ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

EMVS

Πρόσεγγιση στην Μπεϋζιανή Επιλογή Μεταβλητών μέσω του αλγορίθμου ΕΜ

Σταγάκης Γεώργιος

Μεταπτυχιακή Εργασία
που υποβλήθηκε στο ΔΠΜΣ Εφαρμοσμένες Μαθηματικές Επιστήμες του Εθνικού Μετσοβίου Πολυτεχνείου ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος
Αθήνα
Φεβρουάριος 2018

Όνομα: Σταγάκης Γεώργιος
Αριθμός Μητρώου: 09416027
email: georgstag@gmail.com
Επιβλέπων Καθηγητής: Φουσκάκης Δ.
Μέλη Εξεταστικής Επιτροπής: Λουλάκης Μ., Ντζούφρας Ι., Φουσκάκης Δ.
Πρόγραμμα: ΔΠΜΣ Εφαρμοσμένες Μαθηματικές Επιστήμες
Ίδρυμα: ΕΘνικό Μετσόβιο Πολυτεχνείο
2 Μαρτίου 2018

Περιεχόμενα

Πρόλογος

1 Μπεϋζιανά Γραμμικά Παλινδρομικά Μοντέλα **1**
1.1 Η λογική ενός γραμμικού παλινδρομικού μοντέλου 1
1.2 Παλινδρόμηση στην Μπεϋζιανή Στατιστική 3
1.2.1 Καταχρηστική πρότερη για τα β 3
1.2.2 Κανονική πρότερη κατανομή για τα β 4
1.3 Spike-and-Slab 5

2 Επιλογή Μεταβλητών στην Μπεϋζιανή Στατιστική **7**
2.1 Έλεγχοι Υποθέσεων 8
2.2 Παράγοντας Bayes 9
2.3 Εφαρμογή σε Γραμμικά Μοντέλα 9
2.4 Υπέρ και Κατά των μεθόδων στην Μπεϋζιανή Στατιστική 10

3 Ο Αλγόριθμος EM **13**
3.1 Εφαρμογή του EM στην Κλασική Στατιστική 13
3.2 Εφαρμογή του EM στην Μπεϋζιανή Στατιστική 18
3.3 Υπερ και Κατά του Αλγορίθμου 19

4 EM Σύγκλιση **21**
4.1 Η Μονοτονικότητα του Αλγορίθμου 21
4.2 Περιπτώσεις μη επιθυμητής σύγκλισης 23
4.2.1 Σύγκλιση σε τοπικό ακρότατο 24
4.2.2 Σύγκλιση σε σαμαρικό σημείο 28
4.2.3 Πρόβλημα στη σύγκλιση του GEM 31

5 EMVS **35**
5.1 Spike-and-Slab 35
5.1.1 E-step 36
5.1.2 M-step 38
5.2 Συμπερασματολογία για το β 39
5.2.1 Διαγράμματα για τη μελέτη των β 40

6 Εφαρμογή του EMVS **45**
6.1 Εφαρμογή του EMVS σε δείγμα με χαμηλή διακύμανση 45
6.2 Εφαρμογή του EMVS σε δείγμα με κανονική διακύμανση 47
6.3 Πολυσυγγραμικότητα και χαμηλή διακύμανση 49
6.4 Πολυσυγγραμικότητα και κανονική διακύμανση 49

Μεταφορά του EMVS στην R **53**

Βιβλιογραφία **69**

Πρόλογος

Η εργασία αυτή αποτελεί μέρος των υποχρεώσεών μου ως φοιτητής στο ΔΠΜΣ Εφαρμοσμένες Μαθηματικές Επιστήμες του Εθνικού Μετσόβιου Πολυτεχνείου για την απόκτηση του Μεταπτυχιακού Διπλώματος. Το θέμα της είναι η επιλογή μεταβλητών στην Μπεϋζιανή στατιστική μέσω του αλγορίθμου $EMVS$.

Τα τελευταία χρόνια, έχουν προταθεί πολλές τεχνικές και κανόνες απόφασης σχετικά το πως θα πρέπει να επιλέγονται οι μεταβλητές ενός παλινδρομικού μοντέλου. Ο αλγόριθμος $EMVS$ αποτελεί μία από αυτές. Ο αλγόριθμος $EMVS$, χρησιμοποιώντας τη μέθοδο Spike-and-Slab και τον αλγόριθμο ΕΜ, δίνει αποτελέσματα σχετικά με το ποιες μεταβλητές είναι περισσότερο πιθανό να επιδρούν σημαντικά σε ένα γραμμικό παλινδρομικό μοντέλο και ποιες όχι.

Το πρώτο Κεφάλαιο της εργασίας σχετίζεται με γενικές πληροφορίες γύρω από τα Μπεϋζιανά γραμμικά παλινδρομικά μοντέλα. Δίνονται ύστερες κατανομές των παραμέτρων κάτω από συγκεκριμένες πρότερες και γίνεται αναφορά στη μέθοδο Spike-and-Slab.

Το δεύτερο Κεφάλαιο έχει κάποια γενικά στοιχεία σχετικά με την επιλογή μεταβλητών στην Μπεϋζιανή στατιστική. Γίνεται αναφορά σε ελέγχους υποθέσεων, τον παράγοντα $Bayes$ και μεθόδους που χρησιμοποιούνται γενικότερα στην Μπεϋζιανή επιλογή μεταβλητών.

Το τρίτο Κεφάλαιο της εργασίας σχετίζεται τον αλγόριθμο ΕΜ. Διατυπώνονται τα βασικά του βήματα, τι επιτυγχάνει και δίνονται παραδείγματα στην κλασική και Μπεϋζιανή στατιστική.

Το τέταρτο Κεφάλαιο της εργασίας επικεντρώνεται στην σύγκλιση του αλγορίθμου ΕΜ. Ο αλγόριθμος ΕΜ δεν συγκλίνει πάντα στην ποσότητα που απαιτούμε. Δίνεται απόδειξη της μονοτονικότητάς του και παραδείγματα περιπτώσεων μη επιθυμητής σύγκλισης.

Το πέμπτο Κεφάλαιο της εργασίας σχετίζεται με τον αλγόριθμο $EMVS$. Αναφέρεται αναλυτικά πως επιτυγχάνεται η σύστασή των εκτελέσιμων βημάτων του, μέσω του αλγορίθμου ΕΜ, και προτείνονται διαγράμματα που κάνουν ευκολότερη την συμπερασματολογία.

Το έκτο και τελευταίο Κεφάλαιο έχει εφαρμογές του αλγορίθμου $EMVS$. Μέσω κατάλληλων προσομοιωμένων δεδομένων ερευνάται η μηχανική του $EMVS$ σε συγκεκριμένες περιπτώσεις. Ταυτόχρονα με τον $EMVS$ δίνονται αποτελέσματα ως προς την επιλογή μεταβλητών και από το πακέτο της γλώσσας R, BAS.

Στις τελευταίες σελίδες δίνεται αναλυτικός κώδικας στη γλώσσα προγραμματισμού R σχετικά με το πως εκτελείται ο $EMVS$ μέσω υπολογιστή και πως δημιουργούνται τα διαγράμματά του. Ακόμα δίνεται και η βιβλιογραφία που χρησιμοποιήθηκε για τις ανάγκες της εργασίας.

Ευχαριστώ πολύ τον διδάσκοντα για την πολύτιμη βοήθειά του, τόσο στο θεωρητικό κομμάτι της εργασίας όσο και στην επιμέλεια της.

Αθήνα, 2018

Για τις ανάγκες της εργασίας χρησιμοποιούνται οι παρακάτω συμβολισμοί :

Χρήση κεφαλαίων γραμμάτων σε τυχαίες μεταβλητές και πίνακες, πχ X, Y κ.ο.κ.

Χρήση *bold* για διανύσματα, πχ $\boldsymbol{\vartheta}$,$\boldsymbol{\gamma}$ κ.ο.κ.

X^T, ο ανάστροφος πίνακας του X.

$f_{X,Y,Z}(x, y, z)$, η από κοινού πυκνότητα πιθανότητας των X, Y, Z.

$f_{X|Y,Z}(x|y, z)$, η δεσμευμένη πυκνότητα πιθανότητας του X, δοθέντος των $Y = y, Z = z$.

$\pi_X(a)$, η πιθανότητα $X = a$.

$\pi_{X|Y}(a|y)$, η πιθανότητα $X = a$ δοθέντος πως $Y = y$.

π, κάποια πιθανότητα.

$E[\bullet]$, η μέση τιμή του $\bullet$.

$E[\bullet|X]$, η μέση τιμή του $\bullet$ δοθέντος του X.

$E_X[\bullet]$, η μέση τιμή της συνάρτησης $\bullet$ ως προς την τυχαία μεταβλητή X.

$\mathbf{Y}$, η εξαρτημένη τυχαία μεταβλητή των γραμμικών μοντέλων.

$\boldsymbol{\beta}$, η τυχαία μεταβλητή-διάνυσμα των παραμέτρων ενός γραμμικού μοντέλου.

$\boldsymbol{\gamma}$, η τυχαία μεταβλητή $\boldsymbol{\gamma}$ της μεθόδου Spike-and-Slab.

$\widetilde{\boldsymbol{\gamma}}$, το αποτέλεσμα για τα $\boldsymbol{\gamma}$ που δίνει ο $EMVS$.

σ^2, (σταθερή ή τυχαία μεταβλητή) η διακύμανση.

σ, (σταθερή ή τυχαία μεταβλητή) η τυπική απόκλιση.

X, ο πίνακας σχεδιασμού των γραμμικών μοντέλων.

$\boldsymbol{\vartheta}^k$, η ακολουθία που προκύπτει από τον αλγόριθμο EM.

$Q(x; y)$, η ποσότητα του αλγορίθμου EM στην κλασική στατιστική.

$Q(x|y)$, η ποσότητα του αλγορίθμου EM στην Μπεϋζιανή στατιστική.

Κεφάλαιο 1

Μπεϋζιανά Γραμμικά Παλινδρομικά Μοντέλα

Η λειτουργικότητα της επιστήμης των μαθηματικών, και κατά συνέπεια των στατιστικών μεθόδων, γενικότερα βασίζεται στην υπόθεση πως ο κόσμος ακολουθεί μοτίβα (*patterns*). Η παλινδρόμηση (*regression*) είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών, η οποία βασίζεται στην αρχή που αναφέρεται στην προηγούμενη πρόταση. Με την παλινδρόμηση έχουν ασχοληθεί μεγάλοι μαθηματικοί όπως οι *Adrien − Marie Legendre* (1752-1833) και *Carl Friedrich Gauss* (1777-1855), στις αρχές της ανακάλυψής της, αλλά και πιο πρόσφατα οι *George E. P. Box* (1919-2013) και *George C. Tiao* (1933-...) που ασχολήθηκαν και με Μπεϋζιανά παλινδρομικά μοντέλα.

1.1 Η λογική ενός γραμμικού παλινδρομικού μοντέλου

Ξεκινώντας να δουλεύουμε με ένα παλινδρομικό μοντέλο, έχουμε τις παρακάτω μεταβλητές:

- $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ η εξαρτημένη μεταβλητή-διάνυσμα του μοντέλου (ή αλλιώς μεταβλητή απόκρισης, dependent or response variable), διάστασης n.
- X ο πίνακας σχεδιασμού (design matrix) του μοντέλου, διάστασης $n \times p + 1$.
- $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ το διάνυσμα με τις παραμέτρους του μοντέλου, διάστασης $p + 1$.
- $\boldsymbol{\varepsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ το διάνυσμα με τα σφάλματα (ή αλλιώς θορύβους), διάστασης n.
- σ^2 η διακύμανση των σφαλμάτων.

Σε ένα παλινδρομικό μοντέλο υποθέτουμε πως $\forall k \in \{1, 2, ..., n\}$ οι επιμέρους μεταβλητές Y_k του διανύσματος $\mathbf{Y}$ επηρεάζονται από τις k γραμμές $\mathbf{x}^k$ του πίνακα σχεδιασμού X με τον παρακάτω τρόπο,

$$Y_k = \mathbf{x}^k \boldsymbol{\beta} + \epsilon_k.$$

Τα σφάλματα ϵ_k είναι τυχαίες μεταβλητές που ακολουθούν την κανονική κατανομή με μέση τιμή 0 και κάποια διακύμανση σ^2, ανεξάρτητες και ισόνομες μεταξύ τους. Αυτό που εννοείται από την παραπάνω σχέση είναι πως τα Y_k επηρεάζονται γραμμικά από τα $\mathbf{x}^k$ αλλά, επιπλέον, υπάρχει ένας θόρυβος ϵ_k ο οποίος τους προσδίδει μια τυχαία μετατόπιση από την θέση $\mathbf{x}^k\boldsymbol{\beta}$. Επεκτείνοντας την παραπάνω σχέση έχουμε πως,

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Αφού ισχύει πως τα ϵ_k είναι ανεξάρτητα μεταξύ τους και

$$\epsilon_k \sim N(0, \sigma^2),\ \forall k \in \{1, 2, ..., n\}$$

άμεσα έπεται πως

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \sigma^2 I_n),$$

όπου $N_n(\mathbf{0}_n, \sigma^2 I_n)$ η πολυδιάστατη κανονική κατανομή, $\mathbf{0}_n$ διάνυσμα n διάστασης που όλα του τα στοιχεία είναι ίσα με 0 και I_n ο μοναδιαίος πίνακας, διάστασης $n \times n$. Ακόμα, από τα παραπάνω, είναι εμφανές πως

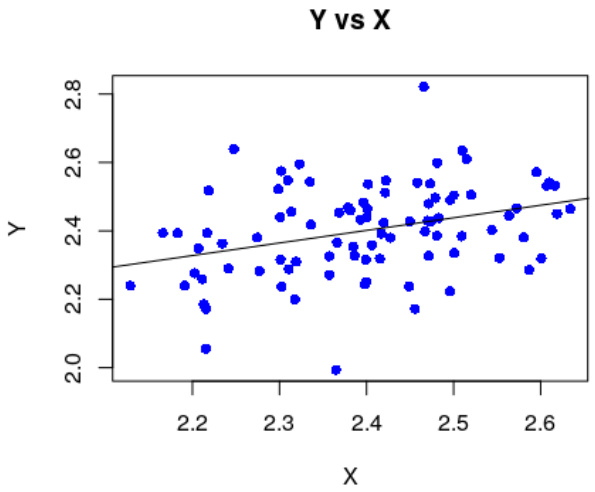
$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n).$$

Οι στήλες του πίνακα σχεδιασμού ονομάζονται ανεξάρτητες μεταβλητές (independent variables). Ο λόγος για τον οποίο δόθηκαν αυτές οι ονομασίες στα $X, \mathbf{Y}$ είναι πως μέσω του X στον οποίο εμείς δίνουμε τιμές, μπορούμε να έχουμε εικόνα για την κατανομή του $\mathbf{Y}$, πράγμα το οποίο σημαίνει πως το $\mathbf{Y}$ εξαρτάται από το X.

Η πρώτη στήλη του πίνακα σχεδιασμού έχει όλες τις τιμές της ίσες με 1. Αυτό γίνεται με σκοπό το β_0 να αντιπροσωπεύει κάποια τιμή της μέσης τιμής του $\mathbf{Y}|\boldsymbol{\beta}, \sigma^2$, που δεν μεταβάλλεται καθώς δίνονται διαφορετικές τιμές στα στοιχεία του X. Το β_0, διαισθητικά, δίνει την μέση τιμή των $Y_k|\boldsymbol{\beta}, \sigma^2$ αν όλες οι άλλες τιμές (εκτός από την πρώτη στήλη) του X είναι μηδέν, ενώ τα υπόλοιπα β_i εκφράζουν την μεταβολή στη μέση τιμή των $Y_k|\boldsymbol{\beta}, \sigma^2$ για μοναδιαία μεταβολή στην τιμή x_{ki} του πίνακα X, κρατώντας σταθερά τα υπόλοιπα στοιχεία $x_{kl},\ l \neq i$, του πίνακα.

Η χρηστικότητα των παλινδρομικών μοντέλων βρίσκεται στο γεγονός ότι μέσω αυτών είναι εύκολο να γίνει προτυποποίηση και εκτίμηση της τιμής ενός ποσοτικοποιημένου φαινομένου το οποίο έχουμε λόγους να πιστεύουμε πως εξαρτάται γραμμικά από άλλες μεταβλητές. Με κατάλληλο δείγμα και αρκετές παραμέτρους β_i μπορεί να περιοριστεί σημαντικά η έκβαση του φαινομένου που βασίζεται σε αστάθμητους παράγοντες, ή αντίστοιχα, η τυπική απόκλιση των σφαλμάτων της παλινδρόμησης να παρατηρηθεί πως είναι πολύ μικρή στην συγκεκριμένη περίπτωση. Επιπλέον, η κανονική κατανομή από το σχήμα της έχει εφαρμογή σε πάρα πολλά παρατηρούμενα ποσοτικά φαινόμενα.

Διάγραμμα 1.1: Σημεία (x, y) που παρατηρήθηκαν για κάποιο φαινόμενο και η ευθεία παλινδρόμησης.



Στο Διάγραμμα 1.1 φαίνεται μια περίπτωση όπου οι παλινδρομικές μέθοδοι έχουν πολύ ικανοποιητικά αποτελέσματα. Έχουμε συλλέξει δισδιάστατο δείγμα στο οποίο θεωρούμε πως για κάθε ζεύγος παρατηρήσεων (x_k, y_k) ισχύει η σχέση

$$y_k = \beta_0 + \beta_1 x_k + \epsilon_k,\ \epsilon_k \sim N(0, \sigma^2).$$

Το δείγμα που συλλέχθηκε απεικονίζεται στο διάγραμμα μέσα από μπλε κουκίδες. Για αυτό το δείγμα εκτιμήθηκε το $\boldsymbol{\beta}$ της γραμμικής συσχέτισης, και μέσω αυτού προστέθηκε στο γράφημα η ευθεία

$$y = \hat{\beta}_0 + \hat{\beta}_1 x,$$

όπου $(\hat{\beta}_0, \hat{\beta}_1)$ οι εκτιμώμενες τιμές του $\boldsymbol{\beta}$, από το δείγμα. Είναι εμφανές πως για οποιαδήποτε έκβαση είχε το y σε κάποια θέση x, το αποτέλεσμα ήταν ικανοποιητικά κοντά στην ευθεία και δοθέντος πως οι συνθήκες θα συνεχίσουν να είναι ίδιες με αυτές που υπήρχαν όταν συλλέχθηκε το δείγμα, τα αποτελέσματα θα εξακολουθήσουν να είναι κοντά στην ευθεία παλινδρόμησης.

Στην επόμενη ενότητα περιγράφεται με περισσότερες λεπτομέρειες το πως εφαρμόζονται παλινδρομικές τεχνικές στην Μπεϋζιανή στατιστική.

1.2 Παλινδρόμηση στην Μπεϋζιανή Στατιστική

Στην παρούσα ενότητα υποθέτουμε πως

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n)$$

και τα $\boldsymbol{\beta}$ και σ^2 είναι τυχαίες μεταβλητές. Στις επόμενες υποενότητες γίνονται υπολογισμοί, μέσω αναλογιών, της ύστερης κατανομής του $\boldsymbol{\beta}$, κάτω από συγκεκριμένες υποθέσεις για την πρότερή κατανομή των παραμέτρων.

1.2.1 Καταχρηστική πρότερη για τα β

Θέλουμε να βρούμε την κατανομή της τυχαίας μεταβλητής $\boldsymbol{\beta}|\mathbf{Y},\sigma^2$. Υποθέτουμε πως η πρότερη κατανομή για το $\boldsymbol{\beta}|\sigma^2$ είναι η καταχρηστική $f_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2) \propto 1, \boldsymbol{\beta} \in \mathbb{R}^{p+1},\ \sigma^2 > 0$. Με τον όρο καταχρηστική ($Improper$) κατανομή εννοείται μία μη αρνητική συνάρτηση για την οποία το ολοκλήρωμα στο πεδίο ορισμού της είναι ίσο με άπειρο. Η συγκεκριμένη πρότερη κατανομή χρησιμοποιείται σε περιπτώσεις όπου δεν υπάρχει εκ των προτέρων πληροφορία για τα $\boldsymbol{\beta}$.

Τότε,

$$\begin{aligned}
f_{\boldsymbol{\beta}|\mathbf{Y},\sigma^2}(\boldsymbol{\beta}|\mathbf{y},\sigma^2) &\propto f_{\mathbf{Y}|\boldsymbol{\beta},\sigma^2}(\mathbf{y}|\boldsymbol{\beta},\sigma^2) f_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2) \\
&\propto exp[-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})] \\
&\propto exp[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^T(X^TX)\boldsymbol{\beta} - \boldsymbol{\beta}^TX^T\mathbf{y} - \mathbf{y}^TX\boldsymbol{\beta})] \\
&\propto exp[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^T(X^TX)(\boldsymbol{\beta} - (X^TX)^{-1}X^T\mathbf{y}) - \mathbf{y}^TX\boldsymbol{\beta})] \\
&\propto exp[-\frac{1}{2\sigma^2}[\boldsymbol{\beta}^T(X^TX)(\boldsymbol{\beta} - (X^TX)^{-1}X^T\mathbf{y}) \\
&\qquad - \mathbf{y}^TX(\boldsymbol{\beta} - (X^TX)^{-1}X^T\mathbf{y} + (X^TX)^{-1}X^T\mathbf{y})]] \\
&\propto exp[-\frac{1}{2\sigma^2}[\boldsymbol{\beta}^T(X^TX)(\boldsymbol{\beta} - (X^TX)^{-1}X^T\mathbf{y}) - \mathbf{y}^TX(\boldsymbol{\beta} - (X^TX)^{-1}X^T\mathbf{y})]] \\
&\propto exp[-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - (X^TX)^{-1}X^T\mathbf{y})^T(X^TX)(\boldsymbol{\beta} - (X^TX)^{-1}X^T\mathbf{y})]. \qquad (1.1)
\end{aligned}$$

Από την (1.1) φαίνεται πως η $\boldsymbol{\beta}|\mathbf{Y} = \mathbf{y}, \sigma^2$ ακολουθεί την πολυδιάστατη κανονική κατανομή $N_{p+1}(\hat{\boldsymbol{\beta}},\Sigma)$, όπου

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y} \qquad (1.2)$$

και

$$\Sigma = \sigma^2(X^TX)^{-1}. \qquad (1.3)$$

Είναι γνωστό πως η πολυδιάστατη κανονική κατανομή έχει μια κορυφή και βρίσκεται στις συντεταγμένες της παραμέτρου της μέσης τιμής. Δηλαδή για το συγκεκριμένο παράδειγμα η κορυφή της $f_{\boldsymbol{\beta}|\mathbf{Y},\sigma^2}(\boldsymbol{\beta}|\mathbf{y},\sigma^2)$ βρίσκεται στις συντεταγμένες $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$.

Για την πρότερη κατανομή της διακύμανσης, συνήθως, κάνουμε την υπόθεση πως ακολουθεί αντίστροφη γάμμα (*inverse gamma*) με παραμέτρους a,δ ($IG(a,\delta)$). Η υπόθεση αυτή γίνεται γιατί η αντίστροφη γάμμα είναι αρκετά ευέλικτη κατανομή που προσαρμόζεται ικανοποιητικά σε πολλές περιπτώσεις και διευκολύνει τους υπολογισμούς στις αναλογίες. Οι παράμετροι a,δ της κατανομής είναι σταθερές, αλλά θα μπορούσαν να είναι και τυχαίες μεταβλητές και να υποθέσουμε για αυτές κάποια πρότερη κατανομή.

Παρακάτω γίνονται πράξεις για την εύρεση της κατανομής του $\sigma^2|\mathbf{Y},\boldsymbol{\beta}$, κάτω από την υπόθεση πως $\sigma^2\sim IG(a,\delta)$,

$$\begin{aligned} f_{\sigma^2|\mathbf{Y},\boldsymbol{\beta}}(\sigma^2|\mathbf{y},\boldsymbol{\beta}) &\propto f_{\mathbf{Y}|\boldsymbol{\beta},\sigma^2}(\mathbf{y}|\boldsymbol{\beta},\sigma^2)f_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2)f_{\sigma^2}(\sigma^2)\\ &\propto \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})}(\sigma^2)^{-a-1}e^{-\frac{\delta}{\sigma^2}}\\ &\propto (\sigma^2)^{-(a+\frac{1}{2})-1}e^{-\frac{1}{\sigma^2}[(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})/2+\delta]}. \end{aligned} \tag{1.4}$$

Από την (1.4) φαίνεται άμεσα πως σε αυτήν την περίπτωση,

$$\sigma^2|\mathbf{Y}=\mathbf{y},\boldsymbol{\beta}\ \sim\ IG(a+\frac{1}{2},(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})/2+\delta).$$

Με επιπλέον πράξεις, μπορούν να βρεθούν και περιθώριες κατανομές. Ενδεικτικά αναφέρεται πως για τις παραπάνω υποθέσεις, η κατανομή της $\boldsymbol{\beta}|\mathbf{Y}$ που προκύπτει είναι πολυδιάστατη κατανομή *student*.

1.2.2 Κανονική πρότερη κατανομή για τα β

Τώρα θέλουμε να βρούμε την κατανομή του $\boldsymbol{\beta}|\mathbf{Y},\sigma^2$ ενώ δεν έχουμε κάποια πρότερη πληροφορία για το $\boldsymbol{\beta}$ αλλά υποθέτουμε πως $\boldsymbol{\beta}|\sigma^2$ ακολουθεί κάποια πολυδιάστατη κανονική κατανομή. Επιπλέον τα β_i είναι ανεξάρτητα ανά δύο. Υποθέτουμε, λοιπόν, κατανομή για το $\boldsymbol{\beta}|\sigma^2$ την κανονική $N_{p+1}(\mathbf{0}_{p+1},\frac{\sigma^2}{\lambda}I_{p+1})$, $\lambda\in\mathbb{R}_+^*$. Όσο μικρότερη η τιμή του λ, τόσο περισσότερο μη πληροφοριακή είναι η προηγούμενη κατανομή.

Όπως και στην (1.1), για τον υπολογισμό της κατανομής του $\boldsymbol{\beta}|\mathbf{Y},\sigma^2$,

$$\begin{aligned} f_{\boldsymbol{\beta}|\mathbf{Y},\sigma^2}(\boldsymbol{\beta}|\mathbf{y},\sigma^2) &\propto f_{\mathbf{Y}|\boldsymbol{\beta},\sigma^2}(\mathbf{y}|\boldsymbol{\beta},\sigma^2)f_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta}|\sigma^2)\\ &\propto e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})}e^{-\frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}}\\ &\propto exp[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^T(X^TX+\lambda I_{p+1})\boldsymbol{\beta}-\boldsymbol{\beta}^TX^T\mathbf{y}-\mathbf{y}^TX\boldsymbol{\beta})]\\ &\propto exp[-\frac{1}{2\sigma^2}[\boldsymbol{\beta}^T(X^TX+\lambda I_{p+1})(\boldsymbol{\beta}-(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y})\\ &\qquad -\mathbf{y}^TX(\boldsymbol{\beta}-(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y}+(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y})]]\\ &\propto exp[-\frac{1}{2\sigma^2}[\boldsymbol{\beta}^T(X^TX+\lambda I_{p+1})(\boldsymbol{\beta}-(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y})\\ &\qquad -\mathbf{y}^TX(\boldsymbol{\beta}-(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y})]]\\ &\propto exp[-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y})^T(X^TX+\lambda I_{p+1})\\ &\qquad (\boldsymbol{\beta}-(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y})]. \end{aligned} \tag{1.5}$$

Άρα τώρα η $\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y},\sigma^2$ ακολουθεί την πολυδιάστατη κανονική κατανομή $N_{p+1}(\hat{\boldsymbol{\beta}}',\Sigma')$, όπου

$$\hat{\boldsymbol{\beta}}'=(X^TX+\lambda I_{p+1})^{-1}X^T\mathbf{y} \tag{1.6}$$

και

$$\Sigma' = \sigma^2(X^TX + \lambda I_{p+1})^{-1}. \tag{1.7}$$

Οι πράξεις στην (1.5) ήταν σχετικά εύκολες γιατί η πρότερη με την ύστερη ήταν συζυγείς κατανομές. Είναι γνωστό πως αν η $B|A$ και η A ακολουθούν κανονικές κατανομές, τότε και η $A|B$ ακολουθεί κανονική κατανομή. Επομένως η κατανομή της $\boldsymbol{\beta}|\mathbf{Y}, \sigma^2$ ήταν γνωστή λόγω των $\mathbf{Y}|\boldsymbol{\beta}, \sigma^2$ και $\boldsymbol{\beta}|\sigma^2$. Το μόνο που έμενε να βρεθεί ήταν οι παράμετροί της.

Αν κανείς ήθελε να δώσει πρότερη κατανομή στο $\boldsymbol{\beta}|\sigma^2$ κανονική (πληροφοριακή ή όχι), μια αρκετά συνηθισμένη άλλη επιλογή είναι η g-πρότερη του *Zellner* (1986). Σύμφωνα με αυτήν, η πρότερη κατανομή του $\boldsymbol{\beta}|\sigma^2$ είναι η $N_{p+1}(\boldsymbol{\mu}_0, g\sigma^2(X^TX)^{-1})$. Τα $\boldsymbol{\mu}_0 \in \mathbb{R}^{p+1}$ και $g \in \mathbb{R}_+^*$ θεωρούνται σταθερές. Τότε η ύστερη $\boldsymbol{\beta}|\mathbf{Y} = \mathbf{y}, \sigma^2$ θα ακολουθεί την $\mathrm{N}_{p+1}(c\hat{\boldsymbol{\beta}} + (1-c)\boldsymbol{\mu}_0, c\sigma^2(X^TX)^{-1})$, όπου $c = \frac{g}{1+g}$ και $\hat{\boldsymbol{\beta}}$ η ποσότητα της σχέσης (1.2). Το g λειτουργεί ως παράμετρος κλίμακας, δηλαδή ανάλογα το μέγεθος του μεγαλώνει ή μικραίνει την διακύμανση της πρότερης κατανομής, κάνοντάς την περισσότερο ή λιγότερο πληροφοριακή. Στο Κεφάλαιο 6 της εργασίας θα ξαναγίνει αναφορά στην g-πρότερη.

Αν η φύση του φαινομένου υποδηλώνει πως η πρότερη κατανομή του $\boldsymbol{\beta}$ έχει πιο βαριά ουρά από αυτήν της κανονικής κατανομής, μια λογική πρότερη που θα μπορούσε να τεθεί είναι η *student*. Επιπλέον, εύκολα αποδεικνύεται (όπως και στην (1.4)) πως αν $\boldsymbol{\beta}|\sigma^2$ ακολουθεί κανονική κατανομή και η πρότερη κατανομή της διακύμανσης είναι η αντίστροφη γάμμα, και πάλι η κατανομή που ακολουθεί το $\sigma^2|\mathbf{Y}, \boldsymbol{\beta}$ θα είναι η αντίστροφη γάμμα.

1.3 Spike-and-Slab

Το Spike-and-Slab είναι μια μέθοδος η οποία εφαρμόζεται στην επιλογή μεταβλητών (*Variable Selection*) ενός γραμμικού Μπεϋζιανού παλινδρομικού μοντέλου. Μέσω αυτού δίνονται αποτελέσματα σχετικά με το ποιες μεταβλητές του πίνακα σχεδιασμού X επιδρούν σημαντικά σε κάποιο γραμμικό μοντέλο και ποιες όχι. Το Spike-and-Slab προτάθηκε αρχικά από τους *Mitchell* και *Beauchamp* (1988). Πάνω σε αυτό δούλεψαν και οι *Madigan* και *Raftery* (1994) και οι *George* και *McCulloch* (1997). Την τελική του μορφή έδωσαν οι *Ishwaran* και *Rao* (2005).

Με τον όρο *spike* (καρφί) ορίζονται οι πυκνότητες πιθανότητας των τυχαίων μεταβλητών β_i που δεν έχουν σημαντική επίδραση στο μοντέλο (έντονα συγκεντρωμένες κοντά στο 0), ενώ *slab* (πλάκα) οι πρότερες κατανομές των β_i (μη πληροφοριακές, σχεδόν απλωμένες). Το όνομά της μεθόδου προήλθε από το σχήμα των παραπάνω κατανομών.

Υποθέτουμε, όπως και στις προηγούμενες ενότητες, πως

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \ \sim \ N_n(X\boldsymbol{\beta}, \sigma^2 I_n).$$

Το β_0 θα αφαιρεθεί από το μοντέλο μιας και δεν σχετίζεται με κάποια από τις μεταβλητές του πίνακα σχεδιασμού και ο στόχος πίσω από το $Spike - and - Slab$ είναι η επιλογή μεταβλητών. Η αφαίρεση του β_0 επιτυγχάνεται κεντράροντας το $\mathbf{Y}$ στο 0. Από εδώ και πέρα υποθέτουμε πως ακολουθείται το παραπάνω και επομένως ισχύει πως, το $\boldsymbol{\beta}$=$(\beta_1, ..., \beta_p)$ είναι p διάστασης και ο πίνακας X, διάστασης $n \times p$, δεν περιέχει τις μονάδες της πρώτης στήλης.

