνακες $\left(\tau ους \ \left(I_q + B^T D^{-1}B\right)^{-1}\right)$.

Επίσης, η ορίζουσα του $\mathrm{BB}^{\mathrm{T}}+D$ μπορεί να υπολογιστεί ως εξής: $\left|\mathrm{BB}^{\mathrm{T}}+D\right|=\left|D\right|/\left|I_q-B^T\left(BB^T+D\right)^{-1}B\right|$.

3.3 Mixtures of factor analyzers

Αντί να εφαρμόσουμε το ίδιο μοντέλο παραγοντικής ανάλυσης πάνω σε όλα τα δεδομένα, το εφαρμόζουμε στα δεδομένα κάθε ομάδας ξεχωριστά. Και αυτό γιατί ένα μόνο μοντέλο παραγοντικής ανάλυσης πάνω σε όλα τα δεδομένα, παρέχει ένα γραμμικό μοντέλο για την αναπαράσταση των δεδομένων σε ένα χώρο μικρότερων διαστάσεων, κάτι το οποίο μπορεί να είναι μη αποδοτικό και να μην ισχύει στην πράξη. Σε αντίθεση, μπορούμε να πάρουμε ένα καθολικό μη γραμμικό μοντέλο αν εφαρμόσουμε την παραγοντική ανάλυση ξεχωριστά μέσα σε κάθε ομάδα. Έτσι λοιπόν, η κάθε παρατήρηση $\mathrm{Y_i}$ μοντελοποιείται πλέον ως: $Y_i=\mu_j+B_jU_{ij}+e_{ij}$, όπου $i=1,...,n$ είναι οι διαφορετικές παρατηρήσεις και $j=1,...,g$ είναι οι διαφορετικές ομάδες. Για τους παράγοντες $U_{1j},...,U_{nj}$ υποθέτουμε ότι ακολουθούν $N\left(0,I_q\right)$, είναι ανεξάρτητοι από τα σφάλματα $\mathrm{e_{ij}}$, τα οποία ακολουθούν $N\left(0,D_j\right)$ κατανομή, όπου $\mathrm{D_j}$ είναι ένας διαγώνιος πίνακας. Κατά συνέπεια το μοντέλο που περιγράφει τα δεδομένα είναι

$$f\left(y_i\mid\Psi\right)=\sum_{j=1}^{g}\pi_j\varphi\left(y_i\mid\mu_j,\Sigma_j\right) \tag{3.3}$$

$$\text{όπου}\quad \Sigma_j=\mathrm{B}_j\mathrm{B}_j^T+D_j,\quad j=1,...,g \tag{3.4}$$

$\mathrm{B_j}$ ο αντίστοιχος $p\times q$ πίνακας επιβαρύνσεων. Το διάνυσμα των παραμέτρων Ψ τώρα αποτελείται από τα στοιχεία των μ_j, $\mathrm{B_j}$, $\mathrm{D_j}$ μαζί με τις πιθανότητες $\pi_j\ \left(j=1,...,g-1\right)$, καθώς $\pi_g=1-\sum_{j=1}^{g-1}\pi_j$.

Το μοντέλο της μίξης των factor analyzers (3.3) παίζει καθοριστικό ρόλο στη μοντελοποίηση high-dimensional δεδομένων με τη χρήση πολυμεταβλητών κανονικών κατανομών. Εάν εφαρμόσουμε ένα μείγμα πολυμεταβλητών κανονικών κατανομών χωρίς να θέσουμε κανένα περιορισμό στους πίνακες διασποράς Σ_j τότε

για τον κάθε ένα πίνακα πρέπει να εκτιμήσουμε $\frac{1}{2}p(p+1)$ παραμέτρους. Αυτό σημαίνει πως όσο ο αριθμός των ομάδων (g) αυξάνεται, ο συνολικός αριθμός των παραμέτρων προς εκτίμηση μπορεί πολύ γρήγορα να γίνει αρκετά μεγάλος σε σχέση με τις παρατηρήσεις (n), οδηγώντας έτσι σε overfitting και κάνοντας το μοντέλο μας να μην είναι ανθεκτικό (robust). Το μοντέλο του μείγματος των factor analyzers παρέχει έναν τρόπο να ελέγξουμε τον αριθμό των παραμέτρων προς εκτίμηση μέσω του "μειωμένου" μοντέλου (3.4) για τους πίνακες διασποράς των ομάδων.

3.4 Ο AECM αλγόριθμος στην προσαρμογή του μείγματος των factor analyzers.

3.4.1 Γενικό πλαίσιο του AECM

Το μοντέλο του μείγματος των factor analyzers (3.3) μπορεί να εφαρμοστεί χρησιμοποιώντας τον αλγόριθμο AECM (alternating expectation conditional maximization) για την εκτίμηση των παραμέτρων. Όπως είδαμε στη 2.11 παράγραφο, ο expectation-conditional maximization (ECM) αλγόριθμος αντικαθιστά το M-βήμα του EM αλγορίθμου από έναν αριθμό πιο υπολογιστικά απλών δεσμευμένων βημάτων (CM-steps). Ο AECM είναι μια επέκταση του ECM, όπου ο προσδιορισμός των complete-data επιτρέπεται να είναι διαφορετικός σε κάθε CM-βήμα. Για να εφαρμόσουμε τον AECM αλγόριθμο στο μοντέλο της μίξης των factor analyzers, διαμερίζουμε το διάνυσμα των αγνώστων παραμέτρων Ψ ως $\left(\Psi_1^{\mathrm{T}}, \Psi_2^{\mathrm{T}}\right)^{\mathrm{T}}$, όπου το διάνυσμα Ψ_1 περιέχει τις πιθανότητες $\pi_j \ \ (j=1,...,g-1)$ και τα στοιχεία των διανυσμάτων των μέσων $\mu_j \ \ (i=1,...,g)$. Το διάνυσμα Ψ_2 περιέχει τα στοιχεία των πινάκων B_j και $D_j \ \ (j=1,...,g)$.

Η log-likelihood του μοντέλου είναι $\log L(\Psi) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{g} \pi_j \varphi\left(y_i \mid \mu_j, \Sigma_j\right) \right\}$, όπου y_i οι παρατηρήσεις μας. Όπως είχαμε κάνει και στο κεφάλαιο 2, για κάθε παρατήρηση y_i θεωρούμε το διάνυσμα z_i το οποίο μας υποδεικνύει από ποια ομάδα προέρχεται η συγκεκριμένη παρατήρηση. Τα z_i όμως δεν τα γνωρίζουμε και κατά συνέπεια τα θεωρούμε ως missing data. Με Z_i συμβολίζουμε τη μεταβλητή που μας υποδεικνύει σε ποια ομάδα ανήκει η i παρατήρηση και με z_i συμβολίζουμε την παρατηρημένη τιμή, την οποία όμως δε γνωρίζουμε. Σε αυτό το πλαίσιο το $z_{ij} = \left(z_i\right)_j$ είναι 1 ή 0, ανάλογα με το αν η παρατήρηση y_i προέκυψε ή όχι από την j ομάδα $\left(i = 1,...,n, \;\; j = 1,...,g\right)$. Η δεσμευμένη αναμενόμενη τιμή της μεταβλητής Z_{ij} δοθέντος της παρατήρησης y_i, είναι η εκ των υστέρων πιθανότητα η i παρατήρηση να προέρχεται από την j ομάδα και δίνεται από τον τύπο: $\tau_{ij} = \tau_j\left(y_i \mid \Psi\right) = P\left\{Z_{ij} = 1 \mid y_i\right\} = \dfrac{\pi_j \varphi\left(y_i \mid \mu_j, \Sigma_j\right)}{\sum_{k=1}^{g} \pi_k \varphi\left(y_i \mid \mu_k, \Sigma_k\right)}$, όπου ο πίνακας Σ_j έχει τη μορφή της (3.4) και $i = 1,...,n, \;\; j = 1,...,g$. Με $\Psi^{(k)} = \left(\Psi_1^{(k)^T}, \Psi_2^{(k)^T}\right)^T$ συμβολίζουμε την τιμή των παραμέτρων Ψ μετά την k επανάληψη του AECM αλγορίθμου. Κάθε επανάληψη του αλγορίθμου αποτελείται από δύο κύκλους, και σε κάθε κύκλο εκτελείται ένα E-βήμα και ένα CM-βήμα. Τα δύο CM-βήματα αντιστοιχούν στα δύο διανύσματα Ψ_1 και Ψ_2, τα οποία συνιστούν το διάνυσμα παραμέτρων Ψ. Δηλαδή, στον πρώτο κύκλο θεωρούμε ως missing-data τα διανύσματα z_i και στο CM-βήμα εκτιμούμε τις παραμέτρους π_j και μ_j του Ψ_1 και στο δεύτερο κύκλο θεωρούμε ως missing-data τα διανύσματα z_i και τους παράγοντες και στο CM-βήμα εκτιμούμε τους πίνακες B_j και D_j του Ψ_2.

3.4.2 Πρώτος κύκλος

Στον πρώτο κύκλο του AECM αλγορίθμου, ορίζουμε ως missing data τα διανύσματα $z_1,\ldots,z_n$, τα οποία μας υποδεικνύουν από ποια ομάδα προέρχεται κάθε

παρατήρηση. Κατά συνέπεια η log-likelihood όλων των δεδομένων (complete data log-likelihood) δίνεται από $\log L_c(\Psi) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij} \log\left\{\pi_j \varphi\left(y_i \mid \mu_j, \Sigma_j\right)\right\}$.

Όπως και στον ΕΜ αλγόριθμο έτσι και εδώ, στο Ε-βήμα ο AECM υπολογίζει τη δεσμευμένη αναμενόμενη τιμή της complete-data log-likelihood δοθέντος των παρατηρημένων δεδομένων y. Πάντα και εδώ ξεκινάμε δίνοντας κάποιες αρχικές τιμές $\Psi^{(0)}$ στις παραμέτρους Ψ. Τότε, στο Ε-βήμα του πρώτου κύκλου της k+1 επανάληψης η δεσμευμένη αναμενόμενη τιμή της complete-data log-likelihood γράφεται ως $Q_1\left(\Psi \mid \Psi^{(k)}\right) = \mathrm{E}_{\Psi^{(k)}}\left\{\log L_c(\Psi) \mid y\right\}$. Το Ε-βήμα επιτυγχάνεται απλά αντικαθιστώντας κάθε z_{ij} με τη δεσμευμένη αναμενόμενη τιμή του, δοθέντος των παρατηρημένων δεδομένων και των εκτιμήσεων των παραμέτρων μετά την k επανάληψη. Δηλαδή αντικαθιστούμε τα z_{ij} με τις εκτιμήσεις τους, $\tau_{ij}^{(k+1/2)} = \tau_j\left(y_i \mid \Psi^{(k)}\right)$.

Το CM-βήμα του πρώτου κύκλου πραγματοποιείται μεγιστοποιώντας την $Q_1\left(\Psi \mid \Psi^{(k)}\right)$ ως προς Ψ, με το Ψ_2 να έχει την τιμή $\Psi_2^{(k)}$. Οπότε στο k+1 βήμα οι εκτιμήσεις των π_j και μ_j που περιέχονται στο διάνυσμα Ψ_1 δίνονται από:

$$\pi_j^{(k+1)} = \sum_{i=1}^{n} \tau_{ij}^{(k+1/2)} / n \tag{3.5}$$

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k+1/2)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k+1/2)}} \quad \text{για } j = 1, \ldots, g \tag{3.6}$$

και όπου

$$\tau_{ij}^{(k+1/2)} = \frac{\pi_j^{(k)} \varphi\left(y_i \mid \mu_j^{(k)}, \Sigma_j^{(k)}\right)}{\sum_{h=1}^{g} \pi_h^{(k)} \varphi\left(y_i \mid \mu_h^{(k)}, \Sigma_h^{(k)}\right)} \tag{3.7}$$

είναι η εκ των υστέρων πιθανότητα η i παρατήρηση να ανήκει στην j ομάδα. Τώρα, μετά το τέλος του πρώτου κύκλου, θέτουμε $\Psi^{(k+1/2)} = \left(\Psi_1^{(k+1)^T}, \Psi_2^{(k)^T}\right)^T$.

3.4.3 Δεύτερος κύκλος

Στο δεύτερο κύκλο, προκειμένου να εκτιμήσουμε το διάνυσμα Ψ_2, το οποίο περιέχει τα στοιχεία των πινάκων B_j και D_j, θεωρούμε ως missing data ξανά τα διανύσματα $z_1,\ldots,z_n$, αλλά καθώς επίσης και τους παράγοντες $u_1,...,u_n$. Τώρα, η complete data log-likelihood δίνεται από $\log L_c(\Psi) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij}\left\{\pi_j \varphi\left(y_i \mid \mu_j + B_j u_{ij}, \Sigma_j\right)\right\}$, όπου μ_j και π_j είναι αυτά που εκτιμήσαμε στον πρώτο κύκλο.

Το Ε-βήμα στο δεύτερο κύκλο στην k+1 επανάληψη, πραγματοποιείται για να υπολογίσουμε την $Q_2\left(\Psi \mid \Psi^{(k+1/2)}\right)$, που είναι η δεσμευμένη αναμενόμενη τιμή της log-likelihood, δοθέντος των παρατηρημένων δεδομένων y. Δηλαδή $Q_2\left(\Psi \mid \Psi^{(k+1/2)}\right) = \mathrm{E}_{\Psi^{(k+1/2)}}\left\{\log L_c(\Psi) \mid y\right\}$. Εδώ, στο Ε-βήμα του 2ου κύκλου στην k+1 επανάληψη υπολογίζουμε ξανά τις εκ των υστέρων πιθανότητες $\tau_{ij}^{(k+1)} = \tau_j\left(y_i \mid \Psi^{(k+1/2)}\right)$, δοθέντος τώρα των εκτιμήσεων των μ_j και π_j που πήραμε από τον πρώτο κύκλο. Επίσης όμως εδώ το Ε-βήμα απαιτεί τον υπολογισμό των δεσμευμένων αναμενόμενων τιμών $E_{\Psi^{(k+1/2)}}\left\{Z_{ij}\left(U_{ij}-\mu_j\right) \mid y_i\right\} = \tau_j\left(y_i \mid \Psi^{(k+1/2)}\right)\gamma_j^{(k)^T}\left(y_i-\mu_j\right)$ και

$E_{\Psi^{(k+1/2)}}\left\{Z_{ij}\left(U_{ij}-\mu_j\right)\left(U_{ij}-\mu_j\right)^T \mid y_i\right\} = \tau_j\left(y_i \mid \Psi^{(k+1/2)}\right)\left\{\gamma_j^{(k)^T}\left(y_i-\mu_j\right)\left(y_i-\mu_j\right)^T\gamma_j^{(k)} + \omega_j^{(k)}\right\}$

όπου $\gamma_j^{(k)} = \left(B_j^{(k)}B_j^{(k)^T} + D_j^{(k)}\right)^{-1} B_j^{(k)}$ και

$\omega_j^{(k)} = I_q - \gamma_j^{(k)^T} B_j^{(k)}$, για $j = 1,...,g$.

Το CM-βήμα σε αυτό το δεύτερο κύκλο πραγματοποιείται μεγιστοποιώντας την $Q_2\left(\Psi \mid \Psi^{(k+1/2)}\right)$ ως προς Ψ, με το Ψ_1 να είναι ίσο με $\Psi_1^{(k+1)}$. Έτσι παίρνουμε εκτιμήσεις των B_j και D_j που περιέχονται στο διάνυσμα Ψ_2 όπως παρακάτω:

$$\mathrm{B}_j^{(k+1)} = V_j^{(k+1)}\gamma_j^{(k)}\left(\gamma_j^{(k)^T}V_j^{(k+1)}\gamma_j^{(k)} + \omega_j^{(k)}\right)^{-1} \tag{3.8}$$

και

$$D_j^{(k+1)} = diag\left(V_j^{(k+1)} - V_j^{(k+1)}\gamma_j^{(k)}B_j^{(k+1)^T}\right) \tag{3.9}$$

όπου

$$V_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k+1)} \left(y_i - \mu_j^{(k+1)}\right)\left(y_i - \mu_j^{(k+1)}\right)^T}{\sum_{i=1}^{n} \tau_{ij}^{(k+1)}}.$$

Αν παραγωγίσουμε άμεσα τη log-likelihood μπορεί να δειχτεί ότι μια εναλλακτική εκτίμηση του πίνακα D_j είναι

$$D_j = diag\left(V_j - B_j B_j^T\right) \tag{3.10}$$

όπου

$$V_j = \frac{\sum_{i=1}^{n} \tau_{ij} \left(y_i - \mu_j\right)\left(y_i - \mu_j\right)^T}{\sum_{i=1}^{n} \tau_{ij}}.$$

Όμως, αν και αυτή η εκτίμηση είναι πολύ ελκυστική εξαιτίας της απλότητάς της, ωστόσο δε χρησιμοποιείται λόγω προβλημάτων στη σύγκλιση του αλγορίθμου (McLachlan et al. 2002).

3.4.4 Παρατηρήσεις

Από την κατασκευή του αλγορίθμου ισχύει ότι $Q_1\left(\Psi^{(k+1/2)} \mid \Psi^{(k)}\right) \geq Q_1\left(\Psi^{(k)} \mid \Psi^{(k)}\right)$ και $Q_2\left(\Psi^{(k+1)} \mid \Psi^{(k+1/2)}\right) \geq Q_2\left(\Psi^{(k+1/2)} \mid \Psi^{(k+1/2)}\right)$ το οποίο μας εξασφαλίζει ότι $L\left(\Psi^{(k+1/2)}\right) \geq L\left(\Psi^{(k)}\right)$ και $L\left(\Psi^{(k+1)}\right) \geq L\left(\Psi^{(k+1/2)}\right)$ αντίστοιχα. Άρα εξασφαλίζεται ότι η (incomplete-data) likelihood $L(\Psi)$ δε μειώνεται μετά από κάθε κύκλο, και άρα μετά από κάθε επανάληψη του AECM αλγορίθμου για οποιεσδήποτε αρχικές τιμές.

Με βάση την (3.10) μπορεί να δειχτεί ότι, αν δεν ταξινομηθούν περισσότερες από q παρατηρήσεις στην j ομάδα του μείγματος πολυμεταβλητών κανονικών κατανομών με βάση τις εκ των υστέρων πιθανότητες τ_{ij}, τότε τα διαγώνια στοιχεία του πίνακα D_j θα είναι κοντά στο μηδέν. Αυτό προκαλεί διάφορα υπολογιστικά προβλήματα στον υπολογισμό της log-likelihood (singularities, spikes). Ένας τρόπος για να

αποφύγουμε αυτά τα προβλήματα είναι να επιβάλουμε έναν κοινό πίνακα $D_j = D \ (j = 1,...,g)$ σε όλες τις ομάδες. Μία άλλη λύση είναι να υιοθετήσουμε κάποια εκ των προτέρων κατανομή (prior) για τους πίνακες D_j, όπως στην μπεϋζιανή προσέγγιση των Fokoué and Titterington (2003).

3.5 Αρχικές τιμές του AECM

Έχει ήδη γίνει μια σύντομη αναφορά στο probabilistic PCA (PPCA) μοντέλο των Tipping και Bishop στην 3.1 παράγραφο. Η προσέγγιση των Tipping και Bishop υιοθετεί ένα μοντέλο το οποίο σχετίζεται στενά με το μοντέλο $Y_j = \mu_i + B_i U_{ij} + e_{ij}$ της παραγοντικής ανάλυσης, και το οποίο είναι το $t = y(x \mid w) + \varepsilon$ (Tipping and Bishop, 1999), όπου t είναι οι παρατηρήσεις, x είναι οι κύριες συνιστώσες, w είναι οι παράμετροι προς εκτίμηση και ε είναι ο θόρυβος. Ανάμεσα σε αυτά τα δύο μοντέλα έχει παρατηρηθεί ότι υπάρχει μια σχέση. Για την ακρίβεια το PPCA μοντέλο είναι μια ειδική περίπτωση του μοντέλου παραγοντικής ανάλυσης $\Sigma_j = B_j B_j^T + D_j \quad (j = 1,...,g),$ το οποίο υποθέτει ομοσκεδαστικότητα των σφαλμάτων e_{ij} μέσα σε κάθε ομάδα, ισχύει δηλαδή $\Sigma_j = B_j B_j^T + d_j I_p \quad (j = 1,...,g)$. Ας θεωρήσουμε την περίπτωση μιας ομάδας.

Στην παραγοντική ανάλυση, ο υπόχωρος που προσδιορίζεται από τις στήλες του πίνακα επιβαρύνσεων B, από τους παράγοντες δηλαδή, δεν αντιστοιχεί γενικά με τον υπόχωρο που παίρνουμε από τις κύριες συνιστώσες της PCA. Ωστόσο, έχει παρατηρηθεί ότι τα στοιχεία του πίνακα B της παραγοντικής ανάλυσης και του πίνακα w της ανάλυσης κύριων συνιστωσών είναι παρόμοια όταν οι εκτιμήσεις των στοιχείων του διαγώνιου πίνακα D είναι σχεδόν ίδιες. Πράγματι, αυτό συμβαίνει όταν ισχύει $D = \sigma^2 \mathrm{I}_p$ (ομοσκεδαστικότητα) και οι $p - q$ μικρότερες ιδιοτιμές του δειγματικού πίνακα διακύμανσης S είναι ακριβώς ίσες. Σε αυτή την ομοσκεδαστική περίπτωση, οι Tipping και Bishop έδειξαν ότι οι εκτιμήσεις μέγιστης πιθανοφάνειας $\hat{B}$ και $\hat{\sigma}^2$ σχετίζονται με την PCA, καθώς $\hat{B}$ είναι ο πίνακας του οποίου οι στήλες περιέχουν τα βαθμιδωτά ιδιοδιανύσματα (αυτά δηλαδή που αντιστοιχούν στις ιδιοτιμές ταξινομημένες από τη μεγαλύτερη προς τη μικρότερη) του δειγματικού

πίνακα διακύμανσης και $\hat{\sigma}^2$ είναι ο μέσος όρος της διασποράς στις διαστάσεις που αφαιρέσαμε. Πιο συγκεκριμένα η log-likelihood του μοντέλου, υπό τον περιορισμό $D = \sigma^2 \mathrm{I}_p$, μεγιστοποιείται από

$$\hat{B} = A\left(\Lambda - \hat{\sigma}^2 \mathrm{I}_q\right)^{1/2} R \tag{3.11}$$

όπου

$A = \left(\underset{\sim}{a}_1, ..., \underset{\sim}{a}_q\right)$, $\Lambda = diag\left(\lambda_1, ..., \lambda_q\right)$, $\hat{\sigma}^2 = \sum_{h=q+1}^{p} \frac{\lambda_h}{p-q}$, με $\underset{\sim}{a}_1, ..., \underset{\sim}{a}_q$ μοναδιαία διανύσματα που αντιστοιχούν στις ιδιοτιμές $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_q$ του δειγματικού πίνακα διακύμανσης και R έναν αυθαίρετο $q \times q$ ορθογώνιο πίνακα.

Το μείγμα των PCAs όπως προτάθηκε από τους Tipping και Bishop έχει και αυτό τη μορφή της (3.3) και όπου τα $\mathrm{D_j}$ τώρα έχουν τη δομή

$$D_j = \sigma_j^2 \mathrm{I}_p \quad \forall \; j = 1, ..., g \tag{3.12}.$$

Με βάση τα προηγούμενα, υπό αυτόν τον ομοσκεδαστικό περιορισμό, το CM-βήμα του AECM για την ανανέωση των εκτιμήσεων των B_j, D_j δεν είναι απαραίτητο, καθώς δοθέντος των παρατηρήσεων που έχουν καταταχθεί στην j ομάδα στην k επανάληψη, βάση του μείγματος των PCAs, τα $B_j^{(k+1)}$ και $\sigma_j^{(k+1)^2}$ δίνονται από τη φασματική ανάλυση του πίνακα $V_j^{(k+1)}$.

