



KAI

BAYESIAN MODEL AVERAGING

$\mu\mu$

μ

$\mu \mu$

μ

μ

μ

μ

$\mu \mu$

2008

μ

μ

μ

μ μ

μ

μ

,

μ.....

μ

μ

μ

μμ

μ μ

μ

:

- ()

-

-

μ

μ μ

μ

μ

μ

.

UNIVERSITY OF PIRAEUS



DEPARTMENT OF STATISTICS AND INSURANCE
SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

BAYESIAN MODEL AVERAGING

by

Emmanouel Parzakonis

MSc Dissertation

submitted to the Department of Statistics and Insurance

Science of the University of Piraeus in partial fulfillment of

the requirements for the degree of Master of Science in

Applied Statistics

Piraeus, Greece

February 2008

“Can't escape this line of best fit”

Death Cab for Cutie - Line of Best Fit

« » μ ... μ
б ...

Abstract

It is common to discard the uncertainty employed by the choice of a single model. This source of variation is highly significant when the purpose of the analysis is the prediction of a quantity of interest (e.g a future observation, the cost of a decision etc.). A way of dealing with the aforementioned problem is the use of weighted top (or all if possible) models among the class of the models the analyst considers. The bayesian paradigm deals with the problem naturally, by simply regarding the models as parameters (therefore assigning them a prior probability) and using the method of Bayesian Model Averaging.

This dissertation aims to present the theoretical background along with practical algorithms in the class of linear models.

Περιεχόμενα

Κατάλογος Πινάκων.....	xv
Κατάλογος Σχημάτων.....	xvii
Κατάλογος Συντομογραφιών.....	xix
1. Εισαγωγή στην αβεβαιότητα της επιλογής μοντέλου.....	1
1.1 Ανασκόπηση της (καθιερωμένης) στατιστικής πρακτικής.....	1
1.2 Προβλήματα των διαδικασιών επιλογής μοντέλου.....	2
1.2.1 Γενικά προβλήματα των διαδικασιών.....	2
1.2.2 Ζητήματα σχετικά με τη φιλοσοφία της επιλογής ενός μοντέλου.....	3
1.3 Επιλογές για την αντιμετώπιση του προβλήματος.....	6
2. BMA και ανασκόπηση του γραμμικού μοντέλου.....	9
2.1 Εισαγωγή στη διαδικασία Bayesian Model Averaging.....	9
2.2 Ανασκόπηση του (Μπεϋζιανού) Γραμμικού Μοντέλου.....	13
3. Υλοποίηση της BMA.....	29
3.1 Κύριες υλοποιήσεις.....	29
3.1.1 HRM υλοποίηση.....	29
3.1.2 Zellner υλοποίηση.....	35
3.2 Το κύριο πρόβλημα.....	36
3.3 Αλγόριθμοι υλοποίησης & Προσεγγιστικές διαδικασίες.....	37
3.3.1 Το Παράθυρο του Οκκάμ.....	38
3.3.2 BMA μέσω Πληροφοριακών Κριτηρίων.....	41
3.3.3 Στοχαστικές αλγοριθμικές μέθοδοι.....	45
3.4 Ένα ολοκληρωμένο παράδειγμα.....	49
4. Περαιτέρω θέματα στη BMA.....	57
4.1 Προσδιορισμός των w_i	57
4.2 Το w ως τυχαία μεταβλητή.....	59

4.3 Εμπειρική μπεϋζιανή εκτίμηση της υπερπαραμέτρου ϕ	62
4.3.1 Τοπικοί Εμπειρικοί εκτιμητές Bayes.....	63
4.3.2 Καθολικοί Εμπειρικοί εκτιμητές Bayes.....	64
4.4. Ένα ακόμα παράδοξο σχετικά με την υπερπαραμέτρο ϕ	67
4.5. Χρήση της εκ των πρότερον κατανομής των Zellner – Siow.....	59
5. Προεκτάσεις και σχολιασμός.....	77
5.1 Πέρα από το γραμμικό μοντέλο	77
5.2 Άλλες διαδικασίες Model Averaging.....	78
5.3 Συλλογισμοί σχετικά με την BMA.....	79
Παραρτήματα.....	81
Α Προσέγγιση με τη μέθοδο του Laplace.....	81
Β Λογισμικό.....	82
Γ Δεδομένα εφαρμογής.....	83
Περίληψη.....	
Abstract.....	
Βιβλιογραφία.....	85

Κατάλογος Πινάκων

Πίνακας 3-1	Οριακή συμπεριφορά σύμφωνα με τον Ορισμό 3.2.1
Πίνακας 3-2	Εύρεση αντιστοιχίας ποινής ενός GIC - εκ των προτέρων πιθανότητας (για γνωστό μέγεθος δείγματος)
Πίνακας 3-3	Αντιστοιχία ποινής – υπερπαραμέτρου της εκ των προτέρων κατανομής Zellner
Πίνακας 3-4	Εκ των υστέρων πιθανότητες συμπερίληψης σε ένα μοντέλο
Πίνακας 3-5	Χαρακτηριστικά των αλγορίθμων
Πίνακας 3-6	Σύγκριση του RMSE για τα μοντέλα HRM, Zellner καθώς και το επιλεγέν με τη stepwise διαδικασία
Πίνακας 3-7	Σημειακές εκτιμήσεις και διακυμάνσεις των ανεξάρτητων μεταβλητών
Πίνακας 3-8	Σημειακή εκτίμηση του μεγέθους των μοντέλων

Κατάλογος Σχημάτων

- Σχήμα 1-1 Πραγματικό ποσοστό κάλυψης για ένα δ.ε 95% ως συνάρτηση της παραμέτρου δ (επιλογή μοντέλου βάσει AIC)
- Σχήμα 2-1 Σύγκριση των κατανομών του BMA – εκτιμητή και του κλασσικού
- Σχήμα 2-2 Κατανομή του BMA - εκτιμητή δοθέντος $\beta_j \neq 0$
- Σχήμα 2-3 Εύρεση των υπερπαραμέτρων της εκ των προτέρων κατανομής της σ^2
- Σχήμα 3-1 Έλεγχος των περιορισμών (3.1.2)
- Σχήμα 3-2 Παράθυρο του Οκκάμ
- Σχήμα 3-3 Εκ των προτέρων πιθανότητες που υποθέτουν τα γνωστά Πληροφοριακά Κριτήρια
- Σχήμα 3-4 Πραγματικές και προβλεπόμενες τιμές για τα μοντέλα HRM, Zellner καθώς και το επιλεγέν με τη stepwise διαδικασία
- Σχήμα 3-5 Εκ των υστέρων κατανομές των συντελεστών
- Σχήμα 3-6 Εκ των υστέρων κατανομή του μεγέθους των μοντέλων που επιλέχθηκαν από τους αλγόριθμους
- Σχήμα 3-7 Σύγκριση των κατανομών των συντελεστών για τους αλγόριθμους RJMCMC/MC³
- Σχήμα 4-1 Προσδιορισμός των εκ των προτέρων πιθανοτήτων μέσω των p-values
- Σχήμα 4-2 Η Βήτα-Διωνυμική κατανομή
- Σχήμα 4-3 Εκ των υστέρων κατανομή του μεγέθους των μοντέλων για σταθερή/τυχαία επιλογή
- Σχήμα 4-4 Εκ των υστέρων κατανομές των συντελεστών για σταθερή/τυχαία επιλογή
- Σχήμα 4-5 Το «παράδοξο της πληροφορίας»
- Σχήμα 4-6 Πραγματικές και προβλεπόμενες τιμές υπό την εκ των προτέρων κατανομή των Zellner-Siow

Κατάλογος Συντομογραφιών

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BMA	Bayesian Model Averaging
EM	Expectation - Maximization
MC ³	Markov Chain Monte Carlo Model Composition
MCMC	Markov Chain Monte Carlo
RIC	Risk Inflation Criterion
RJMCMC	Reversible Jump Markov Chain Monte Carlo
RMSE	Root Mean Square Error
EB(G)	Global - Empirical Bayes
EB(L)	Local - Empirical Bayes

1.

Εισαγωγή στην αβεβαιότητα της επιλογής μοντέλου

Ξεκινώντας, κάνουμε μια ανασκόπηση μιας τυπικής εργασίας ενός στατιστικού, ή γενικότερα ενός ερευνητή, ο οποίος έρχεται αντιμέτωπος με την ανάλυση ενός συνόλου δεδομένων. Στη συνέχεια, δίνονται μερικά αποτελέσματα τα οποία αποκαλύπτουν ορισμένες αδυναμίες αυτής της διαδικασίας.

1.1 Ανασκόπηση της (καθιερωμένης) στατιστικής πρακτικής

Έπειτα από τη συλλογή των δεδομένων, ο ερευνητής θα ασχοληθεί με το δυσκολότερο μέρος της εργασίας του, την μοντελοποίηση. Λογικό είναι να έχει υπόψη του αρκετά μοντέλα παρακινούμενος από προηγούμενες μελέτες ή/και θεμελιωμένες θεωρίες. Όμως και τα ίδια τα δεδομένα μπορούν να υποδείξουν κατάλληλα μοντέλα. Το επόμενο βήμα είναι το σημαντικότερο, η επιλογή ενός μοντέλου. Αυτό το βήμα είναι μια απόφαση, με την στατιστική έννοια του όρου, δηλ. βασισμένο στην ελαχιστοποίηση κάποιας συνάρτησης ζημίας.

Στην περίπτωση των γραμμικών μοντέλων η απόφαση ισοδυναμεί με την επιλογή ενός διανύσματος $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_p)$, όπου $\hat{\gamma}_i = 1(\beta_i \neq 0)$ ενός υποσυνόλου το οποίο περιέχει $\hat{q}_\gamma = \sum \hat{\gamma}_i$ από τις p ανεξάρτητες μεταβλητές το οποίο ελαχιστοποιεί κάποιο κριτήριο. Ο λόγος που οδηγεί στην επιλογή ενός υποσυνόλου είναι το “*Variance – Bias trade-off*”. Καθώς αυξάνεται (αντ. ελαττώνεται) το \hat{q}_γ ανταλλάσσουμε την μειωμένη (αυξημένη) μεροληψία με την αυξημένη (μειωμένη) διακύμανση. Δεν θα επεκταθούμε περισσότερο στους λόγους που επιβάλλουν την επιλογή ενός μοντέλου (Massart, 2003).

Η λήψη της απόφασης μπορεί να εκφραστεί στην ελαχιστοποίηση ενός κριτηρίου. Διαδεδομένα κριτήρια επιλογής αποτελούν τα Κριτήρια Πληροφορίας (Konishi & Kitagawa, 2008) τα οποία είναι της μορφής

$$IC_n(q_\gamma) = 2 \log L - c_{n,q_\gamma}, \quad c_{n,q_\gamma} > 0$$

όπου η c_{n,q_γ} λειτουργεί ως ποινή για τη διάσταση του μοντέλου. Οι συχνότερες επιλογές είναι $2(q_\gamma + 1)$ ή $(q_\gamma + 1) \log n$. Η πρώτη επιλογή αντιστοιχεί στο AIC (Akaike Information Criterion)¹ και η δεύτερη στο BIC (Bayesian Information Criterion).

1.2 Προβλήματα των διαδικασιών επιλογής μοντέλου

Παρ' ό,τι οι διαδικασίες επιλογής μοντέλου αποτελούν τον «πυρήνα» της Εφαρμοσμένης Στατιστικής, δεν είναι απαλλαγμένες από προβλήματα, τα κυριότερα των οποίων παρουσιάζουμε στη συνέχεια.

1.2.1 Γενικά προβλήματα των διαδικασιών

Οι δύο κυριότερες επιθυμητές ιδιότητες μιας διαδικασίας επιλογής είναι η αποδοτικότητα (*efficiency*) και η συνέπεια (*consistency*).

Ορισμός 1.1 Μια διαδικασία επιλογής μοντέλου ονομάζεται αποδοτική όταν ο λόγος της συνάρτησης ζημίας στο επιλεγθέν μοντέλο προς τη συνάρτηση ζημίας στο πραγματικό μοντέλο τείνει στο 1, δηλ.

$$\frac{C_n(\hat{\gamma})}{C_n(\gamma^{TRUE})} \xrightarrow{p} 1,$$

όπου $C_n(\gamma)$ η συνάρτηση ζημίας για το μοντέλο γ .

Πρακτικά, η αποδοτικότητα εγγυάται την εύρεση και συμπερίληψη των ανεξάρτητων μεταβλητών οι οποίες περιλαμβάνονται στο αληθινό μοντέλο.

Ορισμός 1.2 Μια διαδικασία επιλογής μοντέλου ονομάζεται συνεπής όταν το επιλεγθέν μοντέλο τείνει σχεδόν βέβαια στο αληθινό, δηλ.

¹ Η' Another Information Criterion όπως το ονόμασε ο Akaike (1974).

$$\hat{\gamma} \xrightarrow{\sigma, \beta} \gamma^{TRUE},$$

καθώς $n \rightarrow \infty$. Εάν η σύγκλιση είναι κατά πιθανότητα τότε η διαδικασία ονομάζεται ασθενώς συνεπής (*weakly consistent*).

Μια συνεπής διαδικασία δεν περιλαμβάνει (οριακά) ανεξάρτητες μεταβλητές οι οποίες δεν περιλαμβάνονται στο αληθινό μοντέλο.

Το AIC χαρακτηρίζεται ως αποδοτικό, ενώ το BIC είναι συνεπές. Επομένως, επιθυμητό είναι να κατασκευαστεί μια επαρκής και συνεπής διαδικασία. Δυστυχώς, κάτι τέτοιο δεν είναι εφικτό (δες Claeskens & Hjort, 2008, και αναφορές εκεί). Το καλύτερο που μπορούμε να επιτύχουμε είναι να αναζητήσουμε ένα κριτήριο το οποίο ελαχιστοποιεί το αναμενόμενο αριθμό των ανεξάρτητων μεταβλητών οι οποίες δεν περιλαμβάνονται στο μοντέλο που επιλέγει το BIC. Αυτό το επιτυγχάνει το κριτήριο των Hannan & Quinn (1979)

$$HQ(c) = 2 \log L - c \log \log n, \quad c > 1$$

Δυστυχώς, δεν υπάρχει κανόνας για την επιλογή της σταθεράς c .

Τέλος, η δημοφιλέστερη τακτική εύρεσης του καλύτερου μοντέλου είναι η διαδικασία *stepwise* (βηματική), γνωστή και ως αλγόριθμος του Efroymson (Miller, 2002). Παρ' όλη τη δημοφιλία της έχει αρκετά μειονεκτήματα. Οι Altman & Andersen (1989) αναφέρουν ότι τα διαστήματα εμπιστοσύνης είναι υπερβολικά στενά. Εξ άλλου, οι Derksen & Keselman (1992) κατέληξαν ότι η συσχέτιση των ανεξάρτητων μεταβλητών και ο αριθμός τους είναι σημαντικοί παράγοντες οι οποίοι επηρεάζουν τη διαμόρφωση του τελικού, επιλεγέντος μοντέλου. Το μεγαλύτερο πρόβλημα είναι όμως οι διαδοχικοί έλεγχοι. Στα πλαίσια διαδοχικών ελέγχων μοντέλων (*sequential tests*) κάθε έλεγχος φέρει και το σφάλμα Τύπου I του προηγούμενου. Συνεπώς, για μια μακρά ακολουθία τέτοιου είδους ελέγχων, η κατάληξη μπορεί να είναι επισφαλής.

1.2.2 Ζητήματα σχετικά με τη φιλοσοφία της επιλογής ενός μοντέλου

Το σημαντικότερο πρόβλημα στην επιλογή ενός μοντέλου αφορά τη λογική με την οποία αντιμετωπίζεται η απόφαση της επιλογής του. Το στάδιο αυτό της ανάλυσης, αν και είναι καθοριστικό για την εξέλιξή της, δεν θεωρείται ως στάδιο!

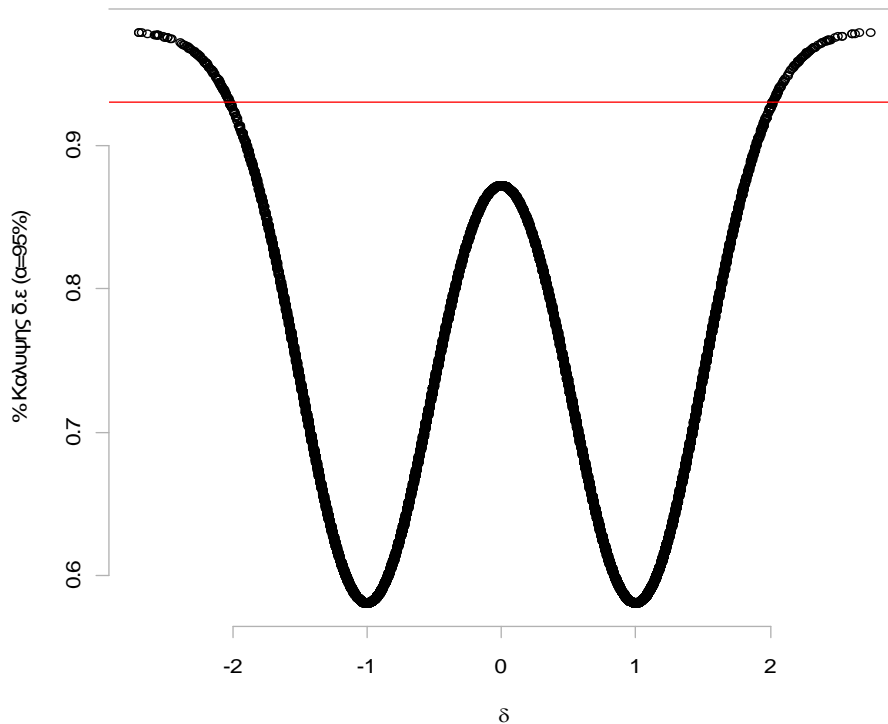
Στην πραγματικότητα το αποτέλεσμα της διαδικασίας θεωρείται ορθό και η ανάλυση συνεχίζεται ωσάν να μην αλλοιώθηκε ο «βαθμός» εγκυρότητας. Όμως, κάθε (στατιστική ή μη) απόφαση εμπεριέχει και αβεβαιότητα. Η καθημερινή στατιστική πρακτική αγνοεί όμως την αβεβαιότητα του μοντέλου (*model uncertainty*). Οι επιπτώσεις είναι σημαντικές καθώς όλα τα αποτελέσματα είναι έγκυρα δοθείσης της εγκυρότητας του επιλεγέντος μοντέλου. Συνεπώς, τα συμπεράσματα της τυπικής ανάλυσης θα αποτελούν μια αισιόδοξη εκτίμηση σε σχέση με την πραγματικότητα. Για παράδειγμα, το τυπικό 95% διάστημα εμπιστοσύνης για μια παράμετρο μ , είναι

$$\hat{\mu}_{\hat{\gamma}} \mp 1.96 \hat{\sigma}_{\hat{\gamma}}(\hat{\mu}_{\hat{\gamma}}) / \sqrt{n},$$

όπου $\hat{\sigma}_{\hat{\gamma}}(\hat{\mu}_{\hat{\gamma}})$ η εκτίμηση της τυπικής απόκλισης της εκτίμησης $\hat{\mu}_{\hat{\gamma}}$ και 1.96 το άνω 2.5% ποσοστιαίο σημείο της $N(0,1)$. Είναι προφανές ότι το διάστημα εμπιστοσύνης καθορίζεται από το μοντέλο το οποίο έχει επιλεγεί.