Ακόμα θεωρούμε διάνυσμα $\boldsymbol{\gamma}$, p διάστασης, το οποίο περιέχει στοιχεία γ_i που εξυπηρετούν σαν δείκτριες για το αν η αντίστοιχη μεταβλητή β_i συγκεντρώνεται έντονα κοντά στο μηδέν ή όχι. Τα γ_i είναι ανεξάρτητα και ισόνομα μεταξύ τους. Αν μια μεταβλητή β_i συγκεντρώνεται έντονα κοντά στο μηδέν θα ήταν καλύτερα να μην εισαχθεί στο μοντέλο, μιας και σε αυτήν την περίπτωση η επίδραση του αντίστοιχου παράγοντα στον οποίο δίνει βάρος το β_i θα ήταν πολύ μικρή στη γραμμική σχέση $\mathbf{x}^i\boldsymbol{\beta}$. Περισσότερα σχετικά με το πότε μια μεταβλητή β_i καλό θα ήταν να μην περιέχεται στο παλινδρομικό μοντέλο φαίνονται στο επόμενο Κεφάλαιο. Το γ_i είναι ίσο με μηδέν αν η αντίστοιχη μεταβλητή β_i συγκεντρώνεται έντονα κοντά στο μηδέν, αλλιώς είναι ίσο με ένα.

Θα μπορούσαμε να θεωρήσουμε πως το γ_i είναι ίσο με μηδέν αν η αντίστοιχη μεταβλητή β_i είναι ίση με το μηδέν, αλλά τότε θα ήταν πιθανό καθώς μελετάμε τις τιμές γ_i ενώ θα βρίσκαμε μια τιμή ίση με 1, η αντίστοιχη συγκέντρωση της β_i να ήταν σε μια τιμή πολύ κοντά στο μηδέν. Είναι ωφέλιμο να αφεθούν έξω από το μοντέλο όχι μόνο οι μεταβλητές β_i που επιδρούν επάνω στο μοντέλο με επίδραση μηδέν, αλλά και αυτές που έχουν αρκετά μικρό μέτρο έτσι ώστε η επίδραση τους στο μοντέλο να μην έχει ιδιαίτερα σημαντική συνεισφορά.

Οι *George* και *McCulloch* (1997) έθεσαν για το Spike-and-Slab πως

$$\begin{aligned}
&\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n)\\
&\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma} \sim N_p(\mathbf{0}_p, D_{\sigma^2,\boldsymbol{\gamma}})\\
&D_{\sigma^2,\boldsymbol{\gamma}} = \sigma^2 diag\{(1-\gamma_i)v_0 + \gamma_i v_1\}\\
&\sigma^2|\boldsymbol{\gamma} \sim IG(v/2, \lambda v/2)\\
&\pi(\boldsymbol{\gamma}|\theta) = \theta^{\sum_{i=0}^{p}\gamma_i}(1-\theta)^{p-\sum_{i=0}^{p}\gamma_i}\\
&\theta \sim Beta(a,b).
\end{aligned} \tag{1.8}$$

Με το *diag* εννοείται ο διαγώνιος πίνακας. Κοιτώντας πιο προσεχτικά τον $D_{\sigma^2,\boldsymbol{\gamma}}$ φαίνεται πως, κάτω από τις δεσμεύσεις, τα β_i που επιδρούν σημαντικά στο μοντέλο έχουν διακύμανση $\sigma^2 v_1$ ενώ αυτά που δεν επιδρούν $\sigma^2 v_0$.

Τα v_0, v_1 θεωρούνται θετικές σταθερές. Για να γίνει καλός διαχωρισμός των β_i καλό θα ήταν το v_0 να έχει μικρή τιμή, ενώ το v_1 μεγάλη. Κάτω από κάποια παραλλαγή του μοντέλου θα μπορούσαν να είναι και τυχαίες μεταβλητές. Συνηθίζεται κάποιες φορές το v_1 να ακολουθεί είτε διπλή εκθετική είτε κατανομή *Cauchy*, λόγω της βαριάς ουράς τους που προσδίδει στο v_1 μεγάλες τιμές με υψηλή πιθανότητα.

Η συνάρτηση $\pi(\boldsymbol{\gamma}|\theta)$ έχει και αυτή συχνά παραλλαγές. Οι *Stingo* και *Vannucci* (2011) έχουν προτείνει σε κάποιες περιπτώσεις να χρησιμοποιείται η *Logistic Regression Product Prior* αντί της παραπάνω (1.8) και οι *Li* και *Zhang* (2010) την *Markov Random Field Prior*.

Οι παράμετροι της τυχαίας μεταβλητής $\sigma^2|\boldsymbol{\gamma}$ ρυθμίζονται ανάλογα της πληροφορίας που έχουμε σχετικά με το πείραμα. Για μη πληροφοριακή πρότερη καλές τιμές είναι οι λ=v=1. Για τις παραμέτρους της πρότερης κατανομής του θ ισχύει το ίδιο. Μια καλή επιλογή για μη πληροφοριακή πρότερη για το θ είναι η a=b=1. Παρόλα αυτά οι *Castillo* και *Van der Vaart* (2012) έδειξαν πως για μεγάλο p θα πρέπει να δοθεί μικρή τιμή στο α και μεγάλη στο b, με καλύτερη επιλογή την a=1 και b=p.

Αναλύοντας τις ύστερες κατανομές των $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$ μπορούμε να βγάλουμε συμπεράσματα για το ποια β_i επιδρούν έντονα στο μοντέλο και ποια όχι. Η μέθοδος Spike-and-Slab αποτελεί βασικό στοιχείο του $EMVS$, όπως θα φανεί στο Κεφάλαιο 5.

Κεφάλαιο 2

Επιλογή Μεταβλητών στην Μπεϋζιανή Στατιστική

Είναι προς όφελος κάθε στατιστικής μελέτης που αφορά παλινδρόμηση, να βρεθεί ποιες μεταβλητές επιδρούν σημαντικά πάνω στο μοντέλο που μελετάται και ποιες όχι. Η επιλογή μοντέλου ως το καλύτερο από όλα τα άλλα, σε σχέση με τις μεταβλητές που χρησιμοποιεί, είναι αρκετά υποκειμενικό θέμα. Τα τελευταία χρόνια έχουν προταθεί αρκετά κριτήρια και κανόνες απόφασης έτσι ώστε να βρίσκονται τα πιο αποδοτικά γραμμικά μοντέλα μέσα από επιμέρους συγκρίσεις, αλλά είναι σχεδόν απίθανο να υπάρξει κάποιο μοντέλο το οποίο θα είναι καλύτερο ως προς κάθε κριτήριο απέναντι σε όλα τα υπόλοιπα μοντέλα.

Το πόσο καλό είναι ένα γραμμικό παλινδρομικό μοντέλο κρίνεται από δύο παράγοντες. Ο πρώτος είναι πόσο καλά προσαρμόζεται στα δεδομένα (goodness of fit criterion). Ο δεύτερος από το πόσο χαμηλό είναι το υπολογιστικό κόστος που απαιτεί ώστε να δώσει αριθμητικά αποτελέσματα (parsimony criterion). Ο πρώτος παράγοντας είναι αρκετά προφανής, όσο καλύτερα προσδιορίζει το μοντέλο τα δεδομένα τόσο καλύτερα θα είναι και τα αποτελέσματά του. Ως προς τον δεύτερο παράγοντα δίνεται το εξής παράδειγμα. Ας υποθέσουμε ότι έχουμε δύο πιθανά γραμμικά μοντέλα και ένα κριτήριο που μας δίνει πόσο καλά εφαρμόζονται τα δεδομένα πάνω τους σε ποσοστό επί τοις εκατό, με το υψηλότερο ποσοστό να σημαίνει την καλύτερη προσαρμογή. Το πρώτο μοντέλο χρησιμοποιεί 50 επεξηγηματικές μεταβλητές και έχει 95% επιτυχία στο να περιγράφει τα δεδομένα, ενώ το δεύτερο χρησιμοποιεί 15 επεξηγηματικές μεταβλητές και έχει ποσοστό επιτυχίας 90%. Σίγουρα το πρώτο μοντέλο είναι καλύτερο ως προς την ακρίβεια. Αυτό φαίνεται και από το κριτήριο προσαρμογής αλλά και από το γεγονός ότι παίρνει πληροφορία από πολύ περισσότερες ποσότητες σχετικά με την έκβαση της μεταβλητής απόκρισης. Παρόλα αυτά κοιτώντας κανείς το Κεφάλαιο 1 εύκολα βλέπει πως το πρώτο μοντέλο μπορεί να έχει αρκετά πολύπλοκες πράξεις λόγω του πολύ μεγαλύτερου πλήθους των β_i αλλά και να είναι αρκετά δύσκολη η συλλογή της πληροφορίας του πίνακα X. Συνεπώς από άποψη «οικονομίας» το δεύτερο μοντέλο θα ήταν καλύτερο. Στην συγκεκριμένη περίπτωση, επιλέγοντας το οικονομικότερο μοντέλο έχουμε και πολύ ικανοποιητική προβλεπτική ικανότητα.

Ο George E. P. Box (1976) έχει γράψει σε μια εργασία του το απόφθεγμα, όλα τα μοντέλα που βγάζουμε είναι λάθος, αλλά κάποια είναι χρήσιμα. Εύκολα καταλαβαίνει κανείς λοιπόν πως ένα καλό γραμμικό μοντέλο έχει καλή προβλεπτική ικανότητα (όχι απαραίτητα την καλύτερη) και ταυτόχρονα χρησιμοποιεί όσο πιο βολικές υποθέσεις γίνεται. Τα κριτήρια goodness of fit και parsimony έχουν φύσει αντίθετη σημασία, καθώς σε λογικά πλαίσια, όσο προσπαθούμε να καλυτερεύσουμε το μοντέλο ως προς το ένα τόσο χειροτερεύει ως προς το άλλο, και αντίστροφα. Όσο προσθέτουμε μεταβλητές β_i, συνήθως, τόσο καλύτερη γίνεται η προβλεπτική ικανότητα του μοντέλου αλλά τόσο ανεβαίνει και το υπολογιστικό κόστος του. Πρέπει να βρεθεί λοιπόν ένας τρόπος που να βρίσκει το καλύτερο μοντέλο, λαμβάνοντας υπόψιν και τα δύο κριτήρια.

2.1 Έλεγχοι Υποθέσεων

Έστω ότι έχουμε δύο υποθέσεις H_0, H_1 και θέλουμε να δούμε ποια είναι πιο πιθανό να ισχύει. Στον χώρο της στατιστικής γενικότερα, τέτοια προβλήματα αντιμετωπίζονται μέσω των ελέγχων υποθέσεων. Στην Μπεϋζιανή στατιστική επικυρώνουμε ή απορρίπτουμε την κάθε υπόθεση βασιζόμενοι στην πιθανότητα του να συμβεί, δοθέντος του δείγματος που συλλέξαμε.

Για να γίνει ένας έλεγχος υπόθεσης, αρχικά, πρέπει να υποθέσουμε πρότερες πιθανότητες για τις υποθέσεις. Έστω ότι $P(H_0) = p_0$ και $P(H_1) = p_1 = 1-p_0$. Αμέσως μετά πρέπει να συλλεχθεί δείγμα $\mathbf{y}$ που αφορά τις υποθέσεις. Η συμπερασματολογία βασίζεται στις πιθανότητες

$$\tilde{p}_0 = P(H_0|\mathbf{Y} = \mathbf{y}) = \frac{f_{\mathbf{Y}}(\mathbf{y}|H_0)P(H_0)}{f_{\mathbf{Y}}(\mathbf{y})} = \frac{f_{\mathbf{Y}}(\mathbf{y}|H_0)p_0}{f_{\mathbf{Y}}(\mathbf{y})} \tag{2.1}$$

και

$$\tilde{p}_1 = P(H_1|\mathbf{Y} = \mathbf{y}) = \frac{f_{\mathbf{Y}}(\mathbf{y}|H_1)P(H_1)}{f_{\mathbf{Y}}(\mathbf{y})} = \frac{f_{\mathbf{Y}}(\mathbf{y}|H_1)(1-p_0)}{f_{\mathbf{Y}}(\mathbf{y})}. \tag{2.2}$$

Πιο συγκεκριμένα, σύμφωνα με τον κανόνα MAP (Maximum A Posteriori) επιλέγουμε την υπόθεση με την μεγαλύτερη ύστερη πιθανότητα ως αληθή. Οι υποθέσεις είναι εξ' ορισμού ξένα σύνολα μεταξύ τους, συνεπώς από τη στιγμή που θα βρεθεί η αληθής αυτόματα ορίζεται και η λανθασμένη.

Παράδειγμα

Ας υποθέσουμε πως η τυχαία μεταβλητή C περιγράφει την ένδειξη ενός πομπού που θα σταλεί σε έναν δέκτη. Οι ενδείξεις που στέλνει ο πομπός είναι:
• 1 με πρότερη πιθανότητα p_0.
• -1 με πρότερη πιθανότητα 1-p_0.
Λόγω εξωτερικών παραγόντων κατά την μετάδοση του σήματος, η ένδειξη που φτάνει στον δέκτη είναι η τυχαία μεταβλητή $R = C + \epsilon$, όπου $\epsilon \sim N(0, \sigma^2)$ ανεξάρτητη του C και σ^2 γνωστό. Θα γίνει έλεγχος υπόθεσης για το αν, για ένδειξη $R = r$ από τον δέκτη, η τιμή του C είναι πιο πιθανό να είναι 1 ή -1, χρησιμοποιώντας τον κανόνα MAP.

Έστω οι υποθέσεις $H_0 = \{C = 1\}$ και $H_1 = \{C = -1\}$. Για την τυχαία μεταβλητή R έχουμε πως,
• $R_0 := R|H_0 = R|C = 1 \ = 1 + \epsilon$,
• $R_1 := R|H_1 = R|C = -1 \ = -1 + \epsilon$.
Όμως, $\epsilon \sim N(0, \sigma^2)$, συνεπώς $R_0 \sim N(1, \sigma^2)$ και $R_1 \sim N(-1, \sigma^2)$. Τότε, από όλα τα παραπάνω,

$$\begin{aligned}
\tilde{p}_0 &= P(C = 1|R = r) = \frac{f_{R_0}(r)p_0}{f_R(r)} = \frac{p_0}{f_R(r)}\frac{1}{\sqrt{2\pi}}e^{-\frac{(r-1)^2}{2\sigma^2}}, \\
\tilde{p}_1 &= P(C = -1|R = r) = \frac{f_{R_1}(r)(1-p_0)}{f_R(r)} = \frac{1-p_0}{f_R(r)}\frac{1}{\sqrt{2\pi}}e^{-\frac{(r+1)^2}{2\sigma^2}}.
\end{aligned} \tag{2.3}$$

Συνοψίζοντας, σύμφωνα με τον κανόνα MAP, από την (2.3), είναι πιο πιθανό η ένδειξη του πομπού να είναι ίση με 1 αν

$$\frac{f_{R_0}(r)p_0}{f_R(r)} > \frac{f_{R_1}(r)(1-p_0)}{f_R(r)} \Leftrightarrow r > \frac{\sigma^2}{2}log(\frac{1-p_0}{p_0}), \tag{2.4}$$

αλλιώς είναι πιο πιθανό να είναι ίση με -1.

2.2 Παράγοντας Bayes

Ο παράγοντας $Bayes$ χρησιμοποιείται, επίσης, στους ελέγχους υποθέσεων στην Μπεϋζιανή στατιστική. Έστω οι πρότερες πιθανότητες $p_0,\ p_1$ για τις αντίστοιχες υποθέσεις $H_0,\ H_1$ και οι ύστερες $\tilde{p}_0,\ \tilde{p}_1$ (όπως στις σχέσεις (2.1) και (2.2)). Τότε ο παράγοντας $Bayes$ υπέρ της H_1 ορίζεται ως,

$$B_{10} := \frac{\tilde{p}_1}{\tilde{p}_0}\frac{p_0}{p_1} = \frac{f_{\mathbf{Y}|H_1}(\mathbf{y}|H_1)}{f_{\mathbf{Y}|H_0}(\mathbf{y}|H_0)}. \tag{2.5}$$

Η δεύτερη ισότητα του παράγοντα $Bayes$ στην (2.5) προκύπτει άμεσα από την πρώτη, μέσω ιδιοτήτων των δεσμευμένων κατανομών. Η φράση «υπέρ της H_1» όπως και ο δείκτης 10 στο σύμβολο B χρησιμεύουν στο να φανεί η θέση της πυκνότητας της κάθε υπόθεσης στο κλάσμα του παράγοντα $Bayes$. Αντίστοιχα, για τον παράγοντα $Bayes$ υπέρ της H_0,

$$B_{01} := \frac{\tilde{p}_0}{\tilde{p}_1}\frac{p_1}{p_0} = \frac{f_{\mathbf{Y}|H_0}(\mathbf{y}|H_0)}{f_{\mathbf{Y}|H_1}(\mathbf{y}|H_1)} = \frac{1}{B_{10}}. \tag{2.6}$$

Για την συμπερασματολογία του παράγοντα $Bayes$ έχουν προταθεί δύο παρεμφερείς τρόποι. Ο πρώτος δόθηκε από τον $Jeffreys$ (1961) και βασίστηκε στην λογαριθμική συνάρτηση με βάση τον αριθμό 10.

Κλίμακα του $Jeffreys$		
$log_{10}B_{10}$	B_{10}	Στοιχεία κατά της H_0
από 0 έως 0.5	από 1 έως 3.2	όχι αρκετά
από 0.5 έως 1	από 3.2 έως 10	ουσιαστικά
από 1 έως 2	από 10 έως 100	ισχυρά
>2	>100	καθοριστικά

Ο δεύτερος δόθηκε από τους $Kass$ και $Raftery$ (1995) και βασίστηκε στον νεπέριο λογάριθμο.

Κλίμακα των $Kass$ και $Raftery$		
$2logB_{10}$	B_{10}	Στοιχεία κατά της H_0
από 0 έως 2	από 1 έως 3	όχι αρκετά
από 2 έως 6	από 3 έως 20	θετικά
από 6 έως 10	από 20 έως 150	ισχυρά
>10	>150	πολύ ισχυρά

Οι κλίμακες φτιάχτηκαν καθαρά διαισθητικά.

2.3 Εφαρμογή σε Γραμμικά Μοντέλα

Έστω ότι έχουμε ένα γραμμικό μοντέλο στο οποίο είναι πιθανό να επιδρούν σημαντικά p διαφορετικοί παράγοντες. Αν προσπαθήσουμε να φτιάξουμε κάθε μοντέλο που μπορεί να φτιαχτεί κάνοντας συνδυασμούς με τους επιμέρους παράγοντες, τότε προκύπτουν 2^p πιθανά μοντέλα. Το πρόβλημα της επιλογής μεταβλητών έχει να κάνει με το να βρούμε ένα από τα παραπάνω πιθανά μοντέλα με καλή προβλεπτική ικανότητα και λογικό αριθμό παραμέτρων β_i.

Υπάρχουν δείκτες οι οποίοι μετρούν το πόσο καλό είναι το μοντέλο σε σχέση με τα δεδομένα. Κάποιοι αρκετά διαδομένοι που υπολογίζονται άμεσα από τα περισσότερα υπολογιστικά πακέτα είναι τα κριτήρια AIC, BIC και DIC. Όσο μικρότερες οι τιμές τους, τόσο καλύτερο είναι το μοντέλο. Μια τεχνική επιλογής μοντέλου, και κατά συνέπεια επιλογής μεταβλητών, θα ήταν να βρεθούν οι δείκτες για κάθε μοντέλο και να κρατήσουμε ως καλύτερα πιθανά μοντέλα αυτά με τις χαμηλότερες τιμές δεικτών.

Μια εναλλακτική μέθοδος επιλογής μοντέλου θα ήταν με το να γίνει χρήση της ποσότητας PO_{10} ($Posterior\ model\ odds\ of\ model\ m_1\ vs\ m_0$),

$$PO_{10} := \frac{\pi(m_1|\mathbf{y})}{\pi(m_0|\mathbf{y})} = \frac{f(\mathbf{y}|m_1)}{f(\mathbf{y}|m_0)} \frac{\pi_{m_1}}{\pi_{m_0}}. \tag{2.7}$$

Στην σχέση (2.7) ορίζεται η ποσότητα PO_{10} για δύο γραμμικά μοντέλα m_0, m_1. Όπως φαίνεται από τον τύπο, συγκρίνουμε τις πιθανότητες των μοντέλων κάτω από την υπόθεση των δεδομένων που συλλέξαμε, όπως και στον κανόνα MAP. Το κλάσμα $\frac{f(\mathbf{y}|m_1)}{f(\mathbf{y}|m_0)}$ ταυτίζεται με τον παράγοντα $Bayes$. Όσο μεγαλύτερο από το 1 είναι το PO_{10}, τόσο πιο ισχυρές ενδείξεις έχουμε υπέρ του μοντέλου m_1 ως καταλληλότερο απέναντι στο m_0 και αντίστροφα για PO_{10} μικρότερο του 1. Τα $f(\mathbf{y}|m_j)$ υπολογίζονται μέσω του θεωρήματος ολικής πιθανότητας ως εξής,

$$f(\mathbf{y}|m_j) = \int_{\boldsymbol{\beta}_j} f(\mathbf{y}|\boldsymbol{\beta}_j, m_j) f(\boldsymbol{\beta}_j|m_j) d\boldsymbol{\beta}_j. \tag{2.8}$$

Στην επιλογή μεταβλητών η ποσότητα PO χρησιμεύει ως εξής, επιλέγουμε ένα μοντέλο m_j και υπολογίζουμε τα $PO_{ij},\ \forall i \neq j$. Θα πρέπει να υποθέσουμε, βέβαια, πρότερες πιθανότητες π_{m_k} για κάθε μοντέλο $m_k,\ k \in \{1, 2, ..., 2^p\}$, σχετικά με το πόσο καλό είναι σε σχέση με τα υπόλοιπα. Προφανώς,

$$\pi_{m_1} + \pi_{m_2} + ... + \pi_{m_{2^p}} = 1.$$

Σε περίπτωση που δεν έχουμε καμιά πρότερη πληροφορία συνήθως θέτουμε

$$\pi_{m_1} = \pi_{m_2} = ... = \pi_{m_{2^p}} = \frac{1}{2^p},$$

έτσι ώστε $\frac{\pi_{m_i}}{\pi_{m_j}} = 1$ και το γινόμενο δεν επηρεάζεται από την πρότερη πληροφορία. Αν

$$PO_{ij} < 1,\ \forall i \neq j,$$

τότε ως καλύτερο μοντέλο επιλέγεται το m_j, αλλιώς το m_b για το οποίο

$$PO_{ij} < PO_{bj},\ \forall i \neq j, b.$$

Ο αλγόριθμος $EMVS$, όπως θα δούμε παρακάτω, είναι πολύ πιο ευέλικτος και στις επιμέρους συγκρίσεις μοντέλων αλλά και στην συμπερασματολογία, σε σχέση με τις μεθόδους που αναφέραμε σε αυτήν την ενότητα. Επιπλέον είναι υπολογιστικά οικονομικότερος σε περιπτώσεις όπου το p είναι αρκετά μεγάλο και σύμφωνα με τα παραπάνω θα πρέπει να γίνουν πράξεις ή να βρεθούν κριτήρια για 2^p πιθανά μοντέλα.

2.4 Υπέρ και Κατά των μεθόδων στην Μπεϋζιανή Στατιστική

Η Μπεϋζιανή προσέγγιση στα γραμμικά μοντέλα γενικότερα αλλά και συγκεκριμένα στην επιλογή μεταβλητών παρουσιάζει και πλεονεκτήματα και μειονέκτημα.

Πλεονεκτήματα:

- Μέσω εφαρμογής μεθόδων MCMC είναι εφικτό να γίνουν πολύπλοκες πράξεις για τις ανάγκες της συμπερασματολογίας και επαρκείς υπολογισμοί για τις ανάγκες της αναζήτησης μοντέλων ($model\ search$).

- Κάτω από κριτήρια και συγκρίσεις είναι εύκολο βρεθεί μοντέλο που χαρακτηρίζεται «καλύτερο» από τα υπόλοιπα.
- Μέσω των εκ των υστέρων πιθανοτήτων είναι εύκολο να γίνει άμεσα και εύκολα κατανοητή συμπερασματολογία σε πολλαπλά μοντέλα.
- Μέσω της Μπεϋζιανής προσέγγισης είναι εφικτό αντί να βρεθεί ένα μοντέλο ως κατάλληλο, να βρεθεί ένα μικρό σύνολο μοντέλων ως κατάλληλα με αντίστοιχα ποσοτικά μέτρα (εκ των υστέρων πιθανότητες) που περιγράφουν πόσο κατάλληλα είναι και δίνεται δυνατότητα περαιτέρω διεργασίας.
- Μέσω της Μπεϋζιανής προσέγγισης είναι εφικτό να γίνουν συγκρίσεις μεταξύ μοντέλων που δομούνται από πολύ διαφορετικούς παράγοντες και υποθέσεις.

Μειονεκτήματα:

- Λόγω των πρότερων κατανομών, είναι πιθανό να υπάρξει μη επιθυμητή ευαισθησία από τις πρότερες πιθανότητες στα ύστερα δεδομένα και τον παράγοντα $Bayes$, και έπειτα παράδοξα στη συμπερασματολογία.
- Οι υπολογισμοί στις ύστερες ποσότητες μπορούν να γίνουν αρκετά σύνθετες, ειδικότερα έξω από την υπόθεση της συζυγίας.
- Το υπολογιστικό κόστος της διαδικασίας αναζήτησης μοντέλων, μπορεί να είναι πολύ υψηλό, ειδικά σε περιπτώσεις όπου το πλήθος των πιθανών μοντέλων είναι πολύ μεγάλο.

Κεφάλαιο 3

Ο Αλγόριθμος ΕΜ

Σε αυτό το μέρος της εργασίας παρουσιάζεται συνοπτικά ο αλγόριθμος ΕΜ. Στόχος αυτού του Κεφαλαίου είναι να δοθεί ο τρόπος με τον οποίο λειτουργεί ο ΕΜ και όχι απόδειξη για το ότι λειτουργεί. Οι λόγοι για τους οποίους ο αλγόριθμος συγκλίνει στην ποσότητα που θέλουμε δίνονται στο επόμενο Κεφάλαιο.

Τα αρχικά ΕΜ προέρχονται από τις αγγλικές λέξεις Expectation-Maximization και έχουν επιλεχθεί ως όνομα του αλγορίθμου γιατί οι αντίστοιχες λέξεις είναι άμεσα συνδεδεμένες με τα βασικά του βήματα. Το όνομα αυτό αποδόθηκε επίσημα από τους Dempster, Laird και Rubin (1977) αλλά η ιδέα πίσω από τον αλγόριθμο και η διαισθητική του χρήση υπήρχε και σε παλιότερα άρθρα. Ο αλγόριθμος ΕΜ αποτελεί ειδική περίπτωση του αλγορίθμου MM (Majorization-Minimization).

3.1 Εφαρμογή του ΕΜ στην Κλασική Στατιστική

Ο ΕΜ είναι ένας επαναληπτικός αλγόριθμος ο οποίος χρησιμοποιείται στην κλασική στατιστική για να βρεθούν οι εκτιμητές μέγιστης πιθανοφάνειας (maximum likelihood estimations, MLE, ΕΜΠ) για τις άγνωστες σταθερές παραμέτρους μιας κατανομής. Στα πλαίσια αυτής της εργασίας μας ενδιαφέρει περισσότερο η εφαρμογή του ΕΜ στην Μπεϋζιανή στατιστική. Στην Μπεϋζιανή στατιστική, όπως φαίνεται στην επόμενη ενότητα, η χρηστικότητα του ΕΜ είναι διαφορετική, μιας και γενικότερα στην Μπεϋζιανή προσέγγιση οι παράμετροι είναι τυχαίες μεταβλητές και δεν μπορούν να οριστούν για αυτές εκτιμητές μέγιστης πιθανοφάνειας. Παρόλα αυτά θα γίνει εισαγωγή του ΕΜ μέσω της κλασικής του προσέγγισης σε αυτήν την ενότητα, και έπειτα θα επεκταθεί στην Μπεϋζιανή στατιστική.

Για τις ανάγκες του ΕΜ θεωρούμε:

- δείγμα που παρατηρήθηκε (ελλιπές), $\mathbf{Y}=(Y_1, Y_2, ..., Y_n)$, διάστασης n,
- δείγμα που δεν καταφέραμε να παρατηρήσουμε, $\mathbf{Z}=(Z_1, Z_2, ..., Z_{m-n})$, διάστασης $m-n$, $n \leq m$,
- πλήρες δείγμα, $\mathbf{W}=(\mathbf{Y},\mathbf{Z})$, διάστασης m,
- σταθερές άγνωστες παραμέτρους, $\boldsymbol{\vartheta}=(\theta_1, \theta_2, ..., \theta_r)$ της κατανομής του δείγματος, διάστασης r,
- την πιθανοφάνεια του $\mathbf{W}$, $L(\boldsymbol{\vartheta}|\mathbf{Y}, \mathbf{Z})$,

- την πιθανοφάνεια του $\mathbf{Y}$, $L(\boldsymbol{\vartheta}|\mathbf{Y})$,
- την επαναληπτική ακολουθία του αλγορίθμου, $\boldsymbol{\vartheta}^k=(\theta_1^k, \theta_2^k, ..., \theta_r^k)$, διάστασης r,
- το σημείο σύγκλισης του αλγορίθμου, $\boldsymbol{\vartheta}^*=(\theta_1^*, \theta_2^*, ..., \theta_r^*)$ διάστασης r.

Το ελλιπές δείγμα αντιπροσωπεύει το δείγμα που έχουμε συλλέξει. Το πλήρες δείγμα αντιπροσωπεύει όλο το θεωρητικό δείγμα ως προς το οποίο θα εφαρμοστεί ο ΕΜ. Το πλήρες δείγμα περιέχει και ένα κομμάτι που δεν καταφέραμε να συλλέξουμε, όπως φαίνεται και παραπάνω. Σε πολλές περιπτώσεις εφαρμογής του ΕΜ, δεν υπάρχει όντως δείγμα που δεν καταφέραμε να συλλέξουμε, αλλά το υποθέτουμε με σκοπό να είναι εφικτό να εφαρμοστεί ο αλγόριθμος.

Ο αλγόριθμος ξεκινά με το να δοθούν αρχικές τιμές. Το αρχικό διάνυσμα που δίνουμε θα το συμβολίζουμε $\boldsymbol{\vartheta}^0$. Το $\boldsymbol{\vartheta}^0$ πρέπει να ανήκει στον παραμετρικό χώρο. Η τιμή του $\boldsymbol{\vartheta}^0$, συνήθως, είναι αυθαίρετη, αλλά όπως θα φανεί στο επόμενο Κεφάλαιο, σχετίζεται άμεσα με την σύγκλιση του αλγορίθμου. Η επαναληπτική ακολουθία του ΕΜ $\boldsymbol{\vartheta}^k$ συγκλίνει στον ΕΜΠ, κάτω από προϋποθέσεις.

Ακολουθεί μια εντολή «Μέχρις ότου» η οποία σταματά μόλις επιτευχθεί η σύγκλιση. Για να ελέγξουμε την σύγκλιση υπάρχουν ποικίλοι τρόποι. Μέσα στην «Μέχρις ότου» επαναλαμβάνονται τα βήματα $E-step$ και $M-step$. Μέσα από κάθε επανάληψη k των βημάτων δημιουργείται μια νέα τιμή $\boldsymbol{\vartheta}^k$ η οποία κάνει όλο και μεγαλύτερη την πιθανοφάνεια $L(\boldsymbol{\vartheta}^k|\mathbf{Y})$ και, υπό συνθήκες, είναι όλο και πιο κοντά στον ΕΜΠ. Στο Διάγραμμα 3.1 φαίνεται ένα διάγραμμα ροής του ΕΜ.

Τα παρακάτω κριτήρια χρησιμοποιούνται για τον έλεγχο της σύγκλισης:

- $DL := \frac{L(\boldsymbol{\vartheta}^{k+1}|\mathbf{Y})-L(\boldsymbol{\vartheta}^k|\mathbf{Y})}{L(\boldsymbol{\vartheta}^k|\mathbf{Y})} < \epsilon$,
- $|\theta_i^{k+1} - \theta_i^k| < \epsilon,\ \forall i \in \{1, 2, ..., r\}$.

Το $\epsilon > 0$ θα είναι μια οριστέα τιμή που πρέπει να είναι πολύ κοντά στο 0. Ο αλγόριθμος σταματά στην επανάληψη $k+1$, όπου είναι η πρώτη φορά που ισχύει η συνθήκη του κριτηρίου που επιλέχθηκε. Το δεύτερο κριτήριο είναι και αυτό που χρησιμοποιείται για τις ανάγκες της εργασίας.