Όπως και ο EM αλγόριθμος έτσι και ο AECM απαιτεί αρχικές τιμές στις παραμέτρους του για να λειτουργήσει. Προκειμένου να βρούμε αρχικές τιμές $\Psi^{(0)}$ για τις παραμέτρους Ψ του AECM αλγορίθμου, κάνουμε χρήση της παραπάνω σχέσης της παραγοντικής ανάλυσης με το PPCA μοντέλο (3.12). Υπό το (3.3) μοντέλο μας, αν μια παρατήρηση $\mathrm{y_i}$ μετασχηματισθεί σε $D_j^{-1/2} y_i$, τότε ο πίνακας διακύμανσης της j ομάδας θα έχει τη μορφή $D_j^{-1/2} \Sigma_j D_j^{-1/2} = \left(D_j^{-1/2} B_j\right)\left(D_j^{-1/2} B_j\right)^T + I_p$ η οποία αντιστοιχεί στο PPCA μοντέλο (3.12) με $\sigma_j^2 = 1 \left(j = 1, ..., g\right)$. Υπενθυμίζεται ότι στο PPCA μοντέλο ο πίνακας διακύμανσης κάθε ομάδας έχει τη μορφή $C = \sigma^2 I + WW^T$, άρα εδώ ο πίνακας $D_j^{-1/2} B_j$ αντιστοιχεί στον W και το $\sigma_j^2 = 1$, οπότε και οι δύο εκφράσεις είναι της ίδιας μορφής. Κατά συνέπεια χρησιμοποιώντας κάποιες αρχικές τιμές $D_j^{(0)}$ και $\Sigma_j^{(0)}$ για τους πίνακες $\mathrm{D_j}$ και Σ_j, μπορούμε να πάρουμε αρχικές τιμές $B_j^{(0)}$ για τους πίνακες B_j εφαρμόζοντας την (3.11) ως εξής:

$$B_j^{(0)} = D_j^{(0)^{1/2}} A_j \left(\Lambda_j - \mathrm{I}_q \right)^{1/2} \qquad j = 1,...,g \tag{3.13}$$

όπου οι q στήλες του πίνακα $\mathrm{A_j}$ είναι τα ιδιοδιανύσματα που αντιστοιχούν στις ιδιοτιμές $\lambda_{j1} \geq \lambda_{j2} \geq ... \geq \lambda_{jq}$ του πίνακα $D_j^{(0)^{-1/2}} \Sigma_j^{(0)} D_j^{(0)^{-1/2}}$ και $\Lambda_j = diag\left(\lambda_{j1},...,\lambda_{jq} \right)$. Καθώς όμως σαν αρχικές τιμές $D_j^{(0)}$ και $\Sigma_j^{(0)}$ χρησιμοποιούμε απλά κάποιες εκτιμήσεις που έχουμε, μπορεί να προκύψει αρνητική τιμή στον πίνακα $\Lambda_j - I_q$ κάτι το οποίο είναι πρόβλημα αφού οι τιμές του πίνακα είναι κάτω από ρίζα. Αυτό μπορεί να συμβεί αν για παράδειγμα $\lambda_{jh} < 1$ για κάποιο $h \geq q^*$ όπου $q^* \leq q$. Για να αποφύγουμε αυτό το πρόβλημα, μπορούμε να περιορίσουμε τον αριθμό των παραγόντων q να είναι μικρότερος από q^*. Αυτό που προτιμάται όμως είναι να αντικαθιστούμε στην (3.13) τον πίνακα $\Lambda_j - I_q$ με τον πίνακα $\Lambda_j - \tilde{\sigma}_j^2 I_q$ που δίνει

$$B_j^{(0)} = D_j^{(0)^{1/2}} A_j \left(\Lambda_j - \tilde{\sigma}_j^2 I_q \right)^{1/2} \quad j = 1,...,g \tag{3.14}$$

όπου το $\tilde{\sigma}_j^2$ δίνεται από

$$\hat{\sigma}_j^2 = \sum_{h=q+1}^{p} \frac{\lambda_{jh}}{p-q}.$$

Για να προσδιορίσουμε τον πίνακα $\Sigma_j^{(0)}$, μπορούμε να διαμερίσουμε αυθαίρετα τα δεδομένα μας σε g ομάδες και να πάρουμε ως $\Sigma_j^{(0)}$ το δειγματικό πίνακα διασποράς-συνδιασποράς της j ομάδας (j=1,…,g). Όσο αφορά την επιλογή του πίνακα $D_j^{(0)}$, μπορούμε να πάρουμε ως $D_j^{(0)}$ το διαγώνιο πίνακα που σχηματίζεται από τα διαγώνια στοιχεία του $\Sigma_j^{(0)}$. Σε αυτή την περίπτωση ο πίνακας $D_j^{(0)^{-1/2}} \Sigma_j^{(0)} D_j^{(0)^{-1/2}}$ έχει τη μορφή ενός πίνακα συσχετίσεων. Αυτή η διαδικασία μπορεί να επαναληφθεί πολλές φορές για να πάρουμε μια ποικιλία διαφορετικών αρχικών τιμών.

Μια άλλη προσέγγιση, αν ο αριθμός των μεταβλητών δεν είναι μεγάλος σε σχέση με τον αριθμό των παρατηρήσεων, ώστε να μην έχουμε προβλήματα με την αντιστροφή των πινάκων διακύμανσης των ομάδων, είναι να χρησιμοποιήσουμε περίπου ίσους πίνακες διασποράς μεταξύ των ομάδων, χρησιμοποιώντας ένα πλήθος αρχικών διαμερίσεων. Δηλαδή, κατατάσσουμε αρχικά τις παρατηρήσεις σε ομάδες με τέτοιον τρόπο ώστε οι πίνακες διασποράς των ομάδων να είναι όσο το δυνατό ίσοι μεταξύ τους. Έτσι όμως υπάρχει ο κίνδυνος μια ομάδα να περιέχει λιγότερες παρατηρήσεις απ' ότι ο αριθμός των μεταβλητών αν στο σύνολο των παρατηρήσεων

το p είναι μεγάλο σε σχέση με το n (όχι αναγκαστικά μεγαλύτερο από το n αλλά απλά μεγάλο). Η εκτίμηση του κοινού πίνακα Σ που παίρνουμε με αυτό τον τρόπο μπορεί να χρησιμοποιηθεί ως αρχική τιμή $\Sigma_j^{(0)}$. Αν το p είναι μεγάλο σε σχέση με το n, μπορούμε να περιορίσουμε τους πίνακες διασποράς να είναι διαγώνιοι (McLachlan and Peel (2000a)).

Ένας εναλλακτικός τρόπος προσδιορισμού των αρχικών τιμών για τους πίνακες B_j είναι να πάρουμε τα στοιχεία τους τυχαία. Για παράδειγμα μπορούμε να πάρουμε τις αρχικές τιμές ως $B_j^{(0)} = W_j \left(\frac{|S|^{1/p}}{q} \right)^{1/2} \quad (j = 1, ..., g)$, όπου κάθε στοιχείο του πίνακα W_j προέρχεται από την τυποποιημένη κανονική κατανομή και S είναι ο δειγματικός πίνακας διακύμανσης των όλων δεδομένων. Επίσης τους πίνακες $D_j^{(0)}$ μπορούμε να τους πάρουμε ως τους διαγώνιους πίνακες που προκύπτουν από τα διαγώνια στοιχεία του δειγματικού S πίνακα. Μόνο που σε αυτή την περίπτωση όλοι οι $D_j^{(0)}$ θα ισούνται μεταξύ τους. Τέλος οι αρχικές τιμές $\mu_j^{(0)}$ για κάθε ομάδα μπορούν να ληφθούν τυχαία από μια πολυμεταβλητή κανονική κατανομή με μέσο το δειγματικό μέσο και πίνακα διασποράς το δειγματικό S.

Ένα επιπλέον ερώτημα που υπάρχει στον AECM αλγόριθμο σε σχέση με τον EM, πέρα από τον αριθμό των ομάδων και το κατάλληλο μοντέλο, (που θα αναπτυχθούν παρακάτω), είναι ο αριθμός των παραγόντων q που πρέπει να χρησιμοποιήσουμε. Μια εκτίμηση για τον αριθμό των παραγόντων δεδομένου του αριθμού των ομάδων μπορούμε να πάρουμε χρησιμοποιώντας το likelihood-ratio test. Η μηδενική υπόθεση είναι η $H_0 : q = q_0$ με εναλλακτική την $H_1 : q = q_0 + 1$. Η στατιστική συνάρτηση $-2\log\lambda$ ακολουθεί ασυμπτωτικά την x^2 κατανομή με $d = g(p - q_0)$ βαθμούς ελευθερίας, και όπου λ είναι ο λόγος των δύο πιθανοφανειών. Ωστόσο, σε περιπτώσεις όπου το n δεν είναι μεγάλο σε σχέση με τον αριθμό των άγνωστων παραμέτρων, τότε προτιμάμε να χρησιμοποιούμε για τον ίδιο σκοπό το BIC κριτήριο. Για να απορρίψουμε τη μηδενική υπόθεση εδώ θα πρέπει να ισχύει $-2\log\lambda > d\log n$.

Κεφάλαιο 4: Οικογένειες μοντέλων για high-dimensional data

4.1 Parsimonious Gaussian Mixture Models (PGMM)

Το μοντέλο (3.3) του μείγματος πολυμεταβλητών κανονικών κατανομών έχει συνολικά προς εκτίμηση $(g-1)+gp+gp(p+1)/2$ παραμέτρους, από τις οποίες οι $gp(p+1)/2$ χρησιμοποιούνται για τον προσδιορισμό των πινάκων διακύμανσης. Ωστόσο, αν αναγκάσουμε τους πίνακες διακύμανσης να είναι ίσοι για όλες τις ομάδες τότε ο αριθμός των παραμέτρων μειώνεται σε $(g-1)+gp+p(p+1)/2$ (EEE μοντέλο). Για την ακρίβεια, στην παράγραφο 2.6 παρουσιάστηκαν διάφοροι περιορισμοί που μπορούν να τεθούν στους πίνακες διακύμανσης Σ_j μέσω της φασματικής ανάλυσης αυτών, οδηγώντας έτσι σε μια οικογένεια μοντέλων (MCLUST) που απαιτεί προς εκτίμηση από έναν ελάχιστο αριθμό παραμέτρων $(g-1)+gp+1$, μέχρι έναν μέγιστο $(g-1)+gp+gp(p+1)/2$. Όμως, όπως αναφέρθηκε στην 3.1 παράγραφο, τα "πλούσια" μοντέλα EEE, EEV, VEV, VVV της οικογένειας MCLUST δεν μπορούν να εφαρμοστούν σε high-dimensional δεδομένα εξαιτίας τόσο των πολλών παραμέτρων που απαιτούν προς εκτίμηση, όσο και εξαιτίας αριθμητικών προβλημάτων που προκύπτουν στον υπολογισμό της log-likelihood. Απ' την άλλη, τα φειδωλά μοντέλα EII, VII, EEI, VEI, EVI, VVI της MCLUST αν και απαιτούν λιγότερες παραμέτρους, δεν είναι πάντα κατάλληλα καθώς δεν παρέχουν μεγάλη ευελιξία.

Το μοντέλο της cluster analysis που αναπτύχθηκε στο κεφάλαιο 3 βάση του AECM αλγορίθμου και βασίζεται στη σχέση (3.4) απαιτεί $(g-1)+gp+g\left[pq+p-q(q-1)/2\right]$ παραμέτρους προς εκτίμηση. Οι McNicholas and Murphy (2008) ανέπτυξαν μια νέα οικογένεια μοντέλων επιβάλλοντας διάφορους περιορισμούς στους πίνακες B_j και D_j. Ανάγκασαν τους πίνακες αυτούς να είναι ίσοι ή διαφορετικοί μεταξύ των ομάδων καθώς επίσης επέβαλαν και περεταίρω περιορισμούς στους πίνακες D_j οδηγώντας έτσι σε μια οικογένεια 8 μοντέλων. Σε αυτά τα μοντέλα οι πίνακες διακύμανσης μπορεί να έχουν από $pq-q(q-1)/2+1$

παραμέτρους προς εκτίμηση, μέχρι $g\left[pq+p-q(q-1)/2\right]$, όπου $q=0,1,2,...,p$. Αυτή η νέα οικογένεια μοντέλων ονομάζεται parsimonious Gaussian mixture models (PGMMs).

Πιο συγκεκριμένα, εκτός από τον περιορισμό οι πίνακες B_j και D_j να είναι ίσοι ή όχι μεταξύ των ομάδων, έθεσαν τον επιπλέον περιορισμό να ισχύει $D_j = d_j I_p$, σε κάθε ομάδα δηλαδή να ισχύει η ομοσκεδαστικότητα. Τα οχτώ μοντέλα που προέκυψαν παρουσιάζονται στον πίνακα 4.1 παρακάτω.

Πίνακας 4.1. Μοντέλα της PGMM οικογένειας

Model ID	$\boldsymbol{B_j = B}$	$\boldsymbol{D_j = D}$	**Isotropic** $\boldsymbol{D_j = d_j I_p}$	**Covariance Structure**	**Covariance Parameters**
CCC	C	C	C	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + dI_p$	$\left[pq-q(q-1)/2\right]+1$
CCU	C	C	U	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + D$	$\left[pq-q(q-1)/2\right]+p$
CUC	C	U	C	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + d_j I_p$	$\left[pq-q(q-1)/2\right]+g$
CUU	C	U	U	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + D_j$	$\left[pq-q(q-1)/2\right]+gp$
UCC	U	C	C	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + dI_p$	$g\left[pq-q(q-1)/2\right]+1$
UCU	U	C	U	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + D$	$g\left[pq-q(q-1)/2\right]+p$
UUC	U	U	C	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + d_j I_p$	$g\left[pq-q(q-1)/2\right]+g$
UUU	U	U	U	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + D_j$	$g\left[pq-q(q-1)/2\right]+gp$

C: constrained, U: unconstrained.

Το γράμμα C δηλώνει την εφαρμογή του εκάστοτε περιορισμού, ενώ το U τη μη εφαρμογή του. Τα ονόματα των μοντέλων προκύπτουν από τα γράμματα που δηλώνουν τον κάθε περιορισμό με τη σειρά που αυτοί εμφανίζονται στον πίνακα 4.1. Έτσι, το CCC μοντέλο έχει σε ισχύ και τους τρεις περιορισμούς. Οι πίνακες επιβαρύνσεων B_j είναι ίσοι για όλες τις ομάδες καθώς επίσης και οι πίνακες D_j. Επιπλέον, τα στοιχεία του πίνακα D επιβάλλεται να είναι ίσα μεταξύ τους. Έτσι, για αυτό το μοντέλο απαιτούνται να εκτιμήσουμε $\left[pq-q(q-1)/2\right]$ παραμέτρους για όλους τους πίνακες B_j (αφού όλοι οι B_j ισούνται με B) και 1 παράμετρο, την d, για όλους τους πίνακες D_j (αφού όλοι οι D_j ισούνται με $D = dI_p$). Άρα, καταλήγουμε σε ένα συνολικό αριθμό $\left[pq-q(q-1)/2\right]+1$ παραμέτρων προς εκτίμηση όπως αποτυπώνεται και στην τελευταία στήλη του πίνακα 4.1.

Στο CCU μοντέλο δεν ισχύει ο περιορισμός $D_j = d_j I_p$, οπότε τα στοιχεία των πινάκων D_j επιτρέπεται να είναι διαφορετικά μεταξύ τους. Άρα, για κάθε πίνακα D_j πρέπει να εκτιμήσουμε p παραμέτρους. Καθώς όμως εδώ όλοι οι πίνακες D_j είναι ίσοι μεταξύ τους, έχουμε να εκτιμήσουμε μόνο p παραμέτρους για όλους τους πίνακες D_j=D. Μαζί με τις $\left[pq - q(q-1)/2\right]$ παραμέτρους που απαιτούνται για όλους τους πίνακες B_j (αφού όλοι οι B_j ισούνται με B) οδηγούμαστε σε ένα συνολικό αριθμό $\left[pq - q(q-1)/2\right] + p$ παραμέτρων προς εκτίμηση. Με παρόμοια λογική προκύπτουν και οι παράμετροι προς εκτίμηση των υπόλοιπων μοντέλων.

Τρία από τα μοντέλα του πίνακα 4.1, τα UCU, UUC και UUU έχουν αναφερθεί και στις προηγούμενες παραγράφους και είχαν εμφανιστεί πριν την ανάπτυξη της οικογένειας PGMM από τους McNicholas and Murphy (2008). Το UCU μοντέλο, που υποθέτει κοινό πίνακα ιδιαιτεροτήτων D_j=D για όλες τις ομάδες, είχε αναπτυχθεί από τους Ghahramani and Hinton (1997) και το συναντήσαμε στις παραγράφους 3.2 και 3.4.4. Το UUU, που συνιστά τη γενικότερη περίπτωση αυτής της οικογένειας μοντέλων, αναπτύχθηκε από τους McLachlan και Peel (2000), και παρουσιάστηκε εκτενώς στην 3.4 παράγραφο. Τέλος, οι Tipping and Bishop (1999) είχαν προτείνει το μοντέλο των probabilistic principal component analyzers το οποίο υποθέτει τον περιορισμό $D_j = d_j I_p$ του UUC μοντέλου, όπως αναφέρθηκε και στην 3.2. Τα υπόλοιπα 5 μοντέλα είναι καινούργια, και πρέπει να σημειωθεί ότι επιβάλλοντας περιορισμούς στους πίνακες επιβαρύνσεων B_j, κάτι το οποίο δεν είχε εφαρμοστεί προηγουμένως, μειώνει δραματικά τον αριθμό των παραμέτρων και κατά συνέπεια παράγει φειδωλά μοντέλα. Επίσης, επιτρέπεται να θέσουμε q=0. Όταν όμως έχουμε q=0, τότε τα μοντέλα της οικογένειας PGMM ισούνται με τα μοντέλα EII, EEI, VII και VVI της MCLUST οικογένειας.

Για την εκτέλεση όλων αυτών των μοντέλων χρησιμοποιείται ο AECM αλγόριθμος. Λεπτομερείς περιγραφές για τον τρόπο με τον οποίο ο αλγόριθμος αυτός εφαρμόζεται στα μοντέλα της οικογένειας PGMM καθώς επίσης και μαθηματικοί υπολογισμοί παρέχονται από τους McNicholas and Murphy (2008). Πάντως, σε όλα τα μοντέλα η διαδικασία εκτίμησης των παραμέτρων είναι η ίδια. Όταν εκτιμούμε τις παραμέτρους π_j και μ_j στον πρώτο κύκλο του AECM σαν missing-data θεωρούμε τα διανύσματα z_i $(i = 1, ..., n)$, που μας υποδεικνύουν σε ποια ομάδα ανήκει κάθε παρατήρηση, και όταν εκτιμούμε τους πίνακες B_j και D_j στο δεύτερο κύκλο του

AECM σαν missing-data θεωρούμε και τα διανύσματα z_i και τους παράγοντες U. Στο τέλος οι εκ των υστέρων πιθανότητες τ_{ij} (παράγραφος 3.4.2) χρησιμοποιούνται για να κατατάξουμε τις παρατηρήσεις σε ομάδες.

Τέλος, αν συγκρίνουμε τα μοντέλα της οικογένειας PGMM με αυτά της MCLUST βλέπουμε ότι ο αριθμός παραμέτρων που χρειάζεται να εκτιμήσουμε για τα πρώτα, είναι της τάξης των $O(p)$ παραμέτρων, ενώ αυτά της MCLUST είναι της τάξης των $O(p^2)$. Δηλαδή, ο αριθμός παραμέτρων προς εκτίμηση στην PGMM αυξάνεται γραμμικά ως προς τον αριθμό μεταβλητών ενώ αυξάνεται τετραγωνικά στην MCLUST. Οπότε γίνεται αντιληπτό το συγκριτικό πλεονέκτημα της οικογένειας PGMM ως προς τον αριθμό παραμέτρων προς εκτίμηση.

4.2 Υπολογιστικά θέματα

4.2.1 Αρχικές τιμές για την οικογένεια PGMM

Στην παράγραφο 3.6 είδαμε μερικούς τρόπους για να δώσουμε αρχικές τιμές στον AECM αλγόριθμο για το UUU μοντέλο. Ωστόσο, μια συνήθης πρακτική για τα μοντέλα της οικογένειας PGMM (εκτός του CCC) είναι να χρησιμοποιούν ως αρχικές τιμές τα αποτελέσματα του CCC μοντέλου. Δηλαδή, για να τρέξουμε για παράδειγμα ένα UUC μοντέλο, πρώτα απαιτείται να τρέξουμε ένα CCC μοντέλο και τα αποτελέσματα αυτού να τα χρησιμοποιήσουμε ως αρχικές τιμές για να τρέξουμε εν τέλει το UUC μοντέλο. Το ερώτημα όμως μετατίθεται σχετικά με το τι αρχικές τιμές να δώσουμε πλέον στο CCC.

Οι McNicholas and Murphy (2008) προτείνουν την παρακάτω προσέγγιση. Ξεκινάμε δίνοντας τυχαίες αρχικές τιμές στα στοιχεία z_{ij} των διανυσμάτων $z_i \ (i = 1,...,n)$ για κάθε παρατήρηση i, έτσι ώστε $\sum_{j=1}^{g} z_{ij} = 1 \ \forall i = 1,...,n$. Με βάση την ταξινόμηση των παρατηρήσεων που γίνεται από αυτή τη διαμέριση, υπολογίζουμε τις αρχικές τιμές για τους μέσους μ_j ως τους δειγματικούς μέσους από όλες τις παρατηρήσεις που ανήκουν στην κάθε ομάδα και τις πιθανότητες π_j ως $\hat{\pi}_j = \sum_{i=1}^{n} \frac{z_{ij}}{n} \ \forall j = 1,...,g$. Οι αρχικές τιμές για τους πίνακες B_j και D_j προκύπτουν

βάση της φασματικής ανάλυσης των πινάκων Σ_j οι οποίοι εκτιμώνται βάση των αρχικών τιμών z_{ij}. Έτσι, οι αρχικές τιμές για τα στοιχεία του πίνακα B_j καθορίζονται ως $b_{ji} = \sqrt{e_i}\rho_{ji}$, όπου e_i είναι η i μεγαλύτερη ιδιοτιμή του πίνακα Σ_j και ρ_{ji} είναι το j μεγαλύτερο στοιχείο του ιδιοδιανύσματος που αντιστοιχεί στην i μεγαλύτερη ιδιοτιμή του Σ_j. Οι πίνακες D_j καθορίζονται κατόπιν ως $D_j = diag\left\{\Sigma_j - B_j B_j^T\right\}$. Με αυτές τις αρχικές τιμές τρέχει το CCC μοντέλο, τα αποτελέσματα του οποίου χρησιμοποιούνται ως αρχικές τιμές στα υπόλοιπα μοντέλα.

4.2.2 Κριτήρια σύγκλισης

Όπως έχει ήδη αναφερθεί, ο AECM αλγόριθμος είναι ένας επαναληπτικός αλγόριθμος ο οποίος εκτελείται μέχρι να επιτευχθεί σύγκλιση. Ωστόσο, η σύγκλιση αυτή καθ' αυτή, δεν εξασφαλίζει ότι η log-likelihood έχει συγκλίνει σε ένα ολικό μέγιστο και όχι σε κάποιο τοπικό. Για αυτό το λόγο, προκειμένου δηλαδή να βρούμε τη λύση του ολικού μεγίστου, απαιτείται να εκτελούμε τον αλγόριθμο πολλές φορές ξεκινώντας από διαφορετικές αρχικές τιμές.

Διάφορα κριτήρια σύγκλισης έχουν προταθεί. Ένα από αυτά είναι το

$$\frac{l^{(k)} - l^{(k-1)}}{l^{(k)}} < \varepsilon \tag{4.1}$$

όπου $l^{(k)}, l^{(k-1)}$ είναι οι log-likelihood στις επαναλήψεις k $\kappa\alpha\iota$ $k-1$, αντίστοιχα. Δηλαδή, ο αλγόριθμος εκτελείται μέχρι ο λόγος $\frac{l^{(k)} - l^{(k-1)}}{l^{(k)}}$ να γίνει μικρότερος από μια πολύ μικρή ποσότητα ε. Ένα άλλο κριτήριο σύγκλισης βασίζεται στο κριτήριο Aitken. Το κριτήριο του Aitken στην k επανάληψη δίνεται από $a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}$, όπου $l^{(k+1)}, l^{(k)}, l^{(k-1)}$ είναι οι log-likelihood στις επαναλήψεις $k+1,\ k,\ k-1$, αντίστοιχα. Η ασυμπτωτική εκτίμηση της log-likelihood στην $k+1$ επανάληψη είναι $l_\infty^{(k+1)} = l^{(k)} + \frac{1}{1-a^{(k)}}\left(l^{(k+1)} - l^{(k)}\right)$. Αυτή είναι η τιμή της log-likelihood στην οποία ο

αλγόριθμος εκτιμάται ότι συγκλίνει, βάση των τριών τελευταίων επαναλήψεων. Έτσι, ο αλγόριθμος μπορεί να τερματιστεί όταν ισχύει

$$l_{\infty}^{(k+1)} - l^{(k+1)} < \varepsilon \tag{4.2}$$

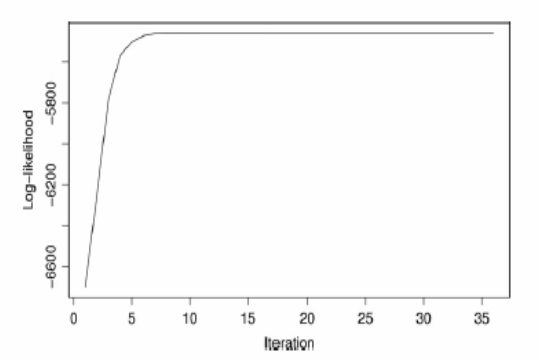
Μια τροποποίηση του Aitken κριτηρίου είναι $l_{\infty}^{(k+1)} - l^{(k)} < \varepsilon$, όπου ε μια πολύ μικρή ποσότητα, όπως για παράδειγμα $\varepsilon = 10^{-2}$.

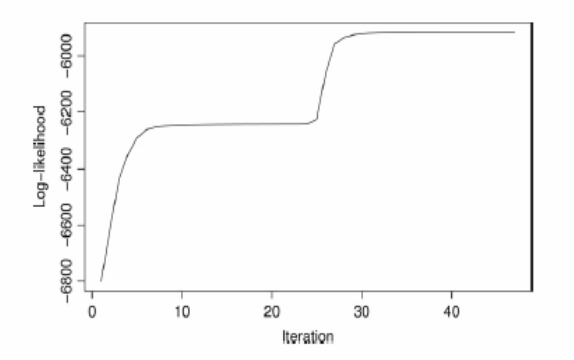
Μερικοί αλγόριθμοι του model-based clustering χρησιμοποιούν τη διαφορά των log-likelihood μεταξύ δύο διαδοχικών βημάτων ως κριτήριο τερματισμού, όπως άλλωστε έχει αναφερθεί στη 2.3 παράγραφο. Σε αυτήν την περίπτωση ο αλγόριθμος θεωρείται ότι έχει συγκλίνει όταν ισχύει

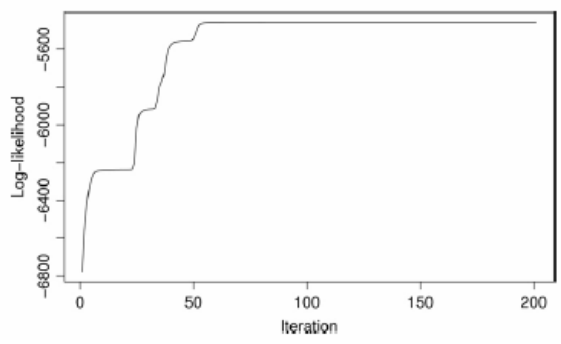
$$l^{(k+1)} - l^{(k)} < \varepsilon \tag{4.3}$$

Το κριτήριο αυτό δεν είναι τόσο κριτήριο σύγκλισης όσο κριτήριο μη αλλαγής κατάστασης από τη μία επανάληψη στην επόμενη. Δηλαδή, η διαμέριση των παρατηρήσεων σε ομάδες παραμένει πρακτικά η ίδια μεταξύ των δύο διαδοχικών επαναλήψεων. Στα παρακάτω γραφήματα φαίνονται τρεις διαφορετικές περιπτώσεις σύγκλισης της log-likelihood συναρτήσει του αριθμού επαναλήψεων.

Γράφημα 4.1. Τρεις διαφορετικές περιπτώσεις σύγκλισης του AECM αλγορίθμου.