Γενικά υπό την επιλογή ενός μοντέλου ισχύουν τα επόμενα (Claeskens & Hjort, 2008): (α) η $\hat{\sigma}_{\hat{\gamma}}(\hat{\mu}_{\hat{\gamma}})$ υποεκτιμά την (πραγματική) $\sigma_{\hat{\gamma}}(\hat{\mu}_{\hat{\gamma}})$ (β) ο εκτιμητής $\hat{\mu}_{\hat{\gamma}}$ είναι σημαντικά μεροληπτικός, και (γ) η κατανομή του εκτιμητή $\hat{\mu}_{\hat{\gamma}}$ δεν είναι κανονική, όπως προβλέπει η θεωρία, αλλά αρκετά περίπλοκη (συνεπώς η τιμή 1.96 ενδέχεται να απέχει από την πραγματικότητα). Τα παραπάνω ισχύουν ακόμα και εάν για κάθε μοντέλο $\gamma \in \Gamma$ η εκτίμηση της τυπικής απόκλισης του $\hat{\mu}_{\gamma}$, $\hat{\sigma}_{\gamma}(\hat{\mu}_{\gamma})$, είναι ικανοποιητική ή εάν η $\sqrt{n}(\hat{\mu}_{\gamma} - \hat{\mu}_{TRUE})$ μπορεί να θεωρηθεί ότι κατανέμεται κανονικά, όπου $\hat{\mu}_{TRUE}$ η εκτίμηση υπό το αληθινό μοντέλο (Leeb, 2006, Potscher & Novak, 1998, Claeskens & Hjort, 2008).

Αντίστοιχα, και οι έλεγχοι υποθέσεων θα έχουν αυξημένο σφάλμα Τύπου I.



Σχήμα 1-1 : Πραγματικό ποσοστό κάλυψης για ένα δ.ε 95% ως συνάρτηση της παραμέτρου δ (επιλογή μοντέλου βάσει AIC)

Το Σχήμα 1-1 αναπαράγει ένα παράδειγμα από τους Claeskens & Hjort (2008) (σελ. 201). Για δύο υποψήφια μοντέλα, ένα «μικρό» (το οποίο περιέχει το διάνυσμα παραμέτρων θ) και ένα «μεγαλύτερο» (το οποίο περιέχει και μία επιπλέον ανεξάρτητη μεταβλητή με παράμετρο θ^*) ορίζουν το local misspecification model. Σύμφωνα με αυτό οι παρατηρήσεις της εξαρτημένης μεταβλητής προέρχονται από μια συνάρτηση πυκνότητας της μορφής $p(y | X, \theta_0, \theta_0^* + \delta / \sqrt{n})$ όπου θ_0^* είναι η πραγματική τιμή της θ^* και θ_0 είναι η πραγματική τιμή του διανύσματος θ . Ο όρος local misspecification αναφέρεται στον όρο δ / \sqrt{n} , ο οποίος ορίζει μια περιοχή γύρω από τη πραγματική τιμή της θ^* . Η διαδικασία επιλογής βασίστηκε στο AIC και για την απλή περίπτωση που περιγράψαμε παραπάνω οι Claeskens & Hjort (2008) υπολόγισαν το πραγματικό ποσοστό κάλυψης (*coverage probability*) ως συνάρτηση της δ για ένα τυπικό διάστημα εμπιστοσύνης 95%. Η απόκλιση από το ονομαστικό επίπεδο είναι σημαντική.

1.3 Επιλογές για την αντιμετώπιση του προβλήματος

Συνειδητοποιώντας τα προβλήματα της υιοθέτησης ενός μοντέλου για την συνέχεια της ανάλυσης, η προφανής λύση είναι να αποφύγουμε τη διαδικασία επιλογής χρησιμοποιώντας το πλήρες μοντέλο. Όμως, όλες οι μεταβλητές δεν θα είναι σημαντικές για τη διαμόρφωση της απόκρισης, όποτε θα μπορούσαμε να χρησιμοποιήσουμε μεθόδους συρρίκνωσης (*shrinkage*) καθώς και παλινδρόμηση ridge, η οποία κατ' ουσία είναι μία μέθοδος συρρίκνωσης των παραμέτρων του μοντέλου.

Η συρρίκνωση ενός εκτιμητή $\hat{\mu}$ ορίζεται στην πιο απλή περίπτωση² ως

$$\hat{\mu}^s = (1 - \delta)\hat{\mu}, \quad \delta \in [0,1).$$

Ο $\hat{\mu}^s$ δεν είναι αμερόληπτος (όπως θεωρητικά είναι ο $\hat{\mu}$) αλλά έχει μικρότερη διακύμανση σε σχέση με τον εκτιμητή $\hat{\mu}$. Υπάρχουν τιμές της δ τέτοιες ώστε το μέσο τετραγωνικό σφάλμα να είναι μικρότερο σε σχέση με του αμερόληπτου εκτιμητή $\hat{\mu}$ (Claeskens & Hjort, 2008). Ενδεικτικά αναφέρουμε την κλάση εκτιμητών συρρίκνωσης των James & Stein

$$\hat{\beta}^s = \left[1 - \frac{p-2}{\hat{\beta}^T X^T X \hat{\beta}} \sigma^2 \right] \hat{\beta}.$$

Αντίστοιχα, η παλινδρόμηση ridge εκτιμά τις παραμέτρους ως

$$\hat{\beta}^R = (X^T X + \delta I_p)^{-1} X^T y,$$

όπου I_p ο $p \times p$ ταυτοτικός πίνακας (η πρώτη στήλη 1,1,...1 αφαιρείται από τον X). Ο

$\hat{\beta}^R$ είναι επίσης μεροληπτικός εκτιμητής του β · παρ' όλ' αυτά υπάρχουν τιμές της

² Για μια ταξινόμηση που περιλαμβάνει 57 εκτιμητές δες Dempster (1977)

παραμέτρου δ τέτοιες ώστε το μέσο τετραγωνικό σφάλμα των $\hat{\beta}^R$ να είναι μικρότερο σε σχέση με του αμερόληπτου εκτιμητή $\hat{\beta}$.

Στην παρούσα εργασία θα περιγράψει μια εναλλακτική διαδικασία η οποία συνδυάζει τόσο τη συρρίκνωση των συντελεστών όσο και την εκτίμηση μέσω παλινδρόμησης ridge ενώ αντιμετωπίζει δομικά και την έννοια της αβεβαιότητας του μοντέλου.

2.

ΒΜΑ και ανασκόπηση του γραμμικού μοντέλου

Στο τέλος του προηγούμενου κεφαλαίου αναφέραμε δύο πιθανές λύσεις. Και οι δύο περιελάμβαναν τη χρήση όλων των ανεξάρτητων μεταβλητών. Επίσης, θα ήταν δυνατό να χρησιμοποιήσουμε και περισσότερα του ενός μοντέλα. Στο παρόν κεφάλαιο θα παρουσιαστούν τα βασικά σημεία της προτεινόμενης διαδικασίας καθώς και μια κατάλληλα προσαρμοσμένη για τη διαδικασία εισαγωγή στο μπεϋζιανό γραμμικό μοντέλο.

2.1 Εισαγωγή στη διαδικασία Bayesian Model Averaging

Έστω ότι η ποσότητα ω είναι η ποσότητα ενδιαφέροντος μας. Εναλλακτικά της χρήσης μεθόδων συρρίκνωσης και παλινδρόμησης ridge μέρος της έρευνας στράφηκε στη μελέτη της χρήσης της ποσότητας,

$$E(\omega | Y) = \sum_{\gamma} E(\omega | Y, \gamma) r_{\gamma}, \quad (2.1.1)$$

για την εκτίμηση της αναμενόμενης τιμής της ω δοθέντων των δεδομένων $Y = y$. Η (2.1.1) σταθμίζει την πρόβλεψη $E(\omega | Y, \gamma)$ για κάθε μοντέλο μέσω του διανύσματος $r = (r_1, r_2, \dots, r_{|\Gamma|})$, όπου $|\Gamma|$ ο πληθάριθμος του συνόλου Γ . Η ποσότητα ω είναι δυνατό να είναι μια μελλοντική παρατήρηση, ένας συντελεστής ενός γραμμικού μοντέλου κ.ο.κ.

Η επιλογή του διανύσματος r είναι αυτή που διαφοροποιεί τα αποτελέσματα των διαφόρων μεθόδων. Στην πράξη (δες Burnham & Anderson, 2003), οι ποσότητες $r_1, r_2, \dots, r_{|\Gamma|}$ αντιμετωπίζονται ως ψευδο-πιθανότητες. Η μπεϋζιανή λογική μάς προσφέρει θεωρητική θεμελίωση στο πρόβλημα αποφεύγοντας ad hoc λύσεις που προτείνει η κλασσική Στατιστική. Συγκεκριμένα, θα χρησιμοποιήσουμε την εκ των

προτέρων αξιολόγηση των μοντέλων καθώς και την εκ των υστερών κατανομή τους για να υπολογίσουμε τα $r_1, r_2, \dots, r_{|E|}$.

Ο λόγος που κάνει την μπεϋζιανή προσέγγιση φυσική στο πρόβλημα μας είναι η δυνατότητα θεώρησης των μοντέλων ως τυχαίων μεταβλητών. Έτσι, εφοδιάζονται με μια συνάρτηση πιθανότητας, με συνέπεια να μπορούμε να χρησιμοποιήσουμε τα δεδομένα μας για να «αναβαθμίσουμε» την αρχική μας αξιολόγηση. Αυτή η «αναβάθμιση», δηλαδή οι εκ των υστερών πιθανότητες, θα χρησιμοποιηθούν ως στάθμες στην (2.1.1).

Η διαδικασία αυτή ονομάζεται Bayesian Model Averaging (BMA). Για να υλοποιήσουμε την εν λόγω διαδικασία χρειάζεται να ορίσουμε τις εξής εκ των προτέρων κατανομές: Μία εκ των προτέρων κατανομή για τα μοντέλα, $\pi(\gamma)$, και μία εκ των προτέρων κατανομή για κάθε παραμετρικό διάνυσμα δοθέντος του μοντέλου γ , $\pi(\theta_\gamma | \gamma)$.

Τώρα, είμαστε έτοιμοι να υπολογίσουμε την εκ των υστερών πιθανότητα κάθε μοντέλου. Από την πιθανόφανεια του μοντέλου παίρνουμε την ολοκληρωμένη πιθανοφάνεια ή εκ των προτέρων κατανομή πρόβλεψης για το μοντέλο γ ,

$$p(y | \gamma) = \int_{\Theta_\gamma} p(y | \theta_\gamma, \gamma) \pi(\theta_\gamma | \gamma) d\theta_\gamma.$$

Επίσης, από απλές σχέσεις δεσμευμένων κατανομών έχουμε

$$p(y | \gamma) = p(\gamma | y) \cdot \frac{p(y)}{\pi(\gamma)}$$

και τελικά,

$$\pi(\gamma | y) = \frac{p(y | \gamma) \pi(\gamma)}{\sum_\gamma p(y | \gamma) \pi(\gamma)}. \quad (2.1.2)$$

Έστω ότι η ποσότητα ω είναι μια μελλοντική παρατήρηση. Από την (2.1.2) είναι άμεσος ο τρόπος με τον οποίο θα εκτιμήσουμε την ω . Αυτό θα γίνει με την (αδέσμευτη ως προς το μοντέλο) αναμενόμενη τιμή

$$\begin{aligned} E(\omega | Y) &= \sum_{\gamma} E(\omega | Y, \gamma) p(\gamma | Y) \\ &= E_{\Gamma} \{E(\omega | Y, \gamma)\}. \end{aligned} \quad (2.1.3)$$

Η πρόβλεψη αυτή είναι βέλτιστη με την έννοια του τετραγωνικού σφάλματος, και συνδυάζει τις επιμέρους βέλτιστες (επίσης με την έννοια του τετραγωνικού σφάλματος) προβλέψεις κάθε μοντέλου.

Στην πραγματικότητα, έχουμε περιγράψει πλήρως τον μηχανισμό της διαδικασίας. Η σχέση (2.1.2) θα μας παράσχει τις εκ των υστέρων πιθανότητες και η (2.1.3) θα αθροίσει για όλα τα μοντέλα τις προβλέψεις λαμβάνοντας υπόψη τις πιθανότητές τους. Όπως θα δούμε αργότερα, αυτή η απλότητα θα έχει το αντιστάθμισμα της!

Ας δούμε τώρα δύο από τα σημαντικότερα επιχειρήματα που διαθέτει η εν λόγω διαδικασία για την υιοθέτησή της στην πράξη.

Πρόταση 2.1.1. (*Βέλτιστη Κατανομή Πρόβλεψης*) Εάν αξιολογήσουμε την προβλεπτική ικανότητα της περιγραφόμενης διαδικασίας με το λογαριθμικό κανόνα του Good (δες τον Ορισμό 2.1.1) τότε ισχύει

$$E \left\{ \log \left[\sum_{\gamma} p(\omega | y, \gamma) \pi(\gamma | y) \right] \right\} \geq E \left\{ \log [g(\omega | y)] \right\}$$

για κάθε πυκνότητα πιθανότητας $g(\omega | y)$ με την αναμενόμενη τιμή να λαμβάνεται ως προς την $\sum_{\gamma} p(\omega | y, \gamma) \pi(\gamma | y)$.

Πρόταση 2.1.2. (*Ανάλυση της Διακύμανσης*) Η διακύμανση της ποσότητας ω μετά τη δειγματοληψία δίνεται από τη σχέση

$$Var(\omega | y) = E_{\Gamma} \{Var(\omega | y, \gamma)\} + Var_{\Gamma} \{E(\omega | y, \gamma)\} \quad (2.1.4)$$

Θα δείξουμε ότι αληθεύει η Πρόταση 2.1.1. Αρχικά θα χρειαστούμε τον επόμενο ορισμό.

Ορισμός 2.1.1 Ο λογαριθμικός κανόνας του Good (*Predictive Log Score*) για ένα μοντέλο $\gamma \in \Gamma$ ορίζεται ως

$$PLS_{\gamma} = -\sum_{y \in Y_v} \log p(y | \gamma, y_b),$$

όπου $Y_b = y_b$, για να καθορίσουμε το μοντέλο μας και τα υπόλοιπα, $Y_v = y_v$, για να αξιολογήσουμε την απόδοση του.

Για τη BMA ο λογαριθμικός κανόνας είναι

$$PLS_{BMA} = -\sum_{y \in Y_v} \log \left[\sum_{\gamma} p(y | \gamma, y_b) \pi(\gamma | y_b) \right].$$

Λαμβάνοντας την αναμενόμενη τιμή της διαφοράς των PLS_{BMA} και PLS_{γ} ως προς την κατανομή πρόβλεψης του BMA, δηλαδή την $\sum_{\gamma} p(y_{pred} | y_b, \gamma) \pi(\gamma | y_b)$,

παίρνουμε

$$E(PLS_{\gamma} - PLS_{BMA}) = -\sum_{y \in Y_v} E \left\{ \log \left[\frac{\sum_{\gamma} p(y | \gamma, y_b) \pi(\gamma | y_b)}{p(y | \gamma, y_b)} \right] \right\}.$$

Η αναμενόμενη τιμή στο δεξιό μέλος είναι η απόκλιση Kullback-Leibler ανάμεσα στην κατανομή πρόβλεψης του BMA και του μοντέλου γ . Συνεπώς, από τις ιδιότητες της απόκλισης Kullback-Leibler¹ έχουμε

¹ Η απόκλιση Kullback-Leibler δύο κατανομών P, Q μιας συνεχούς τ.μ X ορίζεται ως

$$D^{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx.$$

Οι κυριότερες ιδιότητες της απόκλισης Kullback-Leibler είναι οι επόμενες :

- $D^{KL}(P \parallel Q) \geq 0$
- $D^{KL}(P \parallel Q) = 0 \Leftrightarrow P = Q$
- $D^{KL}(P \parallel Q) \neq D^{KL}(Q \parallel P)$

$$E(PLS_{\gamma} - PLS_{BMA}) \geq 0 \Rightarrow \\ E(PLS_{\gamma}) \geq E(PLS_{BMA}), \forall \gamma \in \Gamma$$

Η (2.1.4) είναι γραμμένη στη γνωστή μορφή “διακύμανση μεταξύ” και “διακύμανση εντός” των μοντέλων (Draper, 1995). Ο πρώτος όρος του αθροίσματος είναι η συνήθης εκτίμηση της διακύμανσης όταν υιοθετηθεί μια διαδικασία επιλογής μοντέλου· είναι η αναμενομένη δεσμευμένη (ως προς το μοντέλο) αβεβαιότητα. Ο δεύτερος όρος είναι η διακύμανση η οποία οφείλεται στην αβεβαιότητα για την επιλογή ενός μοντέλου καθώς και στη διαφορετική δομή κάθε γ όπως αυτή εκφράζεται στην $E(\omega | \gamma, y)$. Συνάγεται ότι η διακύμανση υπό τη συνήθη διαδικασία επιλογής ενός μοντέλου είναι μια αισιόδοξη εκτίμηση της πραγματικής καθώς αγνοεί τη μεταβλητότητα μεταξύ των μοντέλων.

2.2 Ανασκόπηση του (Μπεϋζιανού) Γραμμικού Μοντέλου

Τα παραπάνω ισχύουν για γενικές κλάσεις μοντέλων ή ακόμα και για μοντέλα από διαφορετικές κλάσεις μοντέλων (για μερικές αναφορές δεξ κεφ. 5). Στην παρούσα εργασία θα παρουσιαστεί η εφαρμογή της μεθόδου στο γραμμικό μοντέλο.

Το πλήρες μοντέλο είναι το $Y_i = \alpha + \sum X_j \beta_j + \varepsilon_i$ και $Y = (Y_1, Y_2, \dots, Y_n)$ το διάνυσμα των n παρατηρήσεων της εξαρτημένης μεταβλητής και $X_i = (X_1, X_2, \dots, X_j, \dots, X_p)$ οι τιμές των εξαρτημένων μεταβλητών για την i παρατήρηση. Επίσης, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ το διάνυσμα των σφαλμάτων με ε_i ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που κατανέμονται σύμφωνα με τη $N(0, \sigma^2)$. Σε μορφή πινάκων μπορούμε να γράψουμε $Y = \alpha + X\beta + \varepsilon$.

Συνεπώς, στο γραμμικό μοντέλο το παραμετρικό διάνυσμα είναι $(\alpha, \beta, \sigma^2)$. Έστω $\theta = (\alpha, \beta)$ το διάνυσμα το οποίο περιέχει και τη σταθερά.²

Σύμφωνα με τη μπεϋζιανή άποψη, κάποιο μοντέλο «γεννιέται» από την εκ των προτέρων κατανομή,

$$\gamma \sim \pi(\gamma)$$

και στη συνέχεια, δοθέντος του μοντέλου γ , «γεννιούνται» οι παράμετροι

$$\begin{aligned} \sigma^2 | \gamma &\sim \pi(\sigma^2 | \gamma) \\ \theta_\gamma | \sigma^2, \gamma &\sim \pi(\theta_\gamma | \sigma^2, \gamma). \end{aligned}$$

Τέλος, τα δεδομένα προέρχονται από το δεσμευμένο μοντέλο

$$Y | \theta_\gamma, \sigma^2, \gamma \sim N(X_\gamma \theta_\gamma, \sigma^2 I_{q_\gamma+1}),$$

όπου X_γ ο πίνακας ο οποίος έχει πρώτη στήλη το μήκους n διάνυσμα $(1, 1, \dots, 1)^T$ και τις ανεξάρτητες μεταβλητές για τις οποίες $\gamma_i = 1$ και $I_{q_\gamma+1} = \text{diag}\{1, 1, \dots, 1\}$ είναι διάστασης $(q_\gamma + 1) \times (q_\gamma + 1)$.