Κατά το $E-step$, στην k επανάληψη, υπολογίζεται η ποσότητα

$$\begin{aligned} Q(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^k) &= \int log(L(\boldsymbol{\vartheta}|\mathbf{Y}, \mathbf{Z}))P(\mathbf{Z}|\mathbf{Y} = \mathbf{y}; \boldsymbol{\vartheta}^k)d\mathbf{Z} \\ &= E_{\mathbf{Z}|\mathbf{Y};\boldsymbol{\vartheta}^k}[log(L(\boldsymbol{\vartheta}|\mathbf{Y}, \mathbf{Z}))]. \end{aligned}$$

Κατά το $M-step$, στην k επανάληψη, υπολογίζεται η ποσότητα $\boldsymbol{\vartheta}^{k+1}$, η οποία είναι το $\boldsymbol{\vartheta}$ που μεγιστοποιεί την $Q(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^k)$. Συνοπτικά,

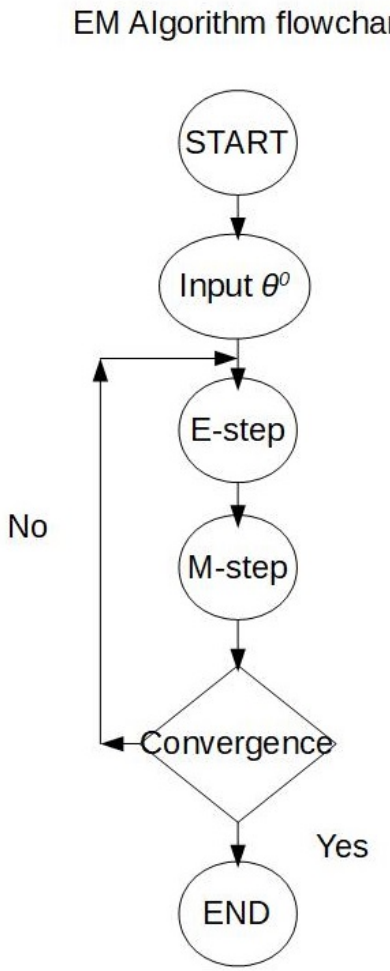
$$\boldsymbol{\vartheta}^{k+1} = argmax_{\boldsymbol{\vartheta}} Q(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^k).$$

Λόγω του ορισμού του $\boldsymbol{\vartheta}^{k+1}$, είναι εμφανές πως σε κάποιες περιπτώσεις παραπάνω από μία τιμές θα ικανοποιούν την προηγούμενη συνθήκη. Σε περίπτωση που δεν έχει επιτευχθεί σύγκλιση, το $\boldsymbol{\vartheta}^{k+1}$ χρησιμοποιείται στην επόμενη επανάληψη του $E-step$ στην ποσότητα

$$Q(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^{k+1}) = \int log(L(\boldsymbol{\vartheta}|\mathbf{Y}, \mathbf{Z}))P(\mathbf{Z}|\mathbf{Y} = \mathbf{y}; \boldsymbol{\vartheta}^{k+1})d\mathbf{Z}$$

και ο αλγόριθμος συνεχίζεται μέχρι την σύγκλιση.

Διάγραμμα 3.1: Διάγραμμα ροής του αλγορίθμου ΕΜ.



Αναφέρεται πως ο αλγόριθμος ΕΜ συγκλίνει σίγουρα στον ΕΜΠ αν η πιθανοφάνεια $L(\boldsymbol{\vartheta}|\mathbf{Y})$ έχει μόνο ένα μέγιστο και είναι κοίλη συνάρτηση ως προς $\boldsymbol{\vartheta}$. Το να έχει μόνο ένα μέγιστο εξυπηρετεί στο γεγονός ότι, καθώς σε κάθε επανάληψη γεννάται ένα $\boldsymbol{\vartheta}^k$ που κάνει όλο και μεγαλύτερη την $L(\boldsymbol{\vartheta}^k|\mathbf{Y})$, το $\boldsymbol{\vartheta}^k$ θα πηγαίνει προς τον ΕΜΠ καθώς k πηγαίνει στο άπειρο. Η δέσμευση του να είναι κοίλη, εξυπηρετεί στο να μην υπάρχουν άλλα στάσιμα σημεία πέραν του μοναδικού μεγίστου. Η επαναληπτική ακολουθία του ΕΜ γενικά συγκλίνει σε στάσιμα σημεία.

Σε περίπτωση που ισχύουν οι συνθήκες για τις οποίες ο αλγόριθμος συγκλίνει στην τιμή που θέλουμε, αν ο στόχος μας είναι μόνο να επιτύχουμε σύγκλιση, κατά το $M-step$ το $\boldsymbol{\vartheta}^{k+1}$ δεν είναι απαραίτητο να μεγιστοποιεί την $Q(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^k)$. Για να συγκλίνει ο αλγόριθμός αρκεί να βρίσκουμε σε κάθε $M-step$ μια τιμή $\boldsymbol{\vartheta}^{k+1}$ μέσα στον παραμετρικό χώρο του $\boldsymbol{\vartheta}$, τέτοια ώστε

$$Q(\boldsymbol{\vartheta}^{k+1};\boldsymbol{\vartheta}^k) \geq Q(\boldsymbol{\vartheta}^k;\boldsymbol{\vartheta}^k). \tag{3.1}$$

Τότε ο αλγόριθμος που χρησιμοποιείται είναι μια επέκταση του ΕΜ, λέγεται GEM (Generalized Expectation-Maximization) και είναι αρκετά χρήσιμος στην περίπτωση που δεν είναι εύκολο να βρεθεί το $\boldsymbol{\vartheta}$ που μεγιστοποιεί την ποσότητα Q αλλά είναι εύκολο να βρεθεί $\boldsymbol{\vartheta}^{k+1}$ που ικανοποιεί την (3.1). Βέβαια πρέπει να σημειωθεί πως στον αλγόριθμο GEM υπάρχουν περισσότερα προβλήματα στην σύγκλιση απ΄ ότι στον ΕΜ, γι΄ αυτό και απαιτείται περισσότερη προσοχή στην χρήση του. Υπάρχουν πολλές παραλλαγές του αλγορίθμου ΕΜ που εξυπηρετούν είτε στην διευκόλυνση του υπολογισμού των $E-step$ και $M-step$, είτε στην επιτάχυνση της σύγκλισης. Μέσα σε αυτές τις παραλλαγές, μια πληθώρα είναι αλγόριθμοι που χρησιμοποιούν μεθόδους Markov Chain Monte Carlo.

Όπως φαίνεται ο αλγόριθμος ΕΜ είναι μια διαδικασία που μπορεί να γίνει αρκετά σύνθετη, από άποψη υπολογισμών και επαναλήψεων. Είναι λογικό σε περιπτώσεις που μπορεί να βρεθεί η μέγιστη πιθανοφάνεια άμεσα, ο αλγόριθμος να παραλείπεται.

Παράδειγμα (Το κλασικό γενετικό πρόβλημα, Fisher (1925))

Έστω σταθερά $p \in (0,1)$ και δείγμα παρατηρήσεων $\mathbf{y}= (y_1, y_2, y_3, y_4)$ από τυχαία μεταβλητή $\mathbf{Y}= (Y_1, Y_2, Y_3, Y_4)$ που ακολουθεί πολυωνυμική κατανομή με πιθανότητες εμφάνισης για τα επιμέρους Y_i,

$$\frac{1}{2}+\frac{p}{4},\ \frac{1}{4}-\frac{p}{4},\ \frac{1}{4}-\frac{p}{4},\ \frac{p}{4}$$

αντίστοιχα. Αναλύω την τυχαία μεταβλητή Y_1 σε $Y_1 = Z_1 + Z_2$ με πιθανότητες εμφάνισης για τα $Z_1,\ Z_2$

$$\frac{1}{2},\ \frac{p}{4}.$$

Επομένως ισχύει πως η τυχαία μεταβλητή $\mathbf{W}=(Z_1, Z_2, Y_2, Y_3, Y_4)$ ακολουθεί την πολυωνυμική κατανομή με πιθανότητες εμφάνισης για τις επιμέρους τυχαίες μεταβλητές

$$\frac{1}{2},\ \frac{p}{4},\ \frac{1}{4}-\frac{p}{4},\ \frac{1}{4}-\frac{p}{4},\ \frac{p}{4},$$

αντίστοιχα. Προφανώς, δοθέντος των Y_1 και Z_1 είναι γνωστή και η Z_2, και ομοίως για την Z_1 δοθέντος των Y_1, Z_2. Ακόμα $\nu := y_1 + y_2 + y_3 + y_4$. Θα χρησιμοποιηθεί ο αλγόριθμος ΕΜ να υπολογιστεί ο εκτιμητής μέγιστης πιθανοφάνειας του p.

Αρχικά θα υπολογιστεί η ποσότητα $Q(p; p^k)$. Λόγω της πολυωνυμικής κατανομής, η πιθανοφάνεια του δείγματος $\mathbf{Y}$ θα είναι η

$$L(p|\mathbf{Y}) = \frac{\nu!}{y_1!y_2!y_3!y_4!}(\frac{1}{2}+\frac{p}{4})^{y_1}(\frac{1}{4}-\frac{p}{4})^{y_2}(\frac{1}{4}-\frac{p}{4})^{y_3}(\frac{p}{4})^{y_4}. \tag{3.2}$$

Θεωρώ πλήρες δείγμα το $\mathbf{W}$ και ελλιπές το $\mathbf{Y}$. Άρα, η πιθανοφάνεια του πλήρους δείγματος θα είναι η

$$L(p|\mathbf{Y},\mathbf{Z}) = \frac{\nu!}{z_1!z_2!y_2!y_3!y_4!}(\frac{1}{2})^{z_1}(\frac{p}{4})^{z_2}(\frac{1}{4}-\frac{p}{4})^{y_2}(\frac{1}{4}-\frac{p}{4})^{y_3}(\frac{p}{4})^{y_4}. \tag{3.3}$$

Για το Q,

$$\begin{aligned} Q(p;p^k) &= E_{\mathbf{Z}|\mathbf{Y};p^k}[logL(\boldsymbol{\vartheta}|\mathbf{Y},\mathbf{Z})] \\ &= \tilde{c} + E_{\mathbf{Z}|\mathbf{Y};p^k}[(Z_2+y_4)log(\frac{p}{4}) + (y_2+y_3)log(\frac{1}{4}-\frac{p}{4})] \\ &= \tilde{c} + (E[Z_2|Y_1=y_1;p^k]+y_4)log(\frac{p}{4}) + (y_2+y_3)log(\frac{1}{4}-\frac{p}{4}), \end{aligned} \tag{3.4}$$

όπου $\tilde{c}$ σταθερά που δεν περιέχει p. Για τον υπολογισμό της μέσης τιμής $E[Z_2|Y_1=y_1;\boldsymbol{\vartheta}^k]$ στην (3.4), αρκεί να γίνει αισθητό πως $Z_2|Y_1=y_1 \sim B(y_1, \frac{\frac{p}{4}}{\frac{1}{2}+\frac{p}{4}})$. Τότε,

$$E[Z_2|Y_1=y_1;p^k] = \frac{y_1\frac{p^k}{4}}{\frac{1}{2}+\frac{p^k}{4}}. \tag{3.5}$$

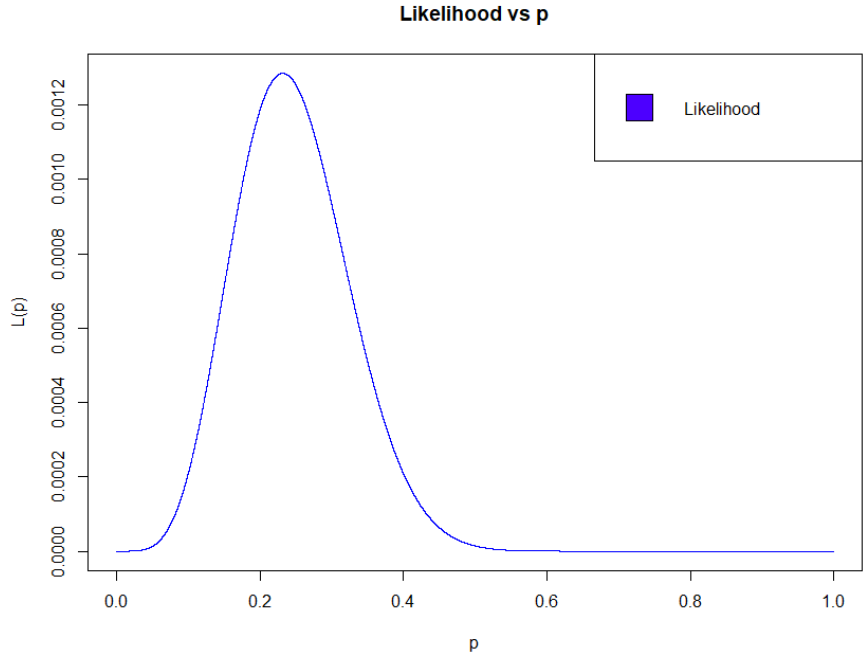
Παραγωγίζοντας ως προς p το Q, εύκολα φαίνεται ότι μεγιστοποιείται για

$$p^{k+1} = \frac{E[Z_2|Y_1=y_1;p^k]+y_4}{E[Z_2|Y_1=y_1;p^k]+y_2+y_3+y_4}. \tag{3.6}$$

Άρα σε αυτήν την περίπτωση η ακολουθία βημάτων του ΕΜ θα είναι:
-ξεκίνα από κάποιο $p^0 \in (0,1)$,
-βάλε στο $p^{k+1} = \frac{\mu^k + y_4}{\mu^k + y_2 + y_3 + y_4}$, όπου μ^k η δεσμευμένη μέση τιμή του Z_2 για παράμετρο p^k,
-επανέλαβε το προηγούμενο βήμα μέχρις ότου επιτευχθεί σύγκλιση.

Αν τώρα υποθέσουμε πως στο παραπάνω παράδειγμα έχουμε δείγμα $\mathbf{y}$=(36,12,17,5), έχουμε τα ακόλουθα αποτελέσματα. Στο Διάγραμμα 3.2 φαίνεται η συνάρτηση πιθανοφάνειας, για το παραπάνω δείγμα. Η συνάρτηση έχει μόνο ένα στάσιμο σημείο στο οποίο βρίσκεται το μέγιστο της κατανομής, όπως φαίνεται και από το διάγραμμα. Επομένως η ακολουθία του ΕΜ θα συγκλίνει στον εκτιμητή μέγιστης πιθανοφάνειας.

Διάγραμμα 3.2: Διάγραμμα της πιθανοφάνειας του παραδείγματος.

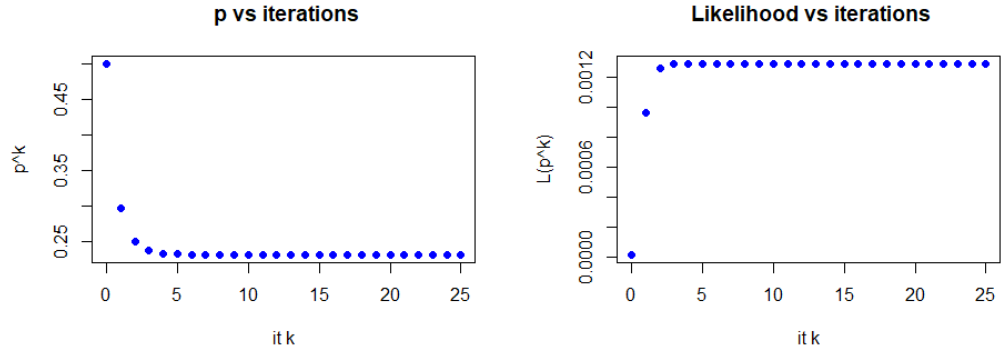


Ο πίνακας 3.1 δείχνει διάφορα μεγέθη που υπολογίζονται καθώς εκτελείται ο αλγόριθμος ΕΜ για το συγκεκριμένο παράδειγμα. Ως αρχική τιμή δόθηκε το 0.5. Σε κάθε επανάληψη φαίνεται πως για κάθε καινούργια τιμή του p η πιθανοφάνεια αυξάνεται, όπως ήταν αναμενόμενο. Στο Διάγραμμα 3.3 φαίνεται η εξέλιξη των $p^k, L(p^k|\mathbf{Y})$ καθώς τρέχει ο αλγόριθμος. Στον άξονα των χ΄χ είναι ο α/α κάθε επανάληψης ενώ στον άξονα των $y'y$ οι αντίστοιχες τιμές των $p^k, L(p^k|\mathbf{Y})$. Ο ΕΜΠ του p θα είναι περίπου ίσος με 0.2314670, η τιμή για την οποία επιτεύχθηκε σύγκλιση.

Πίνακας 3.1: Πίνακας με μεγέθη που προκύπτουν από τις επαναλήψεις του αλγορίθμου ΕΜ.

k	p^k	logL(p^k)	L(p^k)
0	0.5000000	-11.18157	0.0000139285
1	0.2961165	-6.945286	0.0009631649
2	0.2495350	-6.680575	0.0012550562
16	0.2314671	-6.656775	0.0012852841
17	0.2314671	-6.656775	0.0012852841
18	0.2314670	-6.656775	0.0012852841
19	0.2314670	-6.656775	0.0012852841

Διάγραμμα 3.3: Διαγράμματα που δείχνουν την εξέλιξη του αλγορίθμου.



3.2 Εφαρμογή του ΕΜ στην Μπεϋζιανή Στατιστική

Στην προηγούμενη ενότητα έγινε περιγραφή της εφαρμογής του ΕΜ στην κλασική στατιστική. Ο ΕΜ όμως εφαρμόζεται και στην Μπεϋζιανή στατιστική. Στην Μπεϋζιανή στατιστική οι άγνωστες παράμετροι $\boldsymbol{\vartheta}$ δεν είναι σταθερές αλλά τυχαίες μεταβλητές. Ο ΕΜ στην Μπεϋζιανή περίπτωση χρησιμοποιείται για να βρεθεί το σημείο που βρίσκεται η κορυφή της κατανομής των παραμέτρων δοθέντος του δείγματος, δηλαδή για την τυχαία μεταβλητή $\boldsymbol{\vartheta}|\mathbf{Y}$. Το σημείο που συγκλίνει ο αλγόριθμος θα συμβολίζεται και πάλι με $\boldsymbol{\vartheta}^*$. Οι διαφορές στις δύο προσεγγίσεις προκύπτουν από το γεγονός ότι αλλάζει λόγω ορισμού, η φύση της άγνωστης παραμέτρου και συνεπώς ο τρόπος με τον οποίο «βαθμολογούμε» την εμφάνιση των τιμών τους.

Και εδώ, η λογική του ΕΜ είναι ίδια με αυτήν που φαίνεται στο Διάγραμμα ροής (3.1). Είναι ευνόητο, όμως, πως πέρα από το αποτέλεσμα του αλγορίθμου θα υπάρχουν αλλαγές και στο τι πραγματεύονται τα $E-step$ και $M-step$. Για το $E-step$ και πάλι θα πρέπει να υπολογιστεί η ποσότητα Q μόνο που τώρα ορίζεται ως,

$$\begin{aligned} Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) :=& \int log(f_{\boldsymbol{\vartheta}|\mathbf{W}}(\boldsymbol{\vartheta}|\mathbf{Z},\mathbf{y}))P[\mathbf{Z}|\mathbf{Y}=\mathbf{y},\boldsymbol{\vartheta}^k]d\mathbf{Z} \\ =& E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log(f_{\boldsymbol{\vartheta}|\mathbf{W}}(\boldsymbol{\vartheta}|\mathbf{Z},\mathbf{y}))]. \end{aligned} \tag{3.7}$$

Για το $M-step$ πρέπει να βρεθεί, και πάλι, η τιμή του $\boldsymbol{\vartheta}$ για την οποία μεγιστοποιείται η (3.8).

Για να είναι εμφανές το κάτω από ποιες υποθέσεις χρησιμοποιείται ο ΕΜ, στην κλασική στατιστική το διαχωριστικό στο Q θα είναι το ερωτηματικό «;», ενώ στην Μπεϋζιανή η γραμμή «|». Το $f_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})$ είναι η πρότερη κατανομή των παραμέτρων $\boldsymbol{\vartheta}$ και εύκολα φαίνεται πως

$$Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) = Q(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^k) + log f_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}) + \bar{c}, \tag{3.8}$$

από το Θεώρημα $Bayes$. Το $\bar{c}$ είναι σταθερά που δεν περιέχει $\boldsymbol{\vartheta}$.

Και εδώ αναφέρεται πως ο ΕΜ δεν συγκλίνει πάντα στην κορυφή της κατανομής της $\boldsymbol{\vartheta}|\mathbf{Y}$, αλλά υπό προϋποθέσεις. Παρόλα αυτά σε κάθε βήμα το $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y})$ γίνεται όλο και μεγαλύτερο και η σύγκλιση στην κορυφή είναι βέβαιη αν η πυκνότητα πιθανότητας $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}|\mathbf{y})$ είναι μονοκόρυφη και κοίλη ως προς $\boldsymbol{\vartheta}$.

Αν στο παράδειγμα της Ενότητας 3.1 το p ήταν τυχαία μεταβλητή, με πρότερη κατανομή την ομοιόμορφη στο διάστημα (0,1), εύκολα φαίνεται από την (3.8) πως δεν θα άλλαζε κάτι στην λύση σε σχέση με την ακολουθία εντολών του ΕΜ που δόθηκε στην προηγούμενη ενότητα. Η ποσότητα Q στην Μπεϋζιανή προσέγγιση θα διαφέρει κατά κάποια σταθερά $\bar{c}$, που δεν περιέχει p, από την κλασική (3.4) και συνεπώς και στις δύο περιπτώσεις θα μεγιστοποιούνται για την ίδια τιμή p^{k+1}. Επομένως, σε αυτό το παράδειγμα ο αλγόριθμος ΕΜ, για ίδια αρχική τιμή p^0, και στις δύο προσεγγίσεις θα δίνει τα ίδια αποτελέσματα p^k και για οποιαδήποτε τιμή $p^0 \in (0,1)$ θα συγκλίνει στην ποσότητα 0.2314670, το σημείο που βρίσκεται η κορυφή της κατανομής του $p|\mathbf{Y}$.

3.3 Υπερ και Κατά του Αλγορίθμου

Ο ΕΜ είναι μια μέθοδος που εφαρμόζεται ώστε να λύνονται πρακτικά υπολογιστικά προβλήματα. Είναι λογικό σαν μέθοδος να έχει και πλεονεκτήματα και μειονεκτήματα απέναντι στις υπόλοιπες υπολογιστικές μεθόδους.

Πλεονεκτήματα:

- Κάτω από λογικές και όχι πολύ απαιτητικές συνθήκες μπορούμε να διασφαλίσουμε πως ο αλγόριθμος συγκλίνει στην ποσότητα που θέλουμε.
- Είναι συχνά σχετικά εύκολο να υπολογιστούν οι ποσότητες που απαιτούνται για την σύσταση του αλγορίθμου (όπως το Q, (3.5)) και αν όχι υπάρχουν πολλές εναλλατικές μέθοδοι, πέρα από τον άμεσο υπολογισμό, όπου οι ποσότητες αυτές εκτιμούνται αποδοτικά.
- Το υπολογιστικό κόστος του αλγορίθμου είναι συνήθως μικρό.
- Είναι εύκολο να παρακολουθείται η εξέλιξη του αλγορίθμου ως προς την σύγκλιση μέσω διαγραμμάτων, λόγω της μονοτονίας της πιθανοφάνειας.
- Ο αλγόριθμος μπορεί να χρησιμοποιηθεί για να παραχθούν τιμές που προσεγγίζουν τις παρατηρήσεις που δεν καταφέραμε να συλλέξουμε.

Μειονεκτήματα:

- Ο αλγόριθμος δεν συγκλίνει πάντα στην ποσότητα που ψάχνουμε. Στο επόμενο Κεφάλαιο θα δούμε αναλυτικά αυτήν την περίπτωση.
- Δεν είναι πάντα εύκολο να υπολογιστούν οι ποσότητες όπως το Q που απαιτεί ο αλγόριθμος σε κλειστή μορφή.
- Ο αλγόριθμος μπορεί να επιτυγχάνει αργά σύγκλιση, ειδικά σε περιπτώσεις που υπάρχουν πολλά κενά στο πλήρες δείγμα.
- Δεν δίνεται άμεσα από τον αλγόριθμο ο πίνακας διακυμάνσεων-συνδιακυμάνσεων των εκτιμητριών των άγνωστων ποσοτήτων στην κλασική στατιστική.

Κεφάλαιο 4

EM Σύγκλιση

Αυτό το Κεφάλαιο σχετίζεται με την σύγκλιση του αλγορίθμου EM. Στην πρώτη ενότητα αποδεικνύεται η μονοτονικότητα του αλγορίθμου ενώ στην δεύτερη δίνονται παραδείγματα που δεν γίνεται σύγκλιση σε επιθυμητό σημείο.

4.1 Η Μονοτονικότητα του Αλγορίθμου

Οι Dempster, Laird και Rubin (1977) απέδειξαν για την κλασική προσέγγιση του EM πως σε κάθε βήμα του αλγορίθμου γεννιέται μια νέα τιμή για το $\boldsymbol{\vartheta}$, για την οποία η πιθανοφάνεια του ελλιπούς δείγματος γίνεται όλο και μεγαλύτερη. Δηλαδή,

$$L(\boldsymbol{\vartheta}^{k+1}|\mathbf{Y}) \geq L(\boldsymbol{\vartheta}^{k}|\mathbf{Y}) \tag{4.1}$$

για κάθε επανάληψη k. Παρακάτω δίνεται μια παραλλαγή αυτής της απόδειξης για την σύγκλιση στην Μπεϋζιανή στατιστική.

Θα δειχθεί ότι για κάθε νέα τιμή $\boldsymbol{\vartheta}^{k}$ η πυκνότητα πιθανότητα των άγνωστων παραμέτρων $\boldsymbol{\vartheta}$, δοθέντος του ελλιπούς δείγματος $\mathbf{y}$ που παρατηρήθηκε, γίνεται όλο και μεγαλύτερη. Δηλαδή,

$$f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^{k+1}|\mathbf{y}) \geq f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^{k}|\mathbf{y}). \tag{4.2}$$

Ξεκινώντας την απόδειξη, εύκολα φαίνεται πως,

$$f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{w}|\mathbf{y},\boldsymbol{\vartheta}) = \frac{f_{\mathbf{W}|\boldsymbol{\vartheta}}(\mathbf{w}|\boldsymbol{\vartheta})}{f_{\mathbf{Y}|\boldsymbol{\vartheta}}(\mathbf{y}|\boldsymbol{\vartheta})}. \tag{4.3}$$

Τότε,

$$log f_{\mathbf{Y}|\boldsymbol{\vartheta}}(\mathbf{y}|\boldsymbol{\vartheta}) = log f_{\mathbf{W}|\boldsymbol{\vartheta}}(\mathbf{w}|\boldsymbol{\vartheta}) - log f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{w}|\mathbf{y},\boldsymbol{\vartheta}) \tag{4.4}$$

και προσθέτοντας και στις δύο πλευρές της σχέσης (4.4) την ποσότητα «$log f_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}) - log f_{\mathbf{W}}(\mathbf{w}) - log f_{\mathbf{Y}}(\mathbf{y})$»,

$$\begin{aligned} log\frac{f_{\mathbf{Y}|\boldsymbol{\vartheta}}(\mathbf{y}|\boldsymbol{\vartheta}) f_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})}{f_{\mathbf{Y}}(\mathbf{y})} - log f_{\mathbf{W}}(\mathbf{w}) =& log\frac{f_{\mathbf{W}|\boldsymbol{\vartheta}}(\mathbf{w}|\boldsymbol{\vartheta}) f_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})}{f_{\mathbf{W}}(\mathbf{w})} - log f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{w}|\mathbf{y},\boldsymbol{\vartheta}) \\ & - log f_{\mathbf{Y}}(\mathbf{y}) \Rightarrow \end{aligned}$$

$$\begin{aligned} \Rightarrow log f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}|\mathbf{y}) =& log f_{\boldsymbol{\vartheta}|\mathbf{W}}(\boldsymbol{\vartheta}|\mathbf{w}) - log f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{w}|\mathbf{y},\boldsymbol{\vartheta}) \\ & + log\frac{f_{\mathbf{W}}(\mathbf{w})}{f_{\mathbf{Y}}(\mathbf{y})}. \end{aligned} \tag{4.5}$$

Παίρνοντας τώρα στην (4.5) την μέση τιμή ως προς $\mathbf{W}$ δοθέντος του ελλιπούς δείγματος $\mathbf{Y}$ και της παραμέτρου $\boldsymbol{\vartheta}^k$ έχουμε ότι,

$$\begin{aligned} E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}|\mathbf{y})] =& E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log f_{\boldsymbol{\vartheta}|\mathbf{W}}(\boldsymbol{\vartheta}|\mathbf{W})] - E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta})] \\ &+ E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log \frac{f_{\mathbf{W}}(\mathbf{W})}{f_{\mathbf{Y}}(\mathbf{y})}] \Rightarrow \\ \Rightarrow log f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}|\mathbf{y}) =& Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) - H(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) + E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log \frac{f_{\mathbf{W}}(\mathbf{W})}{f_{\mathbf{Y}}(\mathbf{y})}]. \end{aligned} \tag{4.6}$$

Στην (4.6) το Q είναι η κλασική μέση τιμή που υπολογίζεται στον ΕΜ, που φαίνεται και στην σχέση (3.7), και η H ορίζεται ως

$$H(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) := E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta})]. \tag{4.7}$$

Δίνοντας στην (4.6) στην παράμετρο $\boldsymbol{\vartheta}$ τις τιμές $\boldsymbol{\vartheta}^{k+1}$ και $\boldsymbol{\vartheta}^k$ και αφαιρώντας τις δύο εξισώσεις παίρνουμε ότι,

$$\begin{aligned} log f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^{k+1}|\mathbf{y}) - log f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y}) =& Q(\boldsymbol{\vartheta}^{k+1}|\boldsymbol{\vartheta}^k) - H(\boldsymbol{\vartheta}^{k+1}|\boldsymbol{\vartheta}^k) - (Q(\boldsymbol{\vartheta}^k|\boldsymbol{\vartheta}^k) \\ &- H(\boldsymbol{\vartheta}^k|\boldsymbol{\vartheta}^k)) \Rightarrow \\ \Rightarrow log \frac{f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^{k+1}|\mathbf{y})}{f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y})} =& (Q(\boldsymbol{\vartheta}^{k+1}|\boldsymbol{\vartheta}^k) - Q(\boldsymbol{\vartheta}^k|\boldsymbol{\vartheta}^k)) \\ &- (H(\boldsymbol{\vartheta}^{k+1}|\boldsymbol{\vartheta}^k) - H(\boldsymbol{\vartheta}^k|\boldsymbol{\vartheta}^k)). \end{aligned} \tag{4.8}$$

Άρα για να αποδειχθεί η μονοτονία του αλγορίθμου αρκεί να δειχθεί ότι το δεύτερο μέλος της ισότητας (4.8) είναι θετικό ή μηδέν για κάθε k στους φυσικούς αριθμούς.

Το $\boldsymbol{\vartheta}^{k+1}$ είναι η τιμή που μεγιστοποιεί την $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k)$ κατά τον ΕΜ. Επομένως από τον τρόπο με τον οποίο υπολογίζεται το $\boldsymbol{\vartheta}^{k+1}$ στον ΕΜ, η διαφορά των Q στην (4.8) είναι πάντα θετική ή μηδέν. Ακόμα, η διαφορά των Q παραμένει θετική ή μηδέν και κάτω από τον αλγόριθμο GEM.

Ως προς την διαφορά των H, ονομάζεται *Relative Entropy* και για κάθε $\boldsymbol{\vartheta}$ στον παραμετρικό χώρο παραμένει θετική,

$$\begin{aligned} &H(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) - H(\boldsymbol{\vartheta}^k|\boldsymbol{\vartheta}^k) = \\ &= E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta})] - E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta}^k)] \\ &= E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log \frac{f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta})}{f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta}^k)}] \\ &\leq log E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[\frac{f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta})}{f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{W}|\mathbf{y},\boldsymbol{\vartheta}^k)}] \\ &= log \int \frac{f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{w}|\mathbf{y},\boldsymbol{\vartheta})}{f_{\mathbf{W}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{w}|\mathbf{y},\boldsymbol{\vartheta}^k)} f_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{z}|\mathbf{y},\boldsymbol{\vartheta}^k) d\mathbf{z} \\ &= log \int \frac{f_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{z}|\mathbf{y},\boldsymbol{\vartheta})}{f_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{z}|\mathbf{y},\boldsymbol{\vartheta}^k)} f_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{z}|\mathbf{y},\boldsymbol{\vartheta}^k) d\mathbf{z} \\ &= log \int f_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}}(\mathbf{z}|\mathbf{y},\boldsymbol{\vartheta}) d\mathbf{z} = log 1 = 0. \end{aligned} \tag{4.9}$$

Ο λογάριθμος είναι κυρτή συνάρτηση, επομένως η ανισότητα που φαίνεται επάνω προκύπτει από την ανισότητα του *Jensen*. Ακόμα στην τελευταία σειρά των πράξεων ολοκληρώνουμε μια πυκνότητα πιθανότητας στο πεδίο ορισμού της. Αυτό το ολοκλήρωμα προφανώς είναι ίσο με 1. Συνεπώς από τα παραπάνω η (4.9) είναι μικρότερη ή ίση με μηδέν και από την (4.8) έπεται η μονοτονία του αλγορίθμου ΕΜ.