Στην περίπτωση των 2 πρώτων γραφημάτων τα κριτήρια (4.2) και (4.3) θα έδιναν παρόμοια αποτελέσματα. Ωστόσο, στην περίπτωση του τελευταίου γραφήματος τα δύο κριτήρια μπορεί να δώσουν διαφορετικά αποτελέσματα. Εκεί, το κριτήριο (4.3) μπορεί να υποεκτιμήσει τη σωστή τιμή της log-likelihood. Οπότε το κριτήριο (4.2) προτιμάται έναντι του (4.3). Αυτά τα κριτήρια σύγκλισης μπορούν κάλλιστα να εφαρμοστούν και για τον EM αλγόριθμο.

4.2.3 Επιλογή μοντέλου

Η επιλογή μοντέλου όπως την είχαμε δει στην παράγραφο 2.7 είχε δύο πτυχές. Πρώτον, έπρεπε να διαλέξουμε τον κατάλληλο τύπο μοντέλου από την MCLUST οικογένεια και δεύτερον να προσδιορίσουμε τον αριθμό των ομάδων. Στα μοντέλα του μείγματος των factor analyzers υπάρχει και μία τρίτη παράμετρος που θα πρέπει να τη διαλέξουμε σωστά και είναι ο αριθμός των παραγόντων. Η επιλογή του κατάλληλου συνδυασμού των τριών παραπάνω παραμέτρων (μοντέλο, αριθμός ομάδων και αριθμός παραγόντων) αντιμετωπίζεται ως ένα πρόβλημα επιλογής μοντέλου (model selection). Για την επιλογή του κατάλληλου τύπου μοντέλου από την οικογένεια PGMM, του κατάλληλου αριθμού ομάδων αλλά και του κατάλληλου αριθμού παραγόντων χρησιμοποιούμε το BIC κριτήριο.

Υπενθυμίζεται ότι το BIC δίνεται από τον τύπο $BIC \approx 2l\left(y,\hat{\theta}\right) - m\log(n)$, όπου m είναι ο αριθμός των παραμέτρων προς εκτίμηση, n ο αριθμός των παρατηρήσεων, και $l\left(y,\hat{\theta}\right)$ η εκτίμηση μέγιστης πιθανοφάνειας. Οι Fraley and Raftery (1998, 2002) έδειξαν ότι στην πράξη το BIC έχει καλή επίδοση σαν κριτήριο για την επιλογή μοντέλου. Ωστόσο, όπως έχει αναφερθεί και στην 2.7 οι υποθέσεις πάνω στις οποίες βασίζεται το BIC μπορεί να μην ισχύουν για την περίπτωση του μείγματος κατανομών και κατά συνέπεια μπορεί το μοντέλο που θα επιλεχθεί βάση του BIC να μην είναι αυτό που δίνει την καλύτερη ομαδοποίηση (βλ. παράγραφο 5.6). Παρ' όλα αυτά το BIC είναι αυτό που χρησιμοποιείται έως τώρα ευρύτατα στην πράξη. Μπορεί οι υποθέσεις του BIC για την επιλογή μοντέλου και αριθμού ομάδων να μην ισχύουν, ισχύουν όμως για την επιλογή του αριθμού παραγόντων για δοθέν αριθμό ομάδων και μοντέλου (McLachlan and Peel Κεφάλαιο 8, 2000a). Επίσης για την επιλογή του

κατάλληλου αριθμού παραγόντων μπορούμε να εφαρμόσουμε likelihood ratio test όπως στην παράγραφο 3.5.

Το BIC μπορεί να χρησιμοποιηθεί ακόμα και για να διαλέξουμε το σωστό μοντέλο μεταξύ της οικογένειας PGMM και της MCLUST. Ακόμα με βάση το BIC μπορούμε να συγκρίνουμε μοντέλα που έχουν προκύψει από μείγματα διαφορετικών πολυμεταβλητών κατανομών (π.χ. κανονικών και student κατανομών). Ωστόσο υπάρχουν και εναλλακτικά κριτήρια του BIC τα οποία έχουν αναφερθεί στην 2.7.

4.3 Παράδειγμα

Σε αυτό το παράδειγμα θα δούμε μια εφαρμογή των μοντέλων της οικογένειας PGMM και θα συγκρίνουμε τα αποτελέσματα αυτής της οικογένειας με την MCLUST οικογένεια καθώς και με μία άλλη τεχνική ομαδοποίησης την variable selection.

Μερικές φορές η ομαδοποίηση των δεδομένων εξαρτάται σε μεγαλύτερο μέρος από μερικές και όχι από όλες τις μεταβλητές. Δηλαδή, μέσα στο σύνολο των διαθέσιμων μεταβλητών μπορεί να υπάρχουν κάποιες οι οποίες δε συνεισφέρουν πληροφορία στην ομαδοποίηση. Για την ακρίβεια τα πράγματα μπορεί να είναι ακόμα χειρότερα, όπως να υπάρχουν κάποιες μεταβλητές οι οποίες όχι μόνο να μη βοηθάνε στην ομαδοποίηση, αλλά να τη δυσχεραίνουν ή να τη χαλάνε κιόλας. Οπότε μια τεχνική επιλογής του κατάλληλου υποσυνόλου από τις αρχικές μεταβλητές πριν την ομαδοποίηση θα ήταν πολύ χρήσιμη. Αυτό ακριβώς προσπαθεί να επιτύχει η variable selection τεχνική (Raftery and Dean (2006), Maugis et al. (2009a, 2009b)). Δηλαδή, από το σύνολο των διαθέσιμων μεταβλητών επιλέγει ένα υποσύνολο και έπειτα προχωρεί σε model-based clustering. Στη συνέχεια επιλέγει διαφορετικό υποσύνολο μεταβλητών και ξαναεκτελεί ομαδοποίηση, κ.ο.κ.. Οι διαφορετικές ομαδοποιήσεις που προκύπτουν από την επιλογή διαφορετικών μεταβλητών συγκρίνονται μεταξύ τους με το BIC κριτήριο ώστε να βρεθεί το κατάλληλο υποσύνολο μεταβλητών. Τέλος, οι διαμερίσεις που προκύπτουν για διαφορετικά μοντέλα και αριθμό ομάδων αλλά για το ίδιο υποσύνολο μεταβλητών, συγκρίνονται μεταξύ τους με το BIC κριτήριο ώστε να βρεθεί η καλύτερη ομαδοποίηση. Η variable selection τεχνική είναι διαθέσιμη στην R μέσω της βιβλιοθήκης clustvarsel (Dean and Raftery 2006).

Τα δεδομένα του παραδείγματος αφορούν 178 παρατηρήσεις και περιγράφουν 27 χημικές και φυσικές ιδιότητες τριών τύπων κρασιών (Barolo, Grignolino, Barbera) από την Ιταλία. Οι 59 παρατηρήσεις ανήκουν στον τύπο Barolo, οι 71 στον τύπο Grignolino και οι 48 στον τύπο Barbera. 13 από τις μεταβλητές αυτών των δεδομένων είναι διαθέσιμες στη βιβλιοθήκη gclus της R (wine data) ενώ όλες οι μεταβλητές είναι διαθέσιμες στη βιβλιοθήκη pgmm. Τα 8 μοντέλα της οικογένειας PGMM εφαρμόστηκαν στα δεδομένα των 27 μεταβλητών για $g = 1, 2, ..., 6$ ομάδες, $q = 1, 2, ..., 6$ παράγοντες και για 3 διαφορετικές τυχαίες αρχικές ταξινομήσεις για το σύνολο των 27 μεταβλητών. Δηλαδή, συνολικά εκτελέστηκαν και συγκρίθηκαν μεταξύ τους βάση του BIC κριτηρίου $8 \times 6 \times 6 \times 3 = 864$ διαφορετικά μοντέλα.

Πίνακας 4.2: Τα καλύτερα 3 μοντέλα, βάση του BIC, για την PGMM οικογένεια στα wine data.

Model	Number of groups	Number of factors	BIC
CUU	3	4	-11454.11
CUU	3	5	-11457.70
CUU	3	6	-11503.95

Το καλύτερο μοντέλο ήταν το CUU για g=3 ομάδες και q=4 παράγοντες με τιμή $BIC = -11454.11$. Η ομαδοποίηση που έγινε βάση αυτού του μοντέλου φαίνεται στον παρακάτω πίνακα 4.3. Βλέπουμε ότι μόνο 1 παρατήρηση έχει ταξινομηθεί εσφαλμένα!

Πίνακας 4.3: Ταξινόμηση για το καλύτερο μοντέλο της PGMM στα wine data.

	Cluster		
	1	2	3
Barolo	59		
Grignolino		70	1
Barbera			48

Η ομαδοποίηση των δεδομένων βάση της οικογένειας MCLUST για $g = 1, 2, ..., 5$ ομάδες έδωσε ως καλύτερο μοντέλο το VVI για g=3 ομάδες με $BIC = -12119.3$. Η ομαδοποίηση βάση αυτού του μοντέλου φαίνεται στον παρακάτω πίνακα 4.4.

Πίνακας 4.4: Ταξινόμηση για το καλύτερο μοντέλο της MCLUST στα wine data.

	Cluster		
	1	2	3
Barolo	58	1	
Grignolino	4	66	1
Barbera			48

Ενώ και η MCLUST οικογένεια ανέδειξε και αυτή τρεις ομάδες, το σωστό αριθμό ομάδων δηλαδή, είναι φανερό ότι η ομαδοποίηση που έκανε είναι λιγότερο καλή από την ομαδοποίηση που έδωσε η PGMM.

Τέλος, στα δεδομένα αυτά εφαρμόστηκε η variable selection τεχνική ομαδοποίησης για $g = 1, 2, ..., 8$ διαφορετικές ομάδες. Η τεχνική αυτή επέλεξε 19 από τις αρχικές 27 μεταβλητές και με βάση αυτές έδωσε ως καλύτερο αποτέλεσμα g=4 ομάδες για το VVI μοντέλο. Η ομαδοποίηση για αυτό το μοντέλο φαίνεται παρακάτω.

Πίνακας 4.5: Ταξινόμηση για το καλύτερο μοντέλο της variable selection στα wine data.

	Cluster			
	1	2	3	4
Barolo	52	7		
Grignolino		17	54	
Barbera		1		47

Πίνακας 4.6: Rand και Adjusted Rand indices για τα καλύτερα μοντέλα κάθε οικογένειας για τα wine data.

Model	Rand Index	Adjusted Rand Index
CCU (PGMM)	0.99	0.98
VVI (GPCM)	0.95	0.90
Variable Selection	0.91	0.78

Συγκρίνοντας τις τρεις μεθόδους ομαδοποίησης σε αυτά τα δεδομένα βλέπουμε ότι οι δύο πρώτες πετυχαίνουν ικανοποιητική ομαδοποίηση και ανιχνεύουν τη δομή που υπάρχει στα δεδομένα ενώ η variable selection δεν τα καταφέρνει ιδιαίτερα ικανοποιητικά. Αυτό προφανώς συμβαίνει γιατί το υποσύνολο των μεταβλητών που διάλεξε δεν είναι τόσο καλό στο να διαχωρίσει τα κρασιά στα είδη τους όσο το πλήρες σετ των μεταβλητών. Οι οικογένειες MCLUST και PGMM βρίσκουν το σωστό αριθμό ομάδων. Ωστόσο, η PGMM ομαδοποίηση είναι πιο ακριβής καθώς

ταξινομεί μόνο μία παρατήρηση λάθος. Επίσης, με βάση τους δείκτες ταξινόμησης Rand Index και Adjusted Rand Index η PGMM οικογένεια παρουσιάζει καλύτερη ταξινόμηση. Επιπλέον, αν συγκρίνουμε τις τιμές του BIC κριτηρίου για τα δύο καλύτερα μοντέλα, CCU της PGMM οικογένειας και VVI της MCLUST ($BIC = -11454.11$, $BIC = -12119.3$ αντίστοιχα) βλέπουμε ότι η πρώτη τιμή είναι υψηλότερη και άρα το CCU είναι καλύτερο. Το BIC της variable selection μεθόδου δεν μπορούμε να το συγκρίνουμε με τα υπόλοιπα γιατί βασίζεται σε λιγότερες μεταβλητές. Τέλος, ενδιαφέρον είναι το γεγονός ότι το μοντέλο με την μεγαλύτερη BIC τιμή έδωσε και την καλύτερη ομαδοποίηση. Ωστόσο, αυτό δεν είναι βέβαιο ότι θα ισχύει πάντα (βλ. παράγραφο 5.6).

Στη συνέχεια τα 8 μοντέλα της PGMM οικογένειας εφαρμόστηκαν για τις 13 από τις συνολικά 27 μεταβλητές που παρέχονται από τη βιβλιοθήκη gclus της R για αυτά τα δεδομένα. Το καλύτερο μοντέλο ήταν το CUU για g=4 ομάδες και q=2 παράγοντες με BIC=-5294.68. Επίσης εφαρμόστηκαν τα μοντέλα της οικογένειας MCLUST, με καλύτερο το VEI μοντέλο με BIC=-5469.95, καθώς και η variable section τεχνική. Η αξιολόγηση της ομαδοποίησης και με τις τρεις μεθόδους φαίνεται στον ακόλουθο πίνακα.

Πίνακας 4.7: Rand, Adjusted Rand indices και αριθμός των ομάδων για τα καλύτερα μοντέλα κάθε οικογένειας για τα wine data (13 variables).

Model	Rand Index	Adjusted Rand Index	g
CUU (PGMM)	0.91	0.79	4
VEI (MCLUST)	0.80	0.48	8
Variable Selection	0.90	0.78	3

Βλέπουμε από τα παραπάνω αποτελέσματα ότι και οι τρεις μέθοδοι δίνουν χειρότερα αποτελέσματα για τις 13 μεταβλητές σε σχέση με αυτά που έδωσαν για το σύνολο των 27 μεταβλητών καθώς επίσης οι δύο πρώτες αποτυγχάνουν να ανιχνεύσουν και το σωστό αριθμό ομάδων. Ωστόσο, και πάλι η ομαδοποίηση που επιτυγχάνεται μέσω της PGMM οικογένειας είναι η καλύτερη, βάση του adjusted Rand Index, και είναι ικανοποιητική. Επίσης, ξανά το μοντέλο με την μεγαλύτερη τιμή BIC (το CUU) έδωσε και καλύτερα αποτελέσματα ομαδοποίησης με βάση τον adjusted Rand Index. Στον ακόλουθο πίνακα φαίνεται η καλύτερη ομαδοποίηση που έγινε με βάση το CUU μοντέλο.

Πίνακας 4.8: Ταξινόμηση για το καλύτερο μοντέλο της PGMM στα wine data (13 μεταβλητές).

	Barolo			Grignolino							Barbera			
Years	71	73	74	70	71	72	73	74	75	76	74	76	78	79
Cluster 1	19	20	20											
Cluster 2				7	8	4	8	8	2	1				
Cluster 3				2	1	2	1	8	7	10				
Cluster 4						1				1	9	5	29	5

Από τον παραπάνω πίνακα είναι εμφανές ότι το CUU μοντέλο έχει χωρίσει τη δεύτερη ομάδα (Grignolino) σε δύο επιμέρους ομάδες.

4.4 Expanded Parsimonious Gaussian Mixture Models (EPGMM)

Είδαμε στην 4.1 πως οι McNicholas and Murphy (2008) δημιούργησαν διάφορα φειδωλά μοντέλα για τα μείγματα των factor analyzers επιβάλλοντας διάφορους περιορισμούς στους πίνακες επιβαρύνσεων και ιδιαιτεροτήτων (B_j και D_j αντίστοιχα) του μοντέλου $\Sigma_j = B_j B_j^T + D_j \quad (j = 1,...,g)$. Ωστόσο, οι οι McNicholas and Murphy (2010) πρότειναν μια περεταίρω παραμετροποίηση του πίνακα διασποράς του παραγοντικού μοντέλου, γράφοντας τους πίνακες ιδιαιτεροτήτων ως $D_j = \omega_j \Delta_j$, όπου $\omega_j \in R^+$ και $\Delta_j = diag\{\delta_1, \delta_2, ..., \delta_p\}$ έτσι ώστε $|\Delta_j| = 1$, για $j = 1,2,...,g$. Αποτέλεσμα αυτής της νέας παραμετροποίησης είναι ο πίνακας διασποράς των επιμέρους ομάδων να μπορεί να γραφεί ως $\Sigma_j = B_j B_j^T + \omega_j \Delta_j \quad (j = 1,...,g)$. Η νέα δομή του πίνακα διασποράς δίνει τη δυνατότητα παραγωγής νέων μοντέλων επιπλέον των 8 της οικογένειας PGMM. Πιο συγκεκριμένα, τα μοντέλα που προκύπτουν από την παραπάνω παραμετροποίηση είναι τα 8 προϋπάρχοντα μοντέλα της PGMM οικογένειας και 4 καινούργια. Έτσι, συνολικά έχουμε 12 μοντέλα τα οποία απαρτίζουν τη νέα οικογένεια μοντέλων Expanded Parsimonious Gaussian Mixture Models (EPGMM) και συνοψίζονται στον ακόλουθο πίνακα.

Πίνακας 4.9: Μοντέλα της EPGMM οικογένειας, δομή του πίνακα διασποράς, αριθμό παραμέτρων του πίνακα διασποράς και ισοδύναμο μοντέλο της PGMM.

EPGMM Model	$\boldsymbol{B}_j = \boldsymbol{B}$	$\Delta_j = \Delta$	$\boldsymbol{\omega}_j = \boldsymbol{\omega}$	$\Delta_j = \mathrm{I}_p$	**PGMM equivalent**	**Covariance Structure**	**Covariance Parameters**
CCCC	C	C	C	C	CCC	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + \omega I_p$	$\left[pq - q(q-1)/2\right] + 1$
CCUC	C	C	U	C	CUC	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + \omega_j I_p$	$\left[pq - q(q-1)/2\right] + g$
UCCC	U	C	C	C	UCC	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + \omega I_p$	$g\left[pq - q(q-1)/2\right] + 1$
UCUC	U	C	U	C	UUC	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + \omega_j I_p$	$g\left[pq - q(q-1)/2\right] + g$
CCCU	C	C	C	U	CCU	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + \omega\Delta$	$\left[pq - q(q-1)/2\right] + p$
CCUU	C	C	U	U	–	$\Sigma_j = \mathrm{BB}^T + \omega_j\Delta$	$\left[pq - q(q-1)/2\right] + \left[g + (p-1)\right]$
UCCU	U	C	C	U	UCU	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + \omega\Delta$	$g\left[pq - q(q-1)/2\right] + p$
UCUU	U	C	U	U	–	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + \omega_j\Delta$	$g\left[pq - q(q-1)/2\right] + \left[g + (p-1)\right]$
CUCU	C	U	C	U	–	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + \omega\Delta_j$	$\left[pq - q(q-1)/2\right] + \left[1 + g(p-1)\right]$
CUUU	C	U	U	U	CUU	$\Sigma_j = \mathrm{BB}^{\mathrm{T}} + \omega_j\Delta_j$	$\left[pq - q(q-1)/2\right] + gp$
UUCU	U	U	C	U	–	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + \omega\Delta_j$	$g\left[pq - q(q-1)/2\right] + \left[1 + g(p-1)\right]$
UUUU	U	U	U	U	UUU	$\Sigma_j = \mathrm{B}_j\mathrm{B}_j^T + \omega_j\Delta_j$	$g\left[pq - q(q-1)/2\right] + gp$

Σημειώνεται ότι και τα 12 μοντέλα της EPGMM οικογένειας έχουν αριθμό παραμέτρων που αυξάνεται γραμμικά ως προς τον αριθμό μεταβλητών. Αυτό όπως έχουμε δει είναι ιδιαίτερα σημαντικό για εφαρμογές σε high dimensional δεδομένα καθώς δεν απαιτείται η εκτίμηση πάρα πολλών παραμέτρων.

Όπως φαίνεται και από τον Πίνακα 4.8 παραπάνω, η νέα παραμετροποίηση οδηγεί στα 8 προϋπάρχοντα μοντέλα της PGMM οικογένειας αλλά δημιουργούνται επίσης και 4 νέα. Ας πάρουμε πρώτα το CCUU. Αυτό το μοντέλο επιτρέπει τα στοιχεία ω_j να είναι διαφορετικά μεταξύ των ομάδων αλλά οι πίνακες Δ_j να είναι ίσοι. Άρα, πρέπει να εκτιμήσουμε g παραμέτρους για τα στοιχεία ω_j και p-1 παραμέτρους για τους διαγώνιους πίνακες $\Delta_j=\Delta$. Σημειώνεται ότι ενώ οι πίνακες Δ_j περιέχουν p στοιχεία, εμείς πρέπει να εκτιμήσουμε μόνο p-1 καθώς το ένα υπολειπόμενο στοιχείο είναι ίσο με 1, υποδεικνύοντας ότι το στοιχείο που ανήκει σε αυτή τη θέση του πίνακα ιδιαιτεροτήτων D_j ισούται με το ω_j. Άρα για τους πίνακες ιδιαιτεροτήτων πρέπει να εκτιμήσουμε g+p-1 παραμέτρους. Ο πίνακας επιβαρύνσεων κάθε ομάδας απαιτεί τον υπολογισμό $pq - q(q-1)/2$ παραμέτρων. Καθώς όμως οι πίνακες B_j είναι ίσοι μεταξύ των ομάδων, συνολικά πρέπει να εκτιμήσουμε $\left[pq - q(q-1)/2\right] + \left[g + (p-1)\right]$ παραμέτρους. Παρόμοια λογική επικρατεί και για τα υπόλοιπα μοντέλα.

Η εκτίμηση των παραμέτρων των μοντέλων γίνεται μέσω του AECM αλγορίθμου με τρόπο ανάλογο όπως στα μοντέλα της PGMM οικογένειας. Οι εκτιμήσεις των παραμέτρων για τα 8 μοντέλα της EPGMM που προϋπήρχαν στην PGMM, προκύπτουν από τις εκτιμήσεις της PGMM οικογένειας γράφοντας $D_j = \left|D_j\right|^{1/p} \frac{D_j}{\left|D_j\right|^{1/p}}$ και θέτοντας μετά $\omega_j = \left|D_j\right|^{1/p}$ και $\Delta_j = \frac{D_j}{\left|D_j\right|^{1/p}}$. Δηλαδή, όπως έχουμε εκτιμήσει τους πίνακες D_j για τα 8 μοντέλα της PGMM, για να βρούμε τα ω_j και Δ_j των ίδιων μοντέλων στην EPGMM οικογένεια απλά θέτουμε $\omega_j = \left|D_j\right|^{1/p}$ και $\Delta_j = \frac{D_j}{\left|D_j\right|^{1/p}}$. Ωστόσο, για να εκτιμήσουμε τις παραμέτρους μέγιστης πιθανοφάνειας των 4 νέων μοντέλων πρέπει να κάνουμε χρήση της μεθόδου των πολλαπλασιαστών Lagrange. Ο τρόπος εκτίμησης των παραμέτρων αυτών των μοντέλων παρουσιάζεται αναλυτικά από τους McNicholas and Murphy (2010).

4.5 Παράδειγμα

Στο παράδειγμα αυτό θα δούμε την εφαρμογή της EPGMM οικογένειας πάνω σε δεδομένα (leukaemia dataset) που αφορούν εκφράσεις γονιδίων (microarray gene expression study) και θα συγκρίνουμε την επίδοσή της με άλλες μεθόδους ομαδοποίησης. Πιο συγκεκριμένα τα δεδομένα αφορούν 72 ασθενείς, 47 με οξεία λεμφοβλαστική λευχαιμία (ALL, acute lymphoblastic leukaemia) και 25 με οξεία μυελοειδή λευχαιμία (AML, acute myeloid leukaemia) από τους οποίους έχει μελετηθεί η γονιδιακή έκφραση 7129 γονιδίων (7129 μεταβλητές). Ωστόσο, στο παράδειγμα αυτό δε χρησιμοποιούνται όλα τα γονίδια αλλά μόνο τα 2030 καθώς τα υπόλοιπα αφαιρέθηκαν για λόγους μείωσης διαστάσεων. Μέσω model-based clustering επιθυμούμε να ομαδοποιήσουμε τους ασθενείς σε δύο ομάδες, όσες είναι δηλαδή και οι πραγματικές ομάδες (ALL, AML), ώστε να δούμε ποια γονίδια εκφράζονται περισσότερο σε κάθε ομάδα. Εδώ είναι μία ιδανική περίπτωση όπου γνωρίζουμε τις πραγματικές ομάδες. Στην πράξη όμως ποτέ δε γνωρίζουμε τις πραγματικές ομάδες.

Τα 12 μοντέλα της EPGMM οικογένειας εφαρμόστηκαν σε αυτά τα δεδομένα για g=2 ομάδες και $q = 1, 2, ..., 6$ παράγοντες, για 10 διαφορετικές αρχικές τιμές. Η τιμή BIC που αντιστοιχεί στην καλύτερη επιλογή του q για κάθε ένα από τα 12 μοντέλα της EPGMM φαίνεται στον παρακάτω πίνακα.

Πίνακας 4.10: BIC τιμή για τα μοντέλα της EPGMM οικογένειας με το καλύτερο q για τα leukaemia δεδομένα.

Model	q	BIC	Model	q	BIC
CCCC	3	-411646.50	CCUC	3	-411566.29
UCCC	1	-416954.56	UCUC	1	-416803.57
CCCU	4	-414615.22	CCUU*	5	-413207.29
UCCU	1	-423354.79	UCUU*	1	-422089.38
CUCU*	4	-413966.90	CUUU	5	-413978.04
UUCU*	1	-423933.46	UUUU	1	-423532.04

*Ενα από τα 4 νέα μοντέλα.

Το καλύτερο από τα 12 μοντέλα, με βάση το BIC κριτήριο, είναι το CCUC με q=3 παράγοντες. Στο μοντέλο αυτό οι συνδιασπορές των γονιδίων είναι ίσες μεταξύ των ομάδων αλλά οι διασπορές είναι διαφορετικές. Ωστόσο, μέσα σε κάθε ομάδα οι διασπορές όλων των γονιδίων είναι ίσες. Η ταξινόμηση που προκύπτει από αυτό το μοντέλο φαίνεται στον Πίνακα 4.11. Παρατηρούμε ότι μόνο 5 ασθενείς ταξινομήθηκαν λανθασμένα κάτι το οποίο δείχνει πολύ καλή επίδοση του μοντέλου.