Γράφοντας την από κοινού κατανομή των $Y, \theta_\gamma, \sigma^2, \gamma$ ως

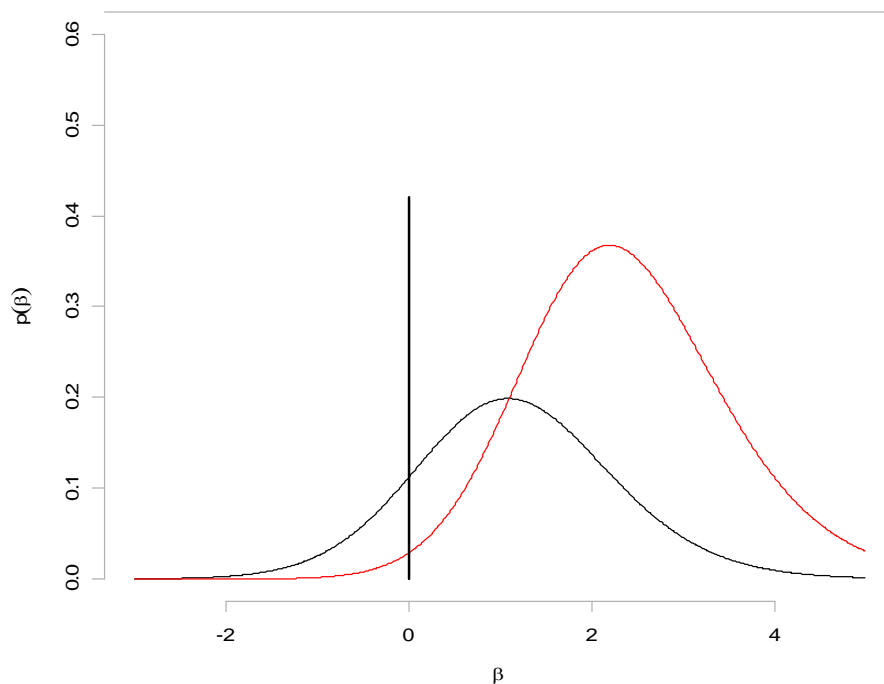
$$p(Y, \theta_\gamma, \sigma^2, \gamma) = p(Y | \theta_\gamma, \sigma^2, \gamma) \cdot [\pi(\theta_\gamma | \sigma^2, \gamma) \pi(\sigma^2 | \gamma)] \cdot \pi(\gamma) \quad (2.2.1)$$

ξεκινώντας από τα δεξιά καταλήγουμε στην παραπάνω περιγραφή «γέννησης».

Επίσης, αν και ευθύ αποτέλεσμα, αναφέρουμε την περίπτωση στην οποία η ποσότητα ω είναι ένας συντελεστής του μοντέλου β_j . Η εκ των υστερών κατανομή του

$$p(\beta_j | Y) = \sum_\gamma p(\beta_j | Y, \gamma) \pi(\gamma | Y)$$

² Όπως θα δούμε αργότερα θα μεταβαίνουμε μεταξύ θ και β ανάλογα με το μοντέλο που θα αναφερόμαστε.



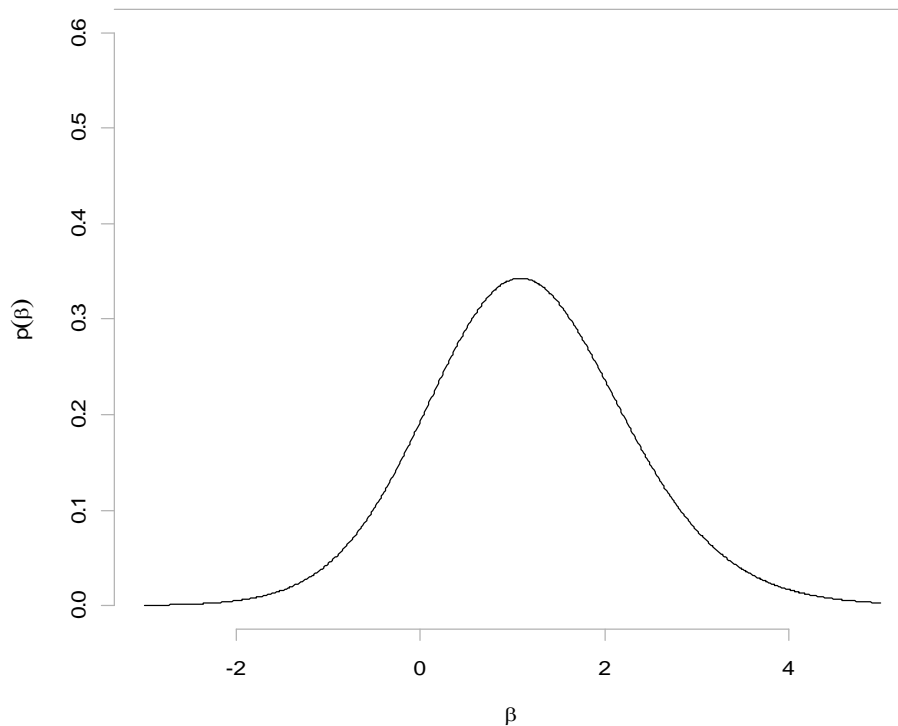
Σχήμα 2-1 Σύγκριση των κατανομών του BMA – εκτιμητή και του κλασσικού

Βλέπουμε ότι η κατανομή του είναι μικτή, θα έχει μάζα πιθανότητας στο 0 και ένα τμήμα της κατανέμεται συνεχώς. (Για αναλυτικά μοντέλα δεξ κεφ. 3).

Η αναμενόμενη τιμή του είναι

$$E(\beta_j | Y) = \sum_{\gamma} E(\beta_j | Y, \gamma) \pi(\gamma | Y).$$

Το Σχήμα 2-1 αναπαριστά την κατανομή ενός συντελεστή β όταν χρησιμοποιηθεί ένα μοντέλο, η οποία είναι μια $t_{n-q, -1}$ κατανομή, με κέντρο τον εκτιμητή μέγιστης πιθανοφάνειας. Η κατανομή η οποία βασίζεται στο ένα μοντέλο θα μας παρείχε (με ασφάλεια!) το συμπέρασμα ότι ο συντελεστής είναι μη μηδενικός. Χρησιμοποιώντας όμως την BMA στο Σχήμα 2-1 βλέπουμε ότι υπάρχει πιθανότητα 40% να είναι μηδενικός ($\Pr(\beta_i = 0 | y) = 0.4$)!



Σχήμα 2-2 Κατανομή του BMA - εκτιμητή δοθέντος $\beta_j \neq 0$

Συχνά, εξετάζουμε ξεχωριστά την πιθανότητα η μεταβλητή X_j να μην περιλαμβάνεται σε κάποιο μοντέλο (ενδεχόμενο το οποίο είναι ισοδύναμο με $\beta_j = 0$) από τη δεσμευμένη κατανομή της συμπερίληψης της X_j στο ίδιο μοντέλο.

Ακολουθώντας τη δυναμική της 2.1.5 θα περιγράψουμε τις κυριότερες επιλογές των εκ των προτέρων κατανομών.

Επιλογές για την $\pi(\gamma)$

Στο άρθρο τους οι Hoeting *et al.* (1999) τονίζουν ότι δεν έχει γίνει πολλή πρόοδος στο θέμα των εκ των προτέρων κατανομών στο χώρο των μοντέλων. Η διαπίστωση αυτή ισχύει και σήμερα.

Η κυρίαρχη επιλογή είναι η κατανομή Bernoulli. Συγκεκριμένα, η κατανομή του μοντέλου γ δίνεται από την

$$\pi(\gamma) = \prod_{j=1}^p w_j^{\gamma_j} (1 - w_j)^{1-\gamma_j} \quad (2.2.2)$$

$$= \left\{ w_1^{\gamma_1} (1 - w_1)^{1-\gamma_1} \right\} \cdot \left\{ w_2^{\gamma_2} (1 - w_2)^{1-\gamma_2} \right\} \cdots \left\{ w_p^{\gamma_p} (1 - w_p)^{1-\gamma_p} \right\},$$

όποτε η X_j περιλαμβάνεται σε κάποιο μοντέλο με πιθανότητα $\Pr(\gamma_j = 1) = w_j = 1 - \Pr(\gamma_j = 0)$. Παράλληλα, η εισαγωγή της X_j είναι ανεξάρτητη από την εισαγωγή ή μη της X_i , $i \neq j$.

Εάν, $w_j = w$ για κάθε j , δηλαδή κάθε μεταβλητή είναι εκ των προτέρων το ίδιο πιθανό να συμπεριληφθεί σε ένα μοντέλο, τότε

$$\pi(\gamma) = w^{\sum \gamma_j} (1 - w)^{p - \sum \gamma_j} = w^{q_\gamma} (1 - w)^{p - q_\gamma}, \quad q_\gamma = 0, 1, \dots, p$$

Η παράμετρος w εκφράζει την πεποίθηση του ερευνητή για το πόσες μεταβλητές περιέχει το μοντέλο. Πράγματι, χρησιμοποιώντας το γεγονός ότι w^p είναι η εκ των προτέρων αναμενόμενη τιμή του αριθμού των μεταβλητών που συμπεριλαμβάνονται στο μοντέλο γ , μπορούμε να θέσουμε $w = \frac{\bar{q}}{p}$, όπου \bar{q} η εκ των προτέρων εκτίμηση του αριθμού των μεταβλητών. Οπότε, το πρόβλημα επιλογής της παραμέτρου w γίνεται ισοδύναμο με την επιλογή του αναμενόμενου μεγέθους του μοντέλου.

Οι Smith *et al.* (1999) χρησιμοποιούν μια περικομμένη στο 0 Poisson με συνάρτηση πιθανότητας

$$\pi(\gamma) = \frac{\lambda^{q_\gamma}}{q_\gamma!} (e^\lambda - 1)^{-1}, \quad q_\gamma = 1, 2, \dots, p.$$

Η περικοπή στο 0 δικαιολογείται από το γεγονός ότι δεν έχει ιδιαίτερο νόημα να εξετάσουμε την περίπτωση του μηδενικού μοντέλου, καθώς θα υποθέτουμε εκ των προτέρων ότι καμμία από τις ανεξάρτητες μεταβλητές μας δεν εξηγεί την εξαρτημένη με θετική πιθανότητα.

Τέλος, συχνά δεν έχουμε επαρκή αιτιολόγηση ώστε να θεωρήσουμε κάποια μοντέλα περισσότερο πιθανά από άλλα, οπότε μια λογική επιλογή είναι $w = 1 / 2$. Τότε,

$$\pi(\gamma) = \frac{1}{2^p}, \quad \forall \gamma \in \Gamma, \quad (2.2.3)$$

δηλαδή η ομοιόμορφη κατανομή στο σύνολο όλων των υποσυνόλων των p ανεξάρτητων μεταβλητών. Η ομοιόμορφη κατανομή είναι απλή και, εκ πρώτης όψεως, είναι μια αντικειμενική αφετηρία για την έρευνα. Εξ άλλου, αποτελεί τη δημοφιλέστερη επιλογή στις εφαρμογές.

Ας σχολιάσουμε λίγο την επιλογή μιας διωνυμικής κατανομής. Ένα μειονέκτημα της (2.3.3) είναι το γεγονός ότι ευνοεί ισοδύναμα, ανταγωνιστικά μοντέλα. Εάν το σύνολο ισοδύναμων μοντέλων λάβει την εκ των προτέρων πιθανότητα που ορίζει η (2.3.3) τότε αυτό το σύνολο θα ευνοηθεί σε σχέση με τα άλλα μοντέλα. Οι Madigan *et al.* (1999) δεν θεωρούν σημαντικό μειονέκτημα το παραπάνω. Για μια εναλλακτική αντιμετώπιση δες George (1999) ο οποίος προτείνει τη χρήση μιας εναλλακτικής εκ των προτέρων κατανομής (*dilution prior*).

Έχοντας ολοκληρώσει την παρουσίαση των επιλογών για τη διαμόρφωση της εκ των προτέρων κατανομής των μοντέλων, μπορούμε να προχωρήσουμε στην παρουσίαση των επιλογών για τις εκ των προτέρων κατανομές των παραμέτρων του γραμμικού μοντέλου.

Επιλογές για τις $\pi(\theta_\gamma | \sigma^2, \gamma)$, $\pi(\sigma^2 | \gamma)$

Οι επιλογές που θα περιγράψουμε βασίζονται στην οικογένεια των συζυγών κατανομών. Αρχικά, θα περιγράψουμε την οικογένεια κατανομών και στην συνέχεια τον τρόπο επιλογής των υπερπαραμέτρων, βασισμένοι σε γενικές αρχές και ιδέες.

Η πιθανοφάνεια είναι

$$\begin{aligned} p(y | \theta_\gamma, \sigma, \gamma) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X_\gamma\theta_\gamma)^T(y - X_\gamma\theta_\gamma)\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}y^T(I - H_\gamma)y\right\} \times \\ &\quad \exp\left\{-\frac{1}{2\sigma^2}(\theta_\gamma - \hat{\theta}_\gamma)^T X_\gamma^T X_\gamma (\theta_\gamma - \hat{\theta}_\gamma)\right\}, \end{aligned}$$

όπου $\bar{y} = \frac{\sum_i y_i}{n}$, $\hat{\theta}_\gamma = (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y$ και $H_\gamma = X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T$ ο γνωστός «hat matrix».

Επιστρέφοντας στην γραφή α, β για τους συντελεστές του γραμμικού μοντέλου και χρησιμοποιώντας τη γραφή $X_{\gamma[-1]}$ για το πίνακα X_γ χωρίς την πρώτη στήλη (η οποία αναφέρεται στην παράμετρο α) ο πίνακας πληροφορίας Fisher είναι ο επόμενος,

$$I^F(\sigma, \alpha, \beta) = \begin{pmatrix} \frac{2n}{\sigma^2} & 0 & 0 \\ 0 & \frac{n}{\sigma^2} & 0 \\ 0 & 0 & \frac{X_{\gamma[-1]}^T X_{\gamma[-1]}}{\sigma^2} \end{pmatrix}.$$

Ας ξεκινήσουμε από την $\pi(\sigma^2 | \gamma)$. Η σ^2 είναι λογικό να υποθέσουμε ότι έχει την ίδια ερμηνεία για κάθε μοντέλο. Πράγματι, η υπόθεση επαληθεύεται. Οι παράμετροι α, σ είναι ορθογώνιες ως προς το διάνυσμα παραμέτρων β , (καθώς ο πίνακας είναι block-διαγώνιος) γεγονός το οποίο μας επιτρέπει να συνάγουμε ότι το ίδιο θα συμβαίνει για κάθε μοντέλο γ . Συνεπώς, οι παράμετροι α, σ θα έχουν την

ίδια ερμηνεία σε κάθε μοντέλο και μπορούμε να ορίσουμε μία εκ των προτέρων κατανομή ανεξαρτήτως του μοντέλου γ (Jeffreys, δεξ και Berger & Pericchi, 2001).

Η συζυγής κατανομή για την σ^2 είναι η αντίστροφη γάμμα

$$\sigma^2 | \gamma \sim IG\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right), \quad \nu > 0, \lambda > 0. \quad (2.2.4)$$

Βάσει των παραπάνω, η δέσμευση ως προς το μοντέλο στην (2.2.4) είναι περιττή και στο εξής θα παραλείπεται. Η ισοδύναμη έκφραση της εκ των προτέρων κατανομής της σ^2 ως χ^2 είναι

$$\frac{\nu\lambda}{\sigma^2} \sim \chi_\nu^2. \quad (2.2.5)$$

Εναλλακτικά, μπορούμε να θεωρήσουμε ως παράμετρο την ακρίβεια (*precision*),

$\tau = \frac{1}{\sigma^2}$ (αντί της διακύμανσης σ^2), της οποίας η εκ των προτέρων κατανομή είναι η γνωστή μας κατανομή γάμμα.

Συχνά, βρισκόμαστε σε συνθήκες έλλειψης ικανοποιητικής εκ των προτέρων πληροφορίας ή, ακόμα χειρότερα, γνώσης σχετικά με το θέμα της ανάλυσης. Τότε είναι επιθυμητό να βρεθούν κάποιες «αντικειμενικές» τιμές των υπερπαραμέτρων. Οι τιμές των υπερπαραμέτρων επηρεάζουν την εκ των υστέρων κατανομή, οπότε λανθασμένες επιλογές τους θα μας οδηγήσουν σε αποτελέσματα αμφισβητούμενης αξιοπιστίας. Η συνειδητοποίηση του γεγονότος αυτού οδήγησε στην αναζήτηση αυτόματων εκ των προτέρων κατανομών για το BMA (και όχι μόνο).³

Ας ξεκινήσουμε με την επιλογή των υπερπαραμέτρων ν, λ . Θα θέλαμε οι επιλογές μας να ικανοποιούν την επόμενη σχέση :

³ Για παράδειγμα, μια από τις γνωστότερες «αντικειμενικές» προτάσεις είναι η εκ των προτέρων κατανομή του Jeffreys, η οποία είναι η $\pi(\theta) \propto |I^F(\theta)|^{1/2}$. Οπότε μια αρχική πρόταση θα ήταν η $\pi(\alpha, \sigma, \beta) \propto (\sigma^2)^{-p}$.

$$\Pr(\sigma_L^2 \leq \sigma^2 \leq \sigma_U^2) = v,$$

όπου το $v \in (0,1]$ είναι μια δεδομένη τιμή. Για τα όρια θα μπορούσαμε να βασιστούμε στην ακόλουθη λογική. Τα γραμμικά μοντέλα θα μας δίνουν μικρότερες εκτιμήσεις της διακύμανσης καθώς προσθέτουμε μεταβλητές, συνεπώς ένα άνω φράγμα θα είναι η δειγματική διακύμανση της μεταβλητής απόκρισης (ισοδύναμα του μηδενικού μοντέλου), s_Y^2 , και ένα κάτω φράγμα η εκτίμηση του πλήρους μοντέλου, s_p^2 .

$$\Pr(s_p^2 \leq \sigma^2 \leq s_Y^2) = v \quad (2.2.6)$$

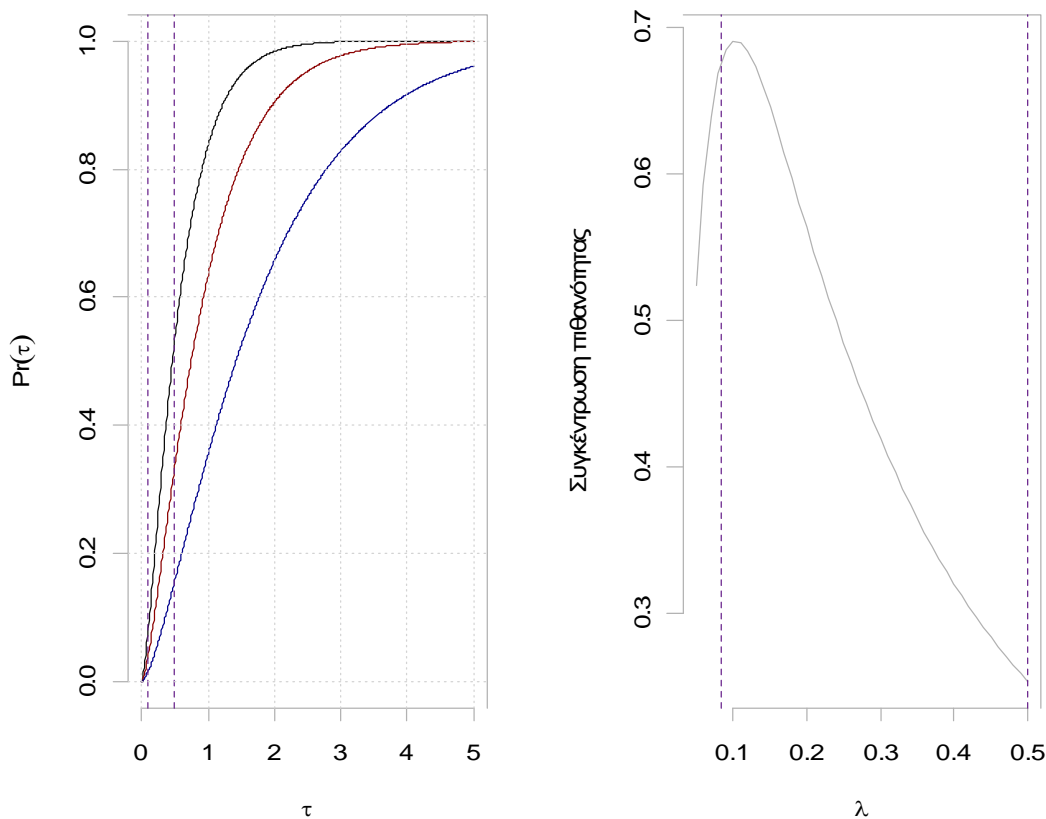
ή, σε όρους του τ ,

$$\Pr(\tau_Y \leq \tau \leq \tau_p) = v$$

Η εξίσωση που θέλουμε να λύσουμε έχει δύο αγνώστους οπότε δεν είναι εφικτό να λυθεί προσδιορίζοντας και τις δύο

$$\int_{1/s_p^2}^{1/s_Y^2} \frac{(\nu\lambda/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} \tau^{\frac{\nu}{2}-1} \exp\{-\nu\lambda\tau/2\} d\tau = v. \quad (2.2.7)$$

Ένα από τα πλεονεκτήματα των συζυγών κατανομών είναι η ερμηνεία τους ως πληροφορία η οποία προέρχεται από ένα δείγμα «ευκολίας» (*convenience/imaginary sample*). Στην περίπτωσή μας, η παράμετρος που μπορεί να θεωρηθεί ότι παριστάνει το μέγεθος του δείγματος είναι το ν . Με άλλα λόγια, μπορούμε να θεωρήσουμε ότι διαθέτουμε ένα δείγμα ν παρατηρήσεων από την $N(0,\lambda)$ οι οποίες μας παρέχουν πληροφορία για το τ (ισοδύναμα, για το σ^2).