Αντίστοιχα έπεται και η μονοτονία του αλγορίθμου GEM. Η μόνη διαφορά των δύο αλγορίθμων, όπως έχει ειπωθεί και στο Κεφάλαιο 3, είναι πως στον GEM η $\boldsymbol{\vartheta}^{k+1}$ δεν μεγιστοποιεί την $Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k)$, αλλά βρίσκουμε ένα $\boldsymbol{\vartheta}^{k+1}$ για το οποίο ισχύει,

$$Q(\boldsymbol{\vartheta}^{k+1}|\boldsymbol{\vartheta}^k) \geq Q(\boldsymbol{\vartheta}^k|\boldsymbol{\vartheta}^k). \tag{4.10}$$

Είναι εμφανές πως στην απόδειξη για την μονοτονία του GEM δεν θα άλλαζε κάτι.

Επομένως η λογική του αλγορίθμου ΕΜ στην Μπεϋζιανή στατιστική είναι να γεννά τιμές που κάνουν όλο και μεγαλύτερη την πυκνότητα πιθανότητας της εκ των υστέρων κατανομής των παραμέτρων δοθέντος του ελλιπούς δείγματος. Επιπλέον αναφέρεται πως ο αλγόριθμος γενικά συγκλίνει σε στάσιμα σημεία της κατανομής του $\boldsymbol{\vartheta}|\mathbf{Y} = \mathbf{y}$.

Είναι λογικό πως αν η εκ των υστέρων είναι μονοκόρυφη και κοίλη τότε κάθε τιμή $\boldsymbol{\vartheta}^k$ θα είναι όλο και πιο κοντά στο $\boldsymbol{\vartheta}^*$ στο οποίο βρίσκεται η κορυφή της κατανομής και μόλις πέσει επάνω στο σημείο, θα συγκλίνει εκεί. Συνεπώς, κάτω από συνθήκες, μέσω του αλγορίθμου επιτυγχάνεται σύγκλιση στο σημείο $\boldsymbol{\vartheta}^*$ στο οποίο βρίσκεται το ολικό μέγιστο της ύστερης πυκνότητας πιθανότητας (*maximum a posteriori*).

Ο αλγόριθμος ΕΜ χρησιμοποιείται με σκοπό να βρεθεί το σημείο που μεγιστοποιείται η πυκνότητα πιθανότητας και όχι ένα οποιοδήποτε στάσιμο σημείο της κατανομής. Αν υπάρχουν παραπάνω από μία κορυφές στην εκ των υστέρων θα φανεί πως είναι πολύ πιθανό να υπάρξει άγνοια σχετικά με το σε ποια κορυφή συγκλίνει ο αλγόριθμος. Το σε ποια κορυφή θα συγκλίνει ο αλγόριθμος εξαρτάται από την αρχική τιμή $\boldsymbol{\vartheta}^0$ που θα δοθεί. Έτσι ο αλγόριθμος μπορεί να συγκλίνει σε τοπικό μέγιστο, ενώ υπάρχει ολικό, που είναι και αυτό που ψάχνουμε.

Συνοψίζοντας, κατά την εκτέλεση του αλγορίθμου, παρόλο που υπάρχει η μονοτονία, είναι πιθανό να εμφανιστούν κάποια προβλήματα ως προς τη σύγκλιση του $\boldsymbol{\vartheta}$. Πιθανά προβλήματα είναι ο αλγόριθμος να συγκλίνει είτε σε $\boldsymbol{\vartheta}^*$ στο οποίο εμφανίζεται τοπικό μέγιστο και όχι ολικό, είτε σε $\boldsymbol{\vartheta}^*$ στο οποίο εμφανίζεται τοπικό ελάχιστο, είτε σε $\boldsymbol{\vartheta}^*$ στο οποίο εμφανίζεται σαγματικό σημείο της κατανομής. Το σαγματικό σημείο (saddle point) είναι ένα σημείο το οποίο εμφανίζεται σε τρισδιάστατες συναρτήσεις και είναι στάσιμο σημείο (σχέση (4.11)) αλλά δεν είναι τοπικό ακρότατο και για τους δύο άξονες των παραμέτρων ταυτόχρονα. Το σημείο της κατανομής που εμφανίζεται στις συντεταγμένες $(v, c) = (20/9, 0)$ στο Διάγραμμα 4.3 αποτελεί σαγματικό σημείο. Στον αλγόριθμο GEM έχει παρατηρηθεί και το φαινόμενο να μην επιτυγχάνεται σύγκλιση των $\boldsymbol{\vartheta}^k$ παρόλο που η κατανομή έχει ολικό μέγιστο και η $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y})$ συγκλίνει κάπου. Αυτά τα προβλήματα θα συζητηθούν στις επόμενες υποενότητες του Κεφαλαίου.

4.2 Περιπτώσεις μη επιθυμητής σύγκλισης

Στην προηγούμενη ενότητα δείχθηκε πως η ακολουθία των τιμών $\boldsymbol{\vartheta}^k$ που γεννάει ο ΕΜ κάνει όλο και μεγαλύτερη την πυκνότητα πιθανότητα $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y})$. Αν η $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}|\mathbf{y})$ είναι άνω φραγμένη ως προς $\boldsymbol{\vartheta}$, τότε για οποιαδήποτε ακολουθία $\boldsymbol{\vartheta}^k$ που παράγεται από τον ΕΜ, η ακολουθία $\{f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y})\}_{k\in\mathbb{N}}$, αφού είναι αύξουσα ως προς k, θα συγκλίνει σε κάποια τιμή $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^*|\mathbf{y})$. Το $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^*|\mathbf{y})$ σχεδόν πάντα είναι στάσιμη τιμή και αυτό σημαίνει πως,

$$\frac{df_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}|\mathbf{y})}{d\boldsymbol{\vartheta}}|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^*} = \mathbf{0}_r. \tag{4.11}$$

Λόγω της μονοτονίας, θα έπρεπε η τιμή $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^*|\mathbf{y})$ να είναι ολικό μέγιστο της πυκνότητας πιθανότητας. Αυτό όμως δεν συμβαίνει πάντα. Όπως αναφέρθηκε και επάνω, το $f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^*|\mathbf{y})$ μπορεί να είναι τοπικό ακρότατο ή σαγματικό σημείο. Το που θα συγκλίνει η $\{f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y})\}_{k\in\mathbb{N}}$ εξαρτάται από το $\boldsymbol{\vartheta}^0$ που δίνεται. Ο ΕΜ έχει το πλεονέκτημα πως αν μετατοπίσουμε κατάλληλα το $\boldsymbol{\vartheta}^0$ μπορούμε να πέσουμε πάνω στο σημείο σύγκλισης που ψάχνουμε, επομένως επιτρέπει δυναμική αναζήτηση της κορυφής.

Για πρακτικές περιπτώσεις, όταν χρησιμοποιείται ο ΕΜ σε θέματα εφαρμοσμένης στατιστικής, αν η εκ των υστέρων πυκνότητα πιθανότητα για την οποία ψάχνουμε ολικό μέγιστο είναι μονοκόρυφη και δεν έχει άλλα στάσιμα σημεία, για οποιοδήποτε $\boldsymbol{\vartheta}^0$ και αν ξεκινήσουμε ο ΕΜ θα συγκλίνει στο σημείο που ψάχνουμε. Στις επόμενες υποενότητες δίνονται παραδείγματα μη επιθυμητής σύγκλισης του ΕΜ.

4.2.1 Σύγκλιση σε τοπικό ακρότατο

Το παρακάτω παράδειγμα βασίζεται σε αυτό που παρουσίασαν οι Arslan, Constable και Kent (1993) για την σύγκλιση του αλγορίθμου ΕΜ σε τοπικό ακρότατο.

Έστω $\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$, όπου Z_i είναι ανεξάρτητα και ισόνομα ανά δύο και

$$Z_i \sim G(\frac{\nu}{2}, \frac{\nu}{2}).$$

Ακόμα $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, όπου $Y_i|Z_i = z_i, \mu$ επίσης ανεξάρτητα και ισόνομα ανά δύο και

$$Y_i|Z_i = z_i, \mu \sim N(\mu, \frac{1}{z_i}).$$

Το μ, ανεξάρτητο του $\mathbf{Z}$, ακολουθεί την καταχρηστική πρότερη $f_\mu(\mu) \propto 1,\ \mu \in \mathbb{R}$. Για την κατανομή των $Y_i|\mu$, από το Θεώρημα Ολικής Πιθανότητας,

$$\begin{aligned} f_{Y_i|\mu}(y_i|\mu) = \int_0^\infty f_{Y_i|Z_i,\mu}(y_i|z,\mu) f_{Z_i}(z)dz &= \int_0^{+\infty} \frac{\sqrt{z}}{\sqrt{2\pi}} e^{-z\frac{(y_i-\mu)^2}{2}} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} z^{\frac{\nu}{2}-1}}{\Gamma(\frac{\nu}{2})} e^{-\frac{\nu}{2}z} dz \\ &= \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\sqrt{2\pi}\Gamma(\frac{\nu}{2})} \int_0^{+\infty} z^{\frac{\nu}{2}-\frac{1}{2}} e^{-(\frac{(y_i-\mu)^2}{2}+\frac{\nu}{2})z} dz \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} (1 + \frac{(y_i-\mu)^2}{\nu})^{-\frac{\nu+1}{2}}, y_i \in \mathbb{R}. \end{aligned} \tag{4.12}$$

Από την (4.12) εύκολα φαίνεται πως

$$Y_i|\mu \sim t_\nu(\mu),\ \forall i \in \{1, 2, ..., n\},$$

όπου $t_\nu(\mu)$ η μετατοπισμένη κατανομή *student* με ν βαθμούς ελευθερίας και μέση τιμή μ. Λόγω της πρότερης κατανομής του μ, ισχύει πως

$$\mu|Y_i = y_i \sim t_\nu(y_i),\ \forall i \in \{1, 2, ..., n\}.$$

Επιπλέον, ισχύει πως για $\forall i \neq j,\ Y_i,\ Z_j$ ανεξάρτητα, και από όλα τα παραπάνω έπεται πως $Y_i|\mu$ επίσης ανεξάρτητα ανά δύο. Για την προηγούμενη πρόταση σχετικά με την ανεξαρτησία των $Y_i|\mu$, για αυθαίρετα $i \neq j$,

$$\begin{aligned} f_{Y_i,Y_j|Z_i,Z_j,\mu}(y_i,y_j|z_i,z_j,\mu) &= f_{Y_i|Y_j,Z_i,Z_j\mu}(y_i|y_j,z_i,z_j,\mu) f_{Y_j|Z_i,Z_j,\mu}(y_j|z_i,z_j,\mu) \Rightarrow \\ f_{Y_i,Y_j|Z_i,Z_j,\mu}(y_i,y_j|z_i,z_j,\mu) &= f_{Y_i|Z_i,\mu}(y_i|z_i,\mu) f_{Y_j|Z_j,\mu}(y_j|z_j,\mu) \Rightarrow \\ f_{Y_i,Y_j|Z_i,Z_j,\mu}(y_i,y_j|z_i,z_j,\mu) f_{Z_i,Z_j}(z_i,z_j) &= f_{Y_i|Z_i,\mu}(y_i|z_i,\mu) f_{Y_j|Z_j,\mu}(y_j|z_j,\mu) f_{Z_i,Z_j}(z_i,z_j) \Rightarrow \\ f_{Y_i,Y_j,Z_i,Z_j|\mu}(y_i,y_j,z_i,z_j|\mu) &= f_{Y_i|Z_i,\mu}(y_i|z_i,\mu) f_{Y_j|Z_j,\mu}(y_j|z_j,\mu) f_{Z_i}(z_i) f_{Z_j}(z_j) \Rightarrow \\ f_{Y_i,Y_j,Z_i,Z_j|\mu}(y_i,y_j,z_i,z_j|\mu) &= f_{Y_i,Z_i|\mu}(y_i,z_i|\mu) f_{Y_j,Z_j|\mu}(y_j,z_j|\mu) \Rightarrow \\ \int\int f_{Y_i,Y_j,Z_i,Z_j|\mu}(y_i,y_j,z_i,z_j|\mu)dz_i dz_j &= \int\int f_{Y_i,Z_i|\mu}(y_i,z_i|\mu) f_{Y_j,Z_j|\mu}(y_j,z_j|\mu)dz_i dz_j \Rightarrow \\ \int\int f_{Y_i,Y_j,Z_i,Z_j|\mu}(y_i,y_j,z_i,z_j|\mu)dz_i dz_j &= \int f_{Y_i,Z_i|\mu}(y_i,z_i|\mu)[\int f_{Y_j,Z_j|\mu}(y_j,z_j|\mu)dz_j]dz_i \Rightarrow \\ f_{Y_i,Y_j|\mu}(y_i,y_j|\mu) &= f_{Y_i|\mu}(y_i|\mu) f_{Y_j|\mu}(y_j|\mu). \end{aligned}$$

Συνεπώς,

$$\begin{aligned} f_{\mu|\mathbf{Y}}(\mu|\mathbf{y}) &\propto f_{\mathbf{Y}|\mu}(\mathbf{y}|\mu) \\ &\propto \prod_{i=1}^{n} f_{Y_i|\mu}(y_i|\mu). \end{aligned} \tag{4.13}$$

Η κατανομή του $\mu|\mathbf{Y}$ για μικρές τιμές των βαθμών ελευθερίας ν έχει παραπάνω από ένα τοπικά ακρότατα, όπως φαίνεται και στο Διάγραμμα 4.1.

Θεωρώ πλήρες δείγμα το $\mathbf{W} = (Y_1, Y_2, ..., Y_n, Z_1, Z_2, ..., Z_n)$, διάστασης $m = 2n$, και ελλιπές το $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, διάστασης n. Θα χρησιμοποιηθεί ο αλγόριθμος EM για την εύρεση της κορυφής της τυχαίας μεταβλητής $\mu|\mathbf{Y}$.

Για το $E-step$ πρέπει να υπολογιστεί η ποσότητα Q. Θα χρησιμοποιηθεί ο τύπος της σχέσης (3.8),

$$\begin{aligned} f_{\mathbf{W}|\mu}(\mathbf{w}|\mu) &= f_{\mathbf{W}|\mathbf{Z},\mu}(\mathbf{w}|\mathbf{z},\mu) f_{\mathbf{Z}}(\mathbf{z}) \\ &= f_{\mathbf{Y}|\mathbf{Z},\mu}(\mathbf{y}|\mathbf{z},\mu) f_{\mathbf{Z}}(\mathbf{z}) \\ &= \prod_{i=1}^{n} f_{Y_i|Z_i,\mu}(y_i|z_i,\mu) f_{Z_i}(z_i) \\ &= \prod_{i=1}^{n} f_{Y_i|Z_i,\mu}(y_i|z_i,\mu) \prod_{i=1}^{n} f_{Z_i}(z_i) \\ &= \frac{\prod_{i=1}^{n}\sqrt{z_i}}{\sqrt{2\pi}} e^{-\sum_1^n z_i \frac{(y_i-\mu)^2}{2}} (\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})})^n (\prod_{i=1}^{n} z_i)^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}\sum_{i=1}^n z_i} \\ &= C \prod_{i=1}^{n} (z_i)^{\frac{\nu}{2}-\frac{1}{2}} e^{-\sum_{i=1}^n (\frac{\nu}{2}+\frac{(y_i-\mu)^2}{2}) z_i}. \end{aligned} \tag{4.14}$$

Με C συμβολίζεται κομμάτι του γινομένου που δεν περιέχει μ. Άρα για τον υπολογισμό του Q,

$$\begin{aligned} Q(\mu|\mu^k) &= E_{\mathbf{Z}|\mathbf{Y},\mu^k}[log(f_{\mathbf{W}|\mu}(\mathbf{W}|\mu))] + log1 + \bar{c} \\ &= E_{\mathbf{Z}|\mathbf{Y},\mu^k}[(\frac{\nu}{2}-\frac{1}{2})\sum_{i=1}^{n} logZ_i - \sum_{i=1}^{n}(\frac{\nu}{2}+\frac{(y_i-\mu)^2}{2})Z_i] + \bar{c}' \\ &= (\frac{\nu}{2}-\frac{1}{2})\sum_{i=1}^{n} E[logZ_i|Y_i=y_i,\mu^k] - \sum_{i=1}^{n}(\frac{\nu}{2}+\frac{(y_i-\mu)^2}{2}) E[Z_i|Y_i=y_i,\mu^k] + \bar{c}' \\ &= -\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2} E[Z_i|Y_i=y_i,\mu^k] + \bar{c}''. \end{aligned} \tag{4.15}$$

Με τα $\bar{c}$, $\bar{c}'$ και $\bar{c}''$ εννοούνται αθροιστικοί όροι του $Q(\mu|\mu^k)$ που δεν περιέχουν μ και παραγωγίζοντας για να βρεθεί το σημείο που μεγιστοποιείται, κατά το $M-step$, θα μηδενιστούν (οπότε δεν υπάρχει λόγος και να υπολογιστούν).

Αν κανείς χρησιμοποιήσει την ιδιότητα,

$$f_{Z_i|Y_i,\mu}(z_i|y_i,\mu) = \frac{f_{Y_i|Z_i,\mu}(y_i|z_i,\mu) f_{Z_i|\mu}(z_i|\mu)}{f_{Y_i|\mu}(y_i|\mu)}, \tag{4.16}$$

οι επιμέρους πυκνότητες πιθανότητας είναι γνωστές και εύκολα φαίνεται πως η πυκνότητα πιθανότητας που προκύπτει για την $Z_i|Y_i=y_i,\mu$ είναι αυτή της κατανομής γάμμα με παραμέτρους $\frac{\nu+1}{2}$ και $\frac{\nu+(y_i-\mu)^2}{2}$.

Άρα τελικά από την (4.15) και την κατανομή της $Z_i|Y_i, \mu$,

$$Q(\mu|\mu^k) = \sum_{i=1}^{n} \frac{(y_i-\mu)^2}{2} \frac{(\nu+1)/2}{(\nu+(y_i-\mu^k)^2)/2} + \bar{c}''$$
$$= \sum_{i=1}^{n} \frac{(y_i-\mu)^2}{2} \frac{\nu+1}{\nu+(y_i-\mu^k)^2} + \bar{c}''. \tag{4.17}$$

Παραγωγίζοντας ως προς μ, εύκολα φαίνεται ότι το Q μεγιστοποιείται για

$$\mu^{k+1} = \frac{\sum_{i=1}^{n} u_i^k y_i}{\sum_{i=1}^{n} u_i^k} \tag{4.18}$$

όπου

$$u_i^k = \frac{\nu+1}{\nu+(y_i-\mu^k)^2}. \tag{4.19}$$

Επομένως αλγοριθμικά ο ΕΜ θα δούλευε ως εξής,
-του δίνεται αρχική τιμή μ^0,
-υπολογίζει τις τιμές u_i^k και μ^{k+1} από τους τύπους (4.18) και (4.19),
-επαναλαμβάνει το προηγούμενο βήμα μέχρι την σύγκλιση.

Αν τώρα δοθεί δείγμα 4 παρατηρήσεων $\mathbf{y} = (1, -15, 1, 2)$ και ν=0.01, μπορεί να βρεθεί εύκολα το διάγραμμα της εκ των υστέρων πυκνότητας πιθανότητας $f_{\mu|\mathbf{Y}}(\mu|\mathbf{y})$ ως προς μ και τα σημεία που παρατηρούνται ακρότατα.

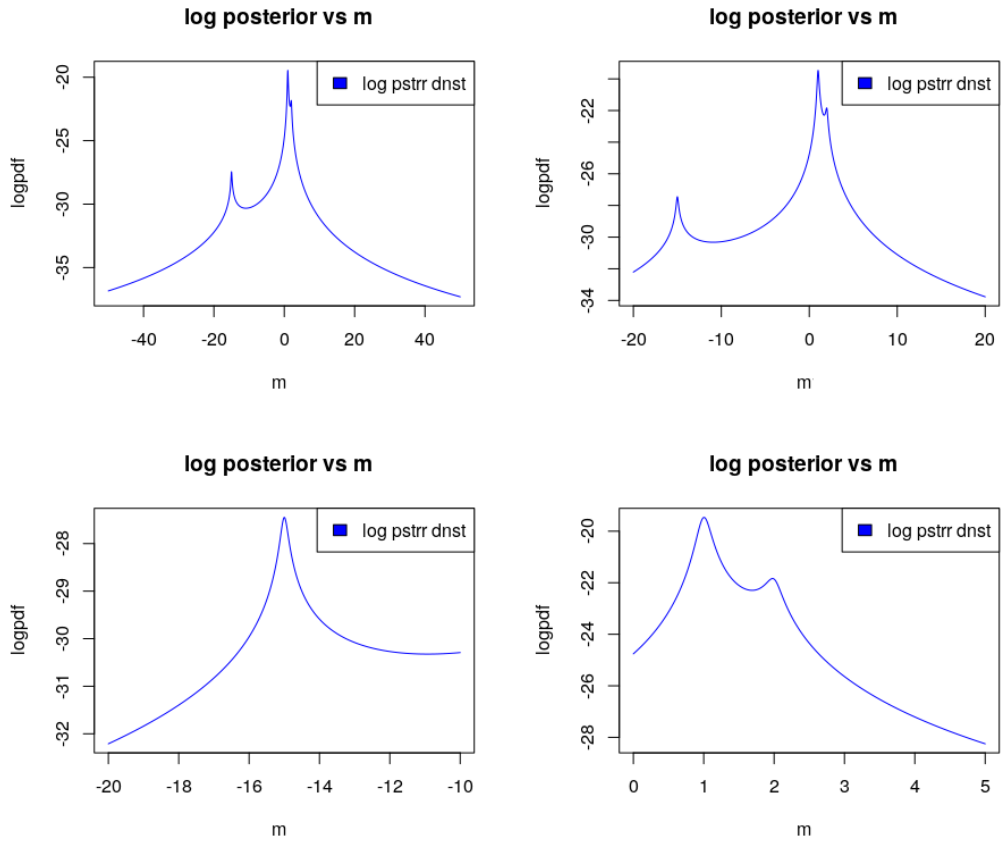
Μετά από μια σύντομη ανάλυση της $f_{\mu|\mathbf{Y}}(\mu|\mathbf{y})$ ως προς μ, φαίνεται πως έχει τοπικά ελάχιστα στα σημεία μ= -10.9227 και 1.68802 και τοπικά μέγιστα στα σημεία στα σημεία μ=-14.9982 και 1.97819. Το ολικό μέγιστο της είναι στο σημείο $\mu = 1.00467$.

Στο Διάγραμμα 4.1 φαίνεται το διάγραμμα της λογαριθμημένης εκ των υστέρων πυκνότητας πιθανότητας $logf_{\mu|\mathbf{Y}}(\mu|\mathbf{y})$ (μείον κάποια σταθερά) ως προς μ, σε διαφορετικά διαστήματα για το μ ανά τις επιμέρους εικόνες. Στις δύο πάνω εικόνες φαίνεται η συνάρτηση γενικότερα, ενώ στις υπόλοιπες γίνεται εστίαση στις περιοχές που εμφανίζονται ακρότατα. Η λογαριθμική συνάρτηση χρησιμοποιείται για να είναι περισσότερο εμφανή τα τοπικά ακρότατα στα διαγράμματα. Η λογαριθμική συνάρτηση είναι γνησίως αύξουσα και συνεπώς η απεικόνισή της ένα προς ένα, άρα στα ίδια σημεία θα εμφανίζονται τα αντίστοιχα ακρότατα και στην $f_{\mu|\mathbf{Y}}(\mu|\mathbf{y})$ σε άλλη κλίμακα (ομοίως και μείον της σταθεράς). Οι δύο αυτοί μετασχηματισμοί αλλάζουν μόνο την κλίμακα της κατανομής στον άξονα των $y'y$ και όχι το σχήμα της.

Αν κανείς ξεκινήσει να εκτελεί τον ΕΜ, για την περίπτωση του παραδείγματος, με αρχική τιμή το $\mu^0 = -10.9227$, φαίνεται πως ο αλγόριθμος «κολλάει» στο στάσιμο σημείο $\mu^* = -10.9227$, το οποίο είναι τοπικό ελάχιστο. Το ίδιο θα συνέβαινε και αν ο αλγόριθμος είχε αρχική τιμή μ^0 με πιο χαμηλή τιμή πυκνότητας πιθανότητας από το τοπικό ελάχιστο και καθώς εκτελούνταν οι επαναλήψεις του αλγορίθμου, οι τιμές μ^k έπεφταν τυχαία στο -10.9227. Βέβαια, με κατάλληλη απομάκρυνση ε>0 από το σημείο, ο αλγόριθμος συνεχίζει την πορεία του προς μέγιστο.

Αν ο αλγόριθμος εκτελεστεί χωρίς συνθήκη σύγκλισης, απλά εκτελώντας τον βρόγχο για μεγάλο πλήθος επαναλήψεων, λόγω της περικομμένης τιμής -10.9227, η οποία απέχει μικρή απόσταση $\delta > 0$ από το σημείο που εμφανίζεται το τοπικό ελάχιστο, το μ^k μετά από αρκετές επαναλήψεις θα ξεπεράσει το στάσιμο σημείο -10.9227 και θα συνεχίσει να παίρνει τιμές μέχρι το στάσιμο σημείο -14.9982, το οποίο είναι τοπικό μέγιστο. Οι συνθήκες σύγκλισης, όμως, για τις οποίες σταματάει ο αλγόριθμος ΕΜ, που δόθηκαν στο Κεφάλαιο 3, θα σταματούσαν τον αλγόριθμο στο τοπικό ελάχιστο. Στο Διάγραμμα 4.2 φαίνεται η εξέλιξη του βρόγχου για προκαθορισμένες 100 επαναλήψεις. Μέσα από το Διάγραμμα γίνεται αισθητό το τι συμβαίνει στον αλγόριθμο αν «κολλήσει»

Διάγραμμα 4.1: Διαγράμματα της $logf_{\mu|\mathbf{Y}}(\mu|\mathbf{y})$ ως προς μ για $\mathbf{y} = (1, -15, 1, 2)$.



σε τοπικό ελάχιστο. Στην αριστερή εικόνα φαίνεται η εξέλιξη του μ^k σε κάθε επανάληψη k, ενώ στην δεξία τα σημεία (μ^k,μ^{k+1}).

Στο Διάγραμμα 4.2, στην αριστερή εικόνα θα έπρεπε να εμφανίζονται 100 τελείες, γιατί τόσα σημεία δόθηκαν. Ο λόγος για τον οποίο φαίνονται το πολύ 12 είναι πως η πλειοψηφία των σημείων βρίσκεται στις θέσεις (-10.9227,-10.9227) και (-14.9982,-14.9982), λόγω των συγκλίσεων.

Αν κανείς ξεκινήσει να εκτελεί τον EM για την περίπτωση του παραδείγματος με αρχική τιμή το $\mu^0 = -15.7444$ το αποτέλεσμα που θα πάρει είναι το στάσιμο σημείο του αλγορίθμου $\mu^* = -14.9982$ το οποίο είναι τοπικό και όχι ολικό μέγιστο.

Επιπλέον, ιδιαίτερο ενδιαφέρον παρουσιάζει το τι συμβαίνει όταν το μ^0 τείνει στο $\pm\infty$. Τότε στο συγκεκριμένο παράδειγμα,

$$u_i^0 \approx \frac{\nu + 1}{\nu + (\mu^0)^2} \tag{4.20}$$

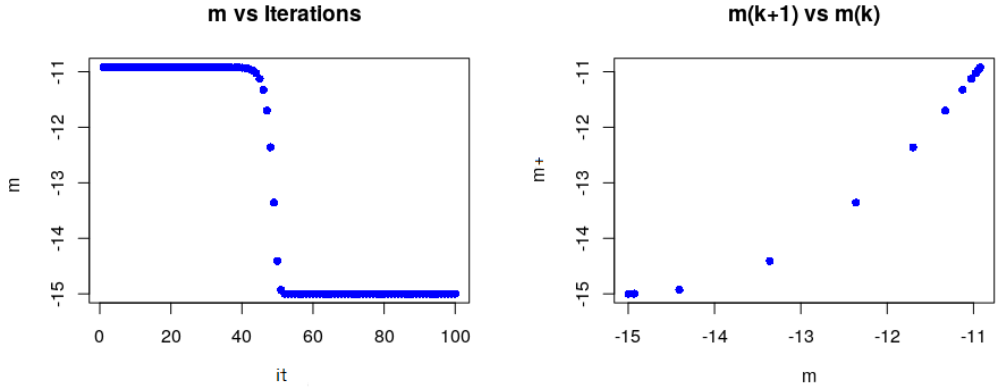
και

$$\mu^1 \approx \sum_{i=1}^{n} \frac{z_i}{n} = \bar{z}. \tag{4.21}$$

Έτσι, όταν δοθεί στον EM, για την συγκεκριμένη περίπτωση, αρχική τιμή μ^0 η οποία είναι πολύ μεγάλη κατά απόλυτη τιμή, στην πρώτη επανάληψη το μ^1 θα μετατοπιστεί πολύ κοντά στη μέση τιμή $\bar{y}$. Αν δοθεί αρχική τιμή $\mu^0 = -1000000$ για παράδειγμα, στην πρώτη επανάληψη του βρόγχου το μ^1 γίνεται -2.7501.

Η *student* με πολύ μικρό βαθμό ελευθερίας είναι μια κατανομή η οποία έχει σπάνια κάποια πρακτική εφαρμογή όπως έθεσαν και οι Arslan, Constable και Kent. Παρόλα αυτά, το παράδειγμα είναι χρήσιμο στο να καταδείξει πιθανά προβλήματα που μπορεί να εμφανιστούν κατά την εκτέλεση του EM.

Διάγραμμα 4.2: Διαγράμματα για την εξέλιξη του αλγορίθμου για αρχική τιμή -10.9222.



4.2.2 Σύγκλιση σε σαγματικό σημείο

Το επόμενο παράδειγμα βασίζεται στο αυτό που έδωσε ο *Murray* (1977). Έστω ότι έχουμε το παρακάτω ζευγαρωτό δείγμα $\mathbf{w}$ (Πίνακας 4.1) με m=24 παρατηρήσεις.

Πίνακας 4.1: Δείγμα $\mathbf{w}$.

w_1	-1	1	-1	1	-2	-2	2	2	;	;	;	;
w_2	1	-1	-1	1	;	;	;	;	-2	2	2	-2

Ισχύει πως κάθε ζεύγος $\mathbf{W}_j|v,c = (W_{1j}, W_{2j})|v,c$ προέρχεται από δισδιάστατη κανονική κατανομή $N_2(\mathbf{0}_2, \Sigma)$, όπου $\Sigma = \begin{bmatrix} v & c \\ c & v \end{bmatrix}$. Τα $\mathbf{W}_j|v,c$ είναι ανεξάρτητα και ισόνομα ανά δύο. Τότε, οι τυχαίες μεταβλητές $W_{ij}|v,c$ ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διακύμανση v. Για σταθερό i, έπεται πως τα $W_{ij}|v,c$ είναι ανεξάρτητα ανά δύο για διαφορετικό j και για σταθερό j, οι $W_{1j}|v,c$, $W_{2j}|v,c$ έχουν συσχέτιση c. Τα v,c είναι ανεξάρτητες τυχαίες μεταβλητές με πρότερες καταχρηστικές κατανομές τις,

$$f_v(v) \propto \frac{1}{v}, \quad v \in \mathbb{R}_+, \tag{4.22}$$

$$f_c(c) \propto 1, \quad c \in \mathbb{R}. \tag{4.23}$$

Θεωρώ πλήρες δείγμα $\mathbf{W}$ τις 24 παρατηρήσεις, ελλιπές δείγμα $\mathbf{Y}$ τις 16 παρατηρήσεις που καταφέραμε να συλλέξουμε και $\mathbf{Z}$ τις 8 παρατηρήσεις δεν καταφέραμε να συλλέξουμε. Θα γίνει μελέτη για το που συγκλίνει ο αλγόριθμος ΕΜ στην πυκνότητα πιθανότητας της $v,c|\mathbf{Y}$ για τις παραπάνω πληροφορίες. Αρχικά θα υπολογιστεί η ποσότητα Q.