Πίνακας 4.11: Ομαδοποίηση βάσει του CCUC για τα leukaemia δεδομένα.

	Cluster	
	1	2
ALL	42	0
AML	5	25

Εκτός από τα μοντέλα της οικογένειας EPGMM, στα δεδομένα αυτά εφαρμόστηκαν και άλλοι αλγόριθμοι ομαδοποίησης προκειμένου να μπορέσουμε να αξιολογήσουμε την επίδοση των EPGMM μοντέλων σε σχέση με άλλα προϋπάρχοντα. Χρησιμοποιήθηκαν ιεραρχικές μέθοδοι ομαδοποίησης βασισμένες στην Ευκλείδεια απόσταση, συγκεκριμένα 3 παραλλαγές της μεθόδου, οι single linkage, complete linkage και average linkage (βλ. Καρλής 2005). Επίσης, χρησιμοποιήθηκε ο αλγόριθμος K-means και K-medoids καθώς και τα μοντέλα της

οικογένειας MCLUST. Τα αποτελέσματα συνοψίζονται στον Πίνακα 4.12 και οι δείκτες Rand Index και Adjusted Rand Index χρησιμοποιούνται για τη σύγκριση αυτών των μοντέλων. Αυτοί οι δείκτες δείχνουν ότι η καλύτερη από τις non-model-based μεθόδους είναι ο K-means αλγόριθμος, με adjusted rand index 0.187, ο οποίος ίσα που ξεπέρασε το καλύτερο μοντέλο VII της MCLUST οικογένειας. Ωστόσο, η υπεροχή του CCUC μοντέλου της EPGMM οικογένειας (για q=3) είναι ξεκάθαρη, καθώς δίνει adjusted rand index ίσο με 0.738.

Πίνακας 4.12: Αποτελέσματα αλγορίθμων ομαδοποίησης στα leukaemia δεδομένα.

	BIC	Rand Index	Adjusted Rand Index
Hierarchical (complete)	–	0.532	0.058
Hierarchical (average)	–	0.525	-0.024
Hierarchical (single)	–	0.532	-0.013
K-means	–	0.593	0.187
K-medoids	–	0.518	0.023
MCLUST (VII)	-416293.2	0.593	0.186
EPGMM (CCUC)	-411566.3	0.869	0.738

4.6 Παρατηρήσεις

Τα μοντέλα της οικογένειας EPGMM είναι ιδιαίτερα χρήσιμα για την ανάλυση high-dimensional δεδομένων και υπάρχουν 3 κύριοι λόγοι γι' αυτό. Πρώτον, ο αριθμός παραμέτρων που πρέπει να εκτιμήσουμε για τον πίνακα διασποράς κάθε ομάδας για κάθε μοντέλο αυξάνεται γραμμικά ως προς τον αριθμό μεταβλητών (και όχι τετραγωνικά όπως στην MCLUST οικογένεια) με αποτέλεσμα να έχουμε αρκετά φειδωλά μοντέλα. Δεύτερον, στα μοντέλα αυτά μπορεί να εφαρμοστεί η ταυτότητα του Woodbury (3.2) ώστε να αποφύγουμε την αντιστροφή πινάκων μεγάλων διαστάσεων $(p \times p)$ που είναι υπολογιστικά απαιτητική, υπολογίζοντας αντίστροφους πίνακες μικρότερων διαστάσεων $(q \times q)$. Τρίτον, αν και γενικά οι αλγόριθμοι υπολογισμού αυτών των μοντέλων είναι αργοί, η φύση των μοντέλων μας παρέχει τη δυνατότητα να χρησιμοποιήσουμε τρόπους για να επιταχύνουμε τη διαδικασία εκτίμησης.

Συνοπτικά παρακάτω παρατίθενται μερικά από τα πλεονεκτήματα και μειονεκτήματα των μοντέλων της οικογένειας EPGMM. Ωστόσο, όλα αυτά αφορούν και την PGMM οικογένεια, αφού αυτή είναι ένα υποσύνολο της EPGMM.

- Ο αριθμός των παραμέτρων προς εκτίμηση των πινάκων διασποράς αυξάνεται γραμμικά ως προς τον αριθμό των μεταβλητών, κάτι το οποίο είναι εξαιρετικά χρήσιμο για την εφαρμογή σε high dimensional δεδομένα. Αντίθετα, στο model-based clustering της οικογένειας MCLUST οι παράμετροι προς εκτίμηση στους μη διαγώνιους πίνακες αυξάνονται τετραγωνικά ως προς τον αριθμό των μεταβλητών. Έτσι, έχουμε πιο φειδωλά μοντέλα αλλά και μοντέλα με μεγαλύτερη ευελιξία τα οποία είναι κατάλληλα για εφαρμογή σε high-dimensional δεδομένα.
- Εμφανίζονται ιδιαίτερα καλά στο να μοντελοποιούν δεδομένα όπου κάποιες από τις μεταβλητές συσχετίζονται έντονα.
- Τα μοντέλα αυτά μπορούμε να τα συγκρίνουμε με άλλα μοντέλα του model based clustering χρησιμοποιώντας το BIC κριτήριο.
- Τα μοντέλα αυτά έχουν πολύ καλή επίδοση στην ομαδοποίηση δεδομένων. Οι ομάδες που σχηματίζονται χρησιμοποιώντας τα, δείχνουν να ανιχνεύουν καλύτερα τη δομή των δεδομένων σε σχέση με άλλες τεχνικές ομαδοποίησης.
- Ένα μειονέκτημα τους είναι ο μεγάλος υπολογιστικός φόρτος. Γενικά απαιτείται μεγάλος αριθμός υπολογισμών και καθώς πρέπει να τρέξουν πολλά και διαφορετικά μοντέλα και να συγκριθούν μεταξύ τους, ο χρόνος υπολογισμού αυξάνει. Ωστόσο, έχουν βρεθεί τρόποι να επιταχυνθεί ο αλγόριθμος υπολογισμού μέσω του συστήματος master-slave (parallelization technique). Με βάση αυτό το σύστημα, ένας κεντρικός επεξεργαστής (ο master) αναθέτει σε επιμέρους επεξεργαστές (τους slaves) να τρέξουν διάφορα μοντέλα της οικογένειας EPGMM και έπειτα ο master παίρνει απόφαση σχετικά με το πιο είναι το καλύτερο μοντέλο. Έτσι, η όλη διαδικασία επιταχύνεται. Περισσότερες πληροφορίες σχετικά παρέχονται από τους McNicholas *et al.* (2010).
- Άλλο μειονέκτημα είναι ότι βασίζονται στις αρχικές τιμές και απαιτείται να ξεκινάμε από πολλές διαφορετικές τυχαίες αρχικές τιμές ώστε να εξασφαλίσουμε ότι θα βρούμε λύση που αντιστοιχεί σε ολικό μέγιστο της πιθανοφάνειας και όχι σε κάποιο τοπικό. Ωστόσο, δεν υπάρχει καμία εγγύηση ότι αυξάνοντας τον αριθμό των αρχικών τιμών θα πετύχουμε και καλύτερα αποτελέσματα στην ομαδοποίηση.

Το unrestricted μοντέλο UUUU της EPGMM οικογένειας είδαμε ότι μειώνει τον αριθμό παραμέτρων προς εκτίμηση σε σχέση με το unrestricted VVV μοντέλο της MCLUST οικογένειας καθώς συνολικά απαιτεί την εκτίμηση $(g-1)+gp+g\left\{pq+p-\frac{1}{2}q(q-1)\right\}$ παραμέτρων. Ωστόσο, παρά αυτή την αρχική μείωση, μπορεί ο αριθμός των παραμέτρων να εξακολουθεί να είναι μεγάλος, ιδιαίτερα αν ο αριθμός των διαστάσεων p είναι μεγάλος ή/και ο αριθμός των ομάδων g είναι μεγάλος. Για αυτό το λόγο κιόλας αναπτύχθηκαν τα πιο φειδωλά μοντέλα της PGMM και EPGMM οικογένειας. Οι Baek et al. (2010) εναλλακτικά των οικογενειών αυτών, πρότειναν το μοντέλο των mixtures of common factor analyzers (MCFA). Πιο συγκεκριμένα, αντί του κλασικού μοντέλου $Y_i=\mu_j+B_jU_{ij}+e_{ij}$, $(i=1,...,n,\ j=1,...,g)$ της παραγοντικής ανάλυσης, πρότειναν το μοντέλο $Y_i=AU_{ij}+e_{ij}$, όπου οι παράγοντες $U_{1j},...,U_{nj}$ ακολουθούν $N(\xi_j,\Omega_j)$, ανεξάρτητα από τα σφάλματα, τα οποία ακολουθούν $N(0,D)$, όπου D διαγώνιος πίνακας. Εδώ ο A είναι ένας $p\times q$ πίνακας επιβαρύνσεων των παραγόντων για τον οποίο ισχύει ο περιορισμός $A^TA=I_q$. Τότε οι μέσοι των ομάδων δίνονται από τη σχέση $\mu_j=A\xi_j$ $(j=1,...,g)$ και οι πίνακες διασπορών από τη σχέση $\Sigma_j=\mathrm{A}\Omega_j\mathrm{A}^{\mathrm{T}}+D$ $(j=1,...,g)$. Έτσι, υπό αυτό το πλαίσιο επιτυγχάνεται επιπλέον μείωση των παραμέτρων προς εκτίμηση, οι οποίοι πλέον είναι $(g-1)+p+q(p+g)+\frac{1}{2}(g-1)q(q+1)$. Επίσης, μπορούν και εδώ να τεθούν διάφοροι περιορισμοί, όπως $D=\sigma^2I_p$, οδηγώντας σε επιπλέον μείωση των παραμέτρων. Περισσότερες πληροφορίες για αυτό το μοντέλο παρέχεται από τους Baek et al. (2010).

Κεφάλαιο 5: Χρήση πολυμεταβλητής t κατανομής

5.1 Πολυμεταβλητή t κατανομή

Για τη μοντελοποίηση πολυμεταβλητών συνεχών δεδομένων ιδιαίτερη έμφαση έχει δοθεί στη χρήση της πολυμεταβλητής κανονικής κατανομής εξαιτίας κυρίως της υπολογιστικής ευκολίας που παρέχει. Αυτό έγινε ιδιαίτερα εμφανές και στην περίπτωση του model-based clustering, όπου όλα τα μοντέλα που παρουσιάστηκαν στα προηγούμενα κεφάλαια βασίζονται στην πολυμεταβλητή κανονική κατανομή ή καλύτερα σε μίξεις πολυμεταβλητών κανονικών κατανομών. Ωστόσο, σε πολλά πρακτικά προβλήματα η κανονική κατανομή δεν είναι η ιδανική για να αναπαραστήσουμε τα δεδομένα. Συχνά η κανονική κατανομή έχει πιο κοντές ουρές απ' ότι υποδεικνύουν τα δεδομένα ότι χρειάζονται για να μοντελοποιηθούν σωστά. Επίσης, οι εκτιμήσεις των παραμέτρων της κανονικής κατανομής μπορεί να επηρεαστούν από ακραίες παρατηρήσεις (outliers). Το πρόβλημα του να αντιμετωπίσουμε τις ακραίες παρατηρήσεις σε πολυμεταβλητά δεδομένα είναι ιδιαίτερα δύσκολο και αυξάνει όσο μεγαλώνουν οι διαστάσεις. Μία τεχνική στο να λάβουμε υπ' όψιν μας καλύτερα τις ακραίες παρατηρήσεις είναι να χρησιμοποιήσουμε μια ομοιόμορφη κατανομή μέσα στο μείγμα των πολυμεταβλητών κανονικών κατανομών για να μοντελοποιήσουμε αυτές τις παρατηρήσεις (Fraley and Raftery 2002). Επίσης, η μοντελοποίηση μέσω της t κατανομής φαίνεται να είναι μια ακόμα λύση στην περίπτωση που έχουμε long tail δεδομένα ή/και ακραίες παρατηρήσεις, καθώς η t κατανομή έχει πιο μακριές ουρές από την κανονική.

Έτσι λοιπόν στο πλαίσιο του model-based clustering αντικαθιστούμε την πολυμεταβλητή κανονική κατανομή κάθε ομάδας με πολυμεταβλητή t κατανομή και θεωρούμε πλέον ότι μια τυχαία παρατήρηση προέρχεται από τη μίξη g πολυμεταβλητών t κατανομών. Με αυτό τον τρόπο το μοντέλο μας γίνεται πιο ανθεκτικό (robust) σε ακραίες παρατηρήσεις και long tail δεδομένα. Ο αριθμός των ακραίων παρατηρήσεων που απαιτείται για να μη δουλεύει σωστά το μοντέλο είναι ο ίδιος και για την κανονική κατανομή και για την t, αλλά στην περίπτωση της t

κατανομής αυτές οι ακραίες παρατηρήσεις πρέπει να είναι μεγαλύτερες, να είναι δηλαδή πιο ακραίες. Αυτό είναι το κέρδος χρησιμοποιώντας την t κατανομή. Ωστόσο, απαιτείται μια επιπλέον παράμετρος, οι βαθμοί ελευθερίας της κατανομής (τους συμβολίζουμε με ν) κάθε ομάδας οι οποίοι ελέγχουν το μήκος των ουρών της t κατανομής. Όσο το ν τείνει στο άπειρο τόσο η t κατανομή προσεγγίζει την κανονική. Κατά συνέπεια οι βαθμοί ελευθερίας μπορούν να ειδωθούν σαν μια παράμετρος που καθορίζει την ανθεκτικότητα (robustness) του μοντέλου. Οι β.ε. μπορεί να είναι προκαθορισμένοι ή μπορεί να εκτιμώνται για κάθε ομάδα από τα δεδομένα.

Πριν εφαρμόσουμε τη μίξη πολυμεταβλητών t κατανομών θα δούμε πως προκύπτει η πολυμεταβλητή t κατανομή για μία ομάδα. Έστω Y μια τυχαία παρατήρηση (διάνυσμα) και μ, Σ η μέση τιμή και ο πίνακας διασποράς αντίστοιχα μιας τυχαίας ομάδας, από την οποία προέρχεται η παρατήρηση. Έστω U μια τυχαία μεταβλητή για την οποία ισχύει $U \sim Gamma\left(\frac{1}{2}\nu, \frac{1}{2}\nu\right)$. Τότε αν ισχύει $Y \mid u \sim N_p\left(\mu, \frac{\Sigma}{u}\right)$, η μη δεσμευμένη κατανομή της Y είναι πολυμεταβλητή t κατανομή με μέσο μ, πίνακα διασποράς $\frac{\nu}{\nu-2}\Sigma$ και βαθμούς ελευθερίας ν. Δηλαδή $Y \sim t_p\left(\mu, \Sigma, \nu\right)$ και η Y έχει σ.π.π.

$$f\left(y \mid \mu, \Sigma, \nu\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)\left|\Sigma\right|^{-1/2}}{\left(\pi\nu\right)^{\frac{p}{2}}\Gamma\left(\frac{\nu}{2}\right)\left\{1+\frac{\delta\left(y,\mu \mid \Sigma\right)}{\nu}\right\}^{\frac{1}{2}(\nu+p)}}, \tag{5.1}$$

όπου $\delta\left(y,\mu \mid \Sigma\right) = \left(y-\mu\right)^T \Sigma^{-1}\left(y-\mu\right)$ είναι η απόσταση Mahalanobis μεταξύ της παρατήρησης y και του μέσου μ. Όσο οι β.ε. ν τείνουν στο άπειρο, η τυχαία μεταβλητή U τείνει στη μονάδα και άρα η Y τείνει να ακολουθεί πολυμεταβλητή κανονική κατανομή.

5.2 Εφαρμογή ΕΜ αλγορίθμου σε μίξεις t πολυμεταβλητών κατανομών

Εφαρμόζοντας τη μίξη πολυμεταβλητών t κατανομών στο πλαίσιο του model-based clustering θεωρούμε ότι μια τυχαία παρατήρηση y προέρχεται από το μείγμα g πολυμεταβλητών t κατανομών με σ.π.π.

$$f\left(y \mid \Psi\right) = \sum_{j=1}^{g} \pi_j f_j\left(y \mid \mu_j, \Sigma_j, \nu_j\right) \tag{5.2}$$

όπου $f_j\left(y \mid \mu_j, \Sigma_j, \nu_j\right)$ είναι η σ.π.π. της t πολυμεταβλητής κατανομής της j ομάδας βάση της (5.1), $\Psi = \left(\pi, \theta, \nu\right)$ με $\pi = \left(\pi_1, ..., \pi_{g-1}\right)$, $\nu = \left(\nu_1, ..., \nu_g\right)$, $\theta = \left(\theta_1, ..., \theta_g\right)$ και όπου το κάθε θ_j περιέχει τα στοιχεία του μ_j και του Σ_j $\forall\ j = 1, 2, ..., g$. Στόχος είναι να εκτιμήσουμε τις παραμέτρους Ψ και για το σκοπό αυτό χρησιμοποιείται ο ΕΜ αλγόριθμος. Όπως και στην περίπτωση του μείγματος πολυμεταβλητών κανονικών κατανομών έτσι και εδώ χρησιμοποιούμε την τεχνική της αύξησης δεδομένων (data augmentation) προκειμένου να εκτιμήσουμε τις παραμέτρους. Θεωρούμε τα διανύσματα $z_1, ..., z_n$ όπου κάθε z_i είναι ένα διάνυσμα διάστασης g, το οποίο μας υποδεικνύει αν η i παρατήρηση προέρχεται ή όχι από την i ομάδα, όπως ακριβώς και στην 2.3 παράγραφο. Αυτά τα z_{ij} δεν τα παρατηρούμε και άρα δεν τα γνωρίζουμε, έτσι τα αντιμετωπίζουμε ως χαμένα δεδομένα (missing data). Οι εκτιμήσεις τ_{ij} που θα πάρουμε για αυτά τα z_{ij} είναι που θα μας υποδείξουν σε ποια ομάδα θα ταξινομηθεί η κάθε παρατήρηση. Εδώ όμως πέραν των στοιχείων z_{ij} θα πρέπει να θεωρήσουμε και άλλα στοιχεία ως missing data προκειμένου να μπορέσουμε να εκτιμήσουμε τις παραπάνω παραμέτρους. Θεωρούμε λοιπόν τα διανύσματα $w_1, ..., w_n$ για τα οποία υποθέτουμε ότι ισχύει $Y_i \mid w_i, z_{ij} = 1 \overset{i.i.d.}{\sim} N_p\left(\mu_j, \Sigma_j / w_i\right)$ και $W_i \mid z_{ij} \sim Gamma\left(\frac{1}{2}\nu_j, \frac{1}{2}\nu_j\right)$, όπου με W_i συμβολίζεται η τ.μ. που αντιστοιχεί στο w_i. Τα w_i θα τα ονομάζουμε βάρη και θα φανεί παρακάτω στις σχέσεις (5.10) & (5.11) γιατί τα ονομάζουμε με αυτό τον τρόπο. Κατόπιν αυτής της προετοιμασίας, με εφαρμογή του ΕΜ αλγορίθμου μπορούμε να εκτιμήσουμε τις παραμέτρους.

Η log-likelihood όλων των δεδομένων (complete-data log-likelihood) $y_c = \left(y_1, ..., y_n, z_1, ..., z_n, w_1, ..., w_n\right)$ μπορεί να γραφεί ως

$\log L_c(\Psi) = \log L_{1c}(\pi) + \log L_{2c}(\nu) + \log L_{3c}(\theta)$ όπου:

$$\log L_{1c}(\pi) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij} \log \pi_j$$

$$\log L_{2c}(\nu) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij} \left\{ -\log\Gamma\left(\frac{1}{2}\nu_j\right) + \frac{1}{2}\nu_j \log\left(\frac{1}{2}\nu_j\right) + \frac{1}{2}\nu_j \left(\log w_i - w_i\right) - \log w_i \right\}$$

$$\log L_{3c}(\theta) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij} \left\{ -\frac{1}{2} p \log(2\pi) - \frac{1}{2}\log\left|\Sigma_j\right| - \frac{1}{2} u_i \left(y_i - \mu_j\right)^T \Sigma_j^{-1} \left(y_i - \mu_j\right) \right\}.$$

Ε-βήμα

Το Ε-βήμα στην $k+1$ επανάληψη του ΕΜ αλγορίθμου υπολογίζει την αναμενόμενη τιμή της complete-data log-likelihood $Q\left(\Psi \mid \Psi^{(k)}\right) = \mathrm{E}_{\Psi^{(k)}}\left\{\log L_c\left(\Psi\right) \mid y\right\}$. Έτσι έχουμε ότι

$$Q\left(\Psi \mid \Psi^{(k)}\right) = Q_1\left(\pi \mid \Psi^{(k)}\right) + Q_2\left(\nu \mid \Psi^{(k)}\right) + Q_3\left(\theta \mid \Psi^{(k)}\right) \tag{5.3}$$

όπου:

$$Q_1\left(\pi \mid \Psi^{(k)}\right) = \sum_{j=1}^{g}\sum_{i=1}^{n} \tau_{ij}^{(k)} \log \pi_j \tag{5.4}$$

$$Q_2\left(\nu \mid \Psi^{(k)}\right) = \sum_{j=1}^{g}\sum_{i=1}^{n} \tau_{ij}^{(k)} \left[-\log\Gamma\left(\frac{1}{2}\nu_j\right) + \frac{1}{2}\nu_j \log\left(\frac{1}{2}\nu_j\right) + \frac{1}{2}\nu_j \left\{ \sum_{i=1}^{n}\left(\log w_{ij}^{(k)} - w_{ij}^{(k)}\right) + \psi\left(\frac{\nu_j^{(k)} + p}{2}\right) - \log\left(\frac{\nu_j^{(k)} + p}{2}\right) \right\} \right] \tag{5.5}$$

$$Q_3\left(\theta \mid \Psi^{(k)}\right) = \sum_{j=1}^{g}\sum_{i=1}^{n} \tau_{ij}^{(k)} \left\{ -\frac{1}{2} p \log(2\pi) - \frac{1}{2}\log\left|\Sigma_j\right| + \frac{1}{2} p \log w_{ij}^{(k)} - \frac{1}{2} w_{ij} \left(y_i - \mu_j\right)^T \Sigma_j^{-1} \left(y_i - \mu_j\right) \right\} \tag{5.6}$$

με $\psi\left(x\right) = \dfrac{\dfrac{\partial\Gamma(x)}{\partial x}}{\Gamma(x)}$. Βάση της $Q\left(\Psi \mid \Psi^{(k)}\right)$ υπολογίζονται τα $\tau_{ij}^{(k+1)}$ και $w_{ij}^{(k+1)}$ ως:

$$\tau_{ij}^{(k+1)} = \frac{\pi_j^{(k)} f_j\left(y_i \mid \mu_j^{(k)}, \Sigma_j^{(k)}, \nu_j^{(k)}\right)}{f\left(y \mid \Psi^{(k)}\right)} \tag{5.7}$$

$$\text{και} \quad w_{ij}^{(k+1)} = \frac{\nu_j^{(k)} + p}{\nu_j^{(k)} + \delta\left(y_i, \mu_j^{(k)} \mid \Sigma_j^{(k)}\right)}, \tag{5.8}$$

όπου $\delta\left(y_i, \mu_j^{(k)} \mid \Sigma_j^{(k)}\right) = \left(y_i - \mu_j^{(k)}\right)^T \Sigma_j^{(k)^{-1}} \left(y_i - \mu_j^{(k)}\right)$.

Μ-βήμα

Στο Μ-βήμα μεγιστοποιούμε την $Q\left(\Psi \mid \Psi^{(k)}\right)$ προκειμένου να πάρουμε εκτιμήσεις για τα π_j, μ_j, Σ_j και ν_j. Από την (5.3) συνεπάγεται ότι τα $\pi^{(k+1)}, \mu^{(k+1)}, \Sigma^{(k+1)}, \nu^{(k+1)}$ μπορούν να υπολογιστούν ανεξάρτητα το ένα απ' τα υπόλοιπα μέσω των σχέσεων (5.4)-(5.6). Οι εκτιμήσεις των $\pi_j^{(k+1)}, \mu_j^{(k+1)}, \Sigma_j^{(k+1)}$ υπάρχουν σε κλειστή μορφή. Μόνο οι εκτιμήσεις των βαθμών ελευθερίας $\nu_j^{(k+1)}$ πρέπει να υπολογιστούν με χρήση επαναληπτικού αλγορίθμου. Έτσι, το Μ-βήμα μας δίνει τις παρακάτω εκτιμήσεις:

$$\pi_j^{(k+1)} = \sum_{i=1}^{n} \frac{\tau_{ij}^{(k+1)}}{n}, \quad (j = 1, 2, \ldots g) \tag{5.9}$$

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k+1)} w_{ij}^{(k+1)} y_i}{\sum_{i=1}^{n} \tau_{ij}^{(k+1)} w_{ij}^{(k+1)}} \tag{5.10}$$

και $$\Sigma_j^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(k+1)} w_{ij}^{(k+1)} \left(y_i - \mu_j^{(k+1)}\right)\left(y_i - \mu_j^{(k+1)}\right)^T}{\sum_{i=1}^{n} \tau_{ij}^{(k+1)}} \tag{5.11}$$

Άρα, το Ε-βήμα ανανεώνει τις τιμές w_{ij}, ενώ το Μ-βήμα χρησιμοποιεί τις w_{ij} ως βάρη για να εκτιμήσει τα $\mu_i^{(k+1)}, \Sigma_i^{(k+1)}$. Από την (5.10) φαίνεται ότι όσο οι βαθμοί ελευθερίας μειώνονται, τόσο μειώνεται το βάρος, άρα και η επιρροή μιας ακραίας παρατήρησης. Αυτό που μένει να εκτιμηθεί είναι οι βαθμοί ελευθερίας. Όπως έχει ήδη αναφερθεί προηγουμένως οι β.ε. μπορεί να είναι προκαθορισμένοι, οπότε και δε χρειάζεται κάποια εκτίμηση. Ωστόσο, όταν δεν είναι προκαθορισμένοι, η εκτίμησή τους δίνεται από τη λύση της εξίσωσης:

$$-\psi\left(\frac{1}{2}\nu_j\right) + \log\left(\frac{1}{2}\nu_j\right) + 1 + \frac{1}{n_j^{(k+1)}} \sum_{i=1}^{n} \tau_{ij}^{(k+1)} \left(\log w_{ij}^{(k+1)} - w_{ij}^{(k+1)}\right) + \psi\left(\frac{\nu_j^{(k)} + p}{2}\right) - \log\left(\frac{\nu_j^{(k)} + p}{2}\right) = 0$$

όπου $n_j^{(k+1)} = \sum_{i=1}^{n} \tau_{ij}^{(k+1)}$. Η εξίσωση αυτή λύνεται επαναληπτικά με χρήση αριθμητικών μεθόδων. Πλήρης παρουσίαση της υπολογιστικής διαδικασίας του ΕΜ αλγορίθμου για την περίπτωση του μείγματος t πολυμεταβλητών κατανομών μαζί με

παραδείγματα από προσομοιωμένα και πραγματικά δεδομένα παρέχεται από τους Peel and McLachlan (2000).