Σχήμα 2-1 : Εύρεση των υπερπαραμέτρων της εκ των προτέρων κατανομής της σ^2

Συγκεκριμένα, η πληροφορία είναι ότι η εκ προτέρων εκτίμηση του σ^2 είναι η λ . Επίσης, η λ ελέγχει τη μέση τιμή του τ . Συνεπώς, εάν θέλουμε να λύσουμε την εξίσωση (2.2.6) επιλέγοντας εξ αρχής κάποια παράμετρο, λογικό είναι να θέσουμε την παράμετρο ν ίση με κάποια «λογική» τιμή. Μια λογική επιλογή θα ήταν $\nu = 2, 3, 4$, δηλαδή θεώρηση ενός μικρού δείγματος.

Το αριστερό γράφημα εξηγεί τη διαδικασία για $\nu = 3$ και $s_y^2 = 12, s_p^2 = 2$. Η λύση της εξίσωσης (2.2.6) για $\nu = 0.70$ είναι $\lambda \cong 0.103$. Το γράφημα αναπαριστά τη συνάρτηση πυκνότητας πιθανότητας για τρεις τιμές της παραμέτρου και συγκεκριμένα για $\lambda = 0.26, 0.42$ και 0.8 . Το δεξιό γράφημα αναπαριστά την διαδικασία αναζήτησης. Παρατηρήστε ότι η λύση είναι κοντά στην εκτίμηση της

διακύμανσης της Y . Πρακτικά, συχνά θεωρούμε $\nu = 2$ ή 3 και λ κοντά στην τιμή της $1 / s_Y^2$, ώστε $E(\sigma^2) \cong s_Y^2$ (Chipman *et al.*, 2001).

Στο επόμενο κεφάλαιο θα παρουσιάσουμε άλλη μια προσέγγιση για τον προσδιορισμό των υπερπαραμέτρων.

Έχοντας προσδιορίσει την εκ των προτέρων κατανομή της σ^2 , μπορούμε να στραφούμε στην εκ των προτέρων κατανομή του διανύσματος β_γ . Για να πάρουμε μια συζυγή κατανομή για το β_γ θα χρειαστεί να δεσμεύσουμε ως προς το σ^2 ,

$$\theta_\gamma | \sigma^2, \gamma \sim N_{q_\gamma+1}(\tilde{\theta}_\gamma, \sigma^2 \Sigma_\gamma) \quad (2.2.8)$$

Είναι εύκολο να δούμε ότι οι (2.7) και (2.9), ο συνδυασμός των οποίων αναφέρεται συχνά ως Normal – Gamma (NG), είναι πράγματι συζυγής για το κανονικό γραμμικό μοντέλο.

Τυποποιώντας τις ανεξάρτητες μεταβλητές έτσι ώστε να ικανοποιείται η $1^T \cdot X = \mathbf{0}$, εξασφαλίζουμε ταυτόσημη ερμηνεία την σταθεράς για κάθε μοντέλο· θα είναι ο μέσος της μεταβλητής απόκρισης. Ταυτόχρονα, καταλήγουμε στο ότι

$$\tilde{\theta}_\gamma = (\bar{y}, 0, 0, \dots, 0).$$

Είναι εύκολο τώρα να βρούμε την (εκ των υστέρων) πιθανοφάνεια του μοντέλου γ

$$p(y | \gamma) = \int l(\theta_\gamma, \sigma^2 | y, \gamma, X) p(\theta_\gamma | \sigma^2, \gamma) p(\sigma^2 | \gamma) d\theta_\gamma d\sigma^2$$

Αντικαθιστώντας την πιθανοφάνεια και τις εκ των πρότερων κατανομές (2.2.4), (2.2.8) έχουμε

$$\begin{aligned}
p(y | \gamma) &= \int (\sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} (y - X_\gamma \theta_\gamma)^T (y - X_\gamma \theta_\gamma) \right] \times \\
&\quad (\sigma^2)^{-1/2} \exp \left[-\frac{1}{2\sigma^2} (\theta_\gamma - \bar{\theta}_\gamma)^T \Sigma_\gamma^{-1} (\theta_\gamma - \bar{\theta}_\gamma) \right] \times \\
&\quad \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \lambda^\nu (\sigma^2)^{-(\nu/2+1)} |\Sigma_\gamma|^{-1/2} \exp \left(-\frac{\nu\lambda}{2\sigma^2} \right) d\theta_\gamma d\sigma^2 =
\end{aligned}$$

Θέτοντας $C = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \lambda^\nu |\Sigma_\gamma|^{-1/2}$ παίρνουμε

$$\begin{aligned}
p(y | \gamma) &= C \int (\sigma^2)^{-(n/2+1/2+\nu/2+1)} \exp \left[-\frac{1}{2\sigma^2} (y - X_\gamma \theta_\gamma)^T (y - X_\gamma \theta_\gamma) \right] \times \\
&\quad \exp \left[-\frac{1}{2\sigma^2} (\theta_\gamma - \bar{\theta}_\gamma)^T \Sigma_\gamma^{-1} (\theta_\gamma - \bar{\theta}_\gamma) \right] \times \\
&\quad \exp \left(-\frac{\nu\lambda}{2\sigma^2} \right) d\theta_\gamma d\sigma^2
\end{aligned}$$

η οποία, χρησιμοποιώντας την σχέση

$$(y - X_\gamma \theta_\gamma)^T (y - X_\gamma \theta_\gamma) = (y - X_\gamma \hat{\theta}_\gamma)^T (y - X_\gamma \hat{\theta}_\gamma) + (\theta_\gamma - \bar{\theta}_\gamma)^T X_\gamma^T X_\gamma (\theta_\gamma - \bar{\theta}_\gamma),$$

γίνεται

$$\begin{aligned}
C \int (\sigma^2)^{-[(n+1+\nu)/2+1]} \exp \left[-\frac{1}{2\sigma^2} \{ (y - X_\gamma \hat{\theta}_\gamma)^T (y - X_\gamma \hat{\theta}_\gamma) + (\theta_\gamma - \bar{\theta}_\gamma)^T X_\gamma^T X_\gamma (\theta_\gamma - \bar{\theta}_\gamma) \} \right] \times \\
\exp \left[-\frac{1}{2\sigma^2} (\theta_\gamma - \bar{\theta}_\gamma)^T \Sigma_\gamma^{-1} (\theta_\gamma - \bar{\theta}_\gamma) \right] \exp \left(-\frac{\nu\lambda}{2\sigma^2} \right) d\theta_\gamma d\sigma^2.
\end{aligned}$$

Συνδυάζοντας τα εκθετικά έχουμε

$$\{ (y - X_\gamma \hat{\theta}_\gamma)^T (y - X_\gamma \hat{\theta}_\gamma) + (\theta_\gamma - \bar{\theta}_\gamma)^T X_\gamma^T X_\gamma (\theta_\gamma - \bar{\theta}_\gamma) \} + (\theta_\gamma - \bar{\theta}_\gamma)^T \Sigma_\gamma^{-1} (\theta_\gamma - \bar{\theta}_\gamma) =$$

$$(\theta_\gamma - \tilde{\theta}_\gamma)^T (\Sigma_\gamma^{-1} + X_\gamma^T X_\gamma) (\theta_\gamma - \tilde{\theta}_\gamma) + (\theta_\gamma - \bar{\theta}_\gamma)^T [\Sigma_\gamma + (X_\gamma^T X_\gamma)^{-1}]^{-1} (\theta_\gamma - \bar{\theta}_\gamma),$$

όπου έχουμε θέσει $\tilde{\theta}_\gamma = (\Sigma_\gamma^{-1} + X_\gamma^T X_\gamma) (\Sigma_\gamma^{-1} \bar{\theta}_\gamma + X_\gamma^T X_\gamma \hat{\theta}_\gamma)$.

Τα ολοκλήρωμα ως προς θ_γ τώρα είναι το ολοκλήρωμα της πυκνότητας πιθανότητας μιας $(q_\gamma + 1)$ -διάστατης κανονικής κατανομής με μέσο διάνυσμα $\tilde{\theta}_\gamma$ και πίνακα συνδιακύμανσης $\sigma^2[\Sigma_\gamma^{-1} + (X_\gamma^T X_\gamma)^{-1}]^{-1}$.

Τότε,

$$p(y | \gamma) = C |\Sigma_\gamma^{-1} + X_\gamma^T X_\gamma|^{-1/2} \int (\sigma^2)^{-(n+\nu)/2+1} \times \\ \exp\left[-\frac{1}{\sigma^2}(y - X_\gamma \hat{\theta}_\gamma)^T (y - X_\gamma \hat{\theta}_\gamma)\right] \times \\ \exp\left[-\frac{1}{\sigma^2}(\theta_\gamma - \bar{\theta}_\gamma)^T (\Sigma_\gamma^{-1} + (X_\gamma^T X_\gamma)^{-1})^{-1}(\theta_\gamma - \bar{\theta}_\gamma) + \nu\lambda\right] d\sigma^2.$$

Τελικά, αναγνωρίζοντας και τον πυρήνα μιας IG έχουμε

$$p(y | \gamma) = C' \lambda^\nu |\Sigma_\gamma|^{-1/2} |\Sigma_\gamma^{-1} + X_\gamma^T X_\gamma|^{-1/2} |\Sigma_\gamma^{-1} + X_\gamma^T X_\gamma|^{1/2} [(y - X_\gamma \hat{\theta}_\gamma)^T (y - X_\gamma \hat{\theta}_\gamma) \\ + (\theta_\gamma - \bar{\theta}_\gamma)^T (\Sigma_\gamma + (X_\gamma^T X_\gamma)^{-1})^{-1}(\theta_\gamma - \bar{\theta}_\gamma) + \nu\lambda], \quad (2.2.9)$$

$$\text{όπου } C' = \frac{\Gamma[(\nu + n)/2]}{\Gamma(\nu/2)} \frac{(\nu/2)^{\nu/2}}{[(\nu + n)/2]^{(\nu+n)/2}}.$$

Η (2.2.9) έχει τον πυρήνα μιας πολυμεταβλητής κατανομής t .

Ας στραφούμε τώρα στην επιλογή του εκ των προτέρων πίνακα συνδιακύμανσης Σ_γ . Γενικά, το εγχείρημα δεν είναι απλό, καθώς θα πρέπει να καθορίσουμε $q_\gamma + 1$ διακυμάνσεις και $q_\gamma(q_\gamma + 1)/2$ συνδιακυμάνσεις. Βολικό θα ήταν να διασπάσουμε το πρόβλημα σε δύο συνιστώσες, γράφοντας $\Sigma_\gamma = \phi_\gamma \cdot A_\gamma^{-1}$, όπου ο A_γ είναι ένας θετικά ορισμένος πίνακας με διάσταση $(q_\gamma + 1) \times (q_\gamma + 1)$. Θα εξετάσουμε δύο επιλογές για τον A_γ : Τον μοναδιαίο πίνακα $I_{q_\gamma+1}$, και τον πίνακα πληροφορίας του μοντέλου γ , $X_\gamma^T X_\gamma$.

Για την τελευταία επιλογή, η κατανομή που προκύπτει,

$$\theta_\gamma | \sigma^2, \gamma \sim N_{q_\gamma+1} \left(\tilde{\theta}_\gamma, \phi_\gamma \sigma^2 (X_\gamma^T X_\gamma)^{-1} \right) \quad (2.2.10)$$

μας θυμίζει την κατανομή του εκτιμητή μέγιστης πιθανοφάνειας του κλασικού γραμμικού μοντέλου. Επίσης, εμφανίζει ομοιότητες με την εκ των προτέρων κατανομή του Zellner, την οποία θα περιγράψουμε στο τέλος της παρούσας ενότητας.

Για να καταλάβουμε τι επάγει η κάθε επιλογή, παρατηρούμε ότι όταν χρησιμοποιήσουμε τον πίνακα πληροφορίας, ο συνδυασμός αυτής της εκ των προτέρων επιλογής με την πιθανοφάνεια θα αναπαραγάγει, εν πολλοίς, την δομή του δείγματος στην εκ των υστέρων κατανομή. Αντίθετα, η χρήση του μοναδιαίου πίνακα θα «χαλαρώσει» τη δομή της διακύμανσης του δείγματος καθώς υποθέτει εκ των πρότερων ανεξαρτησία των συντελεστών β_i, β_j για $i \neq j$.

Τέλος, απομένει ο προσδιορισμός της υπερπαραμέτρου ϕ_γ . Η επιλογή της ϕ_γ είναι καθοριστική για τη διακύμανση της εκ των προτέρων κατανομής, μια και οι λογικές επιλογές για τον A_γ είναι περιορισμένες: η μια βασίζεται στο δείγμα και η άλλη στην ανεξαρτησία. Συνεπώς, θα ήταν επιθυμητό η εκ των προτέρων κατανομή να καλύπτει μεγάλη περιοχή του εύρους των πιθανών τιμών του θ_γ . Αυτό είναι εφικτό επιλέγοντας «μεγάλες» τιμές της υπερπαραμέτρου. Δυστυχώς, υπερβολικά μεγάλες τιμές θα ευνοούν μικρότερα ως προς τη διάσταση τους μοντέλα. Αυτή είναι μια μορφή του παραδόξου των Bartlett-Lindley, ενός από τα παράδοξα της Μπεϋζιανής Στατιστικής (Bartlett, 1957). Συνεπώς, η επιλογή πρέπει να ισορροπεί ανάμεσα στην επιθυμία για κάλυψη μεγάλου εύρους τιμών και του παραδόξου.

Αρκετή συζήτηση έχει γίνει για την επιλογή της υπερπαραμέτρου ϕ_γ . Στις εφαρμογές, χρησιμοποιούνται τα $\phi_\gamma = \{n, p^2\}$. Η επιλογή $\phi_\gamma = n$ έχει προταθεί από τους Kass & Wasserman (1995) στα πλαίσια της UIP (*Unit Information Prior*). Βασίζεται στην ανάθεση στην εκ των πρότερων κατανομή πληροφορίας ίσης με την πληροφορία την οποία προσφέρει μία παρατήρηση. Για το κανονικό γραμμικό μοντέλο η πληροφορία είναι ίση με το πλήθος των παρατηρήσεων, n .⁴

⁴ Για άλλα μοντέλα ο υπολογισμός δεν είναι τόσο απλός (π.χ μοντέλα επιβίωσης με λογοκριμένα δεδομένα)

Η δεύτερη επιλογή προέρχεται από την έρευνα των Foster & George (1994) στις διαδικασίες επιλογής μοντέλου, οι οποίοι βασίστηκαν στο Risk Inflation Criterion (RIC).

Ορισμός 2.2.1 Έστω γ_{TRUE} το αληθινό μοντέλο και $\hat{\beta}_{TRUE}$ το (άγνωστο) διάνυσμα εκτιμήσεων. Για ένα επιλεγέν μοντέλο γ , το Risk Inflation Criterion ορίζεται ως

$$RIC(\gamma) = \sup_{\beta} \{R(\beta, \hat{\beta}_{\gamma}), R(\beta, \hat{\beta}_{TRUE})\},$$

όπου $R(\beta, \hat{\beta}_{\gamma}) := E_{\beta} | X\hat{\beta}_{\gamma} - X\beta |$ ο κίνδυνος πρόβλεψης (*predictive risk*).

Οι Foster & George (1994) υιοθετώντας μια minimax οπτική στην αναζήτησή τους για την βέλτιστη εκ των προτέρων κατανομή, κατέληξαν στην επιλογή $\phi_{\gamma} = p^2$.

Οι Fernandez *et al.* (2001) μελέτησαν διεξοδικά το ζήτημα της επιλογής της υπερπαραμέτρου ϕ_{γ} . Χρησιμοποιώντας δεδομένα προσομοίωσης αλλά και πραγματικά δεδομένα, αξιολόγησαν την απόδοση των ακολούθων επιλογών :

$$n, \frac{n}{q_{\gamma}}, \frac{n}{p^{1/q_{\gamma}}}, \sqrt{n}, \sqrt{\frac{n}{q_{\gamma}}}, (\log n)^{-3}, \frac{\log n}{\log(1+q_{\gamma})}, p^2, \frac{\delta c^{1/q_{\gamma}}}{1-\delta c^{1/q_{\gamma}}}$$

Το συμπέρασμα της εργασίας τους ήταν ότι η ιδανικότερη επιλογή είναι $\phi_{\gamma} = \max\{n, p^2\}$. Η επιλογή αυτή, όταν συνδυάζεται με την εκ των προτέρων κατανομή του Zellner αναφέρεται ως benchmark prior.

Κλείνοντας το κεφάλαιο, παρουσιάζουμε την εκ των προτέρων κατανομή του Zellner, η οποία είναι η δεύτερη δημοφιλέστερη επιλογή εκ των πρότερων κατανομής στο κανονικό γραμμικό μοντέλο. Ο Zellner (1986) θεώρησε ότι οι παράμετροι οι οποίες είναι κοινές μεταξύ των μοντέλων μπορούν να έχουν εκ των πρότερων κατανομή ανεξάρτητη του μοντέλου γ . Στο κανονικό γραμμικό μοντέλο τέτοιες παράμετροι είναι οι α, σ . Σε σχέση με τη NG, το προηγούμενο ισχύει για το σ^2 , όχι όμως και για τον σταθερό όρο α . Επίσης, είναι γνωστό από τη θεωρία η οποία έχει αναπτυχθεί για την επιλογή μοντέλων στη Μπεϋζιανή Στατιστική ότι οι κοινές παράμετροι μεταξύ των μοντέλων είναι δυνατόν να έχουν γενικευμένα εκ των

πρότερων κατανομή (Ghosh, 2008). Ο Zellner πρότεινε μια μη πληροφοριακή κατανομή για τις δύο παραμέτρους α, σ . Συγκεκριμένα χρησιμοποίησε την

$$\pi(\alpha, \sigma^2) \propto \frac{1}{\sigma^2}, \quad \alpha \in \mathbb{R}, \sigma^2 > 0, \quad (2.2.11)$$

και την (2.2.8) για το παραμετρικό διάνυσμα θ_γ . Η κατανομή του Zellner είναι οριακή περίπτωση της NG για $\nu \rightarrow 0$. Συνεπώς σε σχέση με την NG η οποία απαιτεί την επιλογή τεσσάρων υπερπαραμέτρων (εξαιρούμε το διάνυσμα $\tilde{\theta}_\gamma$), η εκ των προτέρων κατανομή του Zellner απαιτεί την επιλογή μόνο μίας, της ϕ_γ !