Η δισδιάστατη κανονική κατανομή με μέση τιμή ίση με (0,0) και πίνακα διακυμάνσεων-συνδιακυμάνσεων $\Sigma = \begin{bmatrix} v & c \\ c & v \end{bmatrix}$ έχει πυκνότητα πιθανότητας,

$$f_{N_2(\mathbf{0}_2,\Sigma)}(x,y) = \frac{1}{2\pi\sqrt{v^2-c^2}} e^{-\frac{vx^2-2cxy+vy^2}{2(v^2-c^2)}}, x,y \in \mathbb{R}. \tag{4.24}$$

Άρα η λογαριθμημένη πυκνότητα πιθανότητας του πλήρους δείγματος $\mathbf{W}|v,c$ θα είναι η,

$$\begin{aligned} logf_{\mathbf{W}|v,c}(\mathbf{w}|v,c) =& -12log(2\pi) - 6log(v^2-c^2) - \frac{\sum_{j=1}^{12}(vw_{1j}^2 - 2cw_{1j}w_{2j} + vw_{2j}^2)}{2(v^2-c^2)} \\ =& -12log(2\pi) - 6log(v^2-c^2) \\ & - \frac{40v + (\sum_{j=5}^{8} w_{2j}^2 + \sum_{j=9}^{12} w_{1j}^2)v}{2(v^2-c^2)} \\ & - \frac{4c(w_{2,5} + w_{2,6} - w_{2,7} - w_{2,8} + w_{1,9} - w_{1,10} - w_{1,11} + w_{1,12})}{2(v^2-c^2)}, \end{aligned} \tag{4.25}$$

και για $\boldsymbol{\vartheta} = (v,c)$,

$$\begin{aligned} Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) =& E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[logf_{v,c|\mathbf{W}}(v,c|\mathbf{W})] = E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[logf_{\mathbf{W}|v,c,}(\mathbf{W}|v,c)] + log1 + log\frac{1}{v} + \bar{c} \\ =& -12log(2\pi) - 6log(v^2-c^2) - log(v) + \bar{c} \\ & - \frac{40v + (\sum_{j=5}^{8} E[W_{2j}^2|W_{1j}=w_{1j},\boldsymbol{\vartheta}^k] + \sum_{j=9}^{12} E[W_{1j}^2|W_{2j}=w_{2j},\boldsymbol{\vartheta}^k])v}{2(v^2-c^2)} \\ & - \frac{c\sum_{j=1}^{12} E[W_{1j}W_{2j}|\mathbf{Y}=\mathbf{y},\boldsymbol{\vartheta}^k]}{v^2-c^2}. \end{aligned} \tag{4.26}$$

Από την (4.26) για να υπολογιστεί το Q μένει να υπολογιστούν οι μέσες τιμές,

$$E[W_{ij}|W_{i'j} = w_{i'j}, \boldsymbol{\vartheta}^k]$$

και

$$E[W_{ij}^2|W_{i'j} = w_{i'j}, \boldsymbol{\vartheta}^k],$$

όπου $i, i' \in \{1,2\}$ και $i \neq i'$. Αν (X,Y) ακολουθούν την $N_2(\mathbf{0}_2, \bar{\Sigma})$, όπου $\bar{\Sigma} = \begin{bmatrix} \sigma_X^2 & \bar{c} \\ \bar{c} & \sigma_Y^2 \end{bmatrix}$, ισχύουν οι γνωστές ιδιότητες (Κούτρας, 2012) (4.27) και (4.28) της δισδιάστατης κανονικής κατανομής,

$$E[Y|X=x] = \frac{\bar{c}}{\sigma_X^2}x, \tag{4.27}$$

$$V[Y|X=x] = \sigma_Y^2 - \frac{\bar{c}^2}{\sigma_X^2}. \tag{4.28}$$

Από αυτές τις ιδιότητες, προκύπτουν άμεσα για το παράδειγμα οι μέσες τιμές που χρειάζεται να υπολογιστούν για τον υπολογισμό του Q,

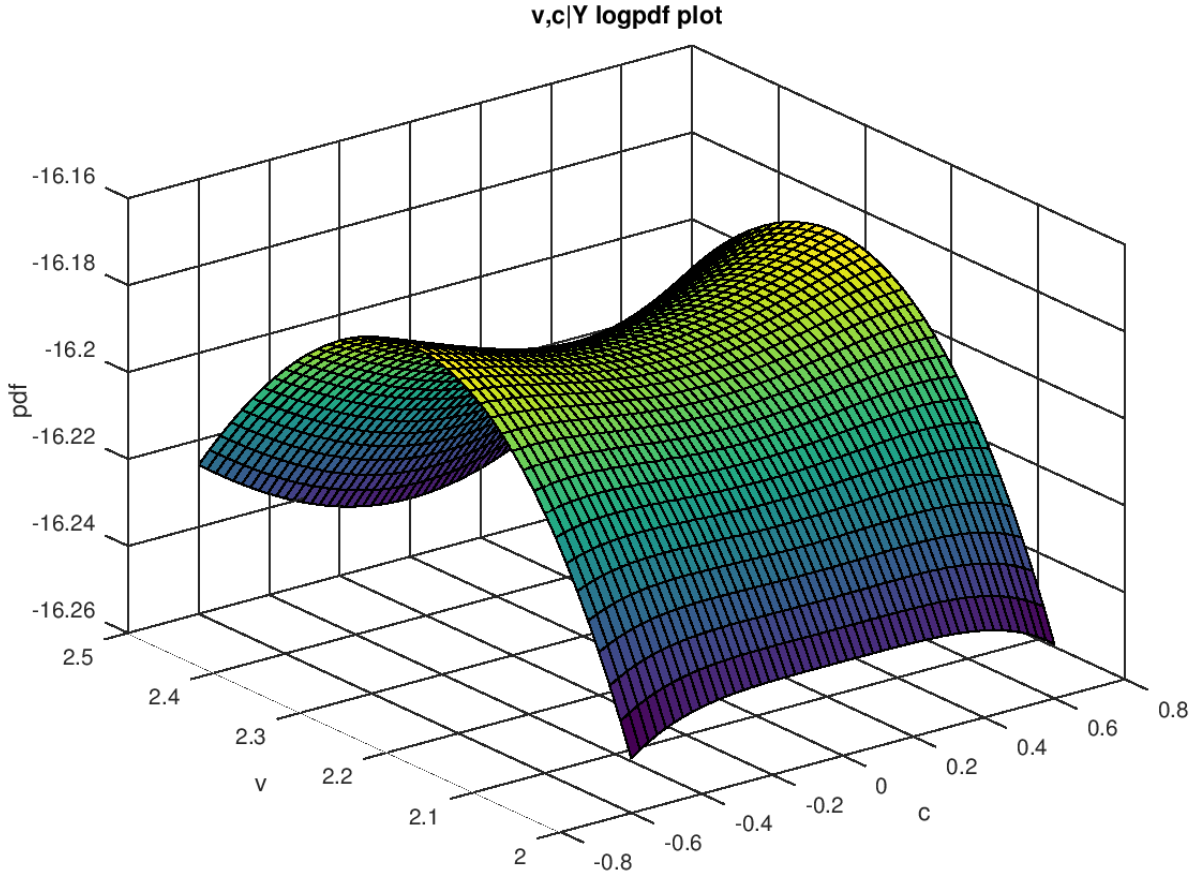
$$E[W_{ij}|W_{i'j} = w_{i'j}, \boldsymbol{\vartheta}^k] = \frac{c^k w_{i'j}}{v^k}, \tag{4.29}$$

$$\begin{aligned} E[W_{ij}^2|W_{i'j} = w_{i'j}, \boldsymbol{\vartheta}^k] &= v^k - \frac{(c^k)^2}{v^k} + (\frac{c^k w_{i'j}}{v^k})^2 \\ &= v^k - \frac{(c^k)^2}{v^k}(1 + \frac{w_{i'j}^2}{v^k}). \end{aligned} \tag{4.30}$$

Άρα τελικά,

$$\begin{aligned} Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) =& -12log(2\pi) - 6log(v^2-c^2) - log(v) + \bar{c} \\ & - \frac{v(40 + \sum_{j=9}^{12}(v^k - \frac{(c^k)^2}{v^k}(1+\frac{w_{1j}^2}{v^k})) + \sum_{j=5}^{8}(v^k - \frac{(c^k)^2}{v^k}(1+\frac{w_{2j}^2}{v^k})))}{2(v^2-c^2)} \\ & + \frac{c(\sum_{j=5}^{8} w_{1j}\frac{c^k w_{1j}}{v^k} + \sum_{j=9}^{12} w_{2j}\frac{c^k w_{2j}}{v^k})}{v^2-c^2} \\ =& -12log(2\pi) - 6log(v^2-c^2) - log(v) + \bar{c} \\ & - \frac{v(40 + 8v^k - 8\frac{(c^k)^2}{v^k}(1+\frac{4}{v^k}))}{2(v^2-c^2)} \\ & + \frac{32c\frac{c^k}{v^k}}{v^2-c^2}. \end{aligned} \tag{4.31}$$

Διάγραμμα 4.3: Διάγραμμα της λογαριθμημένης πυκνότητας πιθανότητας της $v, c|\mathbf{Y}$.



Αν δοθεί τιμή $c^k = 0$ σε κάποια επανάληψη, τότε παραγωγίζοντας την (4.31) εύκολα φαίνεται ότι οι τιμές v^{k+1}, c^{k+1} που μεγιστοποιούν την Q είναι οι

$$c^{k+1} = 0, \tag{4.32}$$

και

$$v^{k+1} = \frac{20 + 4v^k}{13}. \tag{4.33}$$

Με την βοήθεια των σχέσεων (4.32) και (4.33), εκτελώντας τον αλγόριθμο EM για αρχικές τιμές $v^0 = c^0 = 0$, φαίνεται πως η ακολουθία $\boldsymbol{\vartheta}^k$ συγκλίνει στο σημείο $\boldsymbol{\vartheta}^* = (20/9, 0)$. Γενικότερα, για την (4.33), αν κανείς πάρει το όριο της v^k, για k να πηγαίνει στο άπειρο, θα δει ότι ανεξάρτητα του v^0 η ακολουθία συγκλίνει στο 20/9.

Μετά από ανάλυση στην κατανομή της $v, c|\mathbf{Y}$ φαίνεται ότι έχει δύο σημεία που παρουσιάζει μέγιστο και ένα που παρουσιάζει σαγματικό σημείο. Τα σημεία που παρουσιάζει μέγιστο είναι τα $\boldsymbol{\vartheta}_1$=(16/7,2$\sqrt{8}$/7) και $\boldsymbol{\vartheta}_2$=(16/7,-2$\sqrt{8}$/7), ενώ το σαγματικό σημείο βρίσκεται στο $\boldsymbol{\vartheta}_3$=(20/9,0). Στο Διάγραμμα 4.3 φαίνεται η λογαριθμημένη πυκνότητα πιθανότητας της $v, c|\mathbf{Y}$ (μείον κάποια σταθερά). Προφανώς, όπως και στο προηγούμενο παράδειγμα, το διάγραμμα για την πυκνότητα πιθανότητας θα έχει το ίδιο σχήμα, αλλά διαφορετική κλίμακα στον άξονα των $y'y$.

Όπως και φάνηκε, ο αλγόριθμος EM για την συγκεκριμένη περίπτωση, με αρχικές τιμές $\boldsymbol{\vartheta}^0$=(0,0), πέτυχε σύγκλιση σε σαγματικό σημείο. Αν τύχει και η τιμή c^k γίνει ίση με το μηδέν, λοιπόν, η σύγκλιση κατευθύνεται κατευθείαν στο σαγματικό σημείο. Διαισθητικά, για $c^k = 0$ οι τιμές $v^{k+n},\ n \in \mathbb{N}$ «κολλάνε» στον δακτύλιο, γύρω από το σαγματικό σημείο, που εκτείνεται στον άξονα του v. Από την k επανάληψη και μετά η τιμή της πυκνότητας πιθανότητας ασφαλώς και αυξάνεται, αλλά μόνο ως προς τις τιμές του δακτυλίου γύρω από το σαγματικό σημείο. Αν το c^k μετατοπιστεί από το 0 κατάλληλα, η σύγκλιση θα γίνει σε κάποιο από τα μέγιστα.

4.2.3 Πρόβλημα στη σύγκλιση του GEM

Ο *Boyles* (1983) έδωσε το παρακάτω παράδειγμα σχετικά με μη επιθυμητή σύγκλιση του αλγορίθμου GEM. Σύμφωνα με το αποτέλεσμα η ακολουθία του αλγορίθμου $\{f_{\boldsymbol{\vartheta}|\mathbf{Y}}(\boldsymbol{\vartheta}^k|\mathbf{y})\}_{k\in\mathbb{N}}$ συγκλίνει σε κάποια τιμή, αλλά η ακολουθία $\boldsymbol{\vartheta}^k$ δεν επιτυγχάνει σύγκλιση κάπου.

Έχει παρατηρηθεί δισδιάστατο δείγμα $\mathbf{y} = (y_1, y_2)$ που προέρχεται από τυχαία μεταβλητή $\mathbf{Y}|\boldsymbol{\mu}$ που ακολουθεί δισδιάστατη κανονική κατανομή με μέση τιμή $\boldsymbol{\mu}$ και πίνακα διακυμάνσεων-συνδιακυμάνσεων τον ταυτοτικό,

$$\mathbf{Y}|\boldsymbol{\mu} \sim N_2(\boldsymbol{\mu}, I_2).$$

Το $\boldsymbol{\mu}$ είναι τυχαία μεταβλητή που ακολουθεί την πρότερη καταχρηστική κατανομή $f_{\boldsymbol{\mu}}(\mu_1, \mu_2) \propto 1$, $\mu_1, \mu_2 \in \mathbb{R}$.

Σε αυτό το παράδειγμα δεν υπάρχει μη παρατηρούμενη τιμή στο δείγμα, δηλαδή $\mathbf{y} = \mathbf{w}$. Άρα το εσωτερικό της μέσης τιμής του Q είναι σταθερό ως προς τη μέση τιμή,

$$\begin{aligned} Q(\boldsymbol{\mu}|\boldsymbol{\mu}^k) &= log f_{\boldsymbol{\mu}|\mathbf{Y}}(\boldsymbol{\mu}|\mathbf{y}) \\ &= log f_{\mathbf{Y}|\boldsymbol{\mu}}(\mathbf{y}|\boldsymbol{\mu}) + \bar{c} \\ &= -\frac{1}{2}((y_1 - \mu_1)^2 + (y_2 - \mu_2)^2) + \bar{c}^{'}. \end{aligned} \tag{4.34}$$

Τα $\bar{c}$ και $\bar{c}^{'}$ είναι τιμές που δεν περιέχουν τα μ_1, μ_2. Ορίζω τις παρακάτω ακολουθίες,

$$\mu_1^k = y_1 + r^k cos\theta^k, \tag{4.35}$$

$$\mu_2^k = y_2 + r^k sin\theta^k, \tag{4.36}$$

$$r^k = 1 + \frac{1}{k+1}, \tag{4.37}$$

$$\theta^k = \sum_{i=1}^{k} \frac{1}{i+1}. \tag{4.38}$$

Οι ακολουθίες των μ_1^k, μ_2^k (σχέσεις (4.35) και (4.36)) είναι ακολουθίες του αλγορίθμου GEM. Για να δειχθεί αυτό, αρκεί να δειχθεί ότι,

$$Q(\boldsymbol{\mu}^{k+1}|\boldsymbol{\mu}^k) \geq Q(\boldsymbol{\mu}^k|\boldsymbol{\mu}^k),\ \forall k \in \mathbb{N}.$$

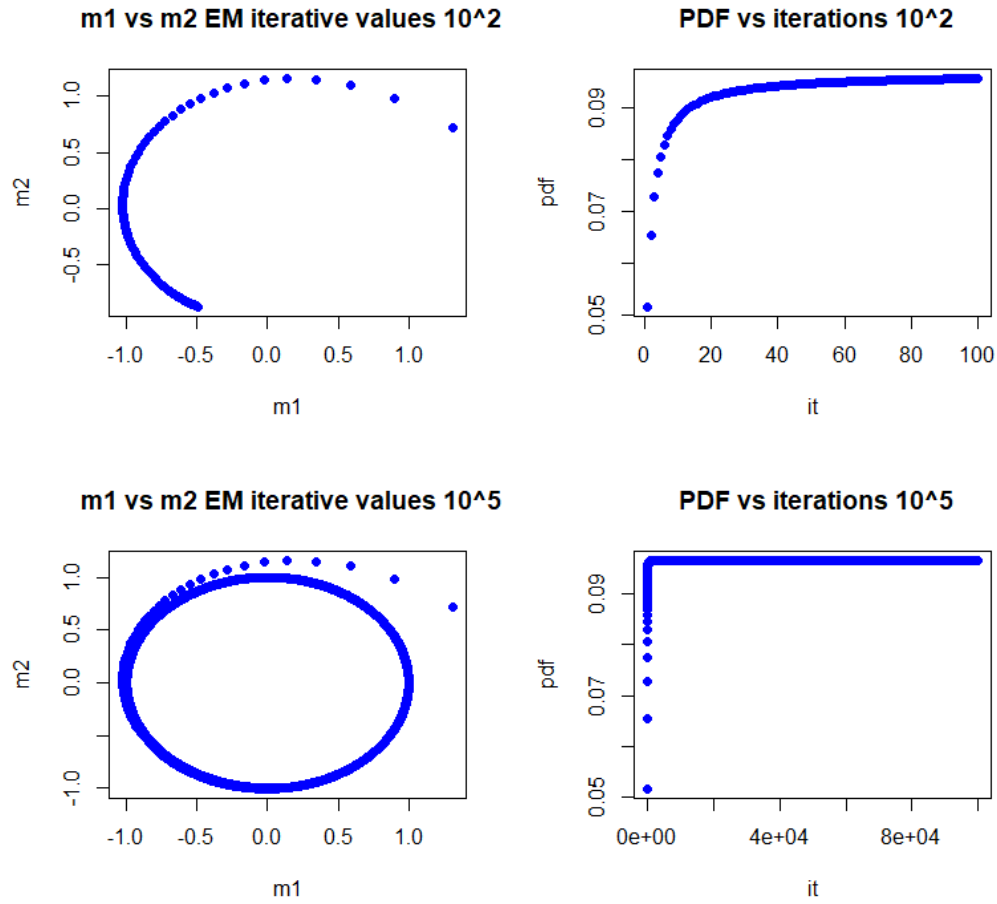
Στην (4.39) αποδεικνύεται το παραπάνω,

$$Q(\boldsymbol{\mu}^{k+1}|\boldsymbol{\mu}^k) - Q(\boldsymbol{\mu}^k|\boldsymbol{\mu}^k) = \frac{1}{2}((r^k)^2 - (r^{k+1})^2)$$
$$= \frac{1}{2}((r^k)^2 - (2 - \frac{1}{(r^k)^2})) \geq 0. \tag{4.39}$$

Η ανισότητα στην (4.37) έπεται από την ιδιότητα $2 - \frac{1}{x} \leq x$, για κάθε $x \in [1, \infty)$. Το r^k είναι εξ' ορισμού μεγαλύτερο ίσο του 1.

Άμα κανείς εκτελέσει τον αλγόριθμο GEM με τα παραπάνω δεδομένα θα δει πως η ακολουθία $f_{\boldsymbol{\mu}|\mathbf{Y}}(\boldsymbol{\mu}^k|\mathbf{y})$ συγκλίνει στην τιμή $\frac{e^{-\frac{1}{2}}}{2\pi}$, ενώ η $\boldsymbol{\mu}^k$ δεν καταφέρνει να επιτύχει σύγκλιση. Αυτό που συμβαίνει είναι πως οι τιμές $\boldsymbol{\mu}^k$ δεν καταφέρνουν να επιτύχουν σύγκλιση σε σημείο, αλλά «κολλάνε» σε ένα σύνολο σημείων που αποτελούν κυκλικό δίσκο με ακτίνα 1 και κέντρο το $\mathbf{y} = (y_1, y_2)$. Για οποιοδήποτε ζεύγος τιμών στον δακτύλιο, η συνάρτηση πυκνότητας πιθανότητας έχει την ίδια τιμή.

Διάγραμμα 4.4: Διαγράμματα για την εξέλιξη του αλγορίθμου.



Στο Διάγραμμα 4.4 φαίνεται τι συμβαίνει κατά την εκτέλεση του ΕΜ στο συγκεκριμένο παράδειγμα για 100 και 10000 επαναλήψεις, κάτω από το δείγμα $y_1 = y_2 = 0$. Όπως και είπαμε δεν επιτυγχάνεται σύγκλιση, για αυτό και για να σταματήσει κάποτε ο βρόγχος προκαθορίζεται ο αριθμός των επαναλήψεων. Τα αριστερά γραφήματα δείχνουν την εξέλιξη των μ_1^k, μ_2^k, ενώ τα δεξιά την ύστερη πιθανοφάνεια που προκύπτει από τα μ_1^k, μ_2^k για την κάθε επανάληψη k.

Λόγω των καταχρηστικών πρότερων στα τρία προηγούμενα παραδείγματα, τα ίδια αριθμητικά αποτελέσματα θα ίσχυαν και σε περίπτωση που υποθέταμε πως οι παράμετροι είναι σταθερές, δηλαδή στην εφαρμογή του αλγορίθμου στην κλασική στατιστική. Εκτελώντας κανείς τον αλγόριθμο

ΕΜ για την επίλυση κάποιου προβλήματος, όπως στην περίπτωση του $EMVS$, καλό είναι να έχει στο μυαλό του τα τρία προηγούμενα παραδείγματα σαν πιθανά ενδεχόμενα, ειδικά σε περιπτώσεις που η σύγκλιση σε ολικό μέγιστο δεν είναι βέβαιη.

Κεφάλαιο 5

EMVS

Ο $EMVS$ προτάθηκε από τους *Rockova* και *George* (2014) και είναι ένας αλγόριθμος που χρησιμοποιείται σε περιπτώσεις επιλογής επεξηγηματικών μεταβλητών σε μοντέλα παλινδρόμησης. Τα αρχικά του προέρχονται από τις λέξεις Expectation Maximization Variable Selection, και όπως φαίνεται από αυτές είναι ένας αλγόριθμος που χρησιμοποιεί τον αλγόριθμο EM (Κεφάλαια 3 και 4) για να δώσει αποτελέσματα σχετικά με το ποιες μεταβλητές επιδρούν σημαντικά πάνω σε ένα γραμμικό μοντέλο, κάνοντας χρήση της μεθόδου Spike-and-Slab (Κεφάλαιο 1, ενότητα 3).

Παλιότερα στο κομμάτι του Variable Selection, για τους υπολογισμούς στη μέθοδο Spike-and-Slab χρησιμοποιούταν μη αιτιοκρατικές προσεγγίσεις, όπως η $SSVS$ που βασίζεται σε μεθόδους Markov Chain Monte Carlo. Ο $EMVS$, απέναντι σε αυτές τις μεθόδους, έχει το πλεονέκτημα πως έχει πολύ χαμηλότερο υπολογιστικό κόστος, ως ντετερμινιστική μέθοδος. Επιπλέον, ένα πλεονέκτημα του αλγορίθμου είναι πως γίνεται να εφαρμοστεί σε μοντέλα όπου το δείγμα είναι μικρότερο από τον αριθμό των άγνωστων παραμέτρων, πράγμα που δεν είναι εφικτό σε περιπτώσεις που γίνεται μελέτη μέσω εκτιμητριών ελαχίστων τετραγώνων.

5.1 Spike-and-Slab

Οι υποθέσεις της μεθόδου Spike-and-Slab, όπως φαίνεται και στο Κεφάλαιο 1, είναι οι παρακάτω,

$$\begin{aligned}
&\mathbf{Y}|\boldsymbol{\beta},\sigma^2 \sim N_n(X\boldsymbol{\beta},\sigma^2 I_n)\\
&\boldsymbol{\beta}|\sigma^2,\boldsymbol{\gamma} \sim N_p(\mathbf{0}_p, D_{\sigma^2,\boldsymbol{\gamma}})\\
&D_{\sigma^2,\boldsymbol{\gamma}} = \sigma^2 diag\{(1-\gamma_i)v_0 + \gamma_i v_1\}\\
&\sigma^2|\boldsymbol{\gamma} \sim IG(v/2, \lambda v/2)\\
&\pi(\boldsymbol{\gamma}|\theta) = \theta^{\sum_{i=0}^{p}\gamma_i}(1-\theta)^{p-\sum_{i=0}^{p}\gamma_i}\\
&\theta \sim Beta(a,b).
\end{aligned} \tag{5.1}$$

Το $\mathbf{Y}$ είναι διάνυσμα n μεταβλητών. Όπως έχει αναφερθεί και στο Κεφάλαιο 1, το β_0 έχει αφαιρεθεί από το μοντέλο, κεντράροντας την κατανομή του $\mathbf{Y}|\boldsymbol{\beta},\sigma^2$ στο 0. Αυτό επιτυγχάνεται υποθέτοντας για το β_0 πρότερη ομοιόμορφη καταχρηστική ($f_{\beta_0}(\beta_0) \propto 1,\ \beta_0 \in \mathbb{R}$) κατανομή και ολοκληρώνοντας ως προς αυτήν,

$$\begin{aligned}
f_{\mathbf{Y}|\boldsymbol{\beta},\sigma^2}(\mathbf{y}|\boldsymbol{\beta},\sigma^2) &= \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\beta_0-X\boldsymbol{\beta})^T(\mathbf{y}-\beta_0-X\boldsymbol{\beta})} d\beta_0\\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}\|\beta_0-\mathbf{y}+X\boldsymbol{\beta}\|_2^2} d\beta_0\\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y}-X\boldsymbol{\beta}\|_2^2}\sigma.
\end{aligned}$$

Στην τελευταία ισότητα, ανοίξαμε τα τετράγωνα της νόρμας και βγάλαμε εκτός του ολοκληρώματος μια ποσότητα που μας επιτρέπει μες το ολοκλήρωμα να κατασκευάσουμε το εκθετικό μέρος της πυκνότητας πιθανότητας της κανονικής κατανομής, ως προς β_0.

Μέσω του αλγορίθμου ΕΜ θα βρεθούν οι τιμές $(\boldsymbol{\beta}^*, \sigma^*, \theta^*)$ στις οποίες μεγιστοποιείται η πυκνότητα πιθανότητα $f_{\boldsymbol{\beta},\sigma,\theta|\mathbf{Y}}(\boldsymbol{\beta}, \sigma, \theta|\mathbf{y})$. Για τις ανάγκες του ΕΜ θεωρούμε πλήρες δείγμα το $\mathbf{W}$=($\mathbf{Y}$,$\boldsymbol{\gamma}$) και ελλιπές το $\mathbf{Y}$. Παρακάτω γίνεται ανάλυση των βασικών βημάτων του αλγορίθμου.

5.1.1 E-step

Κατά το E-step υπολογίζεται η ποσότητα Q. Γενικότερα για τις ανάγκες της εργασίας έχουμε ορίσει πως,

$$Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) = E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log(f_{\boldsymbol{\vartheta}|\mathbf{W}}(\boldsymbol{\vartheta}|\mathbf{Z}, \mathbf{y}))]. \tag{5.2}$$

Οι *Rockova* και *George* (2014) για τις ανάγκες του ΕΜ όρισαν το Q ως εξής,

$$Q(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^k) := E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\vartheta}^k}[log(f_{\boldsymbol{\vartheta},\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\vartheta}, \mathbf{Z}|\mathbf{y}))]. \tag{5.3}$$

Από την (4.5), προσθαφαιρώντας στο δεύτερο μέλος της ισότητας την ποσότητα «$log f_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})$», εύκολα φαίνεται πως και για το Q της (5.3) δεν θα άλλαζε κάτι σχετικά με την μονοτονία του αλγορίθμου. Αυτή η διαφορά στην προσέγγιση του ΕΜ εξυπηρετεί τόσο στη διευκόλυνση των υπολογισμών κατά το $E-step$, όσο και για το $M-step$ στον $EMVS$, όπως θα φανεί και παρακάτω.

Επομένως ισχύει πως,

$$\begin{aligned}
Q(\boldsymbol{\beta}, \sigma, \theta|\boldsymbol{\beta}^k, \sigma^k, \theta^k) =& E_{\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\beta}^k,\sigma^k,\theta^k}[log f_{\boldsymbol{\beta},\theta,\sigma,\boldsymbol{\gamma}|\mathbf{Y}}(\boldsymbol{\beta}, \theta, \sigma, \boldsymbol{\gamma}|\mathbf{y})] \\
=& \bar{c} + E_{\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\beta}^k,\sigma^k,\theta^k}[log f_{\mathbf{Y}|\boldsymbol{\beta},\sigma}(\mathbf{y}|\boldsymbol{\beta}, \sigma) + log f_{\boldsymbol{\beta}|\sigma,\boldsymbol{\gamma}}(\boldsymbol{\beta}|\sigma, \boldsymbol{\gamma}) + log f_{\sigma|\boldsymbol{\gamma}}(\sigma|\boldsymbol{\gamma})] \\
&+ E_{\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\beta}^k,\sigma^k,\theta^k}[log\pi(\boldsymbol{\gamma}|\theta) + log f_\theta(\theta)] \\
=& \bar{c} + Q_1(\boldsymbol{\beta}, \sigma|\boldsymbol{\beta}^k, \sigma^k, \theta^k) + Q_2(\theta|\boldsymbol{\beta}^k, \sigma^k, \theta^k).
\end{aligned} \tag{5.4}$$

Η ποσότητα $\bar{c}$ είναι σταθερά ως προς τα $(\boldsymbol{\beta}, \sigma, \theta)$. Το σ είναι θετικό, συνεπώς η δέσμευση ως προς σ δίνει την ίδια πληροφορία με τη δέσμευση ως προς σ^2 (”1-1” αντιστοιχία). Για την συνάρτηση $f_{\sigma|\boldsymbol{\gamma}}(\sigma|\boldsymbol{\gamma})$ αρκεί να γίνει μετασχηματισμός με ρίζα στην κατανομή $f_{\sigma^2|\boldsymbol{\gamma}}(\sigma|\boldsymbol{\gamma})$ που δίνεται από το $Spike-and-Slab$. Δηλαδή,

$$\begin{aligned}
f_{\sigma|\boldsymbol{\gamma}}(\sigma|\boldsymbol{\gamma}) &= f_{\sigma^2|\boldsymbol{\gamma}}(\sigma^2|\boldsymbol{\gamma})|\frac{d\sigma^2}{d\sigma}| \\
f_{\sigma|\boldsymbol{\gamma}}(\sigma|\boldsymbol{\gamma}) &\propto (\sigma^2)^{-v/2-1} e^{-\frac{v\lambda}{2\sigma^2}} \sigma \\
f_{\sigma|\boldsymbol{\gamma}}(\sigma|\boldsymbol{\gamma}) &\propto (\sigma^2)^{-(v+1)/2} e^{-\frac{v\lambda}{2\sigma^2}}.
\end{aligned}$$

Η $f_\theta(\theta)$ είναι η συνάρτηση πυκνότητας πιθανότητας της κατανομής Βήτα με παραμέτρους a, b. Ο υπολογισμός των Q_1, Q_2 φαίνεται παρακάτω,

$$\begin{aligned}
Q_1(\boldsymbol{\beta},\sigma|\boldsymbol{\beta}^k,\sigma^k,\theta^k) =& E_{\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\beta}^k,\sigma^k,\theta^k}[log f_{\mathbf{Y}|\boldsymbol{\beta},\sigma}(\mathbf{y}|\boldsymbol{\beta},\sigma) + log f_{\boldsymbol{\beta}|\sigma,\boldsymbol{\gamma}}(\boldsymbol{\beta}|\sigma,\boldsymbol{\gamma}) + log f_{\sigma|\boldsymbol{\gamma}}(\sigma|\boldsymbol{\gamma})] \\
=& E_{\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\beta}^k,\sigma^k,\theta^k}[-\frac{(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})}{2\sigma^2} - \frac{n+p+v}{2}log(\sigma^2) \\
& - \frac{v\lambda}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\frac{1}{(1-\gamma_i)v_0+\gamma_i v_1}] + c_1 \\
\\
=& -\frac{(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})}{2\sigma^2} - \frac{n+p+v}{2}log(\sigma^2) - \frac{v\lambda}{2\sigma^2} \\
& - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2 E_{\gamma_i|\boldsymbol{\beta}^k,\sigma^k,\theta^k}[\frac{1}{(1-\gamma_i)v_0+\gamma_i v_1}] + c_1, \qquad (5.5)
\end{aligned}$$

$$\begin{aligned}
Q_2(\theta|\boldsymbol{\beta}^k,\sigma^k,\theta^k) =& E_{\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\beta}^k,\sigma^k,\theta^k}[log\pi(\boldsymbol{\gamma}|\theta) + log f_\theta(\theta)] \\
=& E_{\boldsymbol{\gamma}|\mathbf{Y},\boldsymbol{\beta}^k,\sigma^k,\theta^k}[\sum_{i=1}^{p}\gamma_i log(\frac{\theta}{1-\theta}) + (a-1)log(\theta) + (p+b-1)log(1-\theta)] + c_2 \\
=& \sum_{i=1}^{p} log(\frac{\theta}{1-\theta}) E_{\gamma_i|\boldsymbol{\beta}^k,\sigma^k,\theta^k}[\gamma_i] + (a-1)log(\theta) + (p+b-1)log(1-\theta) + c_2. \qquad (5.6)
\end{aligned}$$

Τα c_1, c_2 είναι σταθερές ως προς τα $\boldsymbol{\beta}, \theta, \sigma$. Για να προχωρήσουμε στο $M-step$ αρκεί να βρεθούν οι μέσες τιμές που αφορούν τα γ_i στις (5.5) και (5.6). Τα γ_i παίρνουν τις τιμές 0 και 1 μόνο, συνεπώς κάτω από οποιεσδήποτε δεσμεύσεις, πάντα θα ακολουθούν κατανομή $Bernoulli$. Άρα,

$$E_{\gamma_i|\boldsymbol{\beta}^k,\sigma^k,\theta^k}[\gamma_i] = P(\gamma_i = 1|\boldsymbol{\beta}^k,\sigma^k,\theta^k) := p_i^*. \qquad (5.7)$$

Από τον πίνακα διακυμάνσεων-συνδιακυμάνσεων και την υπόθεση της κανονικότητας, τα β_i είναι ανεξάρτητα ανά δύο. Από το Θεώρημα $Bayes$, τα p_i^* υπολογίζονται όπως φαίνεται παρακάτω,

$$\begin{aligned}
p_i^* &= \frac{f_{\beta_i|\sigma,\gamma_i}(\beta_i^k|\sigma^k,\gamma_i=1)P(\gamma_i=1|\theta^k)}{f_{\beta_i|\sigma,\gamma_i}(\beta_i^k|\sigma^k,\gamma_i=1)P(\gamma_i=1|\theta^k) + f_{\beta_i|\sigma,\gamma_i}(\beta_i^k|\sigma^k,\gamma_i=0)P(\gamma_i=0|\theta^k)} \\
&= \frac{f_{\beta_i|\sigma,\gamma_i}(\beta_i^k|\sigma^k,\gamma_i=1)\theta^k}{f_{\beta_i|\sigma,\gamma_i}(\beta_i^k|\sigma^k,\gamma_i=1)\theta^k + f_{\beta_i|\sigma,\gamma_i}(\beta_i^k|\sigma^k,\gamma_i=0)(1-\theta^k)}. \qquad (5.8)
\end{aligned}$$

Η πυκνότητα πιθανότητας των μεταβλητών $\beta_i|\sigma^k,\gamma_i$ είναι γνωστή από την (5.1), και σε συνδυασμό με τις σχέσεις (5.7) και (5.8) υπολογίζεται η μέση τιμή που χρειαζόμαστε στην σχέση (5.6).
Για την μέση τιμή της σχέσης (5.5) ισχύει πως,

$$\begin{aligned}
E_{\gamma_i|\boldsymbol{\beta}^k,\sigma^k,\theta^k}[\frac{1}{(1-\gamma_i)v_0+\gamma_i v_1}] =& \frac{1}{(1-1)v_0+1v_1}p_i^* + \frac{1}{(1-0)v_0+0v_1}(1-p_i^*) \\
=& \frac{p_i^*}{v_1} + \frac{1-p_i^*}{v_0}. \qquad (5.9)
\end{aligned}$$

Επομένως τα $Q_1,\ Q_2$ είναι γνωστές ποσότητες κάτω από τις υποθέσεις της (5.1). Επιπλέον, αξίζει να σημειωθεί πως σε κάποια παραλλαγή του $EMVS$, για διαφορετική πρότερη κατανομή της $\boldsymbol{\gamma}|\theta$ και του θ αλλάζει μόνο η ποσότητα Q_2, και όχι η Q_1, και γίνεται,

$$Q_2(\theta|\boldsymbol{\beta}^k,\theta^k,\sigma^k) = E_{\gamma|\mathbf{Y},\boldsymbol{\beta}^k,\theta^k,\sigma^k}[log\pi(\boldsymbol{\gamma}|\theta)] + log\pi(\theta). \qquad (5.10)$$

5.1.2 M-step

Κατά το $E-step$ υπολογίστηκε η ποσότητα Q. Στο $M-step$ θα βρεθούν τα $(\boldsymbol{\beta}^{k+1}, \sigma^{k+1}, \theta^{k+1})$ για τα οποία μεγιστοποιείται το Q ως προς $(\boldsymbol{\beta}, \sigma, \theta)$. Αφού το Q αναλύεται σε Q_1, Q_2 είναι εμφανές πως μπορούν να βρεθούν ξεχωριστά τα $(\boldsymbol{\beta}^{k+1}, \sigma^{k+1})$ και θ^{k+1}.