5.3 Mixtures of Multivariate t-factor analyzers (MMtFA)

Θα θέλαμε σε ένα επόμενο βήμα να εφαρμόσουμε την t πολυμεταβλητή κατανομή σε ένα μοντέλο μίξεων από factor analyzers. Είδαμε ότι η χρήση των factor analyzers ήταν καθοριστική στην ομαδοποίηση high-dimensional δεδομένων και ως εκ τούτου θα ήταν ιδιαίτερα χρήσιμο να εισάγουμε την t-κατανομή σε αυτή την περίπτωση. Έτσι, μπορούμε να κατασκευάσουμε μοντέλα ανθεκτικά σε ακραίες παρατηρήσεις τα οποία ταυτόχρονα θα μπορούν να εφαρμοστούν σε high-dimensional δεδομένα.

Το μοντέλο μας εξακολουθεί να είναι το (5.2). Εδώ όμως ισχύει ότι $\Sigma_j = B_j B_j^T + D_j \quad (j = 1,...,g)$, σύμφωνα με το μοντέλο της παραγοντικής ανάλυσης $Y_i = \mu_j + B_j U_{ij} + e_{ij} \quad (j = 1,...,g)$, όπου U_{ij} οι παράγοντες. Τώρα το διάνυσμα των άγνωστων παραμέτρων Ψ αποτελείται από τις πιθανότητες π_j, τις μέσες τιμές μ_j τους βαθμούς ελευθερίας ν_j κάθε ομάδας, αλλά και από τους πίνακες B_j και D_j. Όπως στην περίπτωση της κανονικής κατανομής (παράγραφος 3.2) είχαμε υποθέσει ότι για μια τυχαία ομάδα το από κοινού διάνυσμα (Y_i, U_i) ακολουθεί πολυμεταβλητή κανονική κατανομή, δηλαδή $\begin{pmatrix} Y_i \\ U_i \end{pmatrix} \sim N_{p+q}(\mu^*, \Sigma)$, εδώ υποθέτουμε ότι $\begin{pmatrix} Y_i \\ U_{ij} \end{pmatrix} | z_{ij} = 1 \sim t_{p+q}(\mu_j^*, \Sigma_j, \nu_j), \quad (j = 1,...,g)$. Δηλαδή για μια τυχαία ομάδα, το από κοινού διάνυσμα (Y_i, U_i) δοθέντος ότι η παρατήρηση προέρχεται από αυτή την ομάδα ακολουθεί πολυμεταβλητή t κατανομή. Αυτό προκύπτει σύμφωνα με τον τρόπο κατασκευής της πολυμεταβλητής t-κατανομής που παρουσιάστηκε στην 5.1 και πιο συγκεκριμένα αν υποθέσουμε ότι μια τυχαία μεταβλητή W_i (βάρη) ακολουθεί

Gamma κατανομή, $W_i \sim Gamma\left(\frac{1}{2}\nu_j, \frac{1}{2}\nu_j\right)$ και αν επιπλέον υποθέσουμε ότι ισχύει

$\begin{pmatrix} Y_i \\ U_{ij} \end{pmatrix} | \, w_i, z_{ij} = 1 \sim N_{p+q}\left(\mu_j^*, \frac{\Sigma_j}{w_i}\right)$ τότε έπεται ότι

$\begin{pmatrix} Y_i \\ U_{ij} \end{pmatrix} | \, z_{ij} = 1 \sim t_{p+q}\left(\mu_j^*, \Sigma_j, \nu_j\right), \ \ (j = 1, ..., g)$. Επίσης ισχύει ότι

$$Y_{ij} \,|\, u_{ij}, w_i, z_{ij} = 1 \sim N_p\left(\mu_j + B_j u_{ij}, \frac{D_j}{w_i}\right).$$

Ως missing data θεωρούμε και εδώ, όπως και στην 5.2, τα στοιχεία z_{ij} και τα βάρη w_{ij} αλλά επιπλέον και τους παράγοντες u_{ij}. Οπότε το διάνυσμα όλων των δεδομένων εδώ είναι το $y_c = \left(y_1, ..., y_n, z_1, ..., z_n, w_1, ..., w_n, u_1, ..., u_n\right)$. Για την εκτίμηση των αγνώστων παραμέτρων π_j, μ_j, B_j, D_j, ν_j κάθε ομάδας χρησιμοποιείται ο AECM αλγόριθμος. Η log-likelihood όλων των δεδομένων (complete-data log-likelihood) είναι $\log L_c\left(\Psi\right) = \sum_{j=1}^{g}\sum_{i=1}^{n} z_{ij} \log a_{ij}$, όπου

$$\alpha_{ij} = \pi_j f_{gamma}\left(w_i \,|\, \frac{1}{2}\nu_j, \frac{1}{2}\nu_j\right)\varphi\left(u_{ij} \,|\, 0, \frac{I_q}{w_i}\right)\varphi\left(y_i \,|\, \mu_j + B_j u_{ij}, \frac{D_j}{w_i}\right).$$

5.3.1 Εφαρμογή AECM αλγορίθμου

Η φιλοσοφία του αλγορίθμου παραμένει η ίδια. Ο αλγόριθμος εξακολουθεί να αποτελείται από 2 κύκλους, με ένα E-βήμα και ένα CM-βήμα ο κάθε ένας, όπως και στην περίπτωση των factor analyzers για την κανονική κατανομή. Το διάνυσμα των παραμέτρων Ψ χωρίζεται σε δύο μέρη Ψ_1 και Ψ_2, $\Psi = \left(\Psi_1, \Psi_2\right)$, όπου το διάνυσμα Ψ_1 περιέχει τις πιθανότητες $\pi_j \ \ (j = 1, ..., g-1)$, τους μέσους $\mu_j \ \ (j = 1, ..., g)$ και τους βαθμούς ελευθερίας $\nu_j \ \ (j = 1, ..., g)$. Το διάνυσμα Ψ_2 περιέχει τα στοιχεία των πινάκων B_j και $D_j \ \ (j = 1, ..., g)$. Στον πρώτο κύκλο θεωρούμε ως missing-data τα στοιχεία z_{ij} και τα βάρη w_i και στο CM-βήμα εκτιμούμε τις παραμέτρους π_j, μ_j και ν_j του Ψ_1, ενώ στο δεύτερο κύκλο ως missing-data παίρνουμε τα z_{ij} τα w_{ij} και τους παράγοντες u_{ij} και στο CM βήμα εκτιμούμε τους πίνακες B_j και D_j του Ψ_2.

Έτσι, μετά το τέλος του πρώτου κύκλου στην $k+1$ επανάληψη παίρνουμε ότι

$$E-\beta\acute{\eta}\mu\alpha\begin{cases}\tau_{ij}^{(k+1/2)}=\tau_j\left(y_i\mid\Psi^{(k)}\right)=\dfrac{\pi_j^{(k)}f_j\left(y_i\mid\mu_j^{(k)},\Sigma_j^{(k)},\nu_j^{(k)}\right)}{\sum_{h=1}^{g}\pi_h^{(k)}f_h\left(y_i\mid\mu_h^{(k)},\Sigma_h^{(k)},\nu_h^{(k)}\right)},\ (j=1,...,g,\ \ i=1,...,n)\\ w_{ij}^{(k+1/2)}=\dfrac{\nu_j^{(k)}+p}{\nu_j^{(k)}+\delta\left(y_i,\mu_j^{(k)}\mid\Sigma_j^{(k)}\right)},\ \text{όπου}\ \delta\left(y_i,\mu_j^{(k)}\mid\Sigma_j^{(k)}\right)=\left(y_i-\mu_j^{(k)}\right)^T\Sigma_j^{(k)^{-1}}\left(y_i-\mu_j^{(k)}\right)\end{cases}$$

$$CM-\beta\acute{\eta}\mu\alpha\begin{cases}\pi_j^{(k+1)}=\sum_{i=1}^{n}\dfrac{\tau_{ij}^{(k+1/2)}}{n}\\ \mu_j^{(k+1)}=\dfrac{\sum_{i=1}^{n}\tau_{ij}^{(k+1/2)}w_{ij}^{(k+1/2)}y_i}{\sum_{i=1}^{n}\tau_{ij}^{(k+1/2)}w_{ij}^{(k+1/2)}}\\ \nu_j^{(k+1)}=...\end{cases}$$

Η εκτίμηση $\nu_j^{(k+1)}$ των βαθμών ελευθερίας δεν υπάρχει σε κλειστή μορφή αλλά μπορεί να βρεθεί από τη λύση της εξίσωσης

$$-\psi\left(\frac{1}{2}\nu_j\right)+\log\left(\frac{1}{2}\nu_j\right)+1+\frac{1}{n_j^{(k+1/2)}}\sum_{i=1}^{n}\tau_{ij}^{(k+1/2)}\left(\log w_{ij}^{(k+1/2)}-w_{ij}^{(k+1/2)}\right)+\psi\left(\frac{\nu_j^{(k)}+p}{2}\right)-\log\left(\frac{\nu_j^{(k)}+p}{2}\right)=0$$

όπου $n_j^{(k+1/2)}=\sum_{i=1}^{n}\tau_{ij}^{(k+1/2)}$. Η εξίσωση αυτή λύνεται επαναληπτικά ως προς ν_j με χρήση αριθμητικών μεθόδων. $\nu_j^{(k)}$ είναι η εκτίμηση των βαθμών ελευθερίας στην προηγούμενη επανάληψη k. Με το τέλος του πρώτου κύκλου, το διάνυσμα $\Psi^{(k)}$ που είχαμε πάρει ως εκτιμήσεις των παραμέτρων από την k επανάληψη ανανεώνεται σε $\Psi^{(k+1/2)}=\left(\Psi_1^{(k+1)},\Psi_2^{(k)}\right)$.

Μετά το τέλος του δεύτερου κύκλου στην $k+1$ επανάληψη παίρνουμε ότι:

$$E-\beta\acute{\eta}\mu\alpha\begin{cases}\tau_{ij}^{(k+1)}=\tau_j\left(y_i\mid\Psi^{(k+1/2)}\right)=\dfrac{\pi_j^{(k+1)}f_j\left(y_i\mid\mu_j^{(k+1)},\Sigma_j^{(k)},\nu_j^{(k+1)}\right)}{\sum_{h=1}^{g}\pi_h^{(k+1)}f_h\left(y_i\mid\mu_h^{(k+1)},\Sigma_h^{(k)},\nu_h^{(k+1)}\right)},\ (j=1,...,g,\ \ i=1,...,n)\\ w_{ij}^{(k+1)}=\dfrac{\nu_j^{(k+1)}+p}{\nu_j^{(k+1)}+\delta\left(y_i,\mu_j^{(k+1)}\mid\Sigma_j^{(k)}\right)},\\ \text{όπου}\ \ \delta\left(y_i,\mu_j^{(k+1)}\mid\Sigma_j^{(k)}\right)=\left(y_i-\mu_j^{(k+1)}\right)^T\Sigma_j^{(k)^{-1}}\left(y_i-\mu_j^{(k+1)}\right)\end{cases}$$

$$CM-\beta\acute{\eta}\mu\alpha\begin{cases}\mathrm{B}_j^{(k+1)}=V_j^{(k+1)}\gamma_j^{(k)}\left(\gamma_j^{(k)^T}V_j^{(k+1)}\gamma_j^{(k)}+\omega_j^{(k)}\right)^{-1}\\ D_j^{(k+1)}=diag\left(V_j^{(k+1)}-V_j^{(k+1)}\gamma_j^{(k)}B_j^{(k+1)^T}\right)\end{cases}$$

όπου $\gamma_j^{(k)}=\left(B_j^{(k)}B_j^{(k)^T}+D_j^{(k)}\right)^{-1}B_j^{(k)}$, $\omega_j^{(k)}=I_q-\gamma_j^{(k)^T}B_j^{(k)}$

και $V_j^{(k+1)}=\dfrac{\sum_{i=1}^{n}\tau_{ij}^{(k+1)}w_{ij}^{(k+1)}\left(y_i-\mu_j^{(k+1)}\right)\left(y_i-\mu_j^{(k+1)}\right)^T}{\sum_{i=1}^{n}\tau_{ij}^{(k+1)}}$. Εδώ, οι νέες εκτιμήσεις των τ_{ij} και w_{ij} έχουν βασιστεί στις εκτιμήσεις που πήραμε από τον πρώτο κύκλο της $k+1$ επανάληψης.

Περισσότερες πληροφορίες για την υπολογιστική διαδικασία του AECM αλγορίθμου για την περίπτωση των t-factor analyzers παρέχεται από τους McLachlan and Peel (2007).

5.4 Extending Mixtures of Multivariate t-factor analyzers (MMtFA)

Είδαμε στην περίπτωση των factor analyzers του μείγματος πολυμεταβλητών κανονικών κατανομών ότι επιβάλλοντας κάποιους περιορισμούς στους πίνακες επιβαρύνσεων (B_j) και ιδιαιτεροτήτων (D_j) πήραμε την οικογένεια μοντέλων EPGMM (παράγραφος 4.4), η οποία απαιτεί την εκτίμηση μικρότερου αριθμού παραμέτρων. Αντίστοιχοι περιορισμοί μπορούν να τεθούν και στην περίπτωση των t-factor analyzers οπότε και προκύπτει μια νέα οικογένεια μοντέλων. Επιπλέον όμως των περιορισμών στους πίνακες επιβαρύνσεων και ιδιαιτεροτήτων, περιορισμοί μπορούν να τεθούν και στον αριθμό των βαθμών ελευθερίας των ομάδων.

Πιο συγκεκριμένα, οι Andrews and McNicholas (2011a) επέκτειναν το μοντέλο της παραγράφου 5.3 θέτοντας τους περιορισμούς $\nu_j=\nu$, $D_j=d_jI_p$ και $B_j=B$. Διάφοροι συνδυασμοί αυτών των περιορισμών έδωσαν μια νέα οικογένεια έξι μοντέλων τα οποία περιγράφονται στον Πίνακα 5.1. Από εδώ και στο εξής αυτά τα έξι μοντέλα θα αναφέρονται ως οικογένεια. Πρέπει να σημειωθεί ότι για αυτά τα έξι μοντέλα ο αριθμός των παραμέτρων προς εκτίμηση αυξάνεται γραμμικά ως προς τον

αριθμό των μεταβλητών, όπως ακριβώς συνέβαινε και στα μοντέλα της EPGMM οικογένειας. Επίσης, το UCU μοντέλο είχε προταθεί και νωρίτερα με την ονομασία mixtures of probabilistic principal t-component analyzers (MPPtCA).

Πίνακας 5.1: Μοντέλα της MMtFA οικογένειας

MMtFA Model	$\mathbf{B}_j = \mathbf{B}$	$\mathbf{D}_j = \mathbf{d}_j\mathbf{I}_\mathrm{p}$	$\boldsymbol{\nu}_j = \boldsymbol{\nu}$	**Covariance and d.f. parameters**
CCC	C	C	C	$\left[pq - q(q-1)/2\right] + g + 1$
CCU	C	C	U	$\left[pq - q(q-1)/2\right] + g + g$
UCC	U	C	C	$g\left[pq - q(q-1)/2\right] + g + 1$
UCU	U	C	U	$g\left[pq - q(q-1)/2\right] + g + g$
UUC	U	U	C	$g\left[pq - q(q-1)/2\right] + gp + 1$
UUU	U	U	U	$g\left[pq - q(q-1)/2\right] + gp + g$

Η ιδέα του να περιορίσουμε τον αριθμό των βαθμών ελευθερίας αρχικά μοιάζει αχρείαστη καθώς η μείωση στον αριθμό των παραμέτρων προς εκτίμηση που επιτυγχάνουμε είναι πολύ μικρή, εκτός και αν ο αριθμός των ομάδων είναι πολύ μεγάλος, κάτι όμως που δεν είναι σύνηθες στην ομαδοποίηση δεδομένων. Όμως, στην πράξη τα μοντέλα με περιορισμένους βαθμούς ελευθερίας δίνουν καλύτερες ομαδοποιήσεις. Η εκτίμηση των παραμέτρων γίνεται και εδώ με τη χρήση του AECM αλγορίθμου.

Όταν θέτουμε τον περιορισμό $\nu_j = \nu \ \ (j = 1, ..., g)$, δηλαδή οι β.ε. να ίδιοι για όλες τις ομάδες, τότε σε κάθε επανάληψη η εκτίμηση των β.ε. δίνεται από τη λύση της εξίσωσης

$$-\psi\left(\frac{\nu^{new}}{2}\right) + \log\left(\frac{\nu^{new}}{2}\right) + 1 + \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n}\tau_{ij}\left(\log w_{ij} - w_{ij}\right) + \psi\left(\frac{\nu^{old}+p}{2}\right) - \log\left(\frac{\nu^{old}+p}{2}\right) = 0$$

ως προς ν^{new}, όπου ν^{old} είναι η εκτίμηση των βαθμών ελευθερίας στην προηγούμενη επανάληψη. Επίσης, οι Andrews and McNicholas (2011a) επέβαλαν επιπλέον οι βαθμοί ελευθερίας να μην υπερβαίνουν τους 200 για κάθε ομάδα έτσι ώστε να μην καθυστερεί η σύγκλιση. Εάν δεν έχουμε επιβάλει άλλους περιορισμούς στους πίνακες επιβαρύνσεων και ιδιαιτεροτήτων, τότε οι εκτιμήσεις των υπόλοιπων μεταβλητών παραμένουν ως έχουν.

Αν θέσουμε τον περιορισμό $D_j = d_j I_p$ τότε η εκτίμηση των παραμέτρων d_j δίνεται από τον τύπο $d_j^{new} = \frac{1}{p} tr\left\{S_j - B_j^{new}\beta_j S_j\right\}$ όπου $S_j = \frac{1}{n_j}\sum_{i=1}^{n}\tau_{ij}w_{ij}\left(y_i - \mu_j\right)\left(y_i - \mu_j\right)^T$ και $\beta_j = B_j^T\left(B_j B_j^T + D_j\right)^{-1}$, $\left(i = 1,...,n,\ j = 1,...,g\right)$. Εάν δεν έχουμε επιβάλει άλλους περιορισμούς στους πίνακες επιβαρύνσεων και στους βαθμούς ελευθερίας, τότε οι εκτιμήσεις των υπόλοιπων παραμέτρων παραμένουν ως έχουν.

Τέλος, ο περιορισμός $B_j = B$ επιβάλει τη μεγαλύτερη μείωση στον αριθμό των παραμέτρων προς εκτίμηση. Υπό αυτό τον περιορισμό, η εκτίμηση του νέου πίνακα B (ένας πίνακας πλέον, κοινός για όλες τις ομάδες) δίνεται από τον τύπο $B^{new} = \left[\sum_{j=1}^{g}\frac{n_j}{d_j}S_j\beta_j^{\mathrm{T}}\right]\left[\sum_{j=1}^{g}\frac{n_j}{d_j}\Theta_j\right]^{-1}$, όπου S_j και β_j όπως προηγουμένως και $\Theta_j = I_p - \beta_j B_j + \beta_j S_j \beta_j^T$, $\left(i = 1,...,n,\ j = 1,...,g\right)$. Όμως, το να επιβάλλουμε ίδιους πίνακες επιβαρύνσεων για όλες τις ομάδες επηρεάζει την εκτίμηση των d_j. Έτσι, οι νέες εκτιμήσεις των d_j δίνονται από τη σχέση $d_j^{new} = \frac{1}{p} tr\left\{S_j - 2B^{new}\beta_j S_j + B^{new}\Theta_j\left(B^{new}\right)^T\right\}$.

5.5 Επιλογή μοντέλου - Αρχικές τιμές - Σύγκλιση

Όπως και στις προηγούμενες ομάδες μοντέλων (MCLUST, PGMM, EPGMM) έτσι και εδώ στην MMtFA εγείρονται ορισμένα ζητήματα όπως, ποιο είναι το κατάλληλο μοντέλο να χρησιμοποιηθεί κάθε φορά, ποιος είναι ο βέλτιστος αριθμός ομάδων g, και ποιος ο κατάλληλος αριθμός παραγόντων q. Απαντήσεις σε αυτά τα ερωτήματα δίνονται και εδώ μέσω των κριτηρίων που έχουν αναφερθεί στην 2.7 παράγραφο με το BIC να είναι ξανά αυτό που χρησιμοποιείται περισσότερο. Επίσης, για την αξιολόγηση της ομαδοποίησης χρησιμοποιούνται και τα ICL και adjusted Rand Index κριτήρια.

Άλλο πρόβλημα που πρέπει να αντιμετωπίσουμε ξανά εδώ είναι οι αρχικές τιμές. Ήδη έχουν αναφερθεί τρόποι για το πώς μπορούμε να δώσουμε αρχικές τιμές στον AECM αλγόριθμο. Παρόμοιοι τρόποι χρησιμοποιούνται και για αυτή την οικογένεια. Οι Andrews and McNicholas (2011a) προτείνουν να χρησιμοποιούμε ως αρχικές τιμές τα αποτελέσματα της ομαδοποίησης μέσω της οικογένειας MCLUST για την περίπτωση που $n > p$, και άρα αυτά τα μοντέλα μπορούν να εφαρμοστούν. Αν έχουμε high dimensional δεδομένα προτείνουν την αρχικοποίηση του αλγορίθμου μέσω της φασματικής ανάλυσης των πινάκων διακύμανσης όπως παρουσιάστηκε στην 4.2.1 παράγραφο για την περίπτωση της PGMM οικογένειας. Στην περίπτωση όμως που ο αριθμός των μεταβλητών είναι εξαιρετικά μεγάλος και άρα ο υπολογισμός των ιδιοτιμών είναι υπολογιστικά δύσκολος και αναξιόπιστος, προτείνουν τη χρήση του k-means αλγορίθμου.

Τέλος, για τον έλεγχο της σύγκλισης του αλγορίθμου οι Andrews and McNicholas (2011a) προτείνουν τη χρήση του τροποποιημένου Aitken κριτηρίου που αναφέρθηκε στην 4.2.2. παράγραφο.

5.6 Παράδειγμα

Στην εργασία των Andrews and McNicholas (2011a) υπάρχουν παραδείγματα που δείχνουν την πολύ καλή επίδοση της MMtFA οικογένειας σε προσομοιωμένα δεδομένα από πολυμεταβλητές κανονικές κατανομές καθώς επίσης γίνεται σύγκριση της MMtFA με την MCLUST με ευνοϊκά αποτελέσματα υπέρ της MMtFA για τα δεδομένα αυτά. Επίσης παραδείγματα εφαρμογής της MMtFA σε πραγματικά δεδομένα παρέχονται και στο Andrews (2010)

Παρακάτω θα δούμε τη εφαρμογή της MMtFA στα wine data δεδομένα που παρουσιάστηκαν στην 4.3 παράγραφο, μόνο για τις 13 μεταβλητές που είναι διαθέσιμες στην gclus βιβλιοθήκη της R και όχι για το σύνολο των 27 μεταβλητών που τα δεδομένα αυτά περιέχουν. Τα 6 μοντέλα της MMtFA εφαρμόστηκαν για $g = 1,..,5$ ομάδες και $q = 1,...,5$ παράγοντες. Τα αποτελέσματα για όλα τα μοντέλα συνοψίζονται στον ακόλουθο πίνακα.

Πίνακας 5.2: Αποτελέσματα κάθε μοντέλου της MMtFA οικογένειας για τα wine data των 13 μεταβλητών.

Model	g	q	BIC	ICL	ν_1	ν_2	ν_3	Adj. Rand Index
UUU	3	2	-5294.4	-5296.3	117.4	8.4	152.6	0.96
UUC	3	2	-5298.8	-5300.6	17.4	17.4	17.4	0.98
UCU	3	2	-5504.0	-5505.8	33.5	6.6	27.9	0.93
UCC	3	2	-5498.1	-5501.5	10.9	10.9	10.9	0.90
CCU	3	4	-5442.0	-5443.3	77.7	7.0	45.7	0.95
CCC	4	4	-5444.2	-5446.9	21.2	21.2	21.2	0.84

Από τον παραπάνω πίνακα φαίνεται ότι το καλύτερο μοντέλο είναι το UUU με $\mathrm{BIC} = -5294.4$ και $\mathrm{ICL} = -5296.3$. Το δεύτερο καλύτερο μοντέλο σύμφωνα και με τα δύο κριτήρια είναι το UUC. Επίσης τα 5 από τα 6 μοντέλα ανιχνεύουν το σωστό αριθμό ομάδων (g=3). Γενικά λοιπόν, και με βάση τον Adjusted Rand Index, η επίδοση της MMtFA οικογένειας είναι πολύ καλή για τα συγκεκριμένα δεδομένα. Για το UUU μοντέλο οι διαφορές που παρατηρούνται στους βαθμούς ελευθερίας μεταξύ των ομάδων είναι σημαντικές. Για την πρώτη και την τρίτη ομάδα, επειδή οι β.ε. είναι πολλοί, φαίνεται η πολυμεταβλητή κανονική κατανομή να είναι κατάλληλη για να περιγράψει τα δεδομένα των ομάδων αυτών. Για τη δεύτερη όμως ομάδα επειδή οι β.ε. είναι λίγοι, η πολυμεταβλητή t κατανομή είναι περισσότερο κατάλληλη. Επιπλέον, αυτό που έχει ιδιαίτερο ενδιαφέρον εδώ είναι ότι παρ' όλο που τα BIC και ICL προτείνουν το UUU ως το καλύτερο μοντέλο, αυτό που ομαδοποιεί πιο σωστά τα δεδομένα είναι το UUC καθώς έχει υψηλότερο adjusted Rand Index. Για αυτό το λόγο τα κριτήρια αυτά δε συνιστούν παρά ενδείξεις για το πιο κατάλληλο μοντέλο και δε θα πρέπει να είμαστε απόλυτοι στη χρήση τους. Στον παρακάτω πίνακα φαίνεται η ομαδοποίηση που πετυχαίνουν τα δύο καλύτερα μοντέλα. Στην ουσία η διαφορά τους αφορά στην ταξινόμηση μίας μόνο παρατήρησης.