3.

Υλοποίηση της BMA

Στο προηγούμενο κεφάλαιο περιγράψαμε σε γενικές γραμμές τη διαδικασία του BMA. Αναφέραμε τις κύριες ιδιότητές του και εισαγάγαμε τη διαδικασία στο κανονικό γραμμικό μοντέλο παρουσιάζοντας τις δύο κυριότερες μορφές μοντελοποίησής του. Στο παρόν κεφάλαιο, το οποίο αποτελεί το κύριο μέρος αυτής της εργασίας, θα παρουσιάσουμε τις κύριες υλοποιήσεις που συναντώνται στην βιβλιογραφία και τις εφαρμογές. Στη συνέχεια θα δούμε ότι καθώς προχωρούμε προς την εφαρμογή της διαδικασίας, θα βρεθούμε αντιμέτωποι με το κύριο εμπόδιο στην πρακτική εφαρμογή του BMA, τον υπολογιστικό φόρτο. Αφού περιγράψουμε τους κυριότερους αλγόριθμους οι οποίοι αντιμετωπίζουν το ζήτημα, θα παρουσιάσουμε ένα ολοκληρωμένο παράδειγμα.

3.1 Κύριες υλοποιήσεις

3.1.1 HRM υλοποίηση

Αναμφισβήτητα, το σημαντικότερο άρθρο στην βιβλιογραφία σχετικά με το BMA είναι των Raftery *et al.* (1997). Στο άρθρο τους περιέγραψαν το πρόβλημα της αβεβαιότητας η οποία περιέχεται στην χρήση ενός μοντέλου, πρότειναν την λύση του BMA, μοντελοποίησαν και προσδιόρισαν το μοντέλο. Το άρθρο συνοδευόταν από λογισμικό, ώστε ο αναγνώστης να έχει τη δυνατότητα να εφαρμόσει την προτεινόμενη διαδικασία στα δικά του δεδομένα. Δεδομένου ότι η BMA είναι ευκολονόητη ως διαδικασία αλλά η εφαρμογή της απαιτεί «προγραμματιστική» παιδεία, η ύπαρξη του λογισμικού έκανε την συγκεκριμένη προσεγγιση κυρίαρχη.

Οι Raftery *et al.* (1997) χρησιμοποίησαν ως εκ των προτέρων κατανομές μία NG για τις θ_γ, σ^2 και ομοιόμορφη κατανομή στο σύνολο των μοντέλων Γ . Οι ανεξάρτητες μεταβλητές μετασχηματίζονται έτσι ώστε να ικανοποιείται η $1^T X = 0$, αυτό επιτυγχάνεται αφαιρώντας από όλες τις μεταβλητές τους δειγματικούς μέσους.

Σε σχέση με τα όσα περιγράψαμε έως τώρα για την NG, ο πίνακας Σ μοντελοποιείται ως

$$\Sigma = \sigma^2 \begin{pmatrix} s_Y^2 & 0 & \dots & 0 \\ 0 & \frac{\phi}{s_{X_1}^2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \frac{\phi}{s_{X_p}^2} \end{pmatrix},$$

όπου s_Y^2 η δειγματική διακύμανση της εξαρτημένης μεταβλητής Y και $s_{X_i}^2$ η δειγματική διακύμανση της ανεξάρτητης μεταβλητής X_i για $i = 1, 2, \dots, p$. Επίσης, καλύπτουν και την περίπτωση μιας κατηγορικής ανεξάρτητης μεταβλητής X_j με $l+1$ επίπεδα ($l \geq 2$). Οι παράμετροι $(\beta_{1j}, \beta_{2j}, \dots, \beta_{(l-1)j})$ οι όποιες σχετίζονται με τις l δείκτριες θα έχουν εκ των πρότερων πίνακα συν διακύμανσης

$$\Sigma_{X_j} = \sigma^2 \phi^2 \left(\frac{A_{X_j}^T A_{X_j}}{n} \right)^{-1},$$

όπου ο A_{X_j} είναι ο $n \times l$ πίνακας με στήλες τις δείκτριες μεταβλητές τυποποιημένες ως προς τον μέσο της κάθε μίας.

Ο πίνακας συνδιακύμανσης στην περίπτωση που η ανεξάρτητη μεταβλητή X_i είναι κατηγορική θα είναι

$$\Sigma = \sigma^2 \begin{pmatrix} s_Y^2 & 0 & 0 & \dots & \dots & 0 \\ 0 & \frac{\phi}{s_{X_1}^2} & & & & \vdots \\ 0 & & \ddots & & & \vdots \\ \vdots & & & \Sigma_{X_j} & & \vdots \\ \vdots & & & & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & \frac{\phi}{s_{X_p}^2} \end{pmatrix}.$$

Ο παραπάνω πίνακας συνδυάζει τόσο τον μοναδιαίο πίνακα όσο και τον πίνακα πληροφορίας που χρησιμοποιεί η εκ των πρότερων κατανομή του Zellner. Γράφοντας και την δειγματική διακύμανση σε μορφή διανυσμάτων έχουμε

$$\Sigma = \sigma^2 \begin{pmatrix} s_Y^2 & 0 & 0 & \dots & \dots & 0 \\ 0 & \phi \left(\frac{x_1^T x_1}{n} \right)^{-1} & & & & \vdots \\ 0 & & \ddots & & & \vdots \\ \vdots & & & \phi^2 \left(\frac{A_{X_j}^T A_{X_j}}{n} \right)^{-1} & & \vdots \\ \vdots & & & & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & \phi \left(\frac{x_p^T x_p}{n} \right)^{-1} \end{pmatrix}$$

$$= \sigma^2 \left(\text{diag} \left\{ s_Y^{-2}, \frac{s_{X_1}^2}{\phi}, \dots, \frac{s_{X_p}^2}{\phi} \right\} \right)^{-1}$$

Αρκετά σημαντικό είναι το γεγονός ότι οι Raftery *et al.* (1997) πρότειναν τιμές των υπερπαραμέτρων βασισμένοι σε ένα συλλογισμό διαφορετικό από αυτό που περιγράψαμε στο προηγούμενο κεφάλαιο. Συγκεκριμένα, επιθυμούν η εκ των πρότερων συνάρτηση πυκνότητας πιθανότητας, $\pi(\theta_p | \sigma^2)$ να είναι «επίπεδη» στο $[-1,1]^p$. Επιλέγουν το διάστημα $[-1,1]^p$ βασισμένοι στο επόμενο (εμπειρικό) γεγονός. Συνέλεξαν τα δεδομένα που χρησιμοποιήθηκαν σε 30 βιβλία Ανάλυσης Παλινδρόμησης τυποποίησαν κάθε εκτίμηση $\hat{\beta}_i$ και κατασκεύασαν την κατανομή τους. Στο ιστόγραμμα (σελ. 182 του άρθρου τους) η μάζα πιθανότητας συγκεντρώνεται στο διάστημα $[-1,1]$. Επίσης, θεωρούν μία θετική τιμή κ αλλά πολύ κοντά στο 0 τέτοιο ώστε η συνάρτηση πυκνότητας πιθανότητας του σ^2 , $\pi(\sigma^2)$, να είναι «επίπεδη» στο $(\kappa, 1]$. Τέλος, σκοπός τους είναι η πιθανότητα $\Pr(\sigma^2 \leq 1)$ να έχει αρκετά μεγάλη τιμή.

Ο προσδιορισμός της συνάρτησης πυκνότητας πιθανότητας, ως επίπεδης συνίσταται στο ότι δεν θα υπάρχουν περιοχές με σημαντική διαφοροποίηση στη πυκνότητα πιθανότητας. Εάν υπάρχει μια τέτοια περιοχή, η πιθανότητα η τιμή της μεταβλητής να βρίσκεται σε αυτήν θα είναι μεγαλύτερη από την πιθανότητα να βρίσκεται σε άλλες, ισομεγέθεις περιοχές. Σε αυτή την περίπτωση η εκ των πρότερων κατανομή δεν θα ήταν αντικειμενική όπως ζητάμε στα μοντέλα που περιγράφουμε σε αυτή την εργασία.

Συνοπτικά, ζητείται η μεγιστοποίηση της εκ των προτέρων πιθανότητας της σ^2 , $\Pr(\sigma^2 \leq 1)$, υπό τους εξής περιορισμούς:

$$\frac{p(\beta_1 = 0, \beta_2 = 0, \dots, \beta_p = 0)}{p(\beta_1 = 1, \beta_2 = 1, \dots, \beta_p = 1)} \leq u_1 \quad (3.1.1)$$

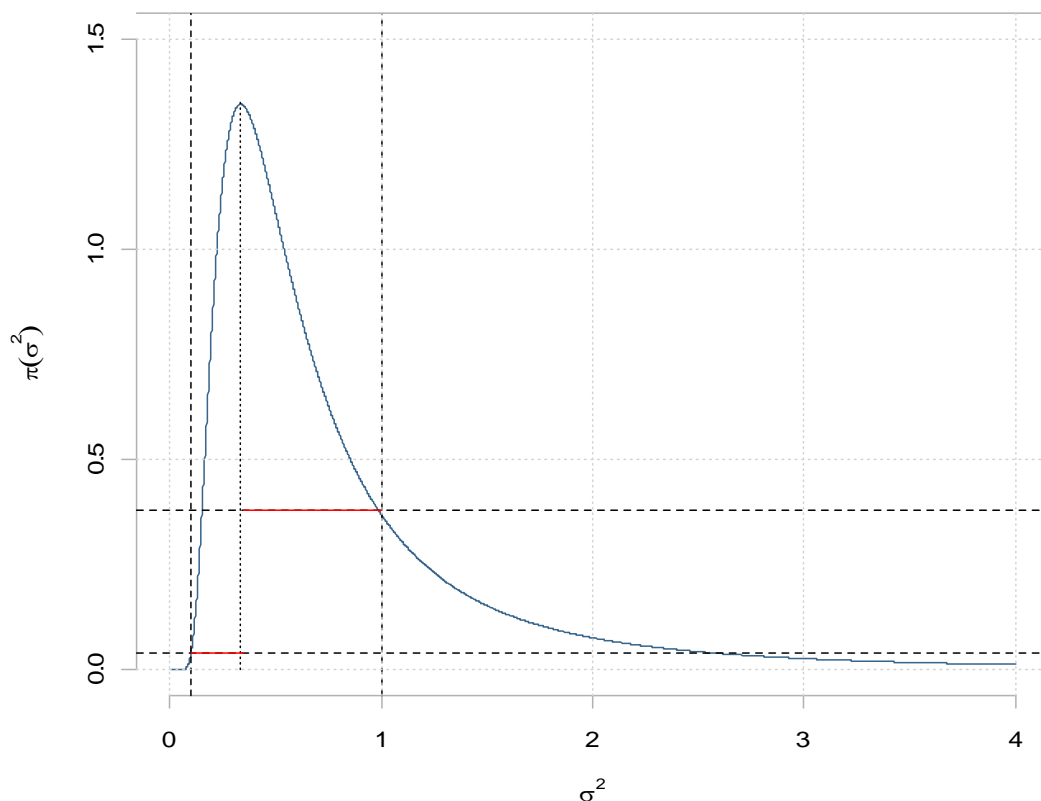
$$\frac{\max_{\alpha \leq \sigma^2 \leq 1} p(\sigma^2)}{p(\sigma^2 = \kappa)} \leq u_2, \quad \frac{\max_{\alpha \leq \sigma^2 \leq 1} p(\sigma^2)}{p(\sigma^2 = 1)} \leq u_2. \quad (3.1.2)$$

Ο πρώτος περιορισμός λόγω της υποτιθεμένης ανεξαρτησίας των $\beta_1, \beta_2, \dots, \beta_p$ μπορεί να γραφεί ως

$$\begin{aligned} \frac{p(\beta_1 = 0, \beta_2 = 0, \dots, \beta_p = 0)}{p(\beta_1 = 1, \beta_2 = 1, \dots, \beta_p = 1)} &\leq u_1 \Leftrightarrow \\ \frac{p(\beta_1 = 0)p(\beta_2 = 0) \cdots p(\beta_p = 0)}{p(\beta_1 = 1)p(\beta_2 = 1) \cdots p(\beta_p = 1)} &\leq u_1 \Leftrightarrow \\ \prod_{i=1}^p \left[\frac{p(\beta_i = 0)}{p(\beta_i = 1)} \right] &\leq u_1, \end{aligned}$$

δηλαδή ως το γινόμενο των λόγων των πυκνοτήτων κάθε παραμέτρου β_i στις δύο ακραίες περιπτώσεις: Όταν δεν περιλαμβάνεται σε ένα μοντέλο και όταν περιλαμβάνεται με την μέγιστη τιμή. Συνεπώς ο περιορισμός αυτός ελέγχει τη δυνατότητα να επιλέξουμε τιμές των υπερπαραμέτρων τέτοιες ώστε τα δύο ακραία μοντέλα να έχουν σημαντική διαφορά πυκνοτήτων.

Με την ίδια λογική λειτουργεί και το ζεύγος περιορισμών (3.1.2). Σκοπός τους είναι να οδηγήσουν σε επιλογές των υπερπαραμέτρων για τις οποίες δεν θα είναι



Σχήμα 3-1 : Έλεγχος των περιορισμών (3.1.2)

υπερβολική η διαφορά των πυκνοτήτων μεταξύ των άκρων του διαστήματος $(\kappa, 1]$ και της μέγιστης τιμής της σε αυτό το διάστημα.

Θεωρώντας τον πρώτο περιορισμό σημαντικότερο, θέτουν $u_2 = u_1^2$ και ακολουθώντας τον Jeffreys (1961) ορίζουν $u_1 = \sqrt{10}$. Τέλος, επιλέγοντας $\kappa = 0.05$ έλυσαν το παραπάνω πρόβλημα μεγιστοποίησης επιτυγχάνοντας $\Pr(\sigma^2 \leq 1) = 0.81$ για τις τιμές $\nu = 2.58$, $\lambda = 0.28$ και $\phi = 2.85$.

Κλείνοντας την παρουσίαση της μεθόδου, το Σχήμα 3-1 δείχνει τη συνάρτηση πυκνότητας μιας αντίστροφης γάμμα με παραμέτρους οι οποίες δεν είναι αποδεκτές υπό τους περιορισμούς. Συγκεκριμένα, ο λόγος της πυκνότητας της μέγιστης τιμής στο $(\kappa, 1]$ προς την πυκνότητα στο κάτω άκρο του $(\kappa, 1]$ είναι μεγαλύτερος του 10.

Η συζυγία την οποία προσφέρει η εκ των προτέρων κατανομή NG κάνει δυνατή την εύρεση της ακριβούς εκ των υστέρων κατανομής του συντελεστή θ_i . Η εκ των

υστερών κατανομή του (δοθείσας της σ^2) είναι κανονική. Ολοκληρώνοντας ως προς την σ^2 , η περιθωριακή κατανομή του συντελεστή θ_i θα είναι (κεντρική) t με $n - (q_\gamma + 1)$ βαθμούς ελευθερίας.

Οι Raftery *et al.* (1997) έδειξαν ότι η πιθανοφάνεια (2.2.9) δίνεται από την

$$p(y | \theta_\gamma, X, \gamma, \sigma^2) = \frac{\Gamma[(n + \nu) / 2]}{\Gamma(\nu / 2)} (\nu\lambda)^{\nu/2} \pi^{n/2} |I + X_\gamma \Sigma X_\gamma^T|^{-1/2} \times \\ [\nu\lambda + (y - X_\gamma \theta_\gamma)^T (I + X_\gamma \Sigma X_\gamma^T)^{-1} (y - X_\gamma \theta_\gamma)]^{-(n+\nu)/2}. \quad (3.1.3)$$

Αναδιάταξη των όρων αποκαλύπτει ότι η (3.1.3) είναι μια n -διάστατη t (\mathcal{T}) με παραμέτρους $X_\gamma \theta_\gamma$, $\lambda(I_n + X_\gamma \Sigma X_\gamma^T)$ και ν .

Η εκ των υστέρων περιθωριακή κατανομή του διανύσματος θ_γ θα είναι $q_\gamma + 1$ -διάστατη t με παραμέτρους $\hat{\mu}_\gamma, \hat{V}_\gamma$ και $n + \nu$

$$\theta_\gamma | y, X \sim \mathcal{T}_{q_\gamma+1}(n + \nu, \hat{\mu}_\gamma, \hat{V}_\gamma),$$

όπου $\hat{\mu}_\gamma = (\Sigma_\gamma^{-1} + X_\gamma^T X_\gamma)^{-1} (X_\gamma^T X_\gamma) \hat{\theta}_\gamma$,

$$\hat{V}_\gamma = \frac{\nu\lambda + (\bar{\theta}_\gamma - \hat{\theta}_\gamma)^T [\Sigma_\gamma + (X_\gamma^T X_\gamma)^{-1}]^{-1} (\bar{\theta}_\gamma - \hat{\theta}_\gamma)}{n + \nu} (\Sigma_\gamma^{-1} + X_\gamma^T X_\gamma)^{-1}.$$

Επίσης, γνωρίζουμε ότι το ακόλουθο:

Λήμμα 3.1.1 Εάν $Y \sim \mathcal{T}_p(n, \mu, V)$ τότε για οποιαδήποτε διαμέριση του διανύσματος ως $Y^T = (Y_{(1)}^T \ Y_{(2)}^T)$ η κατανομή του $Y_{(1)}$ είναι $\mathcal{T}_q(n, \mu_{(1)}, V_{(11)})$, όπου q η διάσταση του

$$Y_{(1)} \text{ και } \mu = (\mu_{(1)}^T \ \mu_{(2)}^T), \quad V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{pmatrix}.$$

Έπεται ότι και κάθε θ_i θα κατανέμεται ως (μονοδιάστατη) t με $q_\gamma + 1 + \nu$ βαθμούς ελευθερίας και παραμέτρους $\hat{\mu}_{\gamma,i}, \hat{v}_{ii}$, δηλαδή το i -στοιχείο του διανύσματος $\hat{\mu}_\gamma$ και το i -διαγώνιο στοιχείο του \hat{V}_γ ,

$$\theta_{\gamma,i} | y, X \sim t_{q_\gamma+1+\nu}(\hat{\mu}_{\gamma,i}, \hat{v}_{ii}).$$

Συνεπώς, επανερχόμενοι στην κατανομή του συντελεστή θ_i υπό τη BMA το συνεχές τμήμα του θα είναι η μίξη των κατανομών $t_{q_\gamma+1+\nu}(\hat{\mu}_{\gamma,i}, \hat{v}_{ii})$ για τα μοντέλα γ τα οποία εκτιμούν $\theta_i \neq 0$. Ακριβώς, αυτή η ιδιότητα των κατανομών υπό τη BMA κάνει την διαδικασία περισσότερο «ανθεκτική» σε σχέση με την χρήση μιας $t_{n-q_\gamma-1}$, όπως γίνεται στο κλασσικό γραμμικό μοντέλο.

3.1.2 Zellner υλοποίηση

Εάν η υλοποίηση των Raftery *et al.* (1997) ξεκίνησε ως η αδιαμφισβήτητη επιλογή για την εφαρμογή του BMA η έρευνα πλέον έχει επικεντρωθεί στην μοντελοποίηση βάσει της εκ των προτέρων κατανομής του Zellner, για παράδειγμα Clyde (1999), Liang *et al.* (2008) και Fernandez *et al.* (2001) Αυτός είναι και ο λόγος που προτιμούμε να την καλούμε ως «υλοποίηση Zellner» αντίθετα με την προηγούμενη υλοποίηση η οποία έχει συνδέσει το όνομά της με τους Hoeting, Raftery και Madigan.