Για το $\boldsymbol{\beta}^{k+1}$ που μεγιστοποιεί την Q_1, ανεξάρτητα του σ, αρκεί να ασχοληθούμε με την (5.11),

$$\frac{-(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})-\boldsymbol{\beta}^T D^*\boldsymbol{\beta}}{2\sigma^2} = \\ -\frac{(\boldsymbol{\beta}-(X^TX+D^*)^{-1}X^T\mathbf{y})^T(X^TX+D^*)(\boldsymbol{\beta}-(X^TX+D^*)^{-1}X^T\mathbf{y})+c}{2\sigma^2}. \tag{5.11}$$

Το δεύτερο μέλος της ισότητας προκύπτει όπως και στην (1.5), παραγοντοποιώντας ως προς $\boldsymbol{\beta}$. Το c είναι μια ποσότητα που δεν περιέχει $\boldsymbol{\beta}$. Το D^* είναι ένας διαγώνιος πίνακας, διαστάσεων $p \times p$. Τα στοιχεία της διαγωνίου είναι τα $d^*_{ii} = \frac{p^*_i}{v_1} + \frac{1-p^*_i}{v_0}$, όπου τα p^*_i είναι αυτά της σχέσης (5.8), στις αντίστοιχες θέσεις i, i του πίνακα. Βλέποντας την (5.11) σαν το εκθετικό μέρος της πυκνότητας πιθανότητας p-διάστατης κανονικής κατανομής, η οποία είναι γνωστό πως έχει ολικό μέγιστο στην μέση τιμή της, εύκολα φαίνεται πως μεγιστοποιείται ως προς $\boldsymbol{\beta}$ για,

$$\boldsymbol{\beta}^{k+1} = (X^TX + D^*)^{-1}X^T\mathbf{y}. \tag{5.12}$$

Η παραπάνω λύση έχει το πλεονέκτημα πως ορίζεται ακόμα και αν ο X^TX δεν είναι αντιστρέψιμος.

Ακόμα, η μέθοδος $Sherman-Morrison-Woodbury$ (Rockova, George (2014)) δίνει διαφορετική λύση για το $\boldsymbol{\beta}^{k+1}$ του Q_1, η οποία έχει καλύτερη εφαρμογή σε περιπτώσεις όπου το μέγεθος του δείγματος είναι μικρότερο από το πλήθος των αγνώστων παραμέτρων. Η λύση αυτή είναι η,

$$\boldsymbol{\beta}^{k+1} = [(D^*)^{-1} - (D^*)^{-1}X^T(I_n + X(D^*)^{-1}X^T)^{-1}X(D^*)^{-1}]X^T\mathbf{y}. \tag{5.13}$$

Από τη στιγμή που το $\boldsymbol{\beta}^{k+1}$ είναι γνωστό, παραγωγίζουμε την Q_1 ως προς σ^2 για να βρεθεί το σημείο που μεγιστοποιείται ως προς σ,

$$\frac{dQ_1(\boldsymbol{\beta}^{k+1}, \sigma|\boldsymbol{\beta}^k, \sigma^k, \theta^k)}{d\sigma^2} = \frac{(\mathbf{y}-X\boldsymbol{\beta}^{k+1})^T(\mathbf{y}-X\boldsymbol{\beta}^{k+1})}{2\sigma^4} - \frac{n+p+v}{2\sigma^2} + \frac{v\lambda}{2\sigma^4} + \frac{(\boldsymbol{\beta}^{k+1})^T D^*(\boldsymbol{\beta}^{k+1})}{2\sigma^4}. \tag{5.14}$$

Άρα, εξισώνοντας με το 0, από την (5.14), το σ^{k+1} που μεγιστοποιεί την (5.5) δίνεται από τον τύπο,

$$\sigma^{k+1} = \sqrt{\frac{(\mathbf{y}-X\boldsymbol{\beta}^{k+1})^T(\mathbf{y}-X\boldsymbol{\beta}^{k+1}) + (\boldsymbol{\beta}^{k+1})^T D^*(\boldsymbol{\beta}^{k+1}) + v\lambda}{n+p+v}}. \tag{5.15}$$

Αν και η παραπάνω παραγώγιση είναι έμμεση, το σημείο σ^{k+1} είναι μέγιστο και το γιατί γίνεται αισθητό από τις σχέσεις (5.16),

$$\frac{df(x)}{d(x^2)} = \frac{df(x)}{dx}\frac{dx}{d(x^2)} = \frac{df(x)}{dx}\frac{1}{2x},\ x>0, \Rightarrow \\ \frac{df(x)}{d(x^2)} = 0 \Leftrightarrow \\ \frac{df(x)}{dx}\frac{1}{2x} = 0 \Leftrightarrow \\ \frac{df(x)}{dx} = 0. \tag{5.16}$$

Η μεγιστοποίηση του σ έγινε με αυτόν τον τρόπο για λόγους που διευκολύνουν τις πράξεις.

Αν τώρα κανείς προσπαθήσει να μεγιστοποιήσει το Q_2 της (5.6) ως προς ϑ, τότε προκύπτει πως,

$$\frac{dQ_2(\theta|\boldsymbol{\beta}^k,\sigma^k,\theta^k)}{d\theta} = \frac{\sum_{i=1}^{p} p_i^* + a - 1}{\theta} - \frac{p - \sum_{i=1}^{p} p_i^* + b - 1}{1-\theta}. \tag{5.17}$$

Άρα, εξισώνοντας με το 0,

$$\theta^{k+1} = \frac{\sum_{i=1}^{p} p_i^* + a - 1}{a + b + p - 2}. \tag{5.18}$$

Από τις (5.12) (ή (5.13)), (5.15) και (5.18) είναι εφικτό να συσταθεί η επαναληπτική ακολουθία του EM. Οι αριθμητικές τιμές που είναι απαραίτητο να δοθούν στον EM για να είναι εκτελέσιμος, σε αυτήν την περίπτωση, είναι το δείγμα ($\mathbf{y}$,X), οι παράμετροι των πρότερων κατανομών της (5.1) και αρχικές τιμές $(\boldsymbol{\beta}^0, \sigma^0, \theta^0)$ για την επαναληπτική ακολουθία.

5.2 Συμπερασματολογία για το β

Παραπάνω φαίνεται αναλυτικά πως στήνεται και εκτελείται ο EM στην περίπτωση του $EMVS$, αλλά δεν έχει αναφερθεί κάτι μέχρι τώρα σχετικά με το πως ο $EMVS$ λύνει το πρόβλημα επιλογής μεταβλητών. Παρακάτω περιγράφεται το πως γίνεται η συμπερασματολογία σχετικά με την ισχύ των $\boldsymbol{\beta}$ στο μοντέλο. Έστω ότι έχει τρέξει ο αλγόριθμος EM και έχουμε πάρει το αποτέλεσμα $(\boldsymbol{\beta}^*, \sigma^*, \theta^*)$. Θα χρησιμοποιήσουμε αυτό το αποτέλεσμα για την συμπερασματολογία.

Ορίζεται,

$$\widetilde{\boldsymbol{\gamma}} := argmax_{\boldsymbol{\gamma}} P(\boldsymbol{\gamma}|\boldsymbol{\beta}^*, \sigma^*, \theta^*), \tag{5.19}$$

και επιπλέον, ως κανόνας απόφασης, ορίζεται πως τα β_i που επιδρούν σημαντικά στο μοντέλο, κάτω από τις υποθέσεις και παραμέτρους της (5.1), είναι αυτά για τα οποία η τιμή του $\widetilde{\boldsymbol{\gamma}}$ στην αντίστοιχη θέση i είναι ίση με 1. Οι υπόλοιπες θεωρούμε πως δεν επιδρούν σημαντικά στο μοντέλο. Η λογική πίσω από τον κανόνα απόφασης είναι πως επιλέγεται ως εκτίμηση του $\boldsymbol{\gamma}$, η τιμή για την οποία μεγιστοποιείται η πιθανότητα, κάτω από την δέσμευση πως οι παράμετροι είναι ίσες με τις τιμές στις οποίες βρίσκεται η κορυφή της ύστερης κατανομής τους.

Κάτω από την υπόθεση της ανεξαρτησίας των γ_i, μπορεί να βρεθεί ξεχωριστά για το καθένα από τα $\widetilde{\gamma}_i$ η τιμή του και συνεπώς να παρθεί η απόφαση για το αν το αντίστοιχο β_i που περιγράφει επιδρά σημαντικά στο μοντέλο ή όχι. Λόγω της ανεξαρτησίας ισχύει πως,

$$\begin{aligned} P(\boldsymbol{\gamma}|\boldsymbol{\beta}^*, \sigma^*, \theta^*) &= \prod_{i=1}^{n} P(\gamma_i|\boldsymbol{\beta}^*, \sigma^*, \theta^*) \\ &= \prod_{i=1}^{n} P(\gamma_i|\beta_i^*, \sigma^*, \theta^*). \end{aligned} \tag{5.20}$$

Όπως έχει αναφερθεί και παραπάνω, τα γ_i παίρνουν τις τιμές 0 και 1 μόνο. Συνεπώς, για να βρεθεί το $\boldsymbol{\gamma}$ που μεγιστοποιεί την (5.20), αρκεί να βρεθεί για κάθε γ_i αν,

$$P(\gamma_i = 1|\boldsymbol{\beta}^*, \sigma^*, \theta^*) \geq P(\gamma_i = 0|\boldsymbol{\beta}^*, \sigma^*, \theta^*),$$

ή όχι. Συνοψίζοντας,

$$\widetilde{\gamma}_i = 1 \Leftrightarrow \frac{P(\gamma_i = 1|\beta_i^*, \sigma^*, \theta^*)}{P(\gamma_i = 0|\beta_i^*, \sigma^*, \theta^*)} \geq 1, \tag{5.21}$$

διαφορετικά $\widetilde{\gamma}_i = 0$. Το $\widetilde{\boldsymbol{\gamma}}$ είναι το διάνυσμα που, προφανώς, περιέχει τα $\widetilde{\gamma}_i$, στις αντίστοιχες θέσεις i. Οι δύο πιθανότητες του κλάσματος της (5.21) υπολογίζονται ομοίως με τα p_i^* στην (5.8).

5.2.1 Διαγράμματα για τη μελέτη των β

Για να δοθεί μια οπτική προσέγγιση σχετικά με τα αποτελέσματα του $EMVS$, προτείνονται τα δύο παρακάτω διαγράμματα. Ο $EMVS$ δίνει τη δυνατότητα για υπολογισμό πολλών αποτελεσμάτων για διαφορετικές τιμές v_0. Αυτήν την δυνατότητα εκμεταλεύονται τα παρακάτω διαγράμματα.

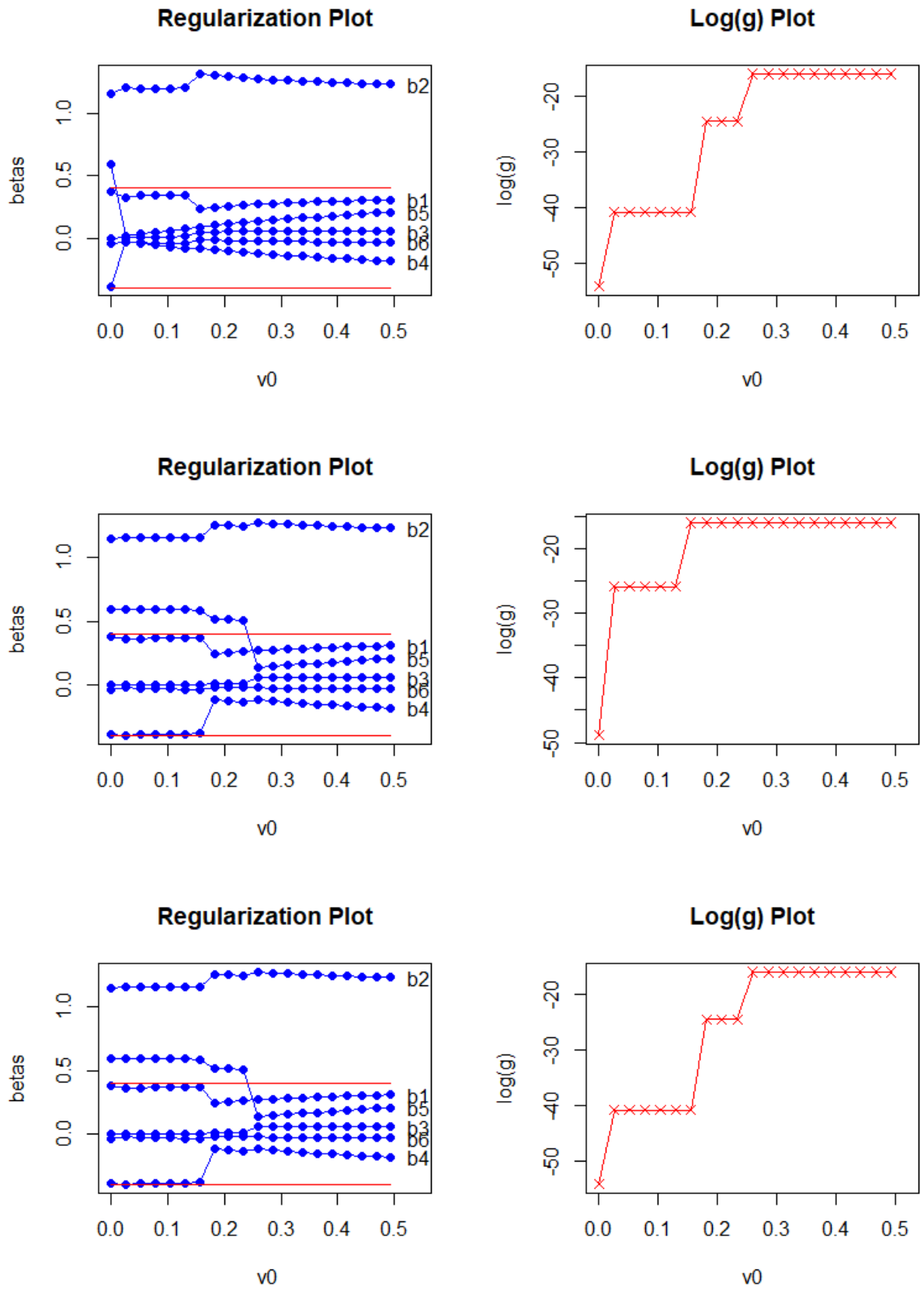
Regularization Plot
Για το Regularization Plot, αρχικά, παίρνουμε ένα καρτεσιανό σύστημα συντεταγμένων (x, y). Στον άξονα των $y'y$ φαίνονται οι τιμές των αποτελεσμάτων του ΕΜ $\boldsymbol{\beta}^*$ ως προς κάθε στοιχείο του, ενώ στον $x'x$ η τιμή που δόθηκε στο v_0 για να τρέξει ο αλγόριθμος. Μέσα από αυτό το διάγραμμα γίνεται εμφανής η εξέλιξη των $\boldsymbol{\beta}^*$, καθώς αυξάνεται το v_0. Όπως έχει αναφερθεί και στο Κεφάλαιο 1, όσο μεγαλύτερη είναι η τιμή του v_0, τόσο πιο πιθανό είναι κάποιο β_i να μην επιδρά ουσιαστικά στο μοντέλο. Το v_0 πρέπει να έχει μια σχετικά μικρή τιμή από τον ορισμό του. Παρόλο που δεν υπάρχει ένα στάνταρ εύρος για το v_0 στο οποίο εξετάζουμε τα $\boldsymbol{\beta}^*$, γενικότερα προτείνεται $v_0 \in [0, 0.5]$ και να παίρνουμε τιμές από αυτό το διάστημα με βήμα 0.01. Ακόμα και αν σε αυτό το διάστημα δεν γίνει καλή απεικόνιση της εξέλιξης των $\boldsymbol{\beta}^*$ ως προς v_0, μπορεί να δοθεί καλή εικόνα για το που πρέπει να εστιαστεί το διάγραμμα. Προφανώς τα β_i^* που βρίσκονται πολύ κοντά στο 0 επιδρούν αμελητέα στο μοντέλο, ενώ αυτά που κρατάνε μια τιμή εμφανώς μετατοπισμένη από το 0 επιδρούν σημαντικά.

Αυτό το διάγραμμα εφαρμόζεται και σε άλλες μεθόδους όπως η *Lasso* και η *Ridge Regression*. Σε αυτές τις μεθόδους βέβαια το $\boldsymbol{\beta}^*$ υπολογίζεται διαφορετικά. Ο $EMVS$ έχει το πλεονέκτημα πως καθώς αυξάνεται το v_0, τα β_i^* πηγαίνουν πιο κοντά στο μηδέν με μικρότερο ρυθμό απ΄ ότι στις παραπάνω μεθόδους.

Log(g) Plot
Στο διάγραμμα Log(g) παίρνουμε και πάλι ένα καρτεσιανό επίπεδο (x, y). Στον άξονα των $y'y$ φαίνονται οι λογαριθμημένες τιμές του $g(\widetilde{\boldsymbol{\gamma}})$, ενώ στον $x'x$ η τιμή που δόθηκε στο v_0 για το αντίστοιχο $\widetilde{\boldsymbol{\gamma}}$ του $EMVS$. Το $g(\widetilde{\boldsymbol{\gamma}})$ είναι μια ποσότητα ανάλογη της πιθανότητας εμφάνισης του $\widetilde{\boldsymbol{\gamma}}$, δοθέντος πως $\mathbf{Y} = \mathbf{y}$, για $v_0 = 0$. Παρακάτω φαίνεται αναλυτικά ο τρόπος με τον οποίο υπολογίζεται. Το $v_0 = 0$ σχετίζεται με το να δοθεί το δεδομένο πως τα β_i, για τα οποία $\widetilde{\gamma}_i = 0$, είναι ίσα με το 0, με πιθανότητα 1, και συνεπώς δεν επιδρούν καθόλου στο μοντέλο. Μέσα από αυτό το διάγραμμα γίνεται σύγκριση μεταξύ των εκ των υστέρων πιθανοτήτων εμφάνισης των $\widetilde{\boldsymbol{\gamma}}$, δοθέντος του δείγματος $\mathbf{y}$ που παρατηρήσαμε και της υπόθεσης πως τα βήτα που απορρίφθηκαν δεν έχουν καμία επίδραση στο μοντέλο, καθώς αυξάνεται το v_0 που χρησιμοποιήθηκε στον $EMVS$ για να πάρουμε την ποσότητα $\widetilde{\boldsymbol{\gamma}}$. Προφανώς όσο μεγαλύτερη η τιμή του $g(\widetilde{\boldsymbol{\gamma}})$, τόσο καλύτερη είναι η αντίστοιχη επιλογή των μεταλβητών για το μοντέλο. Ο λογάριθμος είναι αύξουσα συνάρτηση και συνεπώς δεν αλλάζει κάτι μεταξύ των συγκρίσεων. Η χρήση του λογαρίθμου προτείνεται επειδή, συνήθως, οι αποστάσεις μεταξύ των $g(\widetilde{\boldsymbol{\gamma}})$ είναι πολύ μικρές και δεν είναι εμφανείς οι διαφορές στα γραφήματα. Και πάλι για τα v_0 που θα παρασταθούν στον άξονα των $x'x$ προτείνεται να κυμαίνονται στο [0,0.5] και να επιλεχθούν με βήμα 0.01. Προφανώς βλέποντας το αρχικό διάγραμμα με την παραπάνω πρόταση, μπορεί κανείς να κάνει και δεύτερο διάγραμμα εστιάζοντας περισσότερο στα v_0 με τις μεγαλύτερες τιμές $g(\widetilde{\boldsymbol{\gamma}})$.

Στο Διάγραμμα 5.1 φαίνεται η μορφή των Regularization Plot και Log(g) Plot που προκύπτει από τον κώδικα στο πίσω μέρος της εργασίας. Τα τρία ζεύγη σχημάτων προκύπτουν από το ίδιο δείγμα για διαφορετικές αρχικές τιμές στις παραμέτρους κατά τον ΕΜ. Λεπτές διαφορές μπορεί να οφείλονται σε περιπτώσεις κακής σύγκλισης. Εδώ βέβαια είναι αμελητέες και εύκολα φαίνεται πως το καλύτερο μοντέλο περιέχει μόνο την μεταβλητή X_2.

Διάγραμμα 5.1: Plots για την μελέτη του EMVS.



Υπολογισμός της $\mathbf{g}(\widetilde{\boldsymbol{\gamma}})$

Ισχύει πως,

$$P(\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}}|\mathbf{Y}=\mathbf{y}) = \frac{f_{\mathbf{Y}|\boldsymbol{\gamma}}(\mathbf{y}|\widetilde{\boldsymbol{\gamma}})P(\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}})}{f_{\mathbf{Y}}(\mathbf{y})} \propto f_{\mathbf{Y}|\boldsymbol{\gamma}}(\mathbf{y}|\widetilde{\boldsymbol{\gamma}})P(\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}}). \tag{5.22}$$

Ορίζεται $g(\widetilde{\boldsymbol{\gamma}}) := Cf_{\mathbf{Y}|\boldsymbol{\gamma}}(\mathbf{y}|\widetilde{\boldsymbol{\gamma}})P(\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}})$, όπου C σταθερά ίδια για κάθε v_0. Η σχέση της g με το v_0 προκύπτει, καθώς το g είναι συνάρτηση του $\widetilde{\boldsymbol{\gamma}}$ που υπολογίζεται μέσω κάποιου v_0. Κατά τον υπολογισμό του g γίνεται η υπόθεση πως $v_0=0$, αλλά αυτό δεν σχετίζεται με τον υπολογισμό του $\widetilde{\boldsymbol{\gamma}}$. Όπως φαίνεται και στην (5.17), το $g(\widetilde{\boldsymbol{\gamma}})$ είναι ανάλογο της πιθανότητας εμφάνισης του $\widetilde{\boldsymbol{\gamma}}|\mathbf{Y}=\mathbf{y}$. Για τον υπολογισμό του, αρκεί να υπολογιστούν οι ποσότητες $f_{\mathbf{Y}|\boldsymbol{\gamma}}(\mathbf{y}|\widetilde{\boldsymbol{\gamma}})$ και $P(\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}})$. Για τον υπολογισμό τους θα χρησιμοποιηθούν οι υποθέσεις της (5.1) και θα δοθεί η τιμή $v_0=0$, όπως αναφέρεται και στην προηγούμενη παράγραφο,

$$\begin{aligned}
f_{\mathbf{Y}|\boldsymbol{\gamma}}(\mathbf{y}|\widetilde{\boldsymbol{\gamma}}) &= \int_{\mathbb{R}_+}\int_{\mathbb{R}^p} f_{\mathbf{Y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\widetilde{\boldsymbol{\gamma}})f_{\boldsymbol{\beta}|\sigma^2,\boldsymbol{\gamma}}(\boldsymbol{\beta}|\sigma^2,\widetilde{\boldsymbol{\gamma}})f_{\sigma^2|\boldsymbol{\gamma}}(\sigma^2|\widetilde{\boldsymbol{\gamma}})d\boldsymbol{\beta}d\sigma^2 \\
&= \int_{\mathbb{R}_+}\int_{\mathbb{R}^k} \frac{1}{(2\pi\sigma^2)^{p/2}}exp[-\frac{1}{2\sigma^2}(\mathbf{y}-\tilde{X}\tilde{\boldsymbol{\beta}})^T(\mathbf{y}-\tilde{X}\tilde{\boldsymbol{\beta}})] \\
&\qquad \frac{1}{(2\pi v_1\sigma^2)^{p/2}}exp[-\frac{1}{2v_1\sigma^2}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}}]\frac{(\frac{\lambda v}{2})^{v/2}(\sigma^2)^{-v/2-1}}{\Gamma(\frac{v}{2})}exp(-\frac{\lambda v}{2\sigma^2})d\tilde{\boldsymbol{\beta}}d\sigma^2 \\
&= \int_{\mathbb{R}_+} \frac{1}{(2\pi\sigma^2)^{p/2}}\frac{1}{(2\pi v_1\sigma^2)^{p/2}}\frac{(\frac{\lambda v}{2})^{v/2}(\sigma^2)^{-v/2-1}}{\Gamma(\frac{v}{2})}exp(-\frac{\lambda v}{2\sigma^2}) \\
&\qquad (\int_{\mathbb{R}^k} exp[-\frac{1}{2\sigma^2}(\mathbf{y}-\tilde{X}\tilde{\boldsymbol{\beta}})^T(\mathbf{y}-\tilde{X}\tilde{\boldsymbol{\beta}})-\frac{1}{2v_1\sigma^2}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}}]d\tilde{\boldsymbol{\beta}})\ d\sigma^2.
\end{aligned} \tag{5.23}$$

Λόγω της δέσμευσης $\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}}$ και $v_0=0$, στην (5.18), έχουν αφαιρεθεί από το μοντέλο τα β_i για τα οποία $\widetilde{\gamma}_i=0$. Η προηγούμενη πρόταση φαίνεται άμεσα από την (5.1) για $\boldsymbol{\gamma}=\widetilde{\boldsymbol{\gamma}}$ και $v_0=0$, όπου για αυτές τις υποθέσεις τα β_i παίρνουν μέση τιμή 0 με διακύμανση 0 (σταθερά). Συνεπώς η μέση τιμή της τυχαίας μεταβλητής $\mathbf{Y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma}$ θα είναι η $\tilde{X}\tilde{\boldsymbol{\beta}}$. Τα στοιχεία $\tilde{X}$ και $\tilde{\boldsymbol{\beta}}$ προκύπτουν από τα X, $\boldsymbol{\beta}$, αφαιρώντας τις τιμές που σχετίζονται με τα β_i που φαίνεται να μην επιδρούν σημαντικά στο μοντέλο. Το διάνυσμα $\tilde{\boldsymbol{\beta}}$ είναι k-διάστασης με $k\leq p$, και αντίστοιχα ο $\tilde{X}$, $n\times k$. Ομοίως για την $f_{\boldsymbol{\beta}|\sigma^2,\boldsymbol{\gamma}}(\boldsymbol{\beta}|\sigma^2,\boldsymbol{\gamma})$.

Παρακάτω (σχέση (5.19)) υπολογίζεται το εσωτερικό ολοκλήρωμα της (5.18), ως προς $\tilde{\boldsymbol{\beta}}$. Μεταξύ του πρώτου και δεύτερου βήματος χρησιμοποιείται η ίδια τεχνική με αυτήν που εφαρμόστηκε στην σχέση (1.5) του Κεφαλαίου 1. Το τελικό αποτέλεσμα προκύπτει από το γεγονός ότι ολοκληρώνουμε το εκθετικό μέρος πολυδιάστατης κανονικής κατανομής,

$$\begin{aligned}
&\int_{\mathbb{R}^k} exp[-\frac{1}{2\sigma^2}(\mathbf{y}-\tilde{X}\tilde{\boldsymbol{\beta}})^T(\mathbf{y}-\tilde{X}\tilde{\boldsymbol{\beta}})-\frac{1}{2v_1\sigma^2}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}}]d\tilde{\boldsymbol{\beta}} = \\
&= \int_{\mathbb{R}^k} exp[-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\tilde{X}\tilde{\boldsymbol{\beta}}-(\tilde{X}\tilde{\boldsymbol{\beta}})^T\mathbf{y}+(\tilde{X}\tilde{\boldsymbol{\beta}})^T\tilde{X}\tilde{\boldsymbol{\beta}}+\frac{1}{v_1}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}}]]d\tilde{\boldsymbol{\beta}} \\
&= \int_{\mathbb{R}^k} exp[-\frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\tilde{X}\tilde{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}}^T\tilde{X}^T\mathbf{y}+\tilde{\boldsymbol{\beta}}^T(\tilde{X}^T\tilde{X}+\frac{1}{v_1}I_{k\times k})\tilde{\boldsymbol{\beta}}]]d\tilde{\boldsymbol{\beta}}
\end{aligned} \tag{5.24}$$

$$= \int_{\mathbb{R}^k} exp[-\frac{1}{2\sigma^2}[\tilde{\boldsymbol{\beta}}^T(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})(\tilde{\boldsymbol{\beta}} - (\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y}) - \mathbf{y}^T\tilde{X}(\tilde{\boldsymbol{\beta}} - (\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y} + (\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y}) + \mathbf{y}^T\mathbf{y}]d\tilde{\boldsymbol{\beta}}$$

$$= exp[-\frac{1}{2\sigma^2}[\mathbf{y}^T\tilde{X}(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}]] \int_{\mathbb{R}^k} exp[-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{\beta}} - (\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T)^T(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k}) (\tilde{\boldsymbol{\beta}} - (\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y})]d\tilde{\boldsymbol{\beta}}$$

$$= exp[-\frac{1}{2\sigma^2}[\mathbf{y}^T\tilde{X}(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}]]\sqrt{det(2\pi\sigma^2(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1})}$$

$$= exp[-\frac{1}{2\sigma^2}[\mathbf{y}^T\tilde{X}(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}]](2\pi\sigma^2)^{\frac{k}{2}}\sqrt{det((\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1})}. \quad (5.25)$$

Για τους πίνακες της μορφής X^TX άμεσα αποδεικνύεται πως είναι συμμετρικοί. Ομοίως και για τους πίνακες της μορφής $X^TX + \lambda I_p$, όπου I_p ο ταυτοτικός πίνακας, και ομοίως για τους αντίστροφους πίνακες αυτών.