Πίνακας 5.3: Ομαδοποίηση βάσει των UUU και UUC μοντέλων για τα wine data δεδομένα (13 μεταβλητές).

	UUU			UUC		
	1	2	3	1	2	3
Cluster 1	58	1		58	1	
Cluster 2	1	70			71	
Cluster 3			48			48

Στην 4.3 παράγραφο είχαμε δει την ομαδοποίηση των ίδιων δεδομένων σύμφωνα με τις MCLUST και PGMM οικογένειες καθώς και με την τεχνική της variable selection. Μια σύγκριση όλων των μοντέλων που έχουν εφαρμοστεί μέχρι στιγμής για αυτά τα δεδομένα μπορεί εύκολα να γίνει από τον παρακάτω πίνακα.

Πίνακας 5.4: Σύγκριση μοντέλων για τα wine data (13 μεταβλητές).

Model	**Adjusted Rand Index**
UUC	0.98
UUU	0.96
CCU	0.95
UCU	0.93
UCC	0.90
CCC	0.84
PGMM	0.79
Variable Selection	0.78
MCLUST	0.48

Βλέπουμε ότι όλα τα μοντέλα της MMtFA οικογένειας δίνουν καλύτερες ομαδοποιήσεις για αυτά τα δεδομένα απ' ότι οι PGMM, MCLUST οικογένειες μοντέλων και η variable selection τεχνική. Το γεγονός ότι το UUC μοντέλο έδωσε καλύτερη ομαδοποίηση από το UUU, παρά τη μικρότερη τιμή των BIC και ICL κριτηρίων, μας δείχνει ότι αυτά τα κριτήρια δε διαλέγουν απαραίτητα το μοντέλο με την καλύτερη ομαδοποίηση.

5.7 Παρατηρήσεις

Τα αποτελέσματα της εφαρμογής της MMtFA οικογένειας μοντέλων σε προσομοιωμένα και πραγματικά δεδομένα που πραγματοποιήθηκε από τους Andrews and McNicholas (2011a) έδειξαν ότι τα μοντέλα με περιορισμένους τους βαθμούς ελευθερίας ομαδοποιούν γενικά καλύτερα τα δεδομένα απ' ότι αυτά που επιτρέπουν διαφορετικούς βαθμούς ελευθερίας σε κάθε ομάδα. Αυτό προφανώς οφείλεται στο ότι η εκτίμηση των βαθμών ελευθερίας στην πρώτη περίπτωση είναι πιο αξιόπιστη καθώς βασίζεται σε περισσότερα δεδομένα και αντιστοιχεί θα λέγαμε σε κάτι σαν το μέσο όρο των βαθμών ελευθερίας για όλες τις ομάδες.

Υπάρχει η προοπτική και η δυνατότητα να τεθούν επιπλέον περιορισμοί στους πίνακες διακυμάνσεων οπότε και να προκύψει μια νέα πιο διευρυμένη MMtFA οικογένεια. Αυτή η νέα οικογένεια θα περιέχει 16 μοντέλα. Ωστόσο, το όφελος από μια τέτοια εξέλιξη θα πρέπει να συμβαδίσει με μια ταυτόχρονη εξέλιξη στο τρόπο επιλογής του κατάλληλου μοντέλου. Όπως φάνηκε στο παραπάνω παράδειγμα τα BIC και ICL κριτήρια που χρησιμοποιούνται ως τώρα δε μας δίνουν απαραίτητα το μοντέλο με την καλύτερη ομαδοποίηση. Οπότε θα ήταν ανούσιο να φτιάξουμε μια νέα οικογένεια από 16 μοντέλα χωρίς να έχουμε μια αποτελεσματική τεχνική στο να διαλέγουμε το μοντέλο με την καλύτερη ομαδοποίηση.

Τα μοντέλα της MMtFA οικογένειας είναι κατάλληλα για εφαρμογή σε high-dimensional δεδομένα εξαιτίας της μείωσης των παραμέτρων προς εκτίμηση που απαιτούνται. Για να είναι όμως πιο αποδοτική από πλευράς χρόνου η όλη υπολογιστική διαδικασία είναι προτιμότερο τα μοντέλα της MMtFA να τρέχουν υπό το σύστημα master-slave, το οποίο αναπτύχθηκε από τους McNicholas *et al.* (2010) για την περίπτωση των μοντέλων της PGMM οικογένειας. Ο παράλληλος και ταυτόχρονος υπολογισμός των διαφόρων μοντέλων μπορεί να εφαρμοστεί και σε αυτά τα μοντέλα της MMtFA οικογένειας.

Το model-based clustering που έχουμε δει μέχρι στιγμής είναι μια ειδική περίπτωση του model-based classification. Πιο συγκεκριμένα, μέχρι στιγμής έχουμε δει μοντέλα ομαδοποίησης δεδομένων όπου αγνοούμε εντελώς σε ποια ομάδα ανήκει κάθε μία από τις n παρατηρήσεις. Έστω τώρα η περίπτωση όπου για τις k $(k < n)$ παρατηρήσεις γνωρίζουμε σε ποια ομάδα αυτές πραγματικά ανήκουν. Ο σκοπός είναι να κατατάξουμε τις υπόλοιπες $n-k$ παρατηρήσεις σε ομάδες, χρησιμοποιώντας τη γνώση της ομαδοποίησης των προηγούμενων k παρατηρήσεων. Η κλασική προσέγγιση θα ήταν να χρησιμοποιήσουμε αυτές τις k παρατηρήσεις για να εκτιμήσουμε τις παραμέτρους των ομάδων (μέσες τιμές, πίνακες διασπορών κτλ.) και κατόπιν να χρησιμοποιήσουμε αυτές τις εκτιμήσεις για να κατατάξουμε τις υπόλοιπες $n-k$ παρατηρήσεις. Εναλλακτικά, η model-based προσέγγιση είναι να μοντελοποιούμε μαζί, τόσο τις παρατηρήσεις που γνωρίζουμε όσο και αυτές που δε γνωρίζουμε σε ποιες ομάδες ανήκουν κάτω από ένα κοινό μοντέλο. Οι εκτιμήσεις των παραμέτρων από αυτό το από κοινού μοντέλο μπορούν στη συνέχεια να χρησιμοποιηθούν για να κατατάξουμε τις $n-k$ παρατηρήσεις για τις οποίες δε γνωρίζουμε σε ποιες ομάδες πραγματικά ανήκουν. Η τεχνική αυτή μπορεί να

εφαρμοστεί και για τα μείγματα των factor analyzers της PGMM οικογένειας. Περιγραφή της model-based classification για τα μοντέλα της PGMM οικογένειας καθώς επίσης και παραδείγματα σε πραγματικά δεδομένα παρέχεται από τον McNicholas (2010). Το Model-based classification όμως μπορεί να επεκταθεί και να εφαρμοστεί και στην περίπτωση του μείγματος πολυμεταβλητών t-κατανομών της οικογένειας MMtFA (Andrews and McNicholas (2011b), Andrews et al. (2011)). Αυτό είναι άλλο ένα πλεονέκτημα της MMtFA οικογένειας.

Ένα μειονέκτημα της πολυμεταβλητής κανονικής κατανομής που χρησιμοποιείται ευρέως στο model-based clustering όσο και της πολυμεταβλητής t κατανομής, είναι πως υποθέτουν ότι έχουμε συμμετρικά δεδομένα, κάτι το οποίο είναι αρκετά περιοριστικό και δεν ισχύει απολύτως στην πράξη. Είναι γνωστό ότι ασύμμετρα δεδομένα μπορούν να προσεγγιστούν καλά από ένα μείγμα κατάλληλων κατανομών όπως η κανονική. Ενώ αυτό μπορεί να είναι ιδιαίτερα χρήσιμο για σκοπούς μοντελοποίησης, μπορεί να είναι ιδιαίτερα παραπλανητικό στην περίπτωση της ομαδοποίησης δεδομένων, καθώς κάποια πραγματική ομάδα μπορεί να αναπαρασταθεί από περισσότερες μόνο και μόνο επειδή τα δεδομένα της είναι μη συμμετρικά. Έτσι, θα καταλήξουμε να πάρουμε εσφαλμένα μεγαλύτερο αριθμό ομάδων από αυτόν που πραγματικά υπάρχει. Η συνήθης πρακτική για ασύμμετρα δεδομένα είναι να τα μετασχηματίζουμε έτσι ώστε να γίνουν όσο πιο συμμετρικά γίνεται. Ωστόσο, οι Karlis and Santourian (2009) πρότειναν τη χρήση μείγματος ασύμμετρων κατανομών ώστε να αναπαραστήσουν καλύτερα ασύμμετρα δεδομένα. Συγκεκριμένα, πρότειναν τη χρήση της αντίστροφης πολυμεταβλητής κανονικής κατανομής (multivariate normal inverse Gaussian (MNIG)) η οποία είναι μη συμμετρική. Οι εκτιμήσεις των παραμέτρων του μοντέλου μπορούν να βρεθούν με χρήση του EM αλγορίθμου. Έτσι αυτό το μοντέλο του μείγματος τέτοιων κατανομών προσφέρει ένα πλεονέκτημα στην ομαδοποίηση ασύμμετρων δεδομένων.

Κεφάλαιο 6: Εφαρμογή σε high-dimensional data

6.1 Εφαρμογή και συζήτηση

Στο τελευταίο αυτό κεφάλαιο θα γίνει μια προσπάθεια να εφαρμόσουμε το model-based clustering σε δεδομένα που προέρχονται από την microarray μελέτη των van 't Veer et al. (2002), προκειμένου να δούμε πως συμπεριφέρονται τα μοντέλα που περιγράφηκαν στα προηγούμενα κεφάλαια σε πραγματικά high-dimensional δεδομένα. Πιο συγκεκριμένα, τα δεδομένα αφορούν τη γενετική έκφραση 24.182 γονιδίων (μεταβλητών) όπως αυτή αποτυπώθηκε μέσω microarray μεθόδου για 78 γυναίκες (παρατηρήσεις) με καρκίνο του μαστού.

Οι γυναίκες ασθενείς της μελέτης έπασχαν από πρωτογενές διηθητικό καρκίνο του μαστού, μεγέθους μικρότερο από 5cm, χωρίς να έχουν εμφανίσει κάποια μετάσταση και χωρίς προηγούμενο ιστορικό κακοήθειας. Κατά την ημερομηνία διάγνωσης όλες ήταν μικρότερες των 55 ετών. Όλες οι ασθενείς υποβλήθηκαν σε μαστεκτομή ή σε συντηρητική θεραπεία με εκτομή του μασχαλικού λεμφαδένα ακολουθούμενη από ακτινοθεραπεία και παρακολουθήθηκαν για ένα διάστημα τουλάχιστο 5 ετών.

Από το σύνολο των 78 ασθενών, 34 εμφάνισαν μετάσταση μέσα στα 5 πρώτα έτη παρακολούθησης από την αρχική θεραπεία, ενώ 44 δεν εμφάνισαν. Στόχος της μελέτης ήταν να διερευνηθεί αν οι δύο ομάδες ασθενών, (την πρώτη ομάδα αποτελούν οι γυναίκες που εμφάνισαν μετάσταση είχαν δηλαδή "κακή πρόγνωση" και τη δεύτερη αυτές που παρέμειναν "καθαρές"), διαφέρουν ως προς τη γενετική τους έκφραση. Σε περίπτωση που η γενετική έκφραση διαφέρει, υπάρχουν δηλαδή γονίδια που εκφράζονται με διαφορετικό τρόπο στη μία κατηγορία γυναικών απ΄ ότι στην άλλη, αυτό μπορεί να αποτελέσει πολύ χρήσιμη πληροφορία για μελλοντικές ασθενείς. Απώτερος στόχος είναι μελετώντας το γενετικό προφίλ μελλοντικών ασθενών, να αναγνωρίζεται σε ποια από τις δύο ομάδες ανήκουν, και αν ανήκουν στην ομάδα με κακή πρόγνωση να ενισχύονται με περεταίρω θεραπείες.

Παρακάτω, θα προσπαθήσουμε να ομαδοποιήσουμε τις ασθενείς σε δύο ομάδες μέσω model-based clustering, χρησιμοποιώντας την πληροφορία που έχουμε για τη

γενετική έκφραση των 24.182 γονιδίων τους, αγνοώντας τη γνώση που έχουμε για την πραγματική τους ταξινόμηση. Ωστόσο, η γνώση αυτή θα μας επιτρέψει να ελέγξουμε τα αποτελέσματα της ομαδοποίησης του κάθε μοντέλου. Οι ελλείπουσες τιμές που υπήρχαν αντικαταστάθηκαν από τη διάμεσο κάθε μεταβλητής υποθέτοντας MCAR.

6.1.1 Ομαδοποίηση με 100 τυχαία γονίδια

Σαν μια πρώτη προσέγγιση, εντελώς διερευνητικά, παίρνουμε τυχαία 100 μεταβλητές (γονίδια). Λόγω του πολύ μεγάλου αριθμού μεταβλητών που έχουμε, τα μοντέλα της οικογένειας MCLUST δεν ενδείκνυνται. Έτσι, για αυτά τα δεδομένα (78 παρατηρήσεις και 100 μεταβλητές) εφαρμόζουμε model-based clustering χρησιμοποιώντας τα μοντέλα UUU της PGMM οικογένειας και UUC της MMtFA. Και τα 2 μοντέλα έτρεξαν για μια τυχαία αρχική διαμέριση των παρατηρήσεων στις 2 ομάδες. Οι αρχικές τιμές των πινάκων B και D καθορίστηκαν σύμφωνα με την περιγραφή που δόθηκε στη σελ. 54. Χρησιμοποιήθηκαν 5 παράγοντες και για τα δύο μοντέλα, ενώ για το μοντέλο UUC των t-factor analyzers χρησιμοποιήθηκαν αυθαίρετα 50 βαθμοί ελευθερίας. Ως συνθήκη τερματισμού για το UUC χρησιμοποιήθηκε η $\frac{l^{(k)} - l^{(k-1)}}{l^{(k)}} < \varepsilon$ με $\varepsilon = 10^{-8}$, ενώ για το UUU η $l_{\infty}^{(k+1)} - l^{(k+1)} < \varepsilon$ με $\varepsilon = 0.1$. Τα μοντέλα αυτά, όσο και τα υπόλοιπα που θα χρησιμοποιηθούν παρακάτω, έτρεξαν στο στατιστικό πακέτο R. Για το μοντέλο UUU χρησιμοποιήθηκε η έτοιμη συνάρτηση pgmmEM της pgmm βιβλιοθήκης, ενώ το UUC έτρεξε μέσω του κώδικα του παραρτήματος. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα:

Πίνακας 6.1: Αποτελέσματα ομαδοποίησης των μοντέλων UUU (PGMM) & UUC (MMtFA) για 100 τυχαία γονίδια.

	UUU (PGMM)		UUC (MMtFA)	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Metastasis (0)	11	23	15	19
Disease free (1)	14	30	23	21
Rand Index	0.495		0.497	
Adj. Rand Index	-0.011		-0.007	
BIC	2408.539		2628.368	

Η πρώτη διαπίστωση που κάνει κανείς κοιτώντας τα αποτελέσματα της ομαδοποίησης αυτών των δύο μοντέλων είναι ότι και τα δύο αποτυγχάνουν στο να επιτύχουν μια καλή ομαδοποίηση. Αυτό μπορεί να οφείλεται σε διάφορους λόγους. Μερικοί από αυτούς είναι: α) η τυχαία επιλογή των αρχικών τιμών μπορεί να μην ήταν η κατάλληλη, έχοντας ως αποτέλεσμα η λύση να μην αντιστοιχεί σε ολικό μέγιστο της πιθανοφάνειας αλλά σε τοπικό β) τα μοντέλα UUU και UUC των PGMM και MMtFA οικογενειών αντίστοιχα, μπορεί να μην είναι τα κατάλληλα για να περιγράψουν αυτά τα δεδομένα γ) η αυθαίρετη επιλογή των 5 παραγόντων και για τα 2 μοντέλα και των 50 βαθμών ελευθερίας για το UUC μπορεί να μην είναι η κατάλληλη δ) ίσως τα 100 γονίδια που χρησιμοποιήθηκαν να μην έχουν καλή διακριτική ικανότητα ως προς το διαχωρισμό των παρατηρήσεων στις 2 ομάδες.

Ωστόσο, όπως παρατηρούμε οι ομαδοποιήσεις που επιτυγχάνονται είναι ιδιαίτερα κακές, όπως άλλωστε υποδεικνύεται και από τους Adjusted Rand Index, οδηγώντας μας στο συμπέρασμα ότι οι 3 πρώτοι λόγοι μάλλον δεν ευθύνονται για το αποτέλεσμα καθώς αν πράγματι υπήρχε κάποια δομή στα δεδομένα αυτή θα έπρεπε να αναδειχτεί σε κάποιο έστω μικρό βαθμό παρά την όποια κακή επιλογή στα βήματα α έως γ. Πράγματι, για διαφορετικές εκτελέσεις των αλγορίθμων για διάφορες τυχαίες αρχικές τιμές και διαφορετικό αριθμό παραγόντων και βαθμών ελευθερίας, οι ομαδοποιήσεις που προκύπτουν είναι εξίσου κακές. Έτσι, οδηγούμαστε στο συμπέρασμα ότι αιτία της κακής ομαδοποίησης είναι μάλλον ότι τα 100 γονίδια που επιλέχθηκαν δεν έχουν κάποια πληροφορία σχετικά με το διαχωρισμό των ασθενών στις 2 ομάδες.

Δύο στοιχεία που αξίζει να σημειωθούν επιπλέον για αυτή την ανάλυση είναι: i) παρατηρούμε πόσο διαφορετικά είναι τα αποτελέσματα μεταξύ των δεικτών Rand Index και Adjusted Rand Index. Είναι προφανές πως ο Rand Index στη συγκεκριμένη περίπτωση υπερεκτιμά την επίδοση της ομαδοποίησης των αλγορίθμων και γίνεται αντιληπτό γιατί αυτός ο δείκτης δεν είναι κατάλληλος για τη σύγκριση ομαδοποιήσεων. Αντίθετα, ο Adjusted Rand Index που διορθώνει τα προβλήματα του Rand Index (παράγραφος 2.8) μας δίνει αποτέλεσμα πιο συμβατό στην πραγματικότητα και πιο κοντά στο crosstabulation. ii) Επίσης αυτό που μπορεί να παρατηρήσει κανείς είναι την ελάχιστη υπεροχή που φαίνεται να έχει η χρήση της t κατανομής έναντι της κανονικής, καθώς το UUC μοντέλο οδηγεί σε μεγαλύτερη τιμή BIC κριτηρίου και ελάχιστα υψηλότερης τιμής Adjusted Rand Index. Ωστόσο, η διαφορά στην τιμή του Adjusted Rand Index είναι ελάχιστη και μπορεί αυτή να οφείλεται στην επιλογή των αρχικών τιμών και μόνο. Έτσι, δεν μπορεί κανείς να είναι

σίγουρος για την υπεροχή της t κατανομής για αυτά τα δεδομένα κοιτώντας μόνο τα αποτελέσματα αυτής της παραπάνω ανάλυσης.

Θέλοντας κανείς να προχωρήσει σε ομαδοποίηση πιο σοβαρά, κανονικά δε θα έπρεπε να χρησιμοποιήσει μόνο 100 τυχαία γονίδια, όπως έγινε παραπάνω για εντελώς διερευνητικούς λόγους, αλλά θα έπρεπε να χρησιμοποιήσει όλο το πλήθος των γονιδίων για τα οποία υπάρχει διαθέσιμη πληροφορία. Επειδή όμως το πλήθος των γονιδίων (~25.000), άρα και μεταβλητών, στη συγκεκριμένη περίπτωση είναι πραγματικά υπέρογκο για μια τέτοιου είδους ανάλυση, τεχνικές και τρόποι μείωσης των μεταβλητών που θα χρησιμοποιηθούν φαντάζουν κάτι παραπάνω από απαραίτητες. Μια τέτοια τεχνική περιγράφεται στην εργασία των McLachlan et al. (2002) και αφορά την EMMIX-GENE προσέγγιση όπως έχει αναφερθεί και προηγουμένως στην 3.1 παράγραφο.

Πρέπει να τονιστεί ότι παρά τη μείωση των διαστάσεων που επιτυγχάνεται από τη χρήση των factor analyzers, ένα μέγεθος μεταβλητών της τάξης των 25.000 είναι εξαιρετικά μεγάλο ακόμα και για αυτές τις μεθόδους. Χρησιμοποιώντας ένα τόσο μεγάλο αριθμό μεταβλητών δημιουργούνται άμεσα προβλήματα υπολογιστικής φύσεως (numerical problems). Ένα από αυτά έχει να κάνει με τον υπολογισμό της ορίζουσας του πίνακα D στον τύπο $\left|\mathrm{BB}^{\mathrm{T}}+D\right|=\left|D\right|/\left|I_q-B^T\left(BB^T+D\right)^{-1}B\right|$. Εάν ο αριθμός των μεταβλητών είναι πολύ μεγάλος και τύχει μεγάλος αριθμός των στοιχείων του πίνακα D να είναι μικρότερος της μονάδας (αυτό συμβαίνει κυρίως αν δεν ταξινομηθούν περισσότερες από q παρατηρήσεις στην j ομάδα του μείγματος πολυμεταβλητών κανονικών κατανομών), επειδή ο πίνακας D είναι διαγώνιος και άρα η ορίζουσά του δίνεται ως το γινόμενο των διαγωνίων στοιχείων, επειδή πολλά από αυτά είναι μικρότερα της μονάδας, πολλαπλασιαζόμενα μεταξύ τους δίνουν έναν πολύ πολύ μικρό αριθμό ο οποίος ξεπερνά την ακρίβεια μέτρησης δεκαδικών ψηφίων του στατιστικού πακέτου, μηδενίζοντας έτσι την ορίζουσα. Αυτό όμως έχει ως συνέπεια η log-likelihood να γίνεται άπειρη και ο αλγόριθμος να κολλάει. Για την ακρίβεια, αυτό το πρόβλημα μπορεί να προκύψει για πολύ μικρότερο αριθμό από 25.000 μεταβλητές, όπως της τάξης των 500 ή και παρακάτω μεταβλητών.

Ακόμα όμως και σε περίπτωση που τύχει να μην έχουμε προβλήματα τέτοιας φύσεως, για να τρέξει ένα μοντέλο με τόσες πολλές μεταβλητές απαιτείται τεράστια υπολογιστική ισχύς. Καταλαβαίνει κανείς ότι από τη στιγμή που πρέπει να τρέξουμε ένα μοντέλο για διαφορετικό πλήθος τυχαίων αρχικών τιμών, διαφορετικό αριθμό

παραγόντων, ομάδων ή/και βαθμών ελευθερίας προκειμένου να πετύχουμε την καλύτερη ομαδοποίηση συγκρίνοντας τις BIC τιμές, ένα μεγάλο πλήθος μεταβλητών είναι απαγορευτικό. Έτσι, η ομαδοποίηση high-dimensional δεδομένων, ακόμα και με μοντέλα που μειώνουν τον αριθμό μεταβλητών παραμένει ένα εξαιρετικά δύσκολο πρόβλημα.

6.1.2 Ομαδοποίηση με "κατάλληλα" επιλεγμένα γονίδια

Μετά από όσα αναφέρθηκαν παραπάνω, το να προχωρήσει κανείς σε ομαδοποίηση των δεδομένων του παραδείγματος χρησιμοποιώντας όλο το πλήθος γονιδίων, είναι ανέφικτο. Αντ' αυτού θα επιλέξουμε "κατάλληλα" έναν μικρότερο αριθμό γονιδίων βασιζόμενοι στην EMMIX-GENE διαδικασία των McLachlan et al. (2002) και θα προχωρήσουμε σε ομαδοποίηση με αυτά. Η διαδικασία βάση της οποίας θα επιλέξουμε τα "κατάλληλα" γονίδια περιγράφεται παρακάτω.

Για κάθε γονίδιο, δηλαδή για κάθε μεταβλητή, εφαρμόζουμε model-based clustering χρησιμοποιώντας μονομεταβλητή κανονική κατανομή (μέσω της συνάρτησης Mclust της βιβλιοθήκης mclust της R) και υπολογίζουμε την ποσότητα $-2\log\lambda$, όπου λ ο λόγος πιθανοφανειών για g=1 έναντι g=2 ομάδες. Αν η ποσότητα $-2\log\lambda$ είναι μεγαλύτερη από ένα προκαθορισμένο όριο b_1, $-2\log\lambda > b_1$ (6.1), τότε το γονίδιο επιλέγεται να χρησιμοποιηθεί περεταίρω, υπό την προϋπόθεση ότι $s_{\min} \geq b_2$ (6.2), όπου s_{min} είναι ο αριθμός παρατηρήσεων που περιέχονται στο μικρότερο cluster. Εάν η 6.1 δεν ικανοποιείται το γονίδιο απορρίπτεται. Το γεγονός ότι ισχύει $-2\log\lambda > b_1$ σημαίνει ότι οι 2 ομάδες πλεονεκτούν έναντι της μίας και άρα το γονίδιο είναι κατάλληλο να χρησιμοποιηθεί σε clustering 2 ομάδων. Η $2^{η}$ συνθήκη $s_{\min} \geq b_2$, χρειάζεται διότι μπορεί ο λόγος πιθανοφανειών να υποδεικνύει ομαδοποίηση των δεδομένων του γονιδίου σε 2 ομάδες, αλλά αν η μία ομάδα είναι πολύ μικρή, αυτή η ομάδα μπορεί να αποτελείται αποκλειστικά και μόνο από ακραίες παρατηρήσεις. Δηλαδή, αν οι παρατηρήσεις ενός γονιδίου στην πραγματικότητα συγκροτούν μία μόνο ομάδα, αλλά υπάρχουν και κάποια outliers, τότε ο λόγος πιθανοφανειών θα υποδεικνύει εσφαλμένα ομαδοποίηση σε 2 ομάδες, με τη μία ομάδα να αποτελείται μόνο από τις ακραίες παρατηρήσεις. Έτσι, αυτή η συνθήκη μας προφυλάσσει από αυτήν την περίπτωση. Στο σημείο αυτό πρέπει να τονιστεί ότι η χρήση t κατανομής

θα ήταν πιο ενδεδειγμένη στη συγκεκριμένη περίπτωση, καθώς είναι πιο ανθεκτική σε ακραίες παρατηρήσεις. Ωστόσο, εδώ σε πρώτη φάση θα χρησιμοποιηθεί η κανονική.