Έχουμε τονίσει το γεγονός ότι για τον προσδιορισμό της εκ των προτέρων κατανομής του Zellner χρειάζεται να οριστεί μόλις μία υπερπαραμέτρος, η ϕ . Επί πλέον των γενικών επιλογών τις οποίες αναφέραμε στην προηγούμενη ενότητα, έχουν αναπτυχθεί και εκτιμητές της υπερπαραμέτρου ϕ βασισμένοι σε Εμπειρικές Μπεϋζιανές μεθόδους (δες κεφ. 4).

Γνωρίζοντας ότι η εκ των προτέρων κατανομή του Zellner είναι ειδική περίπτωση της NG, οι υπολογισμοί και τα αποτελέσματα είναι αναμενόμενα.

Η πιθανοφάνεια θα είναι

$$p(y | \beta_\gamma, X, \gamma, \sigma^2) \propto \left(\frac{\phi}{1+\phi} \right)^{-q_\gamma/2} \left\{ \frac{1}{1+\phi} y^T \bar{H}_\gamma y + \frac{\phi}{1+\phi} (y - X_\gamma \bar{\beta}_\gamma)^T (y - X_\gamma \bar{\beta}_\gamma) \right\}^{-(n-1)/2}, \quad (3.1.4)$$

όπου $\bar{H}_\gamma = I - X_\gamma(X_\gamma^T X_\gamma)^{-1} X_\gamma^T$ ο «συμπληρωματικός» πίνακας του H_γ . Η πιθανοφάνεια εξαρτάται από έναν κυρτό συνδυασμό του συνήθους αθροίσματος των σφαλμάτων, $y^T \bar{H}_\gamma y$, και των σφαλμάτων της εκ των προτέρων επιλογής μας, $(y - X_\gamma \tilde{\beta}_\gamma)^T (y - X_\gamma \tilde{\beta}_\gamma)$.¹ Ενσωματώνοντας την εκ των προτέρων γνώση (ή εκτίμησή) μας η εξάρτηση της πιθανοφάνειας από το άθροισμα τετραγώνων των υπολοίπων «συρρικνώνεται» κατά τον παράγοντα $(1 + \phi)^{-1}$.

Η εκ των υστέρων περιθωριακή κατανομή του β_γ θα είναι q_γ -διάστατη t με παραμέτρους $\hat{\mu}_\gamma, \hat{V}_\gamma$ και n όπου

$$\hat{\mu}_\gamma = (\phi X_\gamma^T X_\gamma + X_\gamma^T X_\gamma)^{-1} X_\gamma^T y = [(1 + \phi) X_\gamma^T X_\gamma]^{-1} X_\gamma^T y$$

και

$$\hat{V}_\gamma = \frac{\frac{1}{1 + \phi} y^T H_\gamma y + \frac{\phi}{1 + \phi} (y - X_\gamma \bar{\beta}_\gamma)^T (y - X_\gamma \bar{\beta}_\gamma)}{n - 2} [(1 + \phi) X_\gamma^T X_\gamma]^{-1}.$$

3.2 Το κύριο πρόβλημα

Η εφαρμογή της διαδικασίας σε πρακτικά προβλήματα δεν είναι εύκολο να υλοποιηθεί με τη μέθοδο που έχουμε περιγράψει μέχρι τώρα, δηλαδή να υπολογίσουμε την πιθανοφάνεια κάθε μοντέλου και μετά την εκ των υστέρων πιθανότητα κάθε μοντέλου, όπως μας λέει η (2.1.2). Σε αυτή την ευθεία αντιμετώπιση του προβλήματος γίνεται απαγορευτικό το γεγονός ότι για μεγάλο αριθμό ανεξάρτητων μεταβλητών ο αριθμός των μοντέλων είναι αστρονομικός. Για παράδειγμα, για ένα σύνολο 30 ανεξάρτητων μεταβλητών - χωρίς να υπολογίσουμε πιθανούς όρους αλληλεπίδρασης και όρους υψηλότερης τάξης ($x_i \rightarrow x_i^r$) - θα

¹ Το συνολικό άθροισμα σφαλμάτων (SSY) μπορεί να γραφεί ως $y^T y = y^T \bar{H}_\gamma y + y^T H_\gamma y$, όπου $y^T \bar{H}_\gamma y$ το άθροισμα τετραγώνων των υπολοίπων (SSE) και $y^T H_\gamma y$ το αντίστοιχο άθροισμα «λόγω της παλινδρόμησης» (SSR).

χρειαστεί να υπολογίσουμε 2^{30} φορές την συνάρτηση πιθανότητας $\pi(\gamma | y)$.² Επομένως, είναι αναγκαίο να ελαττώσουμε τον αριθμό των μοντέλων που θα χρησιμοποιήσουμε στην εφαρμογή της διαδικασίας όταν ο αριθμός των ανεξάρτητων μεταβλητών δεν μας επιτρέπει να εξαντλήσουμε το σύνολο Γ . Η ελάττωση του αριθμού των μοντέλων θα μειώσει τους όρους του αθροίσματος $\sum_{\gamma \in \Gamma} p(y | \gamma)\pi(\gamma)$ (και κάθε άλλου αθροίσματος) σε $\sum_{\gamma \in \Gamma'} p(y | \gamma)\pi(\gamma)$, $\Gamma' \subset \Gamma$. Τα μοντέλα που θα παραλείψουμε είναι σημαντικό να είναι μοντέλα που θα είχαν ασήμαντη εκ των υστέρων πιθανότητα. Σε κάθε άλλη περίπτωση, οι εκ των υστέρων πιθανότητες των μοντέλων θα είναι σημαντικά υπερεκτιμημένες.

Στην επόμενη ενότητα παρουσιάζουμε τους πιο διαδεδομένους αλγόριθμους με τους οποίους επιδιώκουμε να καταλήξουμε σε ένα σύνολο Γ' , αρκετά συρρικνωμένο σε σχέση με το αρχικό Γ , με το ελάχιστο κόστος στην εκτίμηση των εκ των υστέρων πιθανοτήτων. Παράλληλα, εξετάζουμε και προσεγγίσεις για τον υπολογισμό των εκ των υστέρων πιθανοτήτων.

3.3 Αλγόριθμοι υλοποίησης & Προσεγγιστικές διαδικασίες

Υπάρχουν αρκετοί αλγόριθμοι οι οποίοι μπορούν να μας βοηθήσουν αποτελεσματικά. Τέτοιοι αλγόριθμοι χρησιμοποιούνται και στην κλασσική Στατιστική, σε προβλήματα όπως αυτά της επιλογής του βέλτιστου υποσυνόλου ανεξάρτητων μεταβλητών (*subset selection*) για ένα μοντέλο (γραμμικό ή μη). Αυτοί οι αλγόριθμοι είναι ντετερμινιστικοί. Η μπεϋζιανή φύση του προβλήματός μας όμως, επιτρέπει να χρησιμοποιήσουμε και στοχαστικούς αλγόριθμους οι οποίοι βασίζονται στη θεωρία των μαρκοβιανών αλυσίδων.

² Με 2^{30} διαφορετικά μοντέλα ακόμη και εάν η ανάλυση κάθε μοντέλου χρειάζεται 10^{-3} δευτερόλεπτα, θα χρειαστούν περίπου 12.5 ημέρες για να ολοκληρώσουμε τον υπολογισμό των εκ των υστέρων πιθανοτήτων!

3.3.1 Το Παράθυρο του Οκκάμ

Ας αρχίσουμε περιγράφοντας τον πρώτο αλγόριθμο ο οποίος βασίστηκε σε λογικές παραδοχές για τον αποκλεισμό ή όχι ενός μοντέλου από το σύνολο-«στόχο». Ο αλγόριθμος αυτός ονομάζεται Παράθυρο του Οκκάμ (*Occam's Window*) και αναπτύχθηκε από τους Madigan & Raftery (1994) αρχικά για γραφικά μοντέλα.

Αλγόριθμος 1: *Occam's Window* (A1)

Βήμα 1. Βρες το μοντέλο με τη μέγιστη εκ των υστέρων πιθανότητα.

Βήμα 2. Σχημάτισε το σύνολο $\mathcal{A} = \left\{ \gamma : \frac{\max_{\gamma'} [\pi(\gamma' | y)]}{\pi(\gamma | y)} \leq u \right\}$ για ένα u τέτοιο

ώστε το σύνολο \mathcal{A} να μην είναι ούτε μικρό ή ούτε υπερβολικά μεγάλο.

Βήμα 3. Σχημάτισε το σύνολο $\mathcal{B} = \left\{ \gamma : \exists \gamma' \in \Gamma, \gamma' \subset \gamma, \frac{\pi(\gamma' | y)}{\pi(\gamma | y)} > 1 \right\}$.

Βήμα 4. Το ζητούμενο σύνολο είναι το $\Gamma' = \mathcal{A} - \mathcal{B}$.

Από τον παραπάνω αλγόριθμο λείπει ένας μηχανισμός ο οποίος θα επιτρέψει τη μετακίνηση μεταξύ των μοντέλων. Ο μηχανισμός αυτός είναι σημαντικό να μην απαιτεί την πλήρη ανάλυση του κάθε μοντέλου (αλλιώς θα βρισκόμασταν στην αρχική κατάσταση). Ευτυχώς, υπάρχει ένας αλγόριθμος με αυτή τη δυνατότητα. Ο αλγόριθμος *branch & bounds* (ή *leaps & bounds*) των Furnival & Wilson (1974) θα αποτελέσει το κύριο εργαλείο για την υλοποίηση του (A1).³ Ο αλγόριθμος επιστρέφει τα r καλύτερα μοντέλα μεγέθους q με βάση κάποιο κριτήριο (π.χ το R^2).⁴ Βασιζόμενος στο γεγονός ότι για δύο εμφωλευμένα μοντέλα γ, γ' ισχύει

³ Δεν θα περιγράψουμε λεπτομερώς τη δομή του αλγορίθμου η οποία βασίζεται σε *sweep* πράξεις πινάκων. Παραπέμπουμε στον Miller (2002) ο οποίος περιγράφει τον αρχικό αλγόριθμο αλλά και παραλλαγές του σε ικανοποιητικό βαθμό.

⁴ Η R μας προσφέρει υλοποίηση του αλγορίθμου στη βιβλιοθήκη *leaps*.

$SSE_\gamma \geq SSE_{\gamma'}$, για $q_\gamma \leq q_{\gamma'}$, η απόρριψη του γ συνεπάγεται και την απόρριψη κάθε γ' . Αυτό είναι το βήμα bounds.

Προφανώς, ο απευθείας υπολογισμός της πιθανοφάνειας είναι απευκταίος, οπότε στρεφόμαστε στην αναζήτηση μιας αξιόπιστης προσέγγισης. Το (ήμισυ) του BIC αποτελεί μια προσέγγιση τάξης $O(1)$ για τον λογάριθμο της πιθανοφάνειας ενός μοντέλου (Schwarz, 1978).

$$\log f(y | \hat{\theta}_\gamma) \cong \frac{1}{2}(BIC_\gamma).$$

Για γραμμικά μοντέλα, το BIC έχει την εξαιρετικά βολική μορφή για τον αλγόριθμο των Furnival & Wilson

$$\begin{aligned} BIC_\gamma &= n \log(1 - R^2) - (q_\gamma + 1) \log n \\ &= n \log \left(1 - \frac{SSE_\gamma}{SSY} \right) - (q_\gamma + 1) \log n, \end{aligned}$$

όπου $SSY = (y - \bar{y})^T (y - \bar{y})$

Συνεπώς, ο υπολογισμός του R^2 μας επιτρέπει να εντοπίσουμε το μοντέλο με τη μέγιστη εκ των υστέρων πιθανότητα.

Τώρα, το βήμα 2 του (A1) γίνεται

$$\frac{\max_\gamma [\pi(\gamma | y)]}{\pi(\gamma | y)} \leq u \Leftrightarrow \frac{\max_\gamma [p(y | \gamma) \pi(\gamma)]}{p(y | \gamma) \pi(\gamma)} \leq u$$

και όταν $\pi(\gamma)$ είναι η ομοιόμορφη κατανομή (2.2.5),

$$\begin{aligned} \frac{\max_{\gamma} [p(y | \gamma) \pi(\gamma)]}{p(y | \gamma) \pi(\gamma)} &\leq u \Leftrightarrow \\ \frac{\max_{\gamma} [p(y | \gamma)]}{p(y | \gamma)} &\leq u \Leftrightarrow \\ \frac{\max_{\gamma} \left[\exp \left(-\frac{1}{2} BIC_{\gamma} \right) \right]}{\exp \left(-\frac{1}{2} BIC_{\gamma} \right)} &\leq u \end{aligned}$$

Τελικά, η συνθήκη την οποία πρέπει να ικανοποιεί ένα μοντέλο για να περιληφθεί στο σύνολο \mathcal{A} είναι

$$BIC_{\hat{\gamma}} \leq 2 \log u + BIC_{\hat{\gamma}},$$

όπου $\hat{\gamma}$ το μοντέλο με την μεγαλύτερη εκ των υστέρων πιθανότητα.

Για τους υπολογισμούς είναι προτιμότερη η χρήση της επόμενης μορφής της συνθήκης,

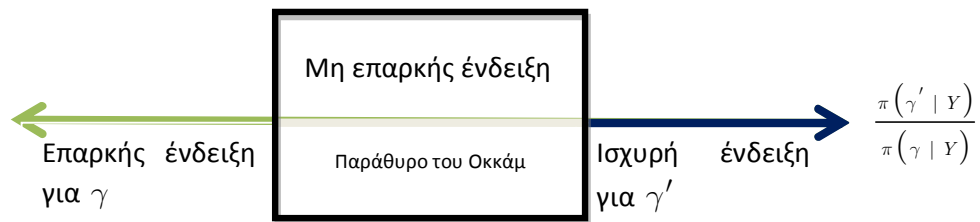
$$\exp \left(\frac{BIC_{\hat{\gamma}} - \max(BIC_{\gamma})}{2} \right) \leq u.$$

Ακριβώς, με την ίδια λογική προκύπτει και η συνθήκη του δεύτερου βήματος.

$$\begin{aligned} \frac{\pi(\gamma' | y)}{\pi(\gamma | y)} > 1 &\Leftrightarrow \frac{p(y | \gamma')}{p(y | \gamma)} > 1 \Leftrightarrow \\ \frac{\exp \left(-\frac{1}{2} BIC_{\gamma'} \right)}{\exp \left(-\frac{1}{2} BIC_{\gamma} \right)} &> 1 \end{aligned}$$

Και πάλι, ενδείκνυται η χρήση της επόμενης

$$\frac{\exp \left\{ -\frac{1}{2} (BIC_{\gamma'} - \max(BIC_{\gamma})) \right\}}{\exp \left\{ -\frac{1}{2} (BIC_{\gamma} - \max(BIC_{\gamma})) \right\}} > 1$$



Σχήμα 3-2 : Παράθυρο του Οκκάμ

Για να καταλάβουμε πώς δουλεύει το παράθυρο, το Σχήμα 3-2 συγκρίνει δύο μοντέλα γ, γ' με $q_\gamma \leq q_{\gamma'}$. Για να απορρίψουμε το «μεγάλο» μοντέλο απαιτείται ασθενής ένδειξη για το γ . Αντίθετα, η απόρριψη του «μικρού» γίνεται μόνο με ισχυρή ένδειξη για το γ' . Τέλος, μεταξύ των δύο περιοχών που ορίζει ο λόγος των εκ των υστέρων πιθανοτήτων (*posterior odds*) υπάρχει μια περιοχή αβεβαιότητας, αυτή είναι το παράθυρο του Οκκάμ. Στο παράθυρο Σχήματος 3-2 διατηρούνται και τα δύο μοντέλα.

Γενικά, εάν απορριφθεί το γ τότε αποκλείονται και όλα τα εμφωλευμένα σε αυτό μοντέλα.

Η διαδικασία που περιγράφηκε είναι η εφαρμογή του παραθύρου του Οκκάμ όπως το υλοποίησαν οι Raftery *et al.* (1997).

3.3.2 BMA μέσω Πληροφοριακών Κριτηρίων

Προφανώς υπάρχουν και άλλες προσεγγίσεις της πιθανοφάνειας. Επεκτείνοντας την προσέγγιση του BIC, μπορούμε να ορίσουμε την κλάση των Γενικευμένων Πληροφοριακών Κριτηρίων (Generalized Information Criteria),

Ορισμός 3.2.1 Η ποσότητα $GIC := n \log(1 - R_\gamma^2) + a q_\gamma$, όπου a η ποινή για το μοντέλο γ είναι ένα γενικευμένο πληροφοριακό κριτήριο εάν $\lim_{n \rightarrow \infty} \frac{a}{n} = 0$ (Nishii, 1984).

Τα γνωστά πληροφοριακά κριτήρια προφανώς ικανοποιούν την παραπάνω συνθήκη, όπως φαίνεται στον Πίνακα 3-1.

	a	$\lim_{n \rightarrow \infty} \frac{a}{n}$
AIC	2	0
BIC	$\log n$	0
HC(c)	$cq_\gamma \log \log n$	0

Πίνακας 3-1 : Οριακή συμπεριφορά σύμφωνα με τον Ορισμό 3.2.1

Η προσέγγιση για την πιθανοφάνεια είναι

$$p(y | \gamma) \propto \exp\left(-\frac{1}{2} GIC_\gamma\right)$$

καθώς και η εκ των υστέρων πιθανότητα του μοντέλου γ θα προσεγγίζεται από την

$$\pi(\gamma | y) \simeq \frac{\exp\left(-\frac{1}{2} GIC_\gamma\right) \pi(\gamma)}{\sum_\gamma \exp\left(-\frac{1}{2} GIC_\gamma\right) \pi(\gamma)}$$

Στην εργασία του Noble (2001) συναντούμε δύο θεωρήματα πρακτικής σημασίας.

Θεώρημα 3.3.1 (*Δυϊκό Θεώρημα Κριτηρίων Πληροφορίας & επιλογής εκ των προτέρων κατανομής*) Η χρήση ενός GIC για την προσέγγιση της περιθωριακής πιθανοφάνειας σε συνδυασμό με την ομοιόμορφη κατανομή στο Γ είναι ισοδύναμη με τη χρήση του BIC και εκ των προτέρων κατανομή στο Γ της μορφής,

$$\pi_a(\gamma) = \frac{\exp\left\{-\frac{1}{2} q_\gamma (a - \log n)\right\}}{\sum_{j=0}^{q_\gamma} \binom{q_\gamma}{j} \exp\left\{-\frac{1}{2} j (a - \log n)\right\}} \dots \quad (3.3.1)$$

Το θεώρημα μας επιτρέπει να εισαγάγουμε μια συγκεκριμένη κλάση εκ των προτέρων κατανομών, όπως αυτή που ορίζεται από την (3.1.3), σε μία προσέγγιση της περιθωριακής πιθανοφάνειας μέσω ενός GIC.

Θεώρημα 3.3.2 Εάν η ανεξάρτητη μεταβλητή X_j , $j = 1, 2, \dots, p$, περιλαμβάνεται στο $\gamma \in \Gamma$ με εκ των προτέρων πιθανότητα w , τότε η εκ των προτέρων κατανομή στο Γ

είναι ισοδύναμη με αυτήν που προκύπτει από την προσέγγιση της περιθωριακής πιθανοφάνειας με ένα GIC για το οποίο η ποιμή είναι ίση με

$$a^* = \log n - 2 \log \left(\frac{w}{1-w} \right)$$

και στο Γ έχουμε επιλέξει την ομοιόμορφη εκ των προτέρων κατανομή.

Το τελευταίο θεώρημα ελαφρύνει αρκετά τον υπολογιστικό φόρτο καθώς στην προσέγγιση της περιθωριακής πιθανοφάνειας περιλαμβάνεται και η εκ των προτέρων μοντελοποίηση στο Γ .