Άρα από τις (5.18) και (5.19),

$$\begin{aligned} f_{\mathbf{Y}|\boldsymbol{\Upsilon}}(\mathbf{y}|\tilde{\boldsymbol{\Upsilon}}) =& (2\pi)^{-\frac{k}{2}}\sqrt{det((\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1})}\frac{(\frac{\lambda v}{2})^{v/2}}{\Gamma(v/2)v_1^{k/2}} \\ & \int_{\mathbb{R}_+}(\sigma^2)^{-\frac{v}{2}-1-\frac{k}{2}}exp[-\frac{v\lambda}{2\sigma^2} - \frac{1}{2\sigma^2}[\mathbf{y}^T\tilde{X}(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}]]d\sigma^2 \\ =& (2\pi)^{-\frac{k}{2}}\sqrt{det((\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1})}\frac{(\frac{\lambda v}{2})^{v/2}}{\Gamma(v/2)v_1^{k/2}}\frac{\Gamma(\frac{v+k}{2})}{\tilde{\lambda}^{\frac{v+k}{2}}}, \end{aligned} \quad (5.26)$$

όπου,

$$\tilde{\lambda} := (v\lambda + \mathbf{y}^T\mathbf{y} + \mathbf{y}^T\tilde{X}(\tilde{X}^T\tilde{X} + \frac{1}{v_1}I_{k\times k})^{-1}\tilde{X}^T\mathbf{y})/2. \quad (5.27)$$

Μπορεί να φαίνεται πως στην (5.20) δεν εμφανίζεται το $\widetilde{\boldsymbol{\gamma}}$, αλλά αυτό δεν είναι αληθές. Η πληροφορία του $\widetilde{\boldsymbol{\gamma}}$ στην παραπάνω σχέση βρίσκεται στον αριθμό k. Ουσιαστικά το k είναι ένας εναλλακτικός τρόπος να γραφεί η ποσότητα $\sum_{i=1}^p \widetilde{\gamma}_i$. Προχωρώντας στην επόμενη ποσότητα,

$$\begin{aligned} P(\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}) &= \int_0^1 P(\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}}|\theta)f_\theta(\theta)d\theta \\ &= \int_0^1 \theta^{\sum_{i=1}^p \widetilde{\gamma}_i}(1-\theta)^{p-\sum_{i=1}^p \widetilde{\gamma}_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}d\theta \\ &= \frac{1}{B(a,b)}\int_0^1 \theta^{\sum_{i=1}^p \widetilde{\gamma}_i + a - 1}(1-\theta)^{p-\sum_{i=1}^p \widetilde{\gamma}_i + b - 1}d\theta \\ &= \frac{B(\sum_{i=1}^p \widetilde{\gamma}_i + a, p - \sum_{i=1}^p \widetilde{\gamma}_i + b)}{B(a,b)} \end{aligned} \quad (5.28)$$

Από τις (5.20), (5.21) και (5.22) υπολογίζεται άμεσα η $g(\widetilde{\boldsymbol{\gamma}})$. Οι ποσότητες της (5.20) στο γινόμενο που δεν περιέχουν το k δεν μας ενδιαφέρουν, ως προς το $g(\widetilde{\boldsymbol{\gamma}})$, λόγω της αναλογίας. Τελικά,

$$g(\widetilde{\boldsymbol{\gamma}}) := (2\pi)^{-\frac{\sum_{i=1}^{p}\widetilde{\gamma_i}}{2}} \frac{\Gamma(\frac{v+\sum_{i=1}^{p}\widetilde{\gamma_i}}{2})}{\widetilde{\lambda}^{\frac{v+\sum_{i=1}^{p}\widetilde{\gamma_i}}{2}} v_1^{\sum_{i=1}^{p}\widetilde{\gamma_i}/2}} B(\sum_{i=1}^{p}\widetilde{\gamma_i}+a, p-\sum_{i=1}^{p}\widetilde{\gamma_i}+b) \sqrt{det((\tilde{X}^T\tilde{X}+\frac{1}{v_1}I)^{-1})}. \tag{5.29}$$

Ο πίνακας $\tilde{X}$ δεν μπορεί να αφαιρεθεί λόγω της αναλογίας, μιας και το σχήμα του αλλάζει για κάθε διαφορετική τιμή $\widetilde{\boldsymbol{\gamma}}$. Ομοίως και η ορίζουσα που φαίνεται στην (5.23). Το $\Gamma(\bullet)$ παραπέμπει στην συνάρτηση γάμμα, ενώ το $\mathrm{B}(\bullet)$ στην συνάρτηση βήτα.

Κλείνοντας το Κεφάλαιο, αξίζει να σημειωθεί πως οι υποθέσεις του $EMVS$ (5.1) μπορούν να επεκταθούν μέσω των υποθέσεων που αναφέρονται στο Κεφάλαιο 1, ενότητα 3, για το Spike-and-Slab, έτσι ώστε να είναι πιο ρεαλιστικές καθώς περιγράφουν κάποιο φαινόμενο. Αναφέρεται ενδεικτικά πως θα μπορούσαμε να χρησιμοποιήσουμε κατανομή με πιο βαριά ουρά για την διακύμανση ή πρότερες κατανομές για τις παραμέτρους του ϑ. Αλλάζοντας τις παραπάνω υποθέσεις, όμως, θα υπάρξουν διαφορές στους υπολογισμούς των $E-step$, $M-step$ και $g(\widetilde{\boldsymbol{\gamma}})$.

Κεφάλαιο 6

Εφαρμογή του EMVS

Στο προηγούμενο Κεφάλαιο φαίνεται αναλυτικά ο τρόπος με τον οποίο λειτουργεί ο $EMVS$, καθώς και πως χρησιμοποιείται με σκοπό την συμπερασματολογία γύρω από την χρησιμότητα των επεξηγηματικών μεταβλητών στο μοντέλο. Ένα πολύ δημοφιλές πακέτο στη γλώσσα προγραμματισμού R σχετικά με την Μπεϋζιανή επιλογή μεταβλητών είναι το BAS ($Liang, Paulo, Molina, Clyde, Berger,$ 2008). Σε αυτό το Κεφάλαιο θα γίνει εφαρμογή του αλγορίθμου πάνω σε αριθμητικά δεδομένα και ταυτόχρονα θα γίνεται σύγκριση με τα αντίστοιχα αποτελέσματα του πακέτου BAS.

Για τις ανάγκες του BAS, υποθέτουμε πως η πρότερη πιθανότητα κάθε μοντέλου να είναι το επικρατέστερο, είναι ίση μεταξύ των μοντέλων, και η πρότερη κατανομή για τα βήτα είναι μία από τις τρεις παρακάτω κατανομές,

- $Zellner's\ g\ prior$, με παράμετρο $g = n$,
- $hyper - g\ prior$, με παράμετρο $a = 3$,
- $hyper - g/n\ prior$, με παράμετρο $a = 3$.

Η πυκνότητα πιθανότητας της $Zellner's\ g\ prior$ φαίνεται στο Κεφάλαιο 1, στην ενότητα 2 (σελίδα 5). Η $hyper - g\ prior$, δοθέντος του g, είναι ίδια με την $Zellner's\ g\ prior$, μόνο που υποθέτουμε πως το g είναι τυχαία μεταβλητή με πρότερη κατανομή την,

$$f_g(g) = \frac{a-2}{2}(1+g)^{-\frac{a}{2}},\ g > 0.$$

Ομοίως για την $hyper - g/n\ prior$, είναι ίδια με την $Zellner's\ g\ prior$, μόνο που υποθέτουμε πως το g είναι τυχαία μεταβλητή με πρότερη κατανομή την,

$$f_g(g) = \frac{a-2}{2n}(1+\frac{g}{n})^{-\frac{a}{2}},\ g > 0.$$

6.1 Εφαρμογή του EMVS σε δείγμα με χαμηλή διακύμανση

Σε αυτήν την ενότητα θα προσομοιωθούν 100 δείγματα $(\mathbf{y}^k, X^k)$, 100 παρατηρήσεων το καθένα, με την ακόλουθη διαδικασία. Ο αύξων αριθμός του κάθε δείγματος θα συμβολίζεται με k. Οι πίνακες σχεδιασμού X^k των δειγμάτων θα είναι διάστασης 100×10 και κάθε επιμέρους στοιχείο τους προκύπτει από προσομοίωση από την κανονική κατανομή με μέση τιμή 0 και διακύμανση 25. Δηλαδή,

$$x_{ij}^k \sim N(0, 25),\ \forall i, k \in \{1, 2, ..., 100\},\ j \in \{1, 2, ..., 10\}.$$

Επιπλέον, θα προσομοιωθούν $\mathbf{y}^k$, τα οποία έχουν ισχυρή γραμμική συσχέτιση με κάποιες από τις στήλες του αντίστοιχου πίνακα σχεδιασμού. Πιο συγκεκριμένα,

$$y_i^k \sim N(c + \beta_1 x_{i1}^k + \beta_2 x_{i2}^k + \beta_3 x_{i3}^k + \beta_4 x_{i4}^k + \beta_5 x_{i5}^k, 1),\ i, k \in \{1, 2, ..., 100\},$$

με τα x_{ij}^k να είναι τα επιμέρους στοιχεία των πινάκων σχεδιασμού. Η επιλογή 1 για την διακύμανση δόθηκε με σκοπό να προσομοιωθούν y_i^k που δεν θα έχουν πολύ μεγάλες αποκλίσεις από την κορυφή της κατανομής. Προφανώς τα $\mathbf{y}^k$ έχουν ισχυρή γραμμική συσχέτιση με τις στήλες 1 έως 5 των πινάκων X^k. Τα $(c, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ είναι σταθερές και οι μέθοδοι θα δώσουν output για το $\boldsymbol{\beta}$=$(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10})$, μία παράμετρος για κάθε στήλη του πίνακα.

Μέσα από τα δείγματα που περιγράφονται παραπάνω, θα δούμε τα αποτελέσματα του $EMVS$ και του BAS, σε σχέση με το ποιες μεταβλητές επιδρούν σημαντικά στα $\mathbf{y}^k$ και ποιες όχι, και θα συγκριθούν οι πιθανότητες για τις επιμέρους μεταβλητές να περιέχονται στο μοντέλο μέσω θηκογραμμάτων. Η πιθανότητα μία μεταβλητή να περιέχεται στο μοντέλο δίνεται στην σχέση (5.7), σύμφωνα με τον $EMVS$, ενώ η αντίστοιχη πιθανότητα για το BAS επιστρέφεται από την γλώσσα R μέσω κατάλληλης εντολής ($bas.lm$).

Στον Πίνακα 6.1 φαίνεται, σε ποσοστό επί τοις εκατό, πόσες φορές η κάθε τεχνική πέτυχε σωστή πρόβλεψη, ανά τα επιμέρους δείγματα. Με το σωστή πρόβλεψη, εννοείται πως το αποτέλεσμα της μεθόδου δείχνει πως το μοντέλο με την μεγαλύτερη πιθανότητα περιέχει τις μεταβλητές που όντως, σύμφωνα με την διαδικασία που ακολουθήσαμε, έπαιξαν ρόλο στην προσομοίωση των $\mathbf{y}^k$, και απέρριψε τις υπόλοιπες. Ο $EMVS$ εκτελέστηκε ξεχωριστά για $v_0 = 0.25$ και $v_0 = 0.5$. Οι δύο αυτές τιμές για το v_0 είναι μια καλή επιλογή, μιας και απορρίπτουν μεταβλητές που έχουν μικρή επίδραση, χωρίς να είναι απαραίτητα μηδενική, στο τελικό μοντέλο, χωρίς να είναι πολύ αυστηρές ως προς το μέγεθος της επιτρεπτής επίδρασης. Όπως φαίνεται και από τον πίνακα καμία τεχνική δεν είχε πρόβλημα να πετύχει το σωστό μοντέλο, σε κανένα από τα δείγματα.

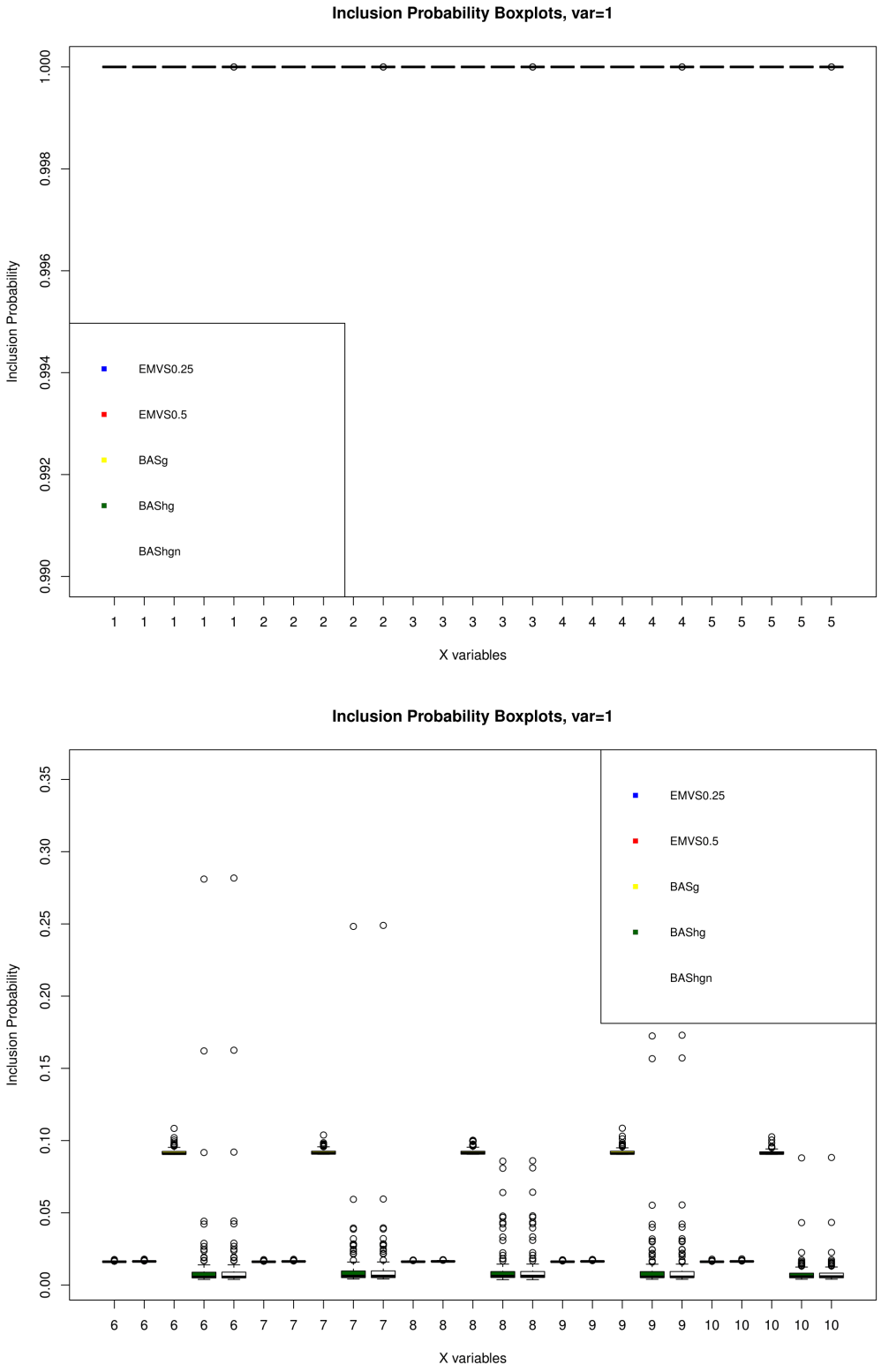
Πίνακας 6.1: Αποτελέσματα για $\sigma^2 = 1$.

Μέθοδος	Ποσοστό Επιτυχίας
EMVS, με $v_0 = 1/4$	100%
EMVS, με $v_0 = 1/2$	100%
BAS "g" prior	100%
BAS "hyper-g" prior	100%
BAS "hyper-g/n" prior	100%

Στο Διάγραμμα 6.1 φαίνονται θηκογράμματα τα οποία προκύπτουν από τις πιθανότητες ένταξης της κάθε επιμέρους μεταβλητής στο μοντέλο, για την κάθε τεχνική ξεχωριστά. Το πρώτο γράφημα δείχνει θηκογράμματα των μεταβλητών που όντως επίδρασαν σημαντικά στα y_i^k, ενώ το δεύτερο θηκογράμματα των υπόλοιπων μεταβλητών.

Όλες οι μέθοδοι δίνουν πιθανότητα ένταξης πολύ έντονα κεντραρισμένη στο 1 για τις μεταβλητές που χρησιμοποιήθηκαν κατά την προσομοίωση. Ακόμα το BAS με $gprior$ δίνει πιθανότητα ένταξης πιο απομακρυσμένη από το 0, σε σχέση με τις άλλες μεθόδους, για τις μεταβλητές που δεν είχαν σχέση με την προσομοίωση.

Διάγραμμα 6.1: Θηκογράμματα των ύστερων πιθανοτήτων ένταξης για τις επιμέρους μεταβλητές.

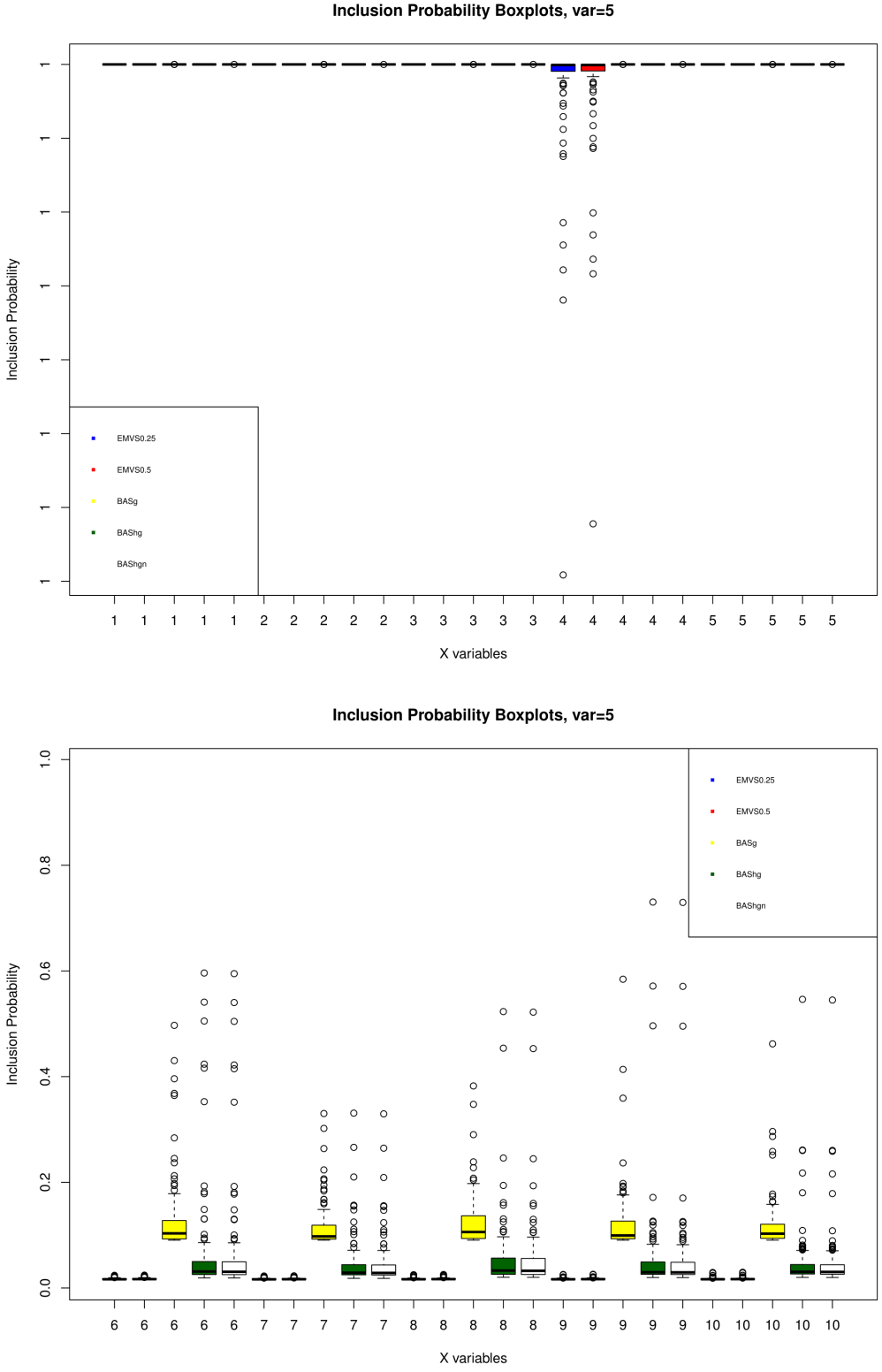


6.2 Εφαρμογή του EMVS σε δείγμα με κανονική διακύμανση

Κάνουμε ότι και στην προηγούμενη ενότητα, μόνο που τώρα η διακύμανση των y_i^k είναι ίση με 25. Έχει ενδιαφέρον να φανεί πως αντιδρούν οι μέθοδοι κάτω από μεγαλύτερη αβεβαιότητα. Παρακάτω

φαίνονται και πάλι τα αποτελέσματα.

Διάγραμμα 6.2: Θηκογράμματα των ύστερων πιθανοτήτων ένταξης για τις επιμέρους μεταβλητές.



Όπως φαίνεται και εδώ, και πάλι οι δύο τεχνικές δεν είχαν κάποιο θέμα στο να εκτιμήσουν τις σωστό αποτέλεσμα (Πίνακας 6.2). Αξίζει να σημειωθεί πως στη συγκεκριμένη περίπτωση το β έχει τιμές αρκετά απομακρυσμένες από το 0. Για χαμηλότερες τιμές κατά απόλυτη τιμή στο β, τα ποσοστά επιτυχίας σε όλες τις μεθόδους δεν θα ήταν τόσο υψηλά. Μόνη διαφορά που μπορεί

Πίνακας 6.2: Αποτελέσματα για $\sigma^2 = 25$.

Μέθοδος	Ποσοστό Επιτυχίας
EMVS, με $v_0 = 1/4$	100%
EMVS, με $v_0 = 1/2$	100%
BAS "g-prior" prior	100%
BAS "hyper-g" prior	100%
BAS "hyper-g-n" prior	100%

να παρατηρήσει κάποιος στα θηκογράμματα (Διάγραμμα 6.2) είναι πως εμφανίζονται περισσότερες ακραίες τιμές, που προφανώς και σχετίζεται με τη μεγαλύτερη διακύμανση.

6.3 Πολυσυγγραμικότητα και χαμηλή διακύμανση

Η κατάσταση η οποία δημιουργείται όταν υπάρχουν ισχυρές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών στην πολλαπλή παλινδρόμηση ονομάζεται πολυσυγγραμικότητα (*multicollinearity*) και είναι ένα συχνό πρόβλημα που εμφανίζεται σε εφαρμογές παλινδρομικών μοντέλων. Η πολυσυγγραμικότητα οδηγεί σε εκτιμήσεις ασταθών παραμέτρων, γεγονός που καθιστά πολύ δύσκολη την εκτίμηση της επίδρασης των ανεξάρτητων μεταβλητών.

Θα κάνουμε και πάλι ότι και στην πρώτη ενότητα, μόνο που τώρα θα προσθέσουμε πολυσυγγραμικότητα στον πίνακα σχεδιασμού, δηλαδή για την τελευταία στήλη του πίνακα σχεδιασμού,

$$x_{i10}^k = 2x_{i2}^k - 3x_{i3}^k + \epsilon_i^k,\ i,k \in \{1,2,...,100\}$$

$$\epsilon_i^k \sim N(0,25).$$

Τα ϵ_i^k είναι ανεξάρτητα ανά δύο. Στον Πίνακα 6.3 φαίνονται και πάλι τα αποτελέσματα των μεθόδων. Καμία μέθοδος δεν κατάφερε να δώσει σωστό αποτέλεσμα σε κανένα δείγμα. Τα αποτελέσματα στο Διάγραμμα 6.3 δεν έχουν κάποια άμεση σχέση με την προσομοίωση, και αυτή η σύγχυση οφείλεται στην πολυσυγγραμικότητα.

Πίνακας 6.3: Αποτελέσματα για $\sigma^2 = 1$ και πολυσυγγραμικότητα.

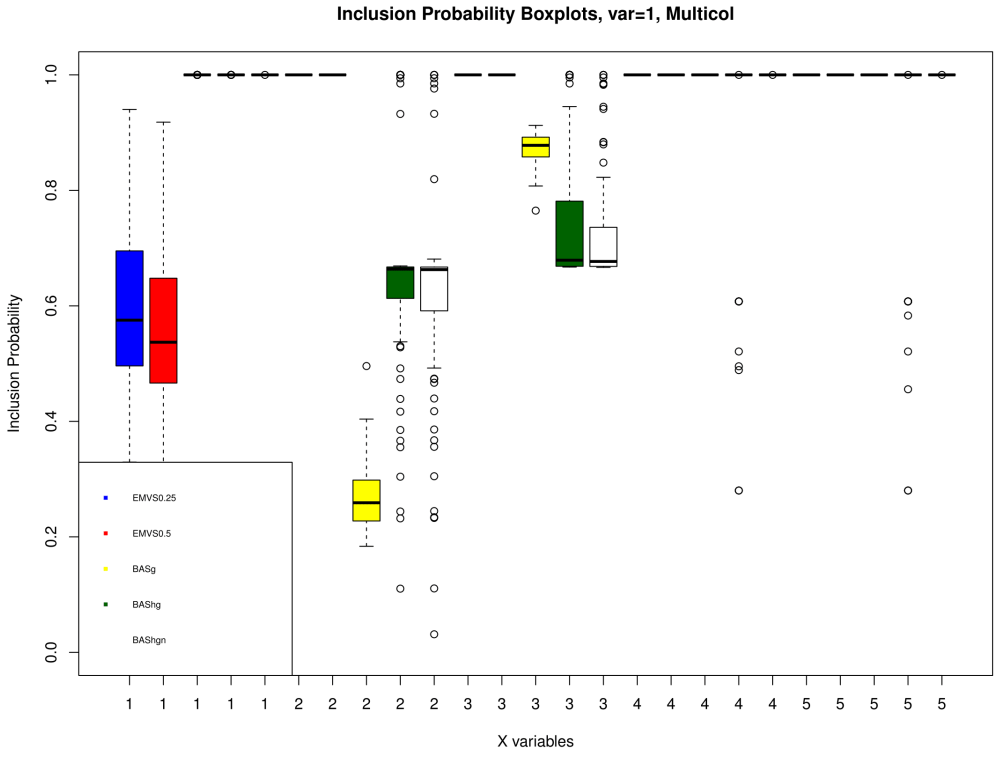
Μέθοδος	Ποσοστό Επιτυχίας
EMVS, με $v_0 = 1/4$	0%
EMVS, με $v_0 = 1/2$	0%
BAS "g-prior" prior	0%
BAS "hyper-g" prior	0%
BAS "hyper-g-n" prior	0%

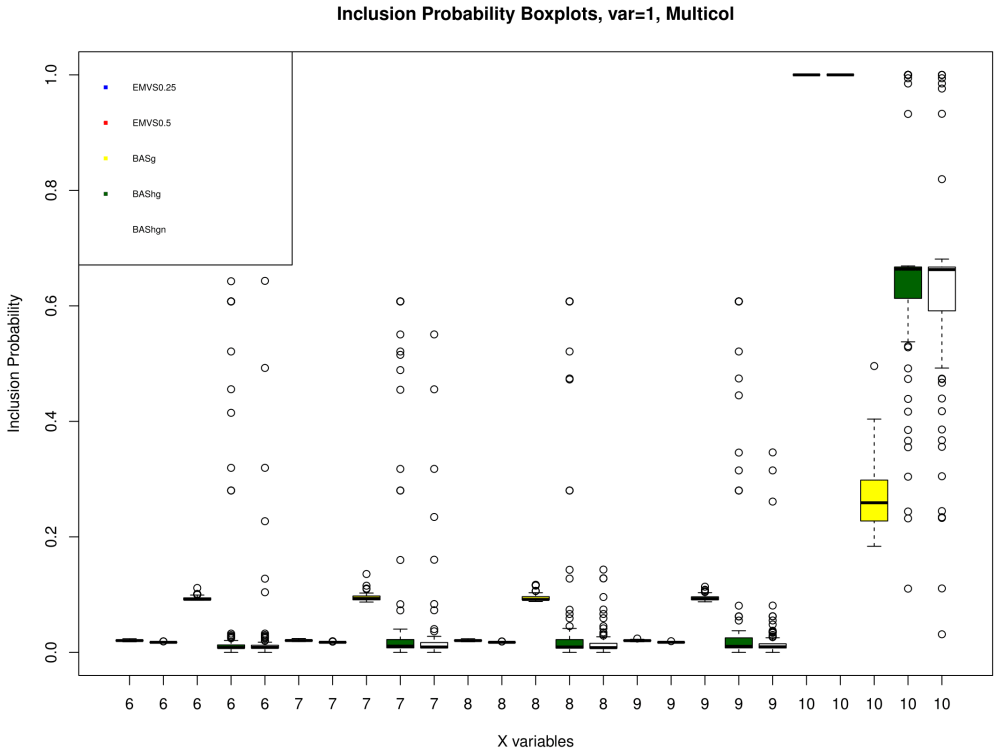
Τα «λάθη» στον $EMVS$, ανεξάρτητα του v_0, βρίσκονται στο ότι δεν απορρίπτεται πουθενά η πολυσυγγραμική μεταβλητή, επιπλέον κάποιες φορές απορρίπτεται και η μεταβλητή X_1. Στο BAS στην περίπτωση της $g\ prior$ ενώ απορρίπτεται η πολυσυγγραμική μεταβλητή, ταυτόχρονα απορρίπτεται και η X_2, ενώ στις περιπτώσεις των άλλων πρότερων είτε δεν απορρίπτεται η X_{10} είτε απορρίπτεται η X_2. Για την περίπτωση της $hyper-g$ παρατηρείται να απορρίπτονται και οι X_4 και X_5 κάποιες φορές.

6.4 Πολυσυγγραμικότητα και κανονική διακύμανση

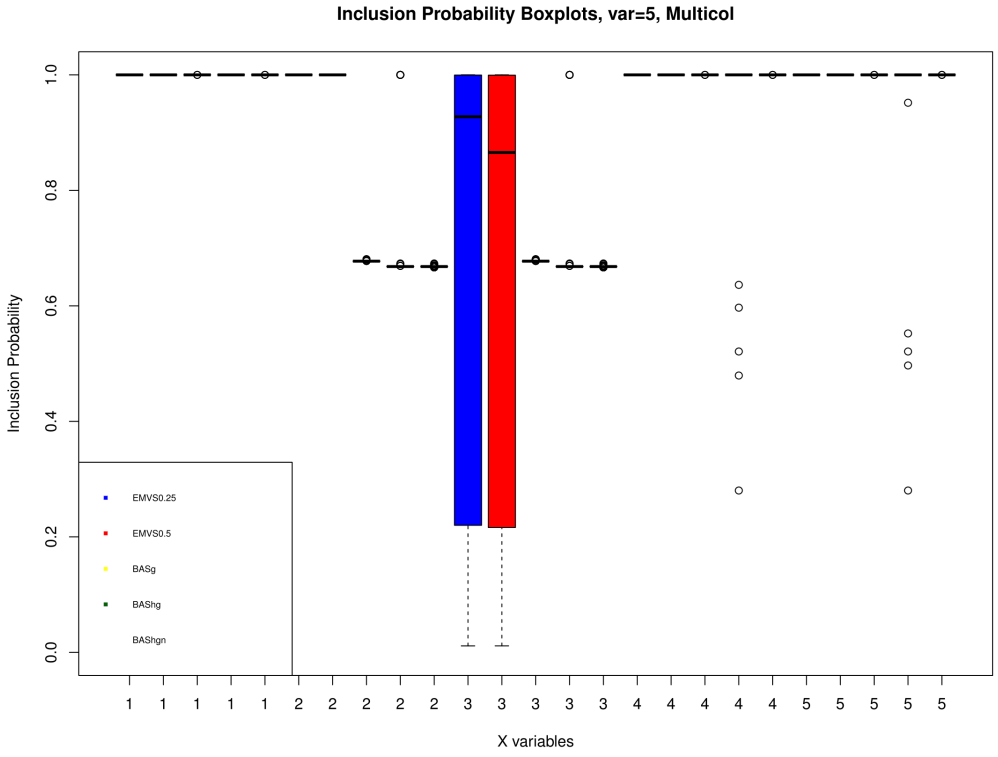
Κάνουμε ότι και στην προηγούμενη ενότητα, μόνο που τώρα η διακύμανση των y_i^k είναι ίση με 25. Παρακάτω φαίνονται τα ποσοστά επιτυχίας και τα θηκογράμματα.