Εάν η 6.1 ισχύει αλλά δεν ισχύει η 6.2 τότε υπολογίζουμε ξανά την ποσότητα $-2\log\lambda > b_1$, αλλά τώρα όπου λ είναι ο λόγος πιθανοφανειών για g=2 έναντι g=3 ομάδων. Αν η 6.1 ισχύει για αυτή την τιμή του $-2\log\lambda$, το γονίδιο επιλέγεται ως "κατάλληλο", αρκεί 2 από τις 3 ομάδες που αφορούν την ομαδοποίηση για g=3 να περιέχουν τουλάχιστο b_2 παρατηρήσεις. Σε διαφορετική περίπτωση το γονίδιο απορρίπτεται.

Παρ' ότι η κατανομή της ποσότητας $-2\log\lambda$ για g=1 vs g=2 δεν είναι ίδια με g=2 vs g=3 θα χρησιμοποιήσουμε το ίδιο όριο b_1. Επίσης η κατανομή της ποσότητας $-2\log\lambda$ για $g=g_0$ vs $g=g_1$ δεν είναι γνωστή για μοντέλα μίξεων κατανομών. Έτσι εδώ αυθαίρετα θα χρησιμοποιήσουμε $b_1 = 10$ και $b_2 = 10$. Είναι προφανές ότι όσο αυξάνονται οι τιμές των b_1 και b_2, δηλαδή όσο πιο αυστηρά κριτήρια θέτουμε τόσο λιγότερα γονίδια επιλέγουμε.

Εκτελώντας λοιπόν την παραπάνω διαδικασία για $b_1 = 10$ και $b_2 = 10$ και υποθέτοντας ίση διασπορά μεταξύ των ομάδων κατά την εκτέλεση της ομαδοποίησης μέσω Mclust (μόνο και μόνο επειδή για κάποια γονίδια, μοντέλα με άνισες διασπορές δε δουλεύουνε) καταλήγουμε σε έναν αριθμό 646 γονιδίων. Με βάση αυτά τα 646 γονίδια προχωράμε σε ομαδοποίηση.

Εφαρμόστηκε model-based clustering για όλα τα μοντέλα της PGMM οικογένειας μέσω της συνάρτησης pgmmEM της pgmm βιβλιοθήκης της R, για g=2 ομάδες, q=1,2,3,4,5,10,15,30 παράγοντες χρησιμοποιώντας ως αρχικές τιμές τα αποτελέσματα του K-means αλγορίθμου. Τα αποτελέσματα του καλύτερου μοντέλου για κάθε περίπτωση συνοψίζονται στον παρακάτω πίνακα.

Πίνακας 6.2: Αποτελέσματα ομαδοποίησης της PGMM οικογένειας για τα 646 "κατάλληλα" επιλεγμένα γονίδια.

		q=1	q=2	q=3	q=4
g=2	Best Model	CCU	UCU	UCU	CUU
	BIC	20655.51	18015.26	16889.74	21955.74*
	Adjusted Rand Index	0.082	0.067	0.067	0.082
		q=5	q=10	q=15	q=30
g=2	Best Model	UCU	CUU	UCU	CCC
	BIC	13898.14	17909.72	-10870.98	-39743.37
	Adjusted Rand Index	0.067	0.082	0.082	0.082

*Μεγαλύτερη τιμή BIC

Από τον πίνακα αυτό βλέπουμε ότι καλύτερο μοντέλο βάση του BIC κριτηρίου είναι το CUU για q=4 παράγοντες. Επίσης, το BIC συμφωνεί με την καλύτερη ομαδοποίηση που επιτυγχάνεται ξανά για q=4 σε ισοδυναμία όμως με q=1,10,15,30. Αυτή η καλύτερη ομαδοποίηση απεικονίζεται παρακάτω.

Πίνακας 6.3: Καλύτερη ομαδοποίηση της PGMM οικογένειας για τα 646 γονίδια.

	CUU, q=4	
	Cluster 1	Cluster 2
Metastasis (0)	22	12
Disease free (1)	39	5

Παρ' όλο που οι δείκτες συμφωνίας αυξήθηκαν σε σχέση με την ανάλυση των 100 τυχαίων γονιδίων, όλοι παραμένουν πολύ μικροί και υποδεικνύουν κακή ομαδοποίηση για όλες τις περιπτώσεις. Ο πιο πιθανός λόγος που συμβαίνει αυτό είναι ότι κατά την παραπάνω διαδικασία επιλογής των "κατάλληλων" γονιδίων δεν επιλέχθηκαν καλά γονίδια, δηλαδή γονίδια με καλή διακριτική ικανότητα ως προς τις 2 ομάδες. Αυτό μπορεί να συνέβη είτε γιατί χρησιμοποιήσαμε την κανονική κατανομή η οποία δεν είναι ανθεκτική σε ακραίες παρατηρήσεις, είτε επειδή η επιλογή των ίσων διασπορών κατά την εκτέλεση της Mclust δεν ισχύει, είτε επειδή η μέθοδος δεν είναι κατάλληλη εκ κατασκευής, είτε επειδή δεν υπάρχουν γονίδια με καλή διακριτική ικανότητα στο δείγμα. Ωστόσο, οδηγηθήκαμε σε καλύτερα αποτελέσματα από την ανάλυση των 100 τυχαίων γονιδίων.

Τέλος, για αυτά τα 646 γονίδια εφαρμόσαμε όλα τα μοντέλα της PGMM για 1 έως 5 ομάδες για q=4 παράγοντες με K-means αρχικές τιμές. Τα αποτελέσματα φαίνονται παρακάτω.

Πίνακας 6.4: Αποτελέσματα ομαδοποιήσεων για τα 646 γονίδια για διάφορους αριθμούς ομάδων.

		g=1	g=2	g=3	g=4	g=5
q=4	Best Model	CCU	CUU	CCU	CCU	CCU
	BIC	9495.513	21955.74*	20530.11	19528.05	17219.28
	Adjusted Rand Index	-	0.082	0.109	0.034	0.061

*Μεγαλύτερη τιμή BIC

Από εδώ διαπιστώνουμε ότι ενώ το clustering αποτυγχάνει βάσει αυτών των 646 γονιδίων, καθώς όλες οι ομαδοποιήσεις που παίρνουμε δεν είναι ικανοποιητικές, ωστόσο επιτυγχάνει στο να αναδείξει το σωστό αριθμό ομάδων, δηλαδή 2.

Εάν για την επιλογή των "κατάλληλων" γονιδίων στη διαδικασία που περιγράφηκε προηγουμένως χρησιμοποιήσουμε την t κατανομή (μέσω της συνάρτησης teigen της βιβλιοθήκης teigen της R), που είναι πιο κατάλληλη καθ' ότι είναι πιο ανθεκτική σε ακραίες παρατηρήσεις και επιτρέψουμε ο αριθμός των βαθμών ελευθερίας να εκτιμάται από το μοντέλο και χρησιμοποιήσουμε τα ίδια κριτήρια b_1 και b_2, τότε καταλήγουμε να επιλέξουμε έναν αριθμό 1904 γονιδίων. Ο αριθμός αυτός είναι αρκετά μεγάλος και τεχνικές περεταίρω μείωσης των διαστάσεων είναι απαραίτητες (π.χ. 2^o στάδιο της EMMIX-GENE διαδικασίας).

6.1.3 Ομαδοποίηση με βάση τα "καλύτερα" γονίδια

Είδαμε προηγουμένως ότι τα 646 γονίδια που επελέγησαν με τον τρόπο που αναφέρθηκε παραπάνω απέτυχαν να δώσουν καλή ομαδοποίηση. Αυτό, εκτός των άλλων (κατάλληλη οικογένεια κατανομών, μοντέλο, αριθμός παραγόντων, αρχικές τιμές), οφείλεται και στο γεγονός ότι τα γονίδια που επελέγησαν δεν είχαν καλή διακριτική ικανότητα. Αν υπήρχε κάποιος πιο αποδοτικός τρόπος να μειώναμε τον αρχικό αριθμό διαστάσεων, επιλέγοντας γονίδια με καλή διακριτική ικανότητα, αν φυσικά αυτά υπάρχουν στο δείγμα μας, τότε τα αποτελέσματα του model-based clustering θα ήταν σίγουρα ανώτερα. Για να το διαπιστώσουμε αυτό ταξινομούμε τα

γονίδια σε φθίνουσα σειρά διακριτικής ικανότητας με βάση το Lamda του Wilks. Αυτό μπορούμε να το κάνουμε καθώς γνωρίζουμε την πραγματική ταξινόμηση των ασθενών. Η διαδικασία αυτή γίνεται απλά για να αξιολογήσουμε την ομαδοποίηση που θα επιτυγχάνονταν αν πραγματικά γνωρίζαμε εκείνα τα γονίδια που συνεισφέρουν περισσότερο στην ομαδοποίηση. Ωστόσο, επειδή αυτά τα γονίδια σε πραγματικές εφαρμογές δεν τα γνωρίζουμε η όλη προσέγγιση δεν έχει κάποιο ρεαλιστικό ενδιαφέρον.

Έτσι, για τις πρώτες 646 καλύτερες μεταβλητές τρέχουμε όλα τα μοντέλα της PGMM οικογένειας (μέσω της pgmmEM συνάρτησης της R), για 2 ομάδες, για q=1 έως 10 παράγοντες και χρησιμοποιώντας K-means αρχικές τιμές. Το καλύτερο μοντέλο βάση του BIC κριτηρίου βρέθηκε ότι είναι το CCU για 5 παράγοντες, το οποίο δίνει Adjusted Rand Index ίσο με 0.157. Ωστόσο, η καλύτερη ομαδοποίηση επιτεύχθηκε για 3 παράγοντες υπό το CUU μοντέλο και είχε Adjusted Rand Index ίσο με 0.31. Δηλαδή, ξανά παρά το ότι χρησιμοποιήσαμε τα 646 "καλύτερα" γονίδια η καλύτερη ομαδοποίηση που επετεύχθη, η οποία απεικονίζεται στον ακόλουθο πίνακα, δεν είναι ικανοποιητική.

Πίνακας 6.5: Αποτελέσματα ομαδοποίησης του CUU (PGMM) για 3 παράγοντες για τα 646 καλύτερα γονίδια βάση του Lamda Wilks.

	CUU, q=3	
	Cluster 1	Cluster 2
Metastasis (0)	4	30
Disease free (1)	31	13
Adjusted Rand Index	0.309	

Επίσης, εδώ παρατηρούμε ότι το BIC κριτήριο δε συμβαδίζει με την καλύτερη ομαδοποίηση, γεγονός που δημιουργεί αμφιβολίες σχετικά με τη χρησιμότητά του στην εύρεση του καταλληλότερου μοντέλου που αντιστοιχεί στην καλύτερη ομαδοποίηση. Για αυτό και κριτήρια όπως το BIC ή το ICL δεν πρέπει να θεωρούνται ότι διαλέγουν πάντα την καλύτερη ομαδοποίηση, όπως άλλωστε υποδεικνύουν και οι Andrews and McNicholas (2011a, page 370).

Τέλος, επαναλαμβάνουμε την παραπάνω ανάλυση για τα 20 έως 80 πρώτα καλύτερα γονίδια. Η καλύτερη ομαδοποίηση βάση του adjusted Rand Index επιτεύχθηκε χρησιμοποιώντας τα 45 πρώτα γονίδια. Το καλύτερο μοντέλο για αυτή την περίπτωση βάση του BIC κριτηρίου βρέθηκε ότι είναι το CCU για 1 παράγοντα, το οποίο έχει Adjusted Rand Index ίσο με 0.473. Βλέπουμε δηλαδή ότι εδώ η

ομαδοποίηση βελτιώθηκε αισθητά. Αυτό σημαίνει ότι τα επιπλέον 601 γονίδια που χρησιμοποιήθηκαν προηγουμένως δεν έχουν καλή διακριτική ικανότητα και εισάγουν μάλλον θόρυβο στην ανάλυση, παρά βοηθούν στην ομαδοποίηση.

Αυτό που γίνεται αντιληπτό από την εφαρμογή αυτή είναι η ανάγκη μείωσης του αρχικού αριθμού γονιδίων σε ένα λογικό αριθμό ο οποίος να μας επιτρέπει να προχωρήσουμε σε model-based clustering ξεπερνώντας υπολογιστικά προβλήματα και προβλήματα υπολογιστικής ισχύος. Ωστόσο, το ποια γονίδια θα επιλεγούν για να χρησιμοποιηθούν στο model-based clustering είναι κρίσιμο και ιδιαίτερα δύσκολο πρόβλημα και καθορίζει σε μεγάλο βαθμό την επιτυχία της ομαδοποίησης από εκεί και πέρα. Φυσικά η οικογένεια μοντέλων που θα χρησιμοποιηθεί, το ίδιο το μοντέλο, ο αριθμός παραγόντων και οι αρχικές τιμές παίζουν και αυτά ρόλο στην καλή ομαδοποίηση.

Σύνοψη

Στα προηγούμενα κεφάλαια περιγράφηκε η μεθοδολογία του model-based clustering χρησιμοποιώντας μίξεις πολυμεταβλητών κανονικών και t κατανομών ως μια εναλλακτική μέθοδος συσταδικής ανάλυσης έναντι των κλασικών distance-based μεθόδων ομαδοποίησης. Τα πλεονεκτήματα του model-based clustering είναι ότι βασίζεται σε πιθανοθεωρητικά μοντέλα τα οποία λαμβάνουν υπ' όψιν τη διακύμανση και συνδιακύμανση των μεταβλητών, σε αντίθεση με τις προσεγγίσεις των distance-based μεθόδων ομαδοποίησης οι οποίες είναι κυρίως μαθηματικές. Ωστόσο, το πιο σημαντικό πλεονέκτημα είναι η δυνατότητα στατιστικής συμπερασματολογίας που παρέχεται, κάτι που δεν είναι εφικτό μέσω των κλασικών αλγορίθμων ομαδοποίησης (hierarchical analysis, K-means) και καθιστά το model-based clustering ιδιαίτερα ελκυστικό.

Ένα άλλο συγκριτικό πλεονέκτημα είναι η δυνατότητα ομαδοποίησης high dimensional δεδομένων μέσω της χρήσης των factor analyzers. Ειδικά στην περίπτωση των high dimensional δεδομένων οι κλασικοί αλγόριθμοι ομαδοποίησης αποτυγχάνουν κυρίως εξαιτίας της τεράστιας υπολογιστικής ισχύος και του χρόνου που απαιτείται για να τρέξουν. Έτσι, στις περιπτώσεις αυτές η ομαδοποίηση μέσω model-based μεθόδων καθίσταται όχι μόνο ελκυστική αλλά αποτελεί μονόδρομο. Επιπλέον των παραπάνω, οι διάφοροι περιορισμοί που μπορούν να τεθούν στους πίνακες διακύμανσης-συνδιακύμανσης των ομάδων μέσω των (2.11) και (3.4) οδηγούν στην παραγωγή ενός πλήθους διαφορετικών μοντέλων. Με τον τρόπο αυτό, παρέχεται κάθε φορά η δυνατότητα επιλογής του κατάλληλου μοντέλου, μέσα από μια οικογένεια διαφορετικών μοντέλων. Πέρα όμως από τη δυνατότητα επιλογής εκείνου του μοντέλου που οδηγεί σε καλύτερα αποτελέσματα ομαδοποίησης, οι περιορισμοί που τίθενται παρέχουν και τη δυνατότητα μείωσης των παραμέτρων προς εκτίμηση, οδηγώντας σε πιο φειδωλά και ανθεκτικά μοντέλα. Αυτό το γεγονός είναι ιδιαίτερα χρήσιμο όταν ο αριθμός των μεταβλητών ή/και των ομάδων είναι πολύ μεγάλος.

Επίσης, στο model-based clustering οι παρατηρήσεις κατατάσσονται σε ομάδες με κάποια πιθανότητα και έτσι μπορούμε να έχουμε ομάδες οι οποίες αλληλοεπικαλύπτονται. Αυτό το αποτελέσμα είναι όχι μόνο ρεαλιστικό αλλά δεν μπορεί να επιτευχθεί μέσω μεθόδων αποστάσεων. Τέλος, παρέχεται η δυνατότητα για

καλύτερη μοντελοποίηση ακραίων παρατηρήσεων μέσω της χρήσης t πολυμεταβλητών κατανομών, κάτι που δεν είναι επίσης εφικτό μέσω των distance-based μεθόδων.

Βάση όλων των ανωτέρω, το model-based clustering εμφανίζεται να είναι ένα ιδιαίτερα ισχυρό εργαλείο στην ομαδοποίηση δεδομένων και κυρίως στην περίπτωση high-dimensional δεδομένων. Ωστόσο, δεν παύει και αυτή η προσέγγιση να έχει μειονέκτηματα καθώς εμφανίζονται αρκετά προβλήματα κατά την εφαρμογή του.

Ένα σοβαρό μειονέκτημα είναι το γεγονός ότι οι αλγόριθμοι EM και AECM οι οποίοι χρησιμοποιούνται για την εκτίμηση των παραμέτρων των μοντέλων εξαρτώνται σε μεγάλο βαθμό από τις αρχικές τιμές. Κατά συνέπεια τα οποιαδήποτε αποτελέσματα είναι ευαίσθητα στις αρχικές τιμές που χρησιμοποιούνται. Για το λόγο αυτό προτείνεται τα μοντέλα να χρησιμοποιούνται πολλές φορές για πλήθος διαφορετικών αρχικών τιμών ώστε να είναι πιο πιθανό να πετύχουμε τη βέλτιση λύση. Ωστόσο, και πάλι δεν εξασφαλίζεται ότι η χρήση πολλών διαφορετικών αρχικών τιμών θα οδηγήσει σε καλύτερα αποτελέσματα ομαδοποίησης. Εξαιτίας αυτής της ευαισθησίας στις αρχικές τιμές έχουν προταθεί διάφοροι τρόποι καθορισμού αυτών. Έτσι, οι αρχικές τιμές μπορεί να βασίζονται στα αποτελέσματα κάποιου ευρετικού αλγόριθμου ομαδοποίησης, στην τυχαιότητα, στη φασματική ανάλυση του πίνακα διασποράς-συνδιασποράς, να ορίζονται μέσω κάποιας προσομοίωσης κ.α. Όλοι όμως αυτοί οι τρόποι είναι αυθαίρετοι και δεν υπάρχει κάποιο κριτήριο που να καθορίζει ποιος είναι ο καλύτερος τρόπος προσδιορισμού των αρχικών τιμών κάθε φορά. Έτσι, εναπόκεινται στον ερευνητή να καθορίσει μόνος του τις αρχικές τιμές ανάλογα με το πρόβλημα που αντιμετωπίζει. Ωστόσο, αν οι ομάδες μέσω των δεδομένων είναι ευδιάκριτες, ο τρόπος καθορισμού των αρχικών τιμών έχει μικρότερη σημασία καθώς αναμένουμε η σωστή δομή των δεδομένων θα αναδειχτεί για οποιεσδήποτε αρχικές τιμές.

Ένα άλλο πρόβλημα του model-based clustering είναι η μεγάλη υπολογιστική ισχύς που απαιτείται ιδιαίτερα στην περίπτωση high dimensional δεδομένων. Όταν οι παρατηρήσεις είναι πολλές, αλλά κυρίως όταν ο αριθμός των μεταβλητών ή/και των ομάδων είναι μεγάλος, άρα και οι παράμετροι προς εκτίμηση είναι πολλοί, τότε ο χρόνος υπολογισμού και η υπολογιστική ισχύς που απαιτείται αυξάνονται δραματικά. Ειδικά στην περίπτωση που πρέπει να εκτελεστεί ο ίδιος αλγόριθμος πολλές φορές για διαφορετικές αρχικές τιμές εκεί το πρόβλημα εντείνεται. Για το λόγο αυτό

απαιτείται η εξεύρεση τρόπων επιτάχυνσης των αλγορίθμων ώστε αυτοί να γίνουν λιγότερο υπολογιστικά απαιτητικοί.

Παρ' ότι το model-based clustering είναι ιδανικό για περιπτώσεις high-dimensional δεδομένων μέσω της χρήσης των factor analyzers, εντούτοις στην περίπτωση που ο αριθμός των μεταβλητών είναι της τάξης των χιλιάδων, δημιουργούνται αριθμητικά προβλήματα (spikes, singularities) κατά τον υπολογισμό της log-likelihood. Ένα παράδειγμα αφορά το μηδενισμό της ορίζουσας του πίνακα ιδιαιτεροτήτων των ομάδων όπως έχει ήδη αναφερθεί και στην παράγραφο 6.1.1. Ακόμα όμως και σε περίπτωση που τύχει να μην έχουμε αριθμητικά προβλήματα, για να τρέξει ένα μοντέλο με τόσες πολλές μεταβλητές απαιτείται τεράστια υπολογιστική ισχύς. Έτσι, είναι απαραίτητο να αναπτυχθούν πρωτογενώς, αποδοτικές τεχνικές μείωσης των αρχικών μεταβλητών σε ένα μικρότερο αριθμό (π.χ. EMMIX-GENE προσέγγιση), όπου η ομαδοποίηση θα μπορεί να επιτευχθεί εν συνεχεία μέσω της χρήσης των factor analyzers.

Πέραν όλων των παραπάνω, το πιο σημαντικό πρόβλημα αφορά την εξεύρεση ενός κριτηρίου το οποίο θα επιλέγει κάθε φορά το καλύτερο μοντέλο, εκείνο δηλαδή που επιτυγχάνει την σωστότερη ομαδοποίηση μέσα από όλα τα μοντέλα της εκάστοτε οικογένειας. Μέχρι στιγμής αν και η επίδοση των κριτηρίων όπως το BIC και το ICL έχει φανεί ικανοποιητική, καθώς τις περισσότερες φορές επιλέγουν σωστά το καταλληλότερο μοντέλο, εντούτοις αυτό δε συμβαίνει πάντα. Για αυτό και η χρήση τους μέχρις στιγμής είναι περισσότερο ενδεικτική παρά απόλυτη. Έτσι, η εξεύρευση κάποιου αποτελεσματικού κριτηρίου το οποίο θα μπορεί να εφαρμοστεί καθολικά για την επιλογή μοντέλου είναι απαραίτητη και αποτελεί αντικείμενο μελλοντικής έρευνας.

Τέλος, ένα άλλο πρόβλημα με το model-based clustering έχει να κάνει με το γεγονός ότι δεν έχει αναπτυχθεί ευρέως κατάλληλο λογισμικό το οποίο να υλοποιεί τα μοντέλα των διαφόρων οικογενειών. Τουλάχιστον προς το παρόν το model-based clustering μπορεί να εφαρμοστεί για συγκεκριμένες οικογένειες μοντέλων μόνο μέσω του στατιστικού πακέτου R, καθώς κανένα άλλο από τα ευρέως χρησιμοποιούμενα στατιστικά πακέτα (Stata, SAS, SPSS) δεν παρέχει ανάλογη δυνατότητα.

Περίληψη

Η παρούσα εργασία αφορά τη μεθοδολογία του model-based clustering ως μια εναλλακτική προσέγγιση στο πρόβλημα ομαδοποίησης δεδομένων έναντι των κλασικών μεθόδων που βασίζονται στην έννοια της απόστασης. Στην εργασία αυτή ιδιαίτερη έμφαση δίνεται στην περίπτωση εφαρμογής του model-based clustering σε high-dimensional δεδομένα καθώς και στη χρήση της πολυμεταβλητής t κατανομής αντί της πολυμεταβλητής κανονικής που χρησιμοποιείται ευρέως στην πράξη. Επίσης παρέχονται αρκετά παραδείγματα και εφαρμογές για την καλύτερη επεξήγηση και κατανόηση των μεθόδων.

Πιο συγκεκριμένα, πριν την περίπτωση των high-dimensional δεδομένων και της t πολυμεταβλητής κατανομής γίνεται μια εισαγωγή στη μεθοδολογία του model-based clustering. Περιγράφεται η θεωρία για μίξεις πολυμεταβλητών κανονικών κατανομών και πως αυτές χρησιμοποιούνται στο πλαίσιο του model-based clustering καθώς επίσης γίνεται και μια εκτενής αναφορά στη χρήση του ΕΜ αλγορίθμου για την εκτίμηση των παραμέτρων των διάφορων μοντέλων. Επιπλέον, περιγράφονται πλήρως τα μοντέλα της GPCM οικογένειας και μελετώνται αμφιλεγόμενα ζητήματα του model-based clustering όπως ο τρόπος επιλογής μοντέλου, σωστού αριθμού ομάδων, κατάλληλων αρχικών τιμών κ.α.

Στη συνέχεια παρουσιάζεται η περίπτωση ομαδοποίησης high-dimensional δεδομένων. Αναφέρονται τα προβλήματα που υπάρχουν στην εφαρμογή της GPCM οικογένειας μοντέλων για αυτή την περίπτωση και εισάγεται η χρήση των factor-analyzers για αυτό το σκοπό. Επίσης, περιγράφεται πλήρως η εφαρμογή του AECM αλγορίθμου για την εκτίμηση των παραμέτρων.

Επιπρόσθετα, παρουσιάζονται δύο οικογένειες μοντέλων (PGMM και EPGMM) κατάλληλες για την ομαδοποίηση high-dimensional δεδομένων, οι οποίες βασίζονται σε μίξεις πολυμεταβλητών κανονικών κατανομών (στην ουσία πρόκειται για μία μόνο οικογένεια καθώς η PGMM είναι υποσύνολο της EPGMM). Παρέχονται παραδείγματα εφαρμογών των μοντέλων αυτών των οικογενειών αλλά ταυτόχρονα αναλύονται τα πλεονεκτήματα και μειονεκτήματά τους.