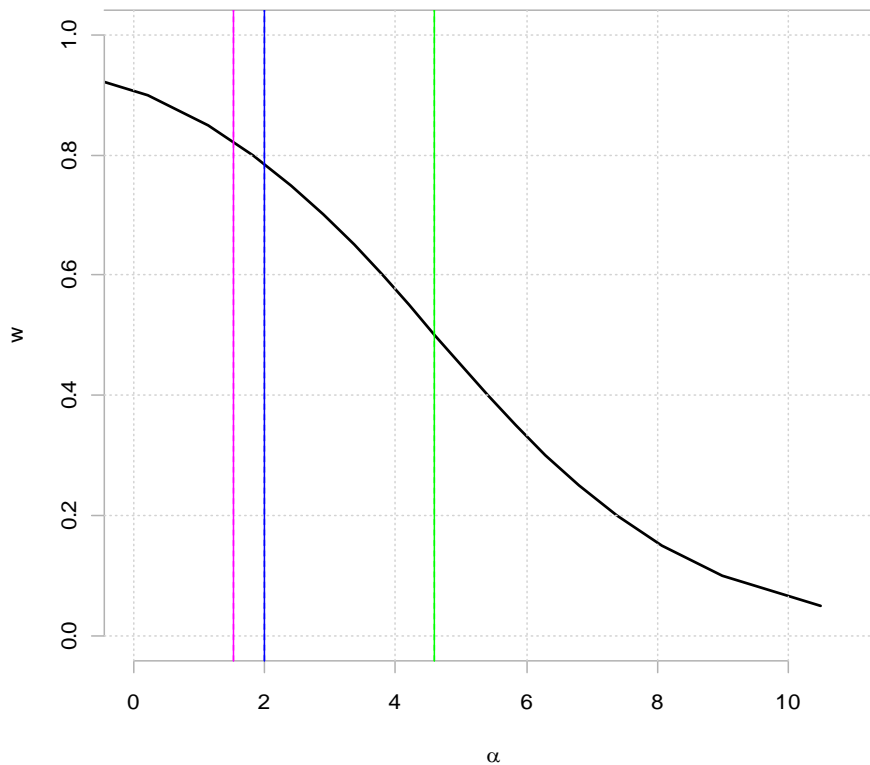
Η εκ των προτέρων πιθανότητα που αντιστοιχίζει σε κάθε X_j , $j = 1, 2, \dots, p$, κάθε ένα από τα γνωστά κριτήρια πληροφορίας. Αυτή θα δίνεται από την

$$w = \frac{\exp \left\{ -\frac{1}{2} (a - \log n) \right\}}{1 + \exp \left\{ -\frac{1}{2} (a - \log n) \right\}}. \quad (3.3.2)$$

Παράδειγμα Ο Ορισμός 3.2.1 ικανοποιείται για κάθε $a \in \mathbb{R}$. Συνεπώς μπορούμε να βρούμε ένα a τέτοιο ώστε η εκ των προτέρων πιθανότητα w για δεδομένο μέγεθος δείγματος n , να είναι η επιθυμητή.

w	n	a	w	n	a
0.2	20	5.76832	0.4	20	3.80666
	50	6.68461		50	4.72295
	100	7.37776		100	5.4161
0.3	20	4.69033	0.65	20	1.75765
	50	5.60662		50	2.67394
	100	6.29977		100	3.36709

Πίνακας 3-2: Εύρεση αντιστοιχίας ποιμής ενός GIC - εκ των προτέρων πιθανότητας (για γνωστό μέγεθος δείγματος)



Σχήμα 3-3 : Εκ των προτέρων πιθανότητες που υποθέτουν τα γνωστά Πληροφοριακά Κριτήρια

Ένα πιθανώς ανεπιθύμητο γεγονός είναι ότι καθώς $n \rightarrow \infty$ για $a = 2$, $cq_\gamma \log \log n$ η εκ των προτέρων πιθανότητα στην (3.3.2) τείνει στο 1. Αντίθετα, για $a = \log n$, η εκ των προτέρων πιθανότητα κάθε μεταβλητής X_j , $j = 1, 2, \dots, p$, είναι 0.5 (σταθερά). Το Σχήμα 3-3 δείχνει την εκ των προτέρων πιθανότητα που συνεπάγεται η χρήση των παραπάνω ποινών για δείγμα 100 παρατηρήσεων.

Αυτή η οριακή συμπεριφορά είναι που δεν επέτρεψε την υιοθέτηση τους σε εφαρμογές με μεγάλα μεγέθη δείγματος, πέρα από το BIC και σπανιότερα το AIC. Σε μικρότερα δείγματα όμως, μπορεί να προσφέρει βελτίωση των αποτελεσμάτων. Αξιοσημείωτη είναι η αυτόματη φύση της προσέγγισης, η οποία δεν απαιτεί προσδιορισμό υπερπαραμέτρων και η υλοποίησή της βασίζεται σε βασικές ποσότητες τις οποίες παρέχει κάθε πακέτο ανάλυσης δεδομένων.

Η χρησιμότητα των κριτηρίων πληροφορίας έχει εκτιμηθεί περισσότερο από τους χρήστες της εκ των προτέρων κατανομής του Zellner. Οι Fernandez *et al.* (2001)

έδειξαν ότι η επιλογή της υπερπαραμέτρου ϕ μπορεί να αναχθεί στην επιλογή ενός κριτηρίου πληροφορίας για την προσέγγιση της (περιθωριακής) πιθανοφάνειας (δες Πίνακα 3-3).

	a	ϕ
AIC	2	0,255
BIC	$\log n$	n
HC(1)	$q_\gamma \log(\log n)$	$\frac{\log n}{\log(q_\gamma + 1)}$
HC(3)	$3q_\gamma \log(\log n)$	$(\log n)^3$

Πίνακας 3-3 : Αντιστοιχία ποινής – υπερπαραμέτρου της εκ των προτέρων κατανομής Zellner

3.3.3 Στοχαστικές αλγοριθμικές μέθοδοι

Ο αλγόριθμος A1 δεν είναι αποδοτικός για μεγάλο αριθμό ανεξάρτητων μεταβλητών. Αυτό οφείλεται κυρίως στους περιορισμούς των αλγορίθμων leaps & bounds. Έτσι, για $p \geq 20$ χρειαζόμαστε μια εναλλακτική λύση. Σε αυτή την περίπτωση, θα μπορούσε να χρησιμοποιηθεί μια προκαταρκτική διαδικασία επιλογής μεταβλητών, π.χ μια stepwise διαδικασία, έτσι ώστε να μειώσουμε τον αριθμό των ανεξάρτητων μεταβλητών. Όμως στη BMA στόχος μας είναι να χρησιμοποιήσουμε κάθε καταγεγραμμένη μεταβλητή η οποία υπάρχει στα δεδομένα μας.

Η λύση βρίσκεται στην προσέγγιση του (ολοένα αυξανόμενων όρων) αθροίσματος μέσω προσομοίωσης. Αυτή η προσέγγιση είναι ιδιαίτερα διαδεδομένη στη Μπεϋζιανή Στατιστική κυρίως εξαιτίας του προβλήματος της ύπαρξης σ.π.π. σε κλειστή μορφή. Οι αλγόριθμοι αυτοί ονομάζονται Markov Chain Monte Carlo (MCMC).

Συγκεκριμένα, στόχος μας είναι να κατασκευάσουμε έναν αλγόριθμο ο οποίος μετακινούμενος στο χώρο των μοντέλων Γ θα μας επιτρέψει να εκτιμήσουμε την εκ των υστέρων πιθανότητα κάθε μοντέλου γ . Αυτό μας το εγγυάται το Εργοδικό Θεώρημα.

Αρχικά, πρέπει να οριστεί η «μετακίνηση» στο χώρο των μοντέλων. Χρειάζεται συνεπώς να προσδιοριστεί μια κατανομή μετάβασης (*transition ή jump distribution*), $j(\gamma' | \gamma)$.

Η πιο απλή κατανομή μετάβασης είναι η επόμενη

$$j(\gamma' | \gamma) = \frac{1}{p}, \quad (3.3.3)$$

όταν ισχύει $\sum_{i=1}^p |\gamma'_i - \gamma_i| = 1$.

Σε αυτή την περίπτωση η μετακίνηση ισοδυναμεί με το επόμενο βήμα :

+ / - (ή Add/Drop): Επιλέγουμε τυχαία μία ανεξάρτητη μεταβλητή i . Προσθέτουμε ή αφαιρούμε την ανεξάρτητη μεταβλητή i ανάλογα με το εάν συμπεριλαμβάνεται ή όχι στο μοντέλο γ . Τα γ', γ διαφέρουν κατά μια μεταβλητή.

Μια άλλη κατανομή μετάβασης η οποία χρησιμοποιείται στις εφαρμογές είναι η επόμενη.

$$j(\gamma' | \gamma) = \frac{1}{(p - q_\gamma) q_\gamma}, \quad (3.3.4)$$

όταν ισχύει $\sum_{i=1}^p |\gamma'_i - \gamma_i| = 2$.

↔ (ή Swap): Επιλέγουμε τυχαία μία ανεξάρτητη μεταβλητή i η οποία δεν περιλαμβάνεται στο μοντέλο γ . Επίσης, επιλέγουμε τυχαία μία ανεξάρτητη μεταβλητή i' η οποία περιλαμβάνεται στο μοντέλο γ . Εναλλάσσουμε την κατάσταση των ανεξάρτητων μεταβλητών i, i' . Τα γ', γ διαφέρουν κατά δύο μεταβλητές.

Ένας γενικός αλγόριθμος ο οποίος χρησιμοποιείται ευρέως σε προβλήματα μεταβλητής διάστασης του παραμετρικού χώρου (όπως η επιλογή μοντέλου ή η BMA) είναι ο επόμενος και οφείλεται στον Green (1995).

Αλγόριθμος 2: RJMCMC (B)

Έστω ότι βρισκόμαστε στο μοντέλο γ , με παραμετρικό διάνυσμα θ_γ .

Βήμα 1. Πρότεινε τη μετακίνηση στο μοντέλο γ' με πιθανότητα $j(\gamma' | \gamma)$.

Βήμα 2. Πάρε ένα διάνυσμα u από την $\varphi(u | \theta_\gamma, \gamma, \gamma')$ διάστασης $|\dim(\gamma) - \dim(\gamma')|$ (ή $|q_\gamma - q_{\gamma'}|$).

Βήμα 3. Θέσε $(\theta_{\gamma'}, u') = h_{\gamma, \gamma'}(\theta_\gamma, u)$, για $h_{\gamma, \gamma'}(\bullet, \bullet)$ μια αντιστρέψιμη συνάρτηση. Τότε $\dim(\theta_\gamma) + \dim(u) = \dim(\theta_{\gamma'}) + \dim(u')$.

Βήμα 4. Αποδέξου την μετακίνηση με πιθανότητα

$$a(\gamma, \gamma') = 1 \wedge \frac{p(y | \theta_{\gamma'}, \gamma') \pi(\theta_{\gamma'} | \gamma') \pi(\gamma') j(\gamma | \gamma')}{p(y | \theta_\gamma, \gamma) \pi(\theta_\gamma | \gamma) \pi(\gamma) j(\gamma' | \gamma)} \times \frac{\varphi(u' | \theta_{\gamma'}, \gamma', \gamma)}{\varphi(u | \theta_\gamma, \gamma, \gamma')} \left| \frac{\partial h_{\gamma, \gamma'}(\theta_\gamma, \gamma)}{d(\theta_\gamma, \gamma)} \right|$$

Βήμα 5. Θέσε $\gamma = \gamma'$ εάν η μετακίνηση έγινε αποδεκτή αλλιώς παράμεινε στο γ .

Η βασική λειτουργία του αλγορίθμου είναι να συμπληρώσει το παραμετρικό διάνυσμα με τη μικρότερη διάσταση «γεννώντας» παραμέτρους από την κατανομή-πρότασης (*proposal density*), $\varphi(u | \theta_\gamma, \gamma, \gamma')$. Στη συνέχεια, μέσω της αντιστοίχισης $h_{\gamma, \gamma'} : \Theta_\gamma \times \mathbb{R}^{q_{\gamma'} - q_\gamma} \rightarrow \Theta_{\gamma'}$, δίνεται το παραμετρικό διάνυσμα του προτεινόμενου μοντέλου, $\theta_{\gamma'}$. Απαιτείται η αντιστοίχιση $h_{\gamma, \gamma'}$ να είναι αντιστρέψιμη ώστε να μπορεί

⁵ Συμβολίζουμε $x \wedge y = \min(x, y)$ και $x \vee y = \max(x, y)$.

να γίνει και η αντίθετη κίνηση, δηλαδή να μετακινείται από μικρότερα σε μεγαλύτερα μοντέλα. Ισχύει $h_{\gamma, \gamma'}^{-1} = h_{\gamma', \gamma}$

Για να εφαρμοστεί ο παραπάνω αλγόριθμος χρειάζεται να προσδιοριστούν οι $j(\gamma' | \gamma)$, $\varphi(u | \theta_\gamma, \gamma, \gamma')$ και $h_{\gamma, \gamma'}$. Για τα μοντέλα που περιγράφηκαν οι μορφές των περιθωριακών πιθανοφανειών μας επιτρέπουν να παραλείψουμε εντελώς το 2^ο και το 3^ο βήμα και να απλοποιήσουμε αρκετά το 4^ο. Συγκεκριμένα εάν χρησιμοποιηθεί ως κατανομή πρότασης η $p(\theta_\gamma | Y, \gamma)$, η ιακωβιανή ορίζουσα είναι ίση με 1 και η πιθανότητα είναι πλέον

$$a(\gamma, \gamma') = 1 \wedge \frac{f(y | \theta_{\gamma'}, \gamma') \pi(\theta_{\gamma'} | \gamma') \pi(\gamma') j(\gamma | \gamma')}{f(y | \theta_\gamma, \gamma) \pi(\theta_\gamma | \gamma) \pi(\gamma) j(\gamma' | \gamma)},$$

(Lopes, 2008). Περαιτέρω απλοποίηση της πιθανότητας αποδοχής είναι δυνατή. Οι κατανομές μετάβασης (3.3.3) και (3.3.4) είναι συμμετρικές, δηλαδή ισχύει $j(\gamma' | \gamma) = j(\gamma | \gamma')$. Τότε,

$$a(\gamma, \gamma') = 1 \wedge \frac{f(y | \theta_{\gamma'}, \gamma') \pi(\theta_{\gamma'} | \gamma') \pi(\gamma')}{f(y | \theta_\gamma, \gamma) \pi(\theta_\gamma | \gamma) \pi(\gamma)},$$

ενώ στην περίπτωση που θεωρηθεί η ομοιόμορφη κατανομή στο Γ , προκύπτει η (πλήρως) απλοποιημένη μορφή

$$\begin{aligned} a(\gamma, \gamma') &= 1 \wedge \frac{f(y | \theta_{\gamma'}, \gamma') \pi(\theta_{\gamma'} | \gamma')}{f(y | \theta_\gamma, \gamma) \pi(\theta_\gamma | \gamma)} \\ &= 1 \wedge \frac{\pi(\gamma' | y)}{\pi(\gamma | y)}. \end{aligned}$$

Τέλος, χρειάζεται να οριστεί η αρχική κατάσταση του γ , δηλαδή ένα αρχικό μοντέλο $\gamma^{(0)}$. Συνήθως, ορίζεται το μηδενικό, $\gamma = \mathbf{0}_p$, ή το πλήρες, $\gamma = \mathbf{1}_p$. Εναλλακτικά, υπάρχει η δυνατότητα να χρησιμοποιήσουμε μια διαδικασία προκαταρκτικού ελέγχου (*pretest procedure*) επιλέγοντας ένα μοντέλο με $q_{\gamma^{(0)}}$ το πλήθος ανεξάρτητες

μεταβλητές και διατηρώντας αυτές που η τιμή της στατιστικής συνάρτησης $T = \sqrt{n} \hat{\beta}_i / s$ για την υπόθεση $\beta_i = 0$ είναι μεγαλύτερη από το 2 (Koop, 2003).

Ο παραπάνω αλγόριθμος περιγράφηκε από την Clyde (1999) και ανάλογα με την επιλογή της κατανομής μετάβασης έχουμε δύο ξεχωριστούς αλγόριθμους. Εάν επιλέξουμε να κάνουμε βήματα $+ / -$ έχουμε τον αλγόριθμο MCMCMC ή MC³ (Markov Chain Monte Carlo Model Composition) των Madigan και York (1995). Αυτός είναι ο αλγόριθμος που χρησιμοποίησαν οι Raftery *et al.* (1997) και αποτελεί τον πιο διαδεδομένο αλγόριθμο στην πράξη. Οι Denison *et al.* (1998) υλοποίησαν τον αλγόριθμο συνδυάζοντας τα $+ / -$ βήματα με βήματα εναλλαγής. Για περισσότερους αλγορίθμους παραπέμπουμε στο άρθρο των Eklund & Karlsson (2007).

Όπως συμβαίνει σε κάθε αλγόριθμο, το κύριο ερώτημα είναι εάν συνέκλινε στην οριακή κατανομή για την οποία σχεδιάστηκε. Η σύγκλιση μπορεί να ελεγχθεί χρησιμοποιώντας ένα προκαθορισμένο υποσύνολο μοντέλων, S , για το οποίο οι εκ των υστέρων πιθανότητες των στοιχείων του υπολογίζονται αναλυτικά. Στη συνέχεια, υπολογίζοντας τη συσχέτιση μεταξύ των αναλυτικών, $\pi(\gamma | y)$, και των εργοδικών εκ των υστέρων πιθανοτήτων, $\hat{\pi}(\gamma | y)$,

$$\rho := \text{corr} \left\{ \pi(\gamma | y), \hat{\pi}(\gamma | y) \right\},$$

αποφασίζουμε για τη σύγκλιση. Προφανώς, καθώς $\rho \rightarrow 1$, τόσο πιο σίγουροι είμαστε για την σύγκλιση. Συνήθως, θεωρούμε ότι το S περιέχει τα k μοντέλα με την μεγαλύτερη εκ των υστέρων πιθανότητα.

3.4 Ένα ολοκληρωμένο παράδειγμα

Στην παρούσα ενότητα θα επιχειρηθεί η επιβεβαίωση των επιχειρημάτων τα οποία προκρίνουν την BMA ως την βέλτιστη διαδικασία πρόβλεψης (Θεώρημα 2.1). Συγκεκριμένα, για τις 15 ανεξάρτητες μεταβλητές του συνόλου δεδομένων USCrime (δες Παράρτημα Δ για το «ιστορικό» των δεδομένων και περιγραφή των μεταβλητών του) η BMA θα εφαρμοστεί στο 50% των συνολικά 47 παρατηρήσεων και οι

υπόλοιπες θα χρησιμοποιηθούν για την αξιολόγηση της προβλεπτικής ικανότητας της διαδικασίας.