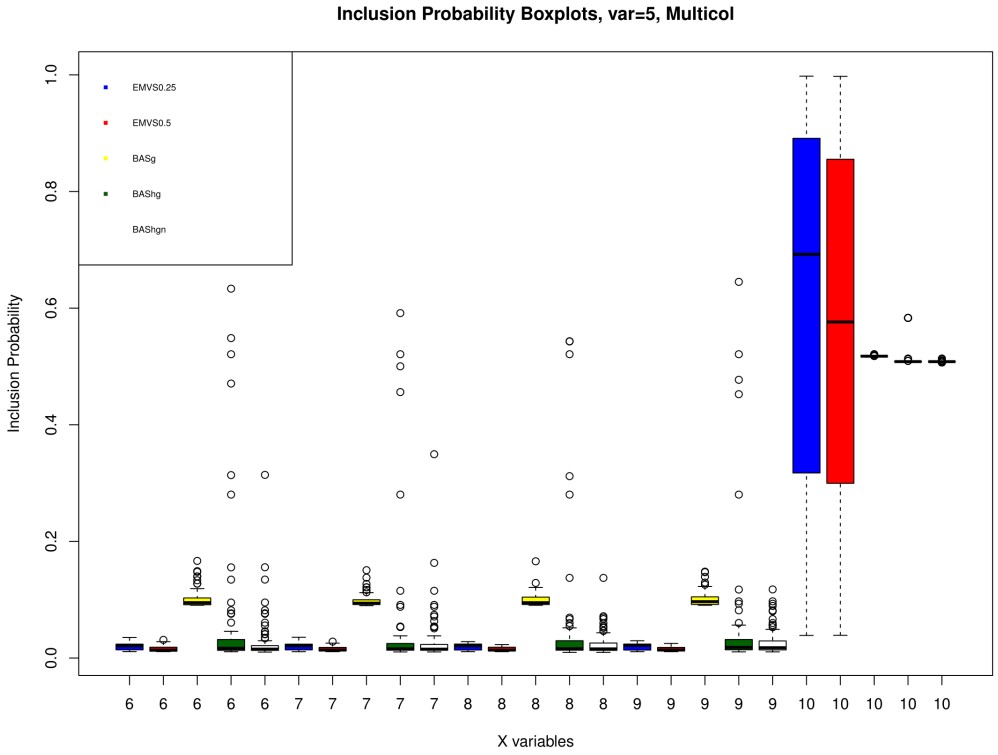
Διάγραμμα 6.3: Θηκογράμματα των ύστερων πιθανοτήτων ένταξης για τις επιμέρους μεταβλητές.





Διάγραμμα 6.4: Θηκογράμματα των ύστερων πιθανοτήτων ένταξης για τις επιμέρους μεταβλητές.





Πίνακας 6.4: Αποτελέσματα για $\sigma^2 = 25$ και πολυσυγγραμικότητα.

Μέθοδος	Ποσοστό Επιτυχίας
EMVS, με $v_0 = 1/4$	37%
EMVS, με $v_0 = 1/2$	42%
BAS "g-prior" prior	26%
BAS "hyper-g" prior	72%
BAS "hyper-g-n" prior	72%

Τα ποσοστά επιτυχίας τώρα, στον Πίνακα 6.4, είναι πολύ καλύτερα από οτι στην προηγούμενη ενότητα. Και εδώ τα «λάθη» για τις μεθόδους είναι πως είτε επιτρέπεται η πολυσυγγραμική μεταβλητή, είτε απορρίπτεται κάποια από τις ισχύουσες στο τελικό μοντέλο. Επιπλέον για την περίπτωση της $hyper - g$ συμβαίνει και να επιτρέπονται και άλλες μη ισχύουσες μεταβλητές από την πολυσυγγραμική. Το να επιτρέπεται η πολυσυγγραμική μεταβλητή συμβαίνει πιο συχνά από το να απορρίπτονται οι ισχύουσες.

Μεταφορά του EMVS στην R

Σε αυτήν την ενότητα δίνονται οι κώδικες που απαιτούνται για τις ανάγκες τις εργασίας στην γλώσσα προγραμματισμού R. Αρχικά δίνονται οι εντολές που φτιάχνουν μια συνάρτηση που εκτελεί την διαδικασία και επιστρέφει τα αποτελέσματα του $EMVS$ και υστέρα, μέσω της συνάρτησης αυτής, οι εντολές για τις οποίες φτιάχνονται τα δύο διαγράμματα που χρησιμοποιούνται στην συμπερασματολογία του $EMVS$. Τέλος δίνεται μια συνάρτηση η οποία συνδυάζει τις προηγούμενες δύο, δηλαδή φτιάχνει και τα δύο γραφήματα σε λιγότερο χρόνο. Και τα διαγράμματα είναι φτιαγμένα υπό μορφή συνάρτησης.

EMVS function

Η συνάρτηση δουλεύει με τις παρακάτω μεταβλητές. Στα Κεφάλαια 1 και 5 μπορούν να φανούν παραπάνω πληροφορίες σχετικά με την σημασία τους.

y="το διάνυσμα-δείγμα y"
X="ο πίνακας $n \times p$ σχεδιασμου X (υπενθυμίζεται χωρίς τη γραμμή με τις μονάδες)"
$v0$ = "η παράμετρος v_0 στην διακύμανση των βήτα "
$v1$ = "η παράμετρος v_1 στην διακύμανση των βήτα "
v = "η παράμετρος v της διακύμανσης δοθέντος των γάμμα "
l = "η παράμετρος λ της διακύμανσης δοθέντος των γάμμα "
a = "η παράμετρος α του ϑ"
b = "η παράμετρος β του ϑ"
$b0$ = "διάνυσμα αρχικών τιμών για τα βήτα στην επαναληπτική διαδικασία"
$s0$ = "αρχική τιμή για το σ στην επαναληπτική διαδικασία"
$theta0$ = "αρχική τιμή για το ϑ στην επαναληπτική διαδικασία"
$error$ = "η τιμή αυτή καθορίζει τον μέγιστο αριθμό μεταβολής μεταξύ των τιμών που προκύπτουν από την επαναληπτική διαδικασία και συνεπώς αν επιτεύχθηκε σύγκλιση ή όχι"
$type$ = "αν δοθεί η τιμή 1 και ισχύει πως $p > n$ χρησιμοποιείται η σχέση (5.11) για τα βήτα αλλιώς χρησιμοποείται η (5.10) "

```
##EMVS function

###libraries needed
library(expm)
library(MASS)
library(matlib)

norm_vec <- function(x){ sqrt(sum(x^2))}

EMVSf<-function(y="sample y", X="design matrix",
                v0="v0 parameter of beta",
                v1="v1 parameter of beta",
                v="v parameter of sigma^2 given gamma",
                l="lambda parameter of sigma^2 given gamma",
                a="a parameter of theta",
                b="beta parameter of theta",
                b0="starter parameter of beta",
                s0="starter parameter of sigma (sd)",
                theta0="starter parameter of theta",
                error="number that arranges if convergence
                          happened by comparing iterative
                          values absolute differences",
                type="special method for p>n")
{

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

beta<-b0      #### algorithm starting values
sigma<-s0
theta<-theta0

## not given information replaced by proposed
if(beta=="starter parameter of beta"){
beta<-matrix(nrow=p,ncol=1)
for(i in 1:p){beta[i]<-rnorm(1,0,1)}}

if(sigma=="starter parameter of sigma (sd)"){
  sigma<-runif(1,0,10^6)}

if(theta=="starter parameter of theta"){theta<-runif(1,0,1)}

if(error=="number that arranges if convergence happened
           by comparing iterative values absolute differences"){
  error<-10^(-10)}

if(type=="special method for p>n"){type<-0}

if(v0=="v0 parameter of beta"){v0<-0.001}
```

```
if(v1=="v1 parameter of beta"){v1<-1000}

if(v=="v parameter of sigma^2 given gamma" ){v<-1}
if(l=="lambda parameter of sigma^2 given gamma"){l<-1}

if(p<10^4){
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-1}
} else {
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-p}
 }


HM<-t(X)%*% X      ## hat matrix

maxdif<-error+1    ## needed to start the loop

if(p>n & type==1){    ## decision for the type
                      ## that is used on beta iterations

 while( maxdif>error ){   ## EMVS loop p>n
                           ## convergence inequality

 ##saving previous iterative values
 previousvalues<-c(beta,sigma,theta)

  pstar<-matrix(nrow=p,ncol=1)    ## p^*
  s1<-sigma*sqrt(v1)
  s0<-sigma*sqrt(v0)
  for(j in 1:p){
   pstar[j]<-(theta*dnorm(beta[j], mean = 0, sd =s1))/
                 (theta*dnorm(beta[j], mean = 0, sd =s1)
                   +(1-theta)*dnorm(beta[j], mean = 0, sd =s0))
   }

 Dstar<-matrix(0,nrow=p,ncol=p)      ## D^*
 for(j in 1:p){
  Dstar[j,j]<-pstar[j]/v1 +(1-pstar[j])/v0
  }

  IDstar<-ginv(Dstar)     ## beta calculation

##ginv function returns inverse matrix through numerical
##method (Moore Penrose)

  ADstar<-diag(n)+X %*% IDstar  %*% t(X)
  IADstar<-ginv(ADstar)
  beta<-( IDstar -  IDstar %*% t(X) %*% IADstar %*% X %*% IDstar )
           %*% t(X) %*% y
```

```
  sqrtDstar<-sqrtm(Dstar)  ## sigma calculation
  sumnorms<-norm_vec(y-X%*%beta) + norm_vec(sqrtDstar %*% beta)
  sigma<-sqrt( (sumnorms+v*l)/(n+p+v) )

  sumpstar<-0   ## theta calculation
  for(j in 1:p){sumpstar<-sumpstar+pstar[j]}
  theta<-(sumpstar+a-1)/(a+b+p-2)

  nowvalues<- c(beta,sigma,theta)
 dif<-vector(length=p+2)
 for(i in 1:(p+2)){dif[i]<-abs( nowvalues[i]-previousvalues[i])}
 ##iterative values absolute difference

 maxdif<-max(dif)  ## maximum of value of dif vector


 }
} else{

  while( maxdif>error ){      ## EMVS loop classic
                               ## convergence inequality

 previousvalues<-c(beta,sigma,theta) ##saving previous
                                      ##iterative values

  pstar<-matrix(nrow=p,ncol=1)      ## p^*
  s1<-sigma*sqrt(v1)
  s0<-sigma*sqrt(v0)
  for(j in 1:p){
    pstar[j]<-(theta*dnorm(beta[j], mean = 0, sd =s1))/
                (theta*dnorm(beta[j], mean = 0, sd =s1)
                  +(1-theta)*dnorm(beta[j], mean = 0, sd =s0))
  }

 Dstar<-matrix(0,nrow=p,ncol=p)      ## D^*
 for(j in 1:p){
  Dstar[j,j]<-pstar[j]/v1 +(1-pstar[j])/v0
  }

  HMDstar<-HM+Dstar    ## beta calculation
  IHMDstar<-ginv(HMDstar)

##ginv function returns inverse matrix through numerical
##method (Moore Penrose)

  beta<-IHMDstar %*% t(X) %*% y

  sqrtDstar<-sqrtm(Dstar)   ## sigma calculation
  sumnorms<-norm_vec(y-X%*%beta) + norm_vec(sqrtDstar %*% beta)
```

```
  sigma<-sqrt( (sumnorms+v*l)/(n+p+v) )

  sumpstar<-0    ## theta calculation
  for(j in 1:p){sumpstar<-sumpstar+pstar[j]}
  theta<-(sumpstar+a-1)/(a+b+p-2)


 nowvalues<- c(beta,sigma,theta)
 dif<-vector(length=p+2)
 for(i in 1:(p+2)){dif[i]<-abs(nowvalues[i]-previousvalues[i])}
 ##iterative values absolute difference

 maxdif<-max(dif)  ## maximum of value of dif vector

 }
}


gamma<-matrix(nrow=p, ncol=1)        ##gamma calculation
for(i in 1:p){
   pcrit<-(theta*dnorm(beta[i], mean = 0, sd =s1))/
            (theta*dnorm(beta[i], mean = 0, sd =s1)
              +(1-theta)*dnorm(beta[i], mean = 0, sd =s0))
  if( pcrit<0.5 ){
      gamma[i]<-0
     }else{
      gamma[i]<-1}
}


return<-c(beta,sigma,theta,gamma)

}
```

Για να τρέξει η συνάρτηση είναι απαραίτητο να δοθούν τουλάχιστον οι μεταβλητές $(\mathbf{y}, X)$. Σε περίπτωση που δεν δοθεί κάποια τιμή από τις υπόλοιπες μεταβλητές, συμπληρώνεται αυτόματα από το τα προτεινόμενα μεγέθη που αναφέρονται στο Κεφάλαιο 1. Στην παραπάνω περίπτωση, που υπάρχουν ελλιπείς τιμές, στις παραμέτρους δίνονται τέτοιες τιμές ώστε οι πρότερες κατανομές να είναι μη πληροφοριακές και οι αρχικές τιμές των ($\boldsymbol{\beta}$,σ,ϑ) προσομοιώνονται από την ομοιόμορφη κατανομή.

Ο αλγόριθμος σταματά όταν για κάθε τιμή των $\boldsymbol{\beta}$,σ,ϑ ισχύει πως η μεταβολή στην $k+1$ επανάληψη της επαναληπτικής διαδικασίας είναι μικρότερη του $error$. Δηλαδή ισχύει πως,

$$|\beta_i^{k+1} - \beta_i^k| < error \ \forall i \in \{1, 2, ..., p\},$$

$$|\sigma_i^{k+1} - \sigma_i^k| < error,$$

$$|\theta_i^{k+1} - \theta_i^k| < error.$$

Το αποτέλεσμα της συνάρτησης ϑα είναι το διάνυσμα $(\boldsymbol{\beta}^*, \theta^*, \sigma^*, \widetilde{\boldsymbol{\gamma}})$.

Plots

Παρακάτω δίνονται οι συναρτήσεις από τις οποίες προκύπτουν τα Regularization Plot και Log(g) Plot αντίστοιχα. Η εντολή επιστροφής τους επιστρέφει τα αντίστοιχα γραφήματα. Και τα δύο βασίζονται στην συνάρτηση $EMVSf$.

Οι συναρτήσεις δουλεύουν με τις παρακάτω μεταβλητές. Στα Κεφάλαια 1 και 5 μπορούν να φανούν παραπάνω πληροφορίες σχετικά με την σημασία τους.

y="το διάνυσμα-δείγμα y"
X="ο πίνακας $n \times p$ σχεδιασμου X (υπενθυμίζεται χωρίς τη γραμμή με τις μονάδες)"
$v0 =$ "η πρώτη παράμετρος v_0 στην διακύμανση των βήτα "
$step =$ " το βήμα με το οποίο εξελίσεται το v_0, δηλαδή το δεύτερο v_0 προκύπτει ως $v_0 + step$, το τρίτο ως $v_0 + 2 * step$ κοκ "
$nv0 =$ "ο αριθμός των v_0 για τα οποία προκύπτουν αποτελέσματα του $EMVS$ που φαίνονται στο γράφημα"
$v1 =$ "η παράμετρος v_1 στην διακύμανση των βήτα "
$v =$ "η παράμετρος v της διακύμανσης δοθέντος των γάμμα "
$l =$ "η παράμετρος λ της διακύμανσης δοθέντος των γάμμα "
$a =$ "η παράμετρος α του ϑ"
$b =$ "η παράμετρος β του ϑ"
$b0 =$ "αρχική τιμή για τα βήτα στην επαναληπτική διαδικασία"
$s0 =$ "αρχική τιμή για το σ στην επαναληπτική διαδικασία"
$theta0 =$ "αρχική τιμή για το ϑ στην επαναληπτική διαδικασία"
$error =$ "η τιμή αυτή καθορίζει τον μέγιστο αριθμό μεταβολής μεταξύ των τιμών που προκύπτουν από την επαναληπτική διαδικασία και συνεπώς αν επιτεύχθηκε σύγκλιση ή όχι"
$type =$ "αν δοθεί η τιμή 1 και ισχύει πως $p > n$ χρησιμοποιείται η σχέση (5.12) για τα βήτα αλλιώς χρησιμοποείται η (5.11) "

Και εδώ για να τρέξει η συνάρτηση είναι απαραίτητο να δοθούν τουλάχιστον οι μεταβλητές $(\mathbf{y}, X)$. Σε περίπτωση που δεν δοθεί κάποια τιμή από τις υπόλοιπες μεταβλητές, συμπληρώνεται αυτόματα από το τα προτεινόμενα μεγέθη που αναφέρονται στο Κεφάλαιο 1. Στην παραπάνω περίπτωση, που υπάρχουν ελλιπείς τιμές στην είσοδο της συνάρτησης, στις παραμέτρους δίνονται τέτοιες τιμές ώστε οι πρότερες κατανομές να είναι μη πληροφοριακές και οι αρχικές τιμές των ($\boldsymbol{\beta}$,σ,ϑ) προσομοιώνονται από την ομοιόμορφη κατανομή.

Στην συνάρτηση $RegPlot$ υπάρχει η επιπλέον μεταβλητή smv η οποιά φτιάχνει δύο κόκκινες ευθείες παράλληλες στον χ΄χ με τις τιμές $\pm smv$ στον άξονα στον $y'y$, πολύ κοντά στο 0. Αυτή η ευθεία χρησιμοποιείται για να φανεί πιο έντονα ποια βήτα είναι πολύ κοντά στο 0 καθώς αυξάνεται το v_0.

Καθώς εκτελούνται οι συναρτήσεις είναι πιθανό να εμφανιστεί η εντολή $warning$. Αυτό είναι λογικό, δεν υπάρχει πρόβλημα στην εκτέλεσή τους και οφείλεται στην αρχικοποίηση των βήτα.

```
##Regularization Plot



RegPlot<-function( y="sample y", X="design matrix",
                   v0="v0 parameter of beta starting value",
                   step="step of v0 sum", nv0="number of v0
                           steps",
                   v1="v1 parameter of beta",
                   v="v parameter of sigma^2 given gamma",
                   l="lambda parameter of sigma^2 given gamma",
                   a="a parameter of theta",
                   b="beta parameter of theta",
                   b0="starter parameter of beta",
                   s0="starter parameter of sigma (sd)",
                   theta0="starter parameter of theta",
                   error="number that arranges if convergence
                            happened by comparing iterative
                            values absolute differences",
                   type="special method for p>n",
                   smv="small value")

{

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

beta<-b0      #### algorithm starting values
sigma<-s0
theta<-theta0

## not given information replaced by proposed

if(beta=="starter parameter of beta"){
beta<-matrix(nrow=p,ncol=1)
for(i in 1:p){beta[i]<-rnorm(1,0,1)}}

if(error=="number that arranges if convergence
            happened by comparing iterative
            values absolute differences"){
  error<-10^(-10)}

if(sigma=="starter parameter of sigma (sd)"){
sigma<-runif(1,0,10^6)}

if(theta=="starter parameter of theta"){theta<-runif(1,0,1)}

if(type=="special method for p>n"){type<-0}
```

```
if(v0=="v0 parameter of beta starting value"){v0<-0.0001}
if(v1=="v1 parameter of beta"){v1<-1000}

if(v=="v parameter of sigma^2 given gamma" ){v<-1}
if(l=="lambda parameter of sigma^2 given gamma"){l<-1}

if(p<10^4){
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-1}
} else {
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-p}
 }

if(step=="step of v0 sum"){step <-0.026}
if(nv0=="number of v0 steps"){nv0 <-20}
if(smv=="small value"){smv<-0.01}

## convergence values matrix
convalues <- matrix( nrow =p, ncol = nv0 )
vecv0 <- vector ( length =nv0)

for (k in 1: nv0 ){ ## convergence values matrix calculation
vecv0[k]<-v0+(k-1)*step
calc<- vecv0[k]
z<- EMVSf(y=y, X=X, v0=calc, v1=v1, v=v, l=l, a=a, b0=beta,
            s0=sigma, theta0=theta, error=error, type=type)

for(l in 1:p){ convalues[l,k]<-z[l]}
}

smvmatrix<-vector(length=nv0)
for(i in 1:nv0){smvmatrix[i]<-smv}


xb1 <-c(v0 ,11*vecv0[nv0]/10+2*v0)
yb1 <-c( min( convalues ),max( convalues ))

return<-{plot (vecv0 , convalues[1 ,1:nv0], type ="o",
                pch =19 , col =" blue ",
                main =" Regularization Plot ", xlab =" v0",
                ylab =" betas ",xlim =xb1 , ylim = yb1 )
text(11*vecv0[nv0]/10+2*v0,convalues[1 ,nv0],"b1")
for (i in 2:p){
lines (vecv0 , convalues[i ,1:nv0 ], type ="o", pch =19 ,
         col =" blue ")
name<-paste("b",toString(i), sep="")
text(11*vecv0[nv0]/10+2*v0,convalues[i ,nv0],name)
lines (vecv0 , -smvmatrix, type ="l", col =" red ")
lines (vecv0 , smvmatrix, type ="l", col =" red ")}  } }
```

```
##Log(g) Plot



LoggPlot<-function( y="sample y", X="design matrix",
                    v0="v0 parameter of beta starting value",
                    step="step of v0 sum", nv0="number of v0
                            steps",
                    v1="v1 parameter of beta",
                    v="v parameter of sigma^2 given gamma",
                    l="lambda parameter of sigma^2 given gamma",
                    a="a parameter of theta",
                    b="beta parameter of theta",
                    b0="starter parameter of beta",
                    s0="starter parameter of sigma (sd)",
                    theta0="starter parameter of theta",
                    error="number that arranges if convergence
                             happened by comparing iterative
                             values absolute differences",
                    type="special method for p>n")
{


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

beta<-b0      #### algorithm starting values
sigma<-s0
theta<-theta0

## not given information replaced by proposed

if(beta=="starter parameter of beta"){
beta<-matrix(nrow=p,ncol=1)
for(i in 1:p){beta[i]<-rnorm(1,0,1)} }

if(sigma=="starter parameter of sigma (sd)"){
sigma<-runif(1,0,10^6)}

if(theta=="starter parameter of theta"){theta<-runif(1,0,1)}

if(error=="number that arranges if convergence
             happened by comparing iterative
             values absolute differences"){
  error<-10^(-10)}

if(type=="special method for p>n"){type<-0}

if(v0=="v0 parameter of beta starting value"){v0<-0.0001}
```

```
if(v1=="v1 parameter of beta"){v1<-1000}

if(v=="v parameter of sigma^2 given gamma" ){v<-1}
if(l=="lambda parameter of sigma^2 given gamma"){l<-1}

if(p<10^4){
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-1}
} else {
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-p}
 }

if(step=="step of v0 sum"){step <-0.026}
if(nv0=="number of v0 steps"){nv0 <-20}


## proportional probabilities vector
aprob <- vector( length = nv0 )

hatgamma <- vector( length =p ) ## gamma x iterations output
vecv0 <- vector( length =nv0) ## v0 step vector

for (k in 1: nv0 ){ ## hatgamma vs iterations matrix
vecv0[k]<-v0+(k-1)*step
calc<- vecv0[k]
 z<- EMVSf( y=y, X=X, v0=calc, v1=v1, v=v, l=l,
               a=a, b0=beta, s0=sigma, theta0=theta,
               error=error, type=type)

 for (i in 1:p){ hatgamma[i]<-z[i+p+2]}
 s <-0
 for(i in 1:p){s<-s+ hatgamma[i]}
 # tilde matrix formation

if(s!=0){    ##possible bug solution
 tildeX <-matrix( nrow =n, ncol =s)

 c<-0
 for (i in 1:p){
   if( hatgamma[i]==1){
   c<-c+1
   for (j in 1:n){ tildeX[j,c]<-X[j,i]}
  }
 }

 ## tilde lambda calculation
 RI<-t( tildeX ) %*% tildeX + diag(s)/ v1
 I<-ginv(RI)
 det<-det(I)
```

```
 tildematrix <- tildeX %*% I %*% t( tildeX )
 tildel <-v*l+t(y) %*% tildematrix %*% y +t(y) %*% y
 ## aprob calculation
 aprob[k]<-(beta(s+a,p-s+b)*gamma((v+s)/2)*(2*pi)^(-s/2))/
              (tildel^((v+s)/2)*v1^(s/2))*det^(1/2)
 }else{
 aprob[k]<-0}

}

aprob<-log(aprob)

xb2<-c(v0,(step*nv0+v0))
yb2<-c(min(aprob),max(aprob))

return<-plot(vecv0,aprob,type="o", pch=4, col="red",
      main="Log(g) Plot",
      xlab="v0",ylab="log(g)",xlim=xb2,ylim=yb2)

}
```

Στην συνάρτηση $LoggPlot$ υπάρχει ένα πρόβλημα. Σε περίπτωση που εμφανιστεί το φαινόμενο όλα τα $\widetilde{\gamma}_i$ να είναι ίσα με το μηδέν, λόγω των πράξεων με πίνακες η συνάρτηση δεν δίνει αποτέλεσμα. Για να λυθεί αυτό το θέμα τέθηκε πως σε αυτήν την περίπτωση η αντίστοιχη τιμή $aprob$ θα είναι ίση με το μηδέν, μιας και θα ήταν κάπως περίεργο να μην επιδρά καμία μεταβλητή στο μοντέλο και λόγω κλίμακας θα ήταν κάπως περίεργο να έχει υψηλή αναλογία πιθανότητας. Σε περίπτωση που στο γράφημα εμφανίζονται πολλές τιμές πολύ χαμηλά στον άξονα των $y'y$ ($log0,\ -\infty$) καλό θα ήταν να εξεταστεί σοβαρά το ενδεχόμενο να μην επιδρά καμία μεταβλητή στο μοντέλο.

Αν κανείς θέλει να φτιάξει και τα δύο προηγούμενα γραφήματα, προτείνεται η επόμενη συνάρτηση. Η παρακάτω συνάρτηση είναι ουσιαστικά μια μίξη των δύο προηγούμενων. Και αυτή στηρίζεται στην συνάρτηση $EMVSf$. Κύριο πλεονέκτημα της απέναντι στο να τρέξει κανείς τις δύο προηγούμενες, είναι ότι δίνει αποτελέσματα πολύ πιο γρήγορα.

Το ότι δίνει πιο γρήγορα αποτελέσματα βασίζεται στο γεγονός ότι τρέχοντας τις δύο προηγούμενες συναρτήσεις, για να πάρει κανείς αποτελέσματα χρειάζεται να πραγματοποιηθούν $2nv0$ συναρτήσεις $EMVSf$, $nv0$ για κάθε γράφημα. Η $both\_plots$ τρέχει μόνο $nv0$ φορές και χρησιμοποιεί τα ίδια αποτελέσματα σε κάθε γράφημα.

Οι εντολές εισόδου είναι οι αντίστοιχες με αυτές της σελίδας 57.

```
##Function that returns the two previous plots



both_plots<-function( y="sample y", X="design matrix",
                      v0="v0 parameter of beta starting value",
                      step="step of v0 sum", nv0="number of v0
                            steps",
                      v1="v1 parameter of beta",
                      v="v parameter of sigma^2 given gamma",
                      l="lambda parameter of sigma^2 given gamma",
                      a="a parameter of theta",
                      b="beta parameter of theta",
                      b0="starter parameter of beta",
                      s0="starter parameter of sigma (sd)",
                      theta0="starter parameter of theta",
                      error="number that arranges if convergence
                               happened by comparing iterative
                               values absolute differences",
                      type="special method for p>n",
                      smv="small value")
{


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


beta<-b0      #### algorithm starting values
sigma<-s0
theta<-theta0


## not given information replaced by proposed
if(beta=="starter parameter of beta"){
beta<-matrix(nrow=p,ncol=1)
for(i in 1:p){beta[i]<-rnorm(1,0,1)} }

if(sigma=="starter parameter of sigma (sd)"){
sigma<-runif(1,0,10^6)}

if(theta=="starter parameter of theta"){theta<-runif(1,0,1)}

if(error=="number that arranges if convergence happened
             by comparing iterative values absolute differences"){
  error<-10^(-10)}

if(type=="special method for p>n"){type<-0}
```

```
if(v0=="v0 parameter of beta starting value"){v0<-0.0001}
if(v1=="v1 parameter of beta"){v1<-1000}

if(v=="v parameter of sigma^2 given gamma" ){v<-1}
if(l=="lambda parameter of sigma^2 given gamma"){l<-1}

if(p<10^4){
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-1}
} else {
if(a=="a parameter of theta"){a<-1}
if(b=="beta parameter of theta"){b<-p}
 }

if(step=="step of v0 sum"){step <-0.026}
if(nv0=="number of v0 steps"){nv0 <-20}
if(smv=="small value"){smv<-0.01}

## convergence values matrix
convalues <- matrix( nrow =p, ncol = nv0 )

## proportional probabilities vector
aprob <- vector( length = nv0 )

hatgamma <- vector( length =p ) ## gamma x iterations output
vecv0 <- vector( length =nv0) ## v0 step vector

for (k in 1: nv0 ){        ## hatgamma vs iterations matrix
vecv0[k]<-v0+(k-1)*step
calc<- vecv0[k]
 z<- EMVSf(y=y, X=X, v0=calc, v1=v1, v=v, l=l, a=a, b0=beta,
             s0=sigma, theta0=theta, error=error, type=type)

for(l in 1:p){ convalues[l,k]<-z[l]}

for (i in 1:p){ hatgamma[i]<-z[i+p+2]}
 s <-0
 for(i in 1:p){s<-s+ hatgamma[i]}
 # tilde matrix formation

 if(s!=0){    ##possible bug solution
 tildeX <-matrix( nrow =n, ncol =s)

 c<-0
 for (i in 1:p){
   if( hatgamma[i]==1){
   c<-c+1
   for (j in 1:n){ tildeX[j,c]<-X[j,i]}
  }
 }
```

```
 ## tilde lambda calculation
 RI<-t( tildeX ) %*% tildeX + diag(s)/ v1
 I<-ginv(RI)
 det<-det(I)
 tildematrix <- tildeX %*% I %*% t( tildeX )
 tildel <-v*l+t(y) %*% tildematrix %*% y +t(y) %*% y
 ## aprob calculation
 aprob[k]<-(beta(s+a,p-s+b)*gamma((v+s)/2)*(2*pi)^(-s/2))/
               (tildel^((v+s)/2)*v1^(s/2))*det^(1/2)
 }else{
 aprob[k]<-0}

}


smvmatrix<-vector(length=nv0)
for(i in 1:nv0){smvmatrix[i]<-smv}

aprob<-log(aprob)

xb1 <-c(v0 ,11*vecv0[nv0]/10+2*v0)
yb1 <-c( min( convalues ),max( convalues ))

xb2<-c(v0,(step*nv0+v0))
yb2<-c(min(aprob),max(aprob))

p2<-plot(vecv0,aprob,type="o", pch=4, col="red",
      main="Log(g) Plot",
      xlab="v0",ylab="log(g)",xlim=xb2,ylim=yb2)


p1<-{plot (vecv0 , convalues[1 ,1:nv0], type ="o", pch =19 ,
   col =" blue ", main =" Regularization Plot ",
   xlab =" v0", ylab =" betas ",xlim =xb1, ylim = yb1 )
text(11*vecv0[nv0]/10,convalues[1 ,nv0],"b1")
for (i in 2:p){
lines (vecv0, convalues[i ,1:nv0 ], type ="o", pch =19 ,
    col =" blue ")
name<-paste("b",toString(i), sep="")
text(11*vecv0[nv0]/10,convalues[i ,nv0],name)
lines (vecv0 , -smvmatrix, type ="l", col =" red ")
lines (vecv0 , smvmatrix, type ="l", col =" red ")}
}

return<-list(p2,p1)


}
```

Βιβλιογραφία

A. P. Dempster, N. M. Laird and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, **39**, 1-38.

Boyles, R.A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society B*, **45**, 47-50.

Castillo, I., Van der Vaart, A (2012). Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences. *The Annals of Statistics*, **40**, 2069–2101.

David Madigan, Adrian E. Raftery (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, **89 (428)**, 1535-1546.

Edward I. George, Robert E. McCulloch (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**, 339-373.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, Jim O Berger (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, **103 (481)**, 410-423.

Geoffrey J. McLachlan, Thriyambakam Krishnan (1997). *The EM Algorithm and Extensions.* John Wiley and Sons , USA.

George E. P. Box (1976). Science and Statistics. *Journal of the American Statistical Association*, **71 (356)**, 791-799.

George E. P. Box, George C. Tjao (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, University of Wisconsin.

H. Jeffreys (1961). *The Theory of Probability (3rd ed.).* Clarendon Press, Oxford.

Hemant Ishwaran, J. Sunil Rao (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, **33**, 730-773.

Hoerl A.E., R.W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

James O. Berger, Luis R. Pericchi, Julia A. Varshavsky (1998). Bayes Factors and Marginal Distributions in Invariant Situations. *Sankhya : The Indian Journal of Statistics*, **60, Series A, Pt. 3**, 307-321.

Li F., Zhang N. R. (2010). Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics. *Journal of the American Statistical Association*, **105**, 1978–2002.

Michiko Watanabe, Kazunori Yamaguchi (2004). *The EM Algorithm and Related Statistical Models.* Marcel Dekker, USA.

Murray G.D. (1977). Contribution to the discussion of paper by A.P Dempster, N.M. Laird, and D.B. Rubin. *Journal of the Royal Statistical Society B*, **39**, 27-28.

Olcay Arslan, Patrick D. L. Constable, John T. Kent (1993). Domains of convergence for the EM algorithm: a cautionary tale in a location estimation problem. *Statistics and Computing*, **3 (3)**, 103–108.

R. A. Fisher (1925). *Statistical Methods for Research Workers, 14th edition.* Oliver and Boyd, Edinburgh.

Robert E. Kass, Adrian E. Raftery (1995). Bayes Factors. *American Statistical Association*, **90 (430)**, 773-795.

Stingo F., Vannucci M. (2011). Variable Selection for Discriminant Analysis With Markov Random Field Priors for the Analysis of Microarray Data. *Bioinformatics*, **27**, 495–501.

T. J. Mitchell, J. J. Beauchamp (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, **83 (404)**, 1023-1032.

Veronika Rockova, Edward I. George (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, **109 (506)**, 828-846.

Zellner A., Goel P. (1986). *Bayesian Inference and Decision Techniques.* Elsevier Science Ltd, Amsterdam.