Κατόπιν, παρουσιάζεται η περίπτωση της χρήσης της t πολυμεταβλητής κατανομής και τα οφέλη που αυτή παρέχει, τόσο για high-dimensional δεδομένα όσο και για μη. Επίσης, περιγράφονται πλήρως οι αλγόριθμοι AECM και ΕΜ που

χρησιμοποιούνται για εκτίμηση των παραμέτρων αντίστοιχα. Επιπλέον, περιγράφεται και η MMtFA οικογένεια μοντέλων, η οποία είναι ανάλογη της EPGMM για την περίπτωση της t κατανομής.

Τέλος, περιγράφεται μια εφαρμογή των μοντέλων των PGMM και MMtFA οικογενειών σε high dimensional δεδομένα από τη μελέτη έκφρασης γονιδίων των van 't Veer et al. (2002). Τα δεδομένα αφορούν τη γενετική έκφραση 24.182 γονιδίων (μεταβλητών) όπως αυτή αποτυπώθηκε μέσω microarray μεθόδου για 78 γυναίκες (παρατηρήσεις) με καρκίνο του μαστού. Αρχικά εφαρμόζονται τα μοντέλα UUU (PGMM) και UUC (MMtFA) για 100 τυχαία γονίδια. Στη συνέχεια επιλέγονται μέσω μιας τεχνικής παρόμοιας της EMMIX-GENE (χρησιμοποιώντας κανονική κατανομή αντί t κατανομής) 646 "κατάλληλα" γονίδια και η ομαδοποίηση προχωρά χρησιμοποιώντας όλα τα μοντέλα της PGMM οικογένειας.

Εν κατακλείδει, το model-based clustering αποτελεί ένα πολύ ισχυρό εργαλείο στην ομαδοποίηση δεδομένων και ειδικά στην περίπτωση high-dimensional δεδομένων. Ωστόσο, υπάρχουν μερικά προβλήματα που πρέπει να ξεπεραστούν με πιο βασικό την εξεύρεση ενός αποδοτικού κριτηρίου για την επιλογή του μοντέλου με την καλύτερη ομαδοποίηση.

Abstract

This thesis concerns the methodology of model-based clustering as an alternative to the classical distance-based clustering techniques. Also, this thesis emphasize to the model-based clustering on high-dimensional data as well as in the use of multivariate t distribution instead of multivariate normal which is broadly in use. Moreover, there are plenty of examples and applications for the best understanding and explanation of the methods.

In more detail, before the case of high-dimensional data and t distribution there is an introduction in the philosophy of model-based clustering. There is some theory about mixtures of multivariate normal distributions and how those mixtures are used in the frame of model-based clustering. An extensive reference in the use of the EM algorithm for the parameters estimation exists as well. Furthermore, the models of the GPCM family are described in detail and there is also a discussion around controversial issues of model-based clustering such as model selection techniques, proper number of groups, initial values etc.

Then, the case of clustering of high-dimensional data is presented. Problems in the use of GPCM family for the case of high-dimensioanl data are discussed also and the use of factor-analyzers is proposed as an alternative. Also, there is a full description of the implementation of the AECM algorithm which is used for parameter estimation in that case.

Furthermore, we present two families models (PGMM και EPGMM) which are appropriate for the clustering of high-dimensional data and are based on mixtures of multivariate normal distributions (more specifically PGMM is nested in EPGMM). Applications of those families are provided and at the same time their advantages and disadantages are discussed as well.

Then, the case of multivariate t distribution and the benefits of its use is presented for the clustering of both high-dimensional data and not. Also, the AECM and EM algorithms respectively are fully described for this case. Moreover, the MMtFA family models is also presented for the case of t factor analyzers accordingly to the EPGMM family for the case of normal factor analyzers.

At the end, there is an application of the PGMM and MMtFA family models on high-dimensional data from the gene expression study of van 't Veer et al. (2002).

Data concern the expression of 24.182 genes (variables) from 78 women (observations) suffered from breast cancer. Initially the UUU (PGMM) and UUC (MMtFA) models are implemented for 100 random genes. Then, we move to model-based clustering using all models of PGMM family based on 646 genes, which were "appropriately" selected through a technique similar to EMMIX-GENE (by using the normal distribution instead of t)

In summary, model-based clustering appears to be a powerful tool in the clustering problem and especially in the case of high-dimensional data. However, there are some problems to overcome and the most basic among them appears to be the development of an efficient criterion for model selection.

Παράρτημα

```
AECM_t<-function(X,p0,m0,S0,q,g,df){

#X είναι ο πίνακας των δεδομένων
#p0 είναι ένα διάνυσμα όπου σε κάθε θέση περιέχει την αρχική πιθανότητα
#μια τυχαία παρατήρηση να ανήκει στην αντίστοιχη ομάδα

#m0 είναι μια λίστα όπου κάθε θέση περιέχει την αρχική μέση τιμή (διάνυσμα)της
#κάθε ομάδας

#S0 είναι μια λίστα όπου κάθε θέση περιέχει τον αντίστοιχο αρχικό πίνακα
#διασποράς-συνδιασποράς της κάθε ομάδας.

#q είναι ο αριθμός των παραγόντων και είναι ίδιος για κάθε ομάδα.
#g είναι ο αριθμός των ομάδων που επιθυμούμε να κατασκευάσουμε.
#df είναι οι βαθμοί ελευθερίας της t κατανομής.

n<-nrow(X)
p<-ncol(X)

#Κατασκευάζω την D0 που είναι μια λίστα όπου σε κάθε θέση θα περιέχει τον αρχικό
#πίνακα ιδιαιτεροτήτων της κάθε ομάδας.
#Υπολογίζω τους πίνακες ιδιαιτεροτήτων ως τους πίνακες που περιέχουν τα διαγώνια
#στοιχεία των αντίστοιχων αρχικών πινάκων διακύμανσης.

D0<-S0
for(i in 1:g){
D0[[i]]<-diag(diag(S0[[i]]))
}

#Παρακάτω υπολογίζω τον πίνακα επιβαρύνσεων B (loadings) για την κάθε ομάδα
#όπως περιγράφηκε στη σελίδα 54.

B0<-as.list(rep(0,g))

for(i in 1:g){     #το i μετράει ομάδες
help1<-matrix(numeric(p^2), nrow=p, byrow=T)
vector<-diag(D0[[i]])
vector<-vector^(-0.5)
help1<-diag(vector)%*%S0[[i]]%*%diag(vector)
A<-as.matrix(eigen(help1)[[2]])
A<-A[,1:q]
eigenvalues<-eigen(help1)[[1]]
Lamda<-diag(eigenvalues)
Lamda<-Lamda[1:q,1:q]
sigma<-sum(eigenvalues[(q+1):p])/(p-q)
I<-diag(rep(1, times=q))
B0[[i]]<-sqrt(D0[[i]])%*%A%*%sqrt(Lamda-sigma*I)
```

```
}


#Στις επόμενες 5 γραμμές υπολογίζεται ο αντίστροφος του πίνακα S για την κάθε
#ομάδα όπου S=BB'+D. Δηλαδή υπολογίζεται ο αντίστροφος του πίνακα BB'+D,
#καθώς θα χρησιμοποιηθεί στη συνέχεια μέσα στη log-likelihood.
#Επίσης υπολογίζεται η ορίζουσα αυτού του πίνακα.


invS<-as.list(rep(0, time=g))   #οι αντίστροφοι πίνακες αποθηκεύονται στη λίστα invS
detS<-numeric(g)                #οι ορίζουσες των πινάκων αποθηκεύονται στο διάνυσμα
#detS

for(j in 1:g){
unique<-diag(rep(1, time=q))
Dsolve<-solve(D0[[j]])
inverse<-solve(unique+t(B0[[j]])%*%Dsolve%*%B0[[j]])
invS[[j]]<-Dsolve-Dsolve%*%B0[[j]]%*%inverse%*%t(B0[[j]])%*%Dsolve

#Παρακάτω υπολογίζεται η ορίζουσα

detS[j]<-det(D0[[j]])/det(unique-t(B0[[j]])%*%invS[[j]]%*%B0[[j]])
}

#Παρακάτω υπολογίζεται η τιμή της log-likelihood για τις αρχικές τιμές.
logl<-numeric(n)

#Το logl είναι ένα διάνυσμα με μήκος όσο ο αριθμός των παρατηρήσεων όπου κάθε
#θέση περιέχει την τιμή την log-likelihood κάθε παρατήρησης.

help2<-numeric(g)  #Σε κάθε θέση αυτού του διανύσματος αποθηκεύεται η
#πιθανοφάνεια μιας παρατήρησης για την κάθε ομάδα.

constant<-lgamma((df+p)/2)-lgamma(df/2)-(p/2)*log(pi*df)    #είναι η σταθερά στον
#υπολογισμό της πιθανοφάνειας μιας παρατήρησης που βγαίνει έξω από το
#λογάριθμο.

for(i in 1:n){   #το i μετράει παρατηρήσεις
for(j in 1:g){   #το j μετράει ομάδες


mahalanobis<-(X[i,]-m0[[j]])%*%invS[[j]]%*%as.matrix(X[i,]-m0[[j]])

numerator<-p0[j]*((detS[j])^(-0.5))
denominator<-(1+mahalanobis/df)^(0.5*(df+p))
help2[j]<-numerator/denominator
}
logl[i]<-log(sum(help2))+constant
}
```

```
logL<-sum(logl)  #logL είναι η συνολική loglikelihood του μοντέλου με βάση τις
#αρχικές τιμές.

#Το διάνυσμα loglikelihood θα περιέχει γενικά στην I θέση τη loglikelihood του
#μοντέλου στο I βήμα και στην i+1 θέση την loglikelihood στο i+1 βήμα. Στην 1η
#θέση υπάρχει το 0 και μετά η loglikelihood που προκύπτει από τις αρχικές τιμές.

loglikelihood<-c(0,logL)

tol<-0.00000001    #tol είναι η ανοχή και καθορίζει πότε θα σταματήσει ο
#αλγόριθμος.
test<-abs((loglikelihood[2]-loglikelihood[1])/(loglikelihood[2]))

#test είναι η συνάρτηση τερματισμού.


#Wij είναι ένας πίνακας στον οποίο θα φαίνεται σε ποια ομάδα θα ανήκει κάθε
#παρατήρηση. Θα αποτελείται από n γραμμές (παρατηρήσεις) και g στήλες (ομάδες).
#Σε κάθε γραμμή θα υπάρχουν οι πιθανότητες η i παρατήρηση (γραμμή) να ανήκει
#στην j ομάδα.

Wij<-matrix(numeric(n*g), nrow=n, byrow=T)

#Varoi_ij είναι ένας πίνακας στον οποίο θα αποθηκεύονται τα βάρη κάθε
#παρατήρησης για κάθε ομάδα.
Varoi_ij<-matrix(numeric(n*g), nrow=n, byrow=T)

iter<-0    #δείχνει τον αριθμό επαναλήψεων του αλγορίθμου

while(test>tol){

iter<-iter+1

#Ξεκινάμε με τον πρώτο κύκλο της επανάληψης.

for(i in 1:n){   #το i μετράει παρατηρήσεις

#Το help3 θα περιέχει σε κάθε θέση την πιθανοφάνεια (εκτός της σταθερής
#ποσότητας) της i παρατήρησης για την κάθε ομάδα.

help3<-numeric(g)

for(j in 1:g){   #το j μετράει ομάδες
mahalanobis<-(X[i,]-m0[[j]])%*%invS[[j]]%*%as.matrix(X[i,]-m0[[j]])
Varoi_ij[i,j]<-(df+p)/(df+mahalanobis)
numerator<-p0[j]*((detS[j])^(-0.5))
denominator<-(1+mahalanobis/df)^(0.5*(df+p))
help3[j]<-numerator/denominator
}
```

```
for(j in 1:g){
Wij[i,j]<-help3[j]/sum(help3)
}
} #κλείνει το for που μετράει παρατηρήσεις

#Παρακάτω υπολογίζεται το νέο διάνυσμα πιθανοτήτων p0(new)
for(j in 1:g){
p0[j]<-sum(Wij[,j])/n
}

#Παρακάτω υπολογίζονται οι νέες μέσες τιμές κάθε ομάδας.

for(j in 1:g){          #το j μετράει ομάδες
numerator1<-numeric(p)
if(sum(Wij[,j])!=0){    #Αυτό χρειάζεται γιατί αν όλες οι παρατηρήεις καταταχτούν
 #σε μία ομάδα μόνο τότε για τις άλλες ομάδες η μέση τιμή είναι 0/0, δηλαδή δεν
#ορίζεται και δημιουργείται πρόβλημα στον αλγόριθμο. Οπότε αν σε μια ομάδα δεν
#καταταχτεί καμία παρατήρηση αφήνω το διάνυσμα των μέσων αυτής της ομάδας
#όπως είναι.

for(i in 1:n){
numerator1<-numerator1+Wij[i,j]*Varoi_ij[i,j]*X[i,]
}
denominator1<-sum(Wij[,j]*Varoi_ij[,j])
m0[[j]]<-numerator1/denominator1
}      #κλείνει το if
}

#Εδώ ξεκινά ο 2ος κύκλος της επανάληψης
#Πρέπει εδώ να υπολογιστούν οι νέες τιμές των πινάκων B & D.

for(i in 1:n){
#Για την κάθε παρατήρηση υπολογίζεται η πιθανότητα να ανήκει στην κάθε ομάδα
#αλλά για τις νέες τιμές των m0 και p0 πλέον.

#Το help4 θα περιέχει σε κάθε θέση την πιθανοφάνεια της i παρατήρησης για την
#κάθε ομάδα.
help4<-numeric(g)

for(j in 1:g){
mahalanobis<-(X[i,]-m0[[j]])%*%invS[[j]]%*%as.matrix(X[i,]-m0[[j]])
Varoi_ij[i,j]<-(df+p)/(df+mahalanobis)
numerator<-p0[j]*((detS[j])^(-0.5))
denominator<-(1+mahalanobis/df)^(0.5*(df+p))
help4[j]<-numerator/denominator
}

for(j in 1:g){
Wij[i,j]<-help4[j]/sum(help4)
}
```

```
}

#Παρακάτω υπολογίζεται ο πίνακας V. Ο V έχει διάσταση pxp.
#Για κάθε ομάδα έχω διαφορετικό πίνακα V.

V<-as.list(rep(0,g))
for(j in 1:g){              #το j μετράει ομάδες
for(i in 1:n){              #το i μετράει παρατηρήσεις
V[[j]]<-V[[j]]+Wij[i,j]*Varoi_ij[i,j]*(as.matrix(X[i,]-m0[[j]])%*%(X[i,]-m0[[j]]))
}
V[[j]]<-V[[j]]/sum(Wij[,j])
}

#Παρακάτω υπολογίζονται οι πίνακες gamma και wmega και έπειτα οι πίνακες B και
#D. Κάθε ομάδα έχει το δικό της πίνακα gamma και wmega.

#Δημιουργούνται όλοι οι gamma πίνακες
gamma<-as.list(rep(0,g))
for(i in 1:g){
gamma[[i]]<-invS[[i]]%*%B0[[i]]
}

#Δημιουργούνται όλοι οι wmega πίνακες
wmega<-as.list(rep(0,g))
for(i in 1:g){
wmega[[i]]<-diag(rep(1,time=q))-t(gamma[[i]])%*%B0[[i]]
}

#Παρακάτω υπολογίζονται οι νέοι πίνακες B για την κάθε ομάδα.
for(i in 1:g){
solve<-solve(t(gamma[[i]])%*%V[[i]]%*%gamma[[i]]+wmega[[i]])
B0[[i]]<-V[[i]]%*%gamma[[i]]%*%solve
}

# Παρακάτω υπολογίζονται οι νέοι πίνακες D για την κάθε ομάδα.
for(i in 1:g){
D0[[i]]<-V[[i]]-V[[i]]%*%gamma[[i]]%*%t(B0[[i]])
D0[[i]]<-diag(diag(D0[[i]]))
}

#Παρακάτω υπολογίζεται η νέα loglikelihood. Πρώτα όμως υπολογίζονται οι νέοι
#πίνακες inv(BB'+D) καθώς χρησιμοποιούνται στη log-likelihood.

for(j in 1:g){
unique<-diag(rep(1, time=q))
Dsolve<-solve(D0[[j]])
inverse<-solve(unique+t(B0[[j]])%*%Dsolve%*%B0[[j]])
invS[[j]]<-Dsolve-Dsolve%*%B0[[j]]%*%inverse%*%t(B0[[j]])%*%Dsolve

detS[j]<-det(D0[[j]])/det(unique-t(B0[[j]])%*%invS[[j]]%*%B0[[j]])
```

```
}

#Παρακάτω υπολογίζεται η loglikelihood

help5<-numeric(g)
logl<-numeric(n)

constant<-lgamma((df+p)/2)-lgamma(df/2)-(p/2)*log(pi*df)     #είναι η σταθερά στον
#υπολογισμό της πιθανοφάνειας μιας παρατήρησης που βγαίνει έξω από το
#λογάριθμο.

for(i in 1:n){
for(j in 1:g){
mahalanobis<-(X[i,]-m0[[j]])%*%invS[[j]]%*%as.matrix(X[i,]-m0[[j]])

numerator<-p0[j]*((detS[j])^(-0.5))
denominator<-(1+mahalanobis/df)^(0.5*(df+p))
help5[j]<-numerator/denominator
}
logl[i]<-log(sum(help5))+constant
}

#logL είναι η loglikelihood του μοντέλου σε αυτή την επανάληψη.

logL<-sum(logl)

#Παρακάτω ανανεώνω την τιμή της loglkelihood. Αρχικά αυξάνω το μήκος του
#διανύσματος loglikelihood κατά μία θέση η οποία παίρνει την τιμή μηδέν (option
#length.out). Σε αυτή τη θέση αποθηκεύω τη νέα τιμή της loglikelihood.

mikos<-length(loglikelihood)
loglikelihood<-rep(loglikelihood, length.out=(mikos+1))
loglikelihood[mikos+1]<-logL

#Παρακάτω υπολογίζεται η τιμή της συνάρτησης τερματισμού για αυτό το βήμα.
mikos<-length(loglikelihood)
test<-abs((loglikelihood[mikos]-loglikelihood[mikos-1])/(loglikelihood[mikos]))

#Ο αλγόριθμος πρέπει σε κάθε επανάληψη να δίνει μεγαλύτερη loglikelihood. Οπότε
#φτιάχνω την παρακάτω προειδοποίηση.

if(loglikelihood[mikos]<loglikelihood[mikos-1]){
print(paste("Error: the log-likelihood in the",iter,"iteration is lower than the previous
one"))
break
}

}  #κλείνει το  while

#Παρακάτω τυπώνονται τα αποτελέσματα.
```

```
#Πρώτα τυπώνεται ένας πίνακας, ο result με 2 στήλες. Η 1η περιέχει τον αριθμό
#παρατήρησης και η 2η τον αριθμό της ομάδας στην οποία έχει ταξινομηθεί.

omada<-numeric(n)
for(i in 1:n){
omada[i]<-which.max(Wij[i,])
}
obs<-1:n
result<-cbind(obs,omada)

#Παρακάτω υπολογίζονται οι πίνακες διασποράς των ομάδων όπως έχουν
#ταξινομηθεί οι παρατηρήσεις.

for(j in 1:g){
S0[[j]]<-B0[[j]]%*%t(B0[[j]])+D0[[j]]
}

#Παρακάτω υπολογίζεται το BIC.

m<-g*(p*q-0.5*q*(q-1))+g*p+g*p+g-1   #m είναι ο αριθμός των ελεύθερων
#παραμέτρων στο μοντέλο.

BIC<-2*loglikelihood[mikos]-m*log(n)

names(p0)<-c(paste(rep("p",g),1:g))   #Δίνονται ετικέτες στο διάνυσμα p0
names(test)<-c("Apotelesma synthikis termatismou")
names(iter)<-c("number of iterations")
names(BIC)<-c("BIC criterion")
names(df)<-c("Degrees of freedom")

return(list(result,p0,m0,S0,test,iter,BIC,loglikelihood[2:mikos],df))

}     #κλείνει η function
```

Βιβλιογραφία

Andrews J.L. (2010) Model-based clustering using mixtures of t-factor analyzers: A food authenticity example. Presentation for the 10th conference on sensometrics, Rotterdam, the Netherlands, July 2010.

Andrews J.L. and McNicholas P.D. (2011a) Extending mixtures of multivariate t-factor analyzers. Statistics and Computing, 21(3), 361–373.

Andrews J.L. and McNicholas P.D. (2011b) Mixtures of modified t-factor analyzers for model-based clustering, classification, and discriminant analysis. Journal of Statistical Planning and Inference, 141(4), 1479–1486.

Andrews J.L., McNicholas P.D., Subedi S. (2011) Model-based classification via mixtures of multivariate t-distributions. Computational Statistics and Data Analysis, 55(1), 520–529.

Baek J., McLachlan G.J., Flack L.K. (2010) Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 1298–1309.

Banfield J.D. and Raftery A.E. (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803-821.

Biernacki C. and Govaert G. (1999) Choosing models in model-based clustering and discriminant analysis. Journal of Statistical Computation and Simulation, 64, 49-71.

Biernacki C., Celeux G., Govaert G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7), 719–725.

Bouveyron C., Girard S., Schmid C. (2007) High-dimensional data clustering. Computational Statistics and Data Analysis, 52(1), 502–519.

Dean N. and Raftery A.E. (2006) The clustvarsel package: R package version 0.2-4

Dempster A.P., Laird N.M., Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society, Series B, 39(1), 1–38.

Fokoué E. and Titterington D.M. (2003) Mixtures of Factor Analysers. Bayesian Estimation and Inference by Stochastic Simulation. Machine Learning, 50(1-2), 73-94.

Fowlkes E.B. (1979) Some methods for studying the mixture of two normal (lognormal) distributions. Journal of the American Statistical Association 74, 561-575.

Fraley C. and Raftery A.E. (2006 September) MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Department of Statistics, University of Washington. Minor revisions January 2007 and November 2007.

Fraley C. and Raftery A.E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal, 41(8), 578-588.

Fraley C. and Raftery A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97, 611–631.

Fraley C. and Raftery A.E. (2007) Model-based Methods of Classification: Using the mclust Software in Chemometrics. Journal of Statistical Software, 18(6), 1-13.

Ghahramani Z. and Hinton G.E. (1997) The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Department of Computer Science, University Of Toronto, Toronto.

Hubert L. and Arabie P. (1985) Comparing partitions. Journal of Classification, 2, 193–218.

Karlis D. and Santourian A. (2009) Model-based clustering with nonelliptically contoured distributions. Statistics and Computing, 19(1), 73–83.

Maugis C., Celeux G., Martin-Magniette M.-L. (2009a) Variable selection for clustering with Gaussian mixture models. Biometrics, 65(3), 701–709.

Maugis C., Celeux G., Martin-Magniette M.-L. (2009b) Variable selection in model-based clustering: A general variable role modeling. Computational Statistics and Data Analysis, 53(11), 3872–3882.

Meng X.L. and Rubin D.B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrica, 80, 267-278.

McLachlan G.J. and Peel D. (2000a) Finite Mixture Models. John Wiley & Sons, New York.

McLachlan G.J. and Peel D. (2000b) Mixtures of Factor Analyzers, In Proceedings of the Seventeenth International Conference on Machine Learning. P. Langley (Ed.). San Francisco: Morgan Kaufmann, 599-606.

McLachlan G.J., Peel D., Bean R.W. (2002) Modelling high-dimensional data by mixtures of factor analyzers. Computational Statistics and Data Analysis, 41, 379-388.

McLachlan G.J., Peel D., Bean R.W. (2002) A mixture model-based approach to the clustering of microarray expression data. Bioinformatics, 18, 412–422.

McLachlan G.J., Bean R.W., Jones L.B.-T. (2007) Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. Computational Statistics and Data Analysis, 51(11), 5327–5338.

McLachlan G.J. and Krishnan T. (2008) The EM Algorithm and Extensions. 2nd eds. Wiley, New York.

McNicholas P.D. and Murphy T.B. (2008) Parsimonious Gaussian mixture models. Statistics and Computing, 18, 285–296.

McNicholas P.D. and Murphy T.B. (2010) Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics, 22, 2705-2712.

McNicholas P.D. (2010) Model-based classification using latent Gaussian mixture models, Journal of Statistical Planning and Inference 140(5), 1175-1181.

McNicholas P.D., Murphy T.B., McDaid A.F., Frost D. (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. Computational Statistics and Data Analysis, 54, 711–723.

McNicholas P.D. (2011) On Model-Based Clustering, Classification, and Discriminant Analysis, Journal of the Iranian Statistical Society, 10(2), 181-199.

Ng S.K., Krishnan T., McLachlan G.J. (2004) The EM algorithm. In Handbook of Computational Statistics, J. Gentle, W. Hardle, Y. Mori (Eds). New York: Springer-Verlag, 1, 137-168.

Peel D. and McLachlan G.J. (2000) Robust mixture modelling using the t distribution. Statistics and Computing, 10(4), 339–348.

Raftery A.E. and Dean N. (2006) Variable selection for model-based clustering. Journal of the American Statistical Association, 101(473), 168–178.

Tipping T.E. and Bishop C.M. (1999) Mixtures of probabilistic principal component analysers. Neural Computation, 11, 443–482.

Yeung K.Y. and Walter L.R. (2001) Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data ". Bioinformatics, 17(9), 763-774.

Yeung K.Y., Haynor D.R., Ruzzo W.L. (2001a) Validating clustering for gene expression data. Bioinformatics, 17, 309-318.

Yeung K.Y., Fraley C., Murua A., Raftery A.E., Ruzzo W.L. (2001b) Model-based clustering and data transformations for gene expression data. Bioinformatics, 17, 977–987.

van 't Veer L.J., Dai H., van de Vijver M.J., He Y.D., Hart A.A., Mao M., Peterse H.L., van der Kooy K., Marton M.J., Witteveen A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R., Friend S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530-535.

Καρλής Δ. (2005) Πολυμεταβλητή Στατιστική Ανάλυση, Αθ. Σταμούλης, Αθήνα.