	$\Pr(\beta_i \neq 0 Y)$				Stepwise
	HRM		Zellner		
	MC ³	Occam's Window	MC ³	RJMCMC	
M	0.65	0.92	0.65	0.64	*
So	0.17	0.04	0.13	0.12	
Ed	0.85	0.99	0.80	0.79	*
Po1	0.64	0.53	0.56	0.55	
Po2	0.59	0.47	0.47	0.48	
LF	0.10	0.03	0.11	0.11	
M.F	0.10	0.07	0.11	0.11	
Pop	0.22	0.36	0.19	0.19	*
NW	0.36	0.59	0.25	0.25	*
U1	0.10	0.12	0.10	0.11	
U2	0.25	0.64	0.24	0.24	*
GDP	0.22	0.18	0.12	0.12	*
Ineq	0.97	1.00	0.96	0.96	*
Prob	0.61	0.91	0.56	0.56	*
Time	0.14	0.28	0.12	0.12	

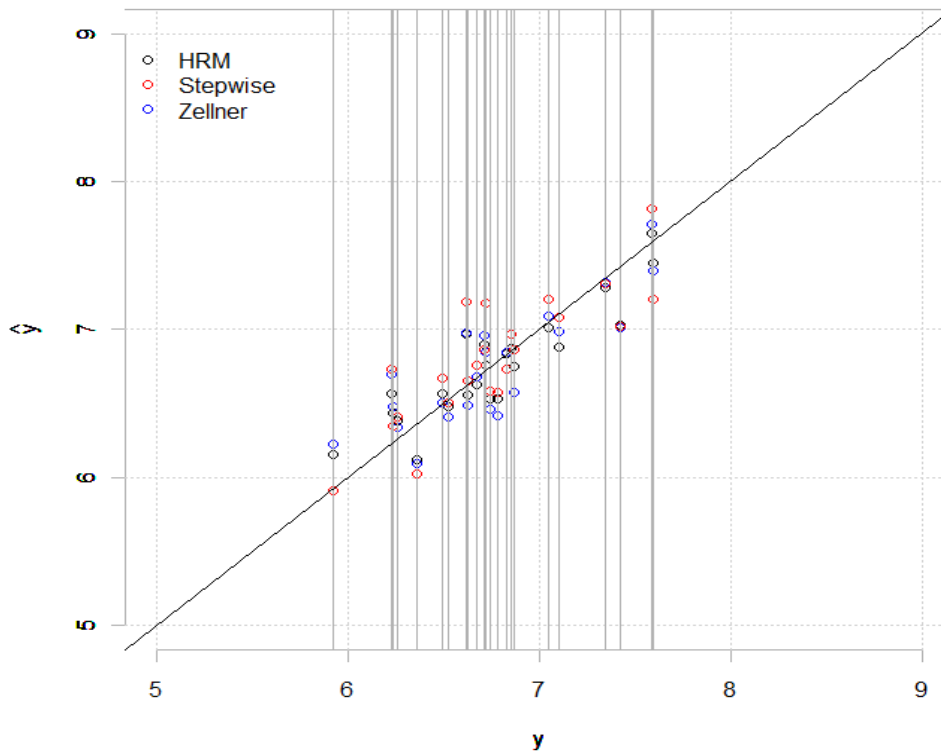
*Εμφάνιση σε 2 σημαντικά ψηφία ** 50:50

Πίνακας 3-4 : Εκ των υστέρων πιθανότητες συμπερίληψης σε ένα μοντέλο

Για τους αλγορίθμους έχουμε τα επόμενα στοιχεία. Το κυριότερο αποτέλεσμα είναι η σύγκλιση του αλγόριθμου (σύμφωνα με το κριτήριο της συσχέτισης).

	HRM		Zellner	
	MC ³	Occam's Window	MC ³	RJMCMC
Επαναλήψεις	800.000	-	800.000	800.000
% Επισκέψεων	4,55	-	16,36	12,85
Επιλεγέντα Μοντέλα	-	102	-	-
ρ	> 0,95	-	> 0,95	> 0,95

Πίνακας 3-5 : Χαρακτηριστικά των αλγορίθμων



Σχήμα 3-4 : Πραγματικές και προβλεπόμενες τιμές για τα μοντέλα HRM, Zellner καθώς και το επιλεγέν με τη stepwise διαδικασία

Ακολουθώντας τους Raftery *et al.* (1997), οι προβλέψεις βασίζονται στα αποτελέσματα του Παραθύρου του Οκκάμ καθώς και στο βασικό αλγόριθμο MC^3 για το μοντέλο Zellner.

Για την αξιολόγηση των προβλέψεων χρησιμοποιείται η ρίζα του Μέσου Τετραγωνικού Σφάλματος (MSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_v} (\hat{Y}_i - Y_i)^2}{n_v}},$$

όπου n_v το μέγεθος του δείγματος το οποίο χρησιμοποιείται για την αξιολόγηση του μοντέλου. Επίσης, για σύγκριση δίνεται και το αντίστοιχο RMSE για το μοντέλο το οποίο επιλέγει η stepwise διαδικασία.

	HRM	Zellner	Stepwise
RMSE	0.1875916	0.2312784	0.2537807

Πίνακας 3-6 : Σύγκριση του RMSE για τα μοντέλα HRM, Zellner καθώς και το επιλεγέν με τη stepwise διαδικασία

Παρατηρούμε ότι και τα δύο μοντέλα υπερέχουν του μοντέλου το οποίο έχει επιλεγεί από τη stepwise διαδικασία.

Στο πνεύμα των calibration plots των Raftery *et al.* (1997) το Σχήμα 3-4 είναι κατ' ουσία ένα διάγραμμα διασποράς το οποίο δείχνει πόσο κοντά βρίσκεται η πραγματική και η προβλεπόμενη τιμή για κάθε μοντέλο (HRM, Zellner, Stepwise) παρέχοντας και μια εποπτική αίσθηση των αποτελεσμάτων της σύγκρισης με το RMSE.

Το Σχήμα 3-5 αναπαριστά τις πυκνότητες πιθανότητας των συντελεστών του γραμμικού μοντέλου. Θεωρητικά, σύμφωνα με το κεφ. 2, αυτές είναι μείξεις κατανομών t δεδομένου ότι $\beta_i \neq 0$.

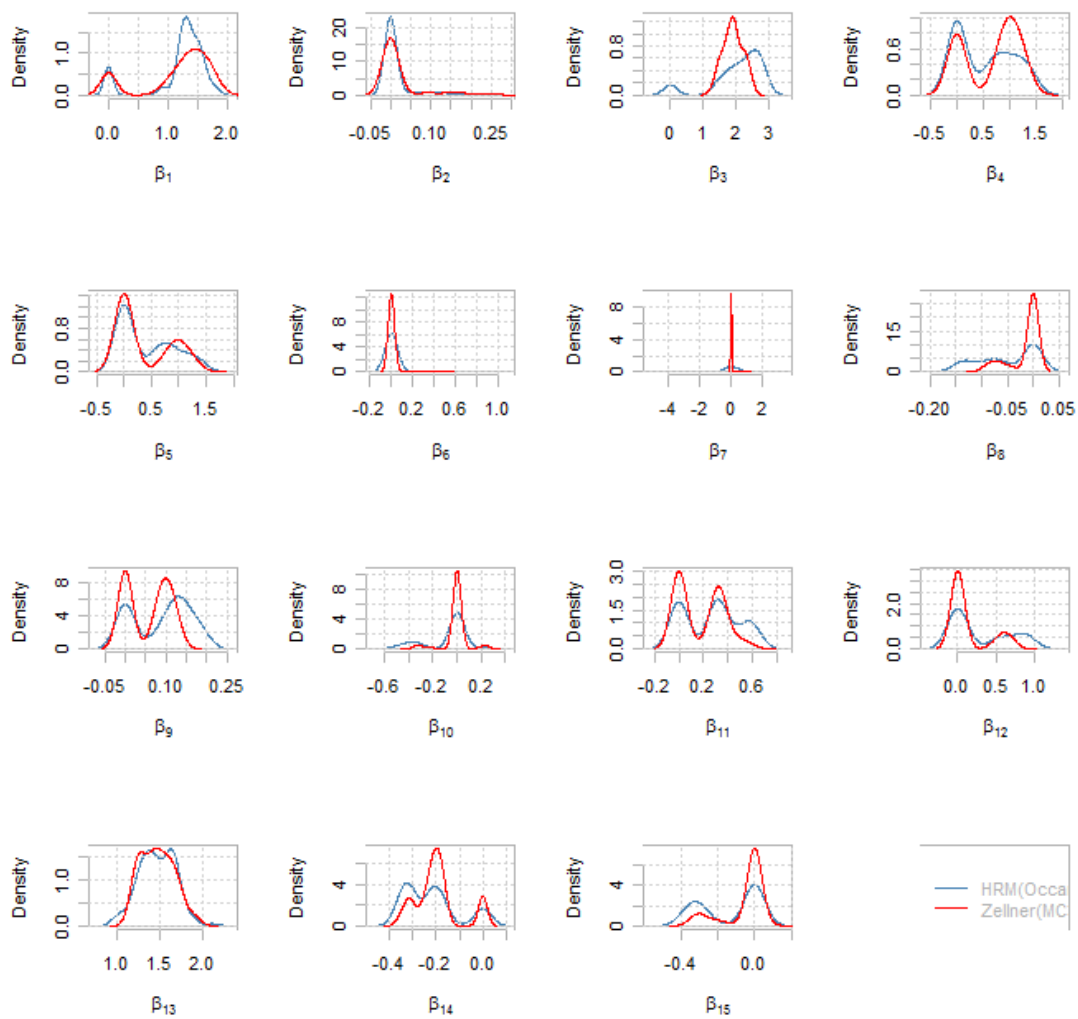
Από τα αποτελέσματα των αλγορίθμων είναι δυνατός ο υπολογισμός των σημειακών εκτιμήσεων καθώς και της τυπικής απόκλισής τους.

	HRM (Occam's Window)*		Zellner (MC ³)*		Πλήρες Μοντέλο*
	$E(\beta_i Y)$	$\sqrt{Var(\beta_i Y)}$	$E(\beta_i Y)$	$\sqrt{Var(\beta_i Y)}$	
M	1.30	0.68	1.09	0.73	2.52**
So	0.06	0.04	0.03	0.08	0.14
Ed	2.14	0.72	1.84	0.62	2.14**
Po1	0.54	0.56	0.68	0.54	0.55
Po2	0.48	0.55	0.35	0.52	0.22
LF	0.20	0.16	0.03	0.21	1.27**
M.F	-0.20	0.89	0.06	0.58	-4.03**
Pop	-0.03	0.05	-0.02	0.04	0.04
NW	0.07	0.07	0.05	0.06	0.05
U1	-0.02	0.16	-0.01	0.12	-0.09
U2	0.22	0.23	0.16	0.20	0.46**
GDP	0.12	0.32	0.12	0.31	0.71**
Ineq	1.47	0.41	1.45	0.35	1.68**
Prob	-0.21	0.12	-0.19	0.12	-0.45**
Time	-0.09	0.18	-0.05	0.13	-0.73**

*Εμφάνιση σε 2 σημαντικά ψηφία

** Σημαντικό σε ε.σ 5%

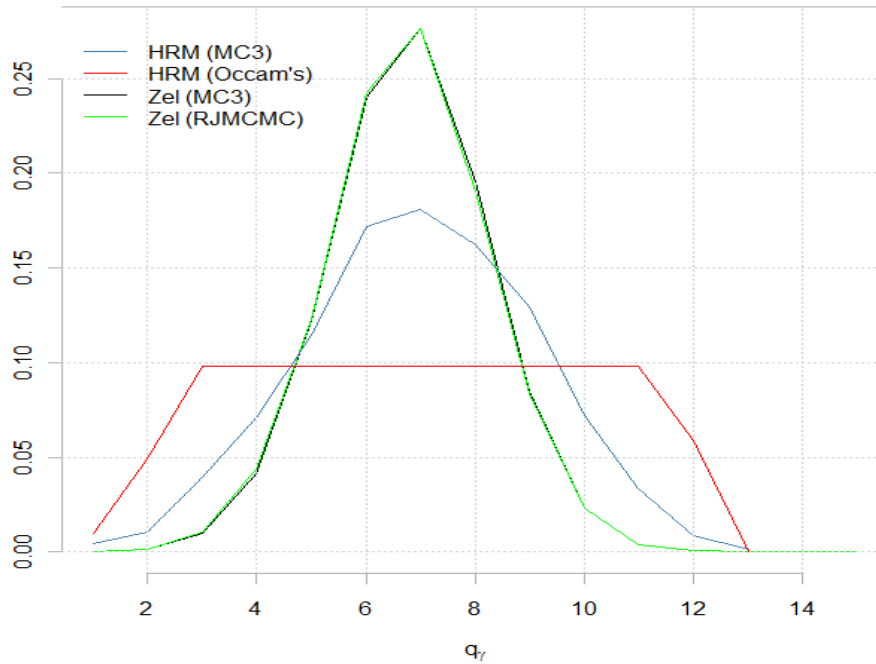
Πίνακας 3-7 : Σημειακές εκτιμήσεις και διακυμάνσεις των ανεξάρτητων μεταβλητών



Σχήμα 3-5 : Εκ των υστέρων κατανομές των συντελεστών

Στον Πίνακα 3-7 είναι εμφανής η εμφάνιση του shrinkage των σημειακών εκτιμητών στη BMA (εκτός των περιπτώσεων όπου η στρογγυλοποίηση δεν το επιτρέπει). Και σε αυτές τις περιπτώσεις όμως, ο εκτιμητής έχει «τραβηχτεί», έστω και ελάχιστα, προς το 0.

Οι σημειακές εκτιμήσεις περιέχουν όλη την πληροφορία από τους αλγόριθμους για τις εκ των υστέρων πιθανότητες των μοντέλων για το σχηματισμό των προβλέψεων.



Σχήμα 3-6 : Εκ των υστέρων κατανομή του μεγέθους των μοντέλων που επιλέχθηκαν από τους αλγόριθμους

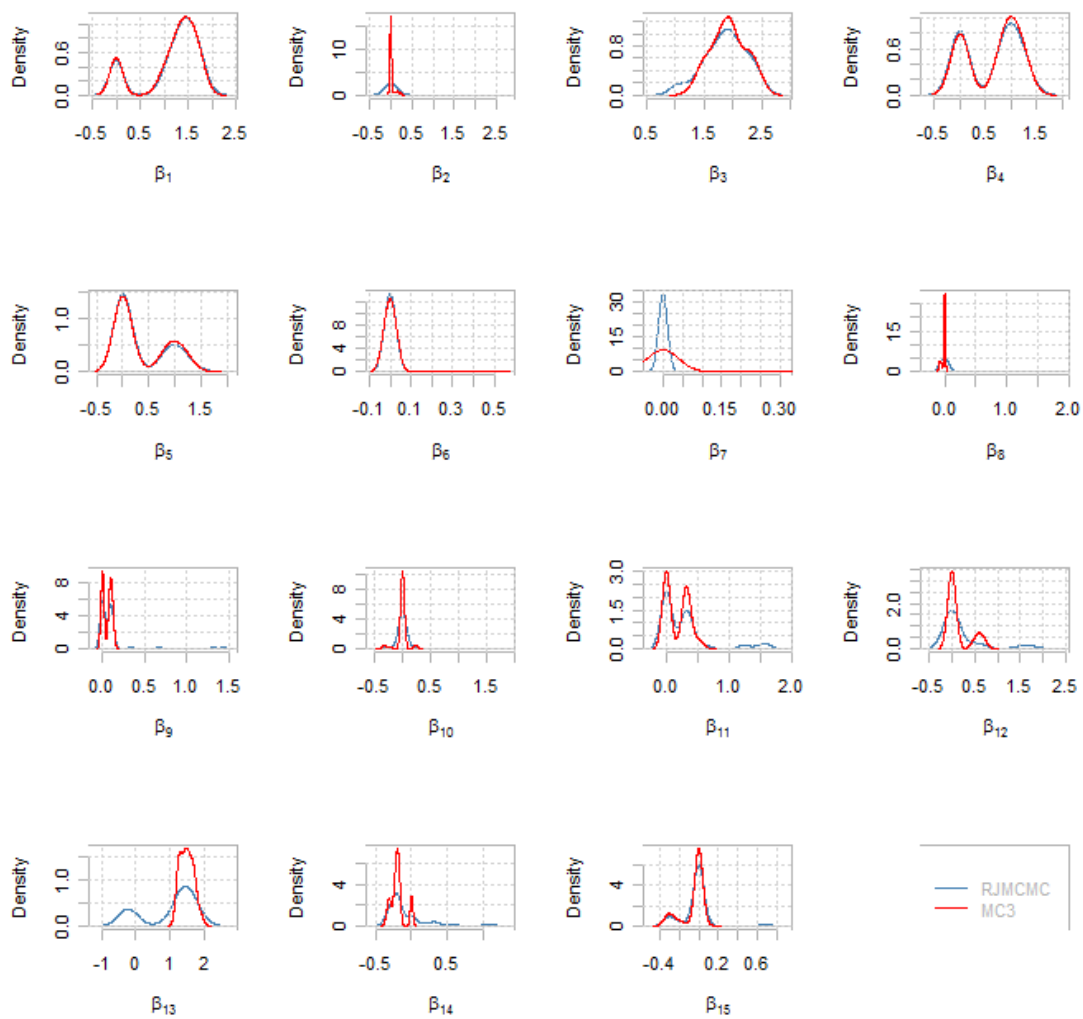
Οι προβλέψεις (σύμφωνα με την πρωταρχική σχέση (2.1.1)) δίνονται από την επόμενη

$$\begin{aligned}\hat{Y}_i &= a + \sum_{i=1}^p E(\beta_i | Y) X_i \\ &= a + \sum_{i=1}^p \left(\sum_{\gamma} E(\beta_i | Y, \gamma) \pi(\gamma | Y) \right) X_i.\end{aligned}$$

Ενδιαφέρουσα είναι και η εκ των υστέρων κατανομή του μεγέθους των μοντέλων που επιλέχθηκαν από τους αλγόριθμους. Ως εκ των προτέρων μέγεθος του μοντέλου είχε επιλεγεί το $p / 2$, δηλαδή 7.5 (ή $(p + 1) / 2 = 8.5$ εάν συμπεριλάβουμε την σταθερά η οποία περιλαμβάνεται σε όλα τα μοντέλα).

Η εκ των υστέρων κορυφή των κατανομών έχει μετακινηθεί (περίπου) στο 7 από το 7.5 και όπως αναμενόταν το σημαντικότερο ποσοστό πιθανότητας είναι συγκεντρωμένο στα μοντέλα με μέγεθος περίπου $p / 2$. Επίσης, αξιοσημείωτη είναι η κατανομή του Παράθυρου του Οκκάμ: είναι ομοιόμορφη για τις τιμές 3 έως 11.

Οι σημειακές εκτιμήσεις δίνονται στον Πίνακα 3-8.



Σχήμα 3-7 : Σύγκριση των κατανομών των συντελεστών για τους αλγόριθμους RJMCMC/MC³

Τέλος, αν και πιθανότητες που έδωσαν τόσο όσο ο MC³ όσο και ο RJMCMC είναι σχεδόν όμοιες, χρήσιμη θα ήταν η σύγκριση των κατανομών τους. Ποσοτικά δεν φαίνεται κάποια διαφοροποίηση, όμως εμφανίζονται διαφορές στα γραφήματα για κάποιους συντελεστές, π.χ. για τον συντελεστή της ανεξάρτητης μεταβλητής $\ln e_q$ (β_{13}).

	HRM		Zellner	
	MC ³	Occam's Window	MC ³	RJMCMC
$E(q, Y)$	6.97	6.29	6.80	6.78

Πίνακας 3-8 : Σημειακή εκτίμηση του μεγέθους των μοντέλων

Σε αυτή την ενότητα δείξαμε την υπεροχή της BMA ως προβλεπτικής διαδικασίας έναντι της παραδοσιακής επιλογής ενός μοντέλου. Επίσης, παρουσιάστηκε η ιδιότητα της συρρίκνωσης των εκτιμητών. Άξια αναφοράς είναι η μέτρια απόδοση της υλοποίησης του Zellner.

4.

Περαιτέρω θέματα στη ΒΜΑ

Στο προηγούμενο κεφάλαιο παρουσιάσαμε την «κλασσική» μορφή της ΒΜΑ, δηλαδή έτσι όπως διαμορφώθηκε έως το 2001 οπότε και δημοσιεύτηκε το άρθρο των Fernandez *et al.* Στο παρόν κεφάλαιο θα παρουσιάσουμε μερικά θέματα τα οποία δεν έχουν καθιερωθεί ακόμα στις εφαρμογές.

4.1 Προσδιορισμός των w_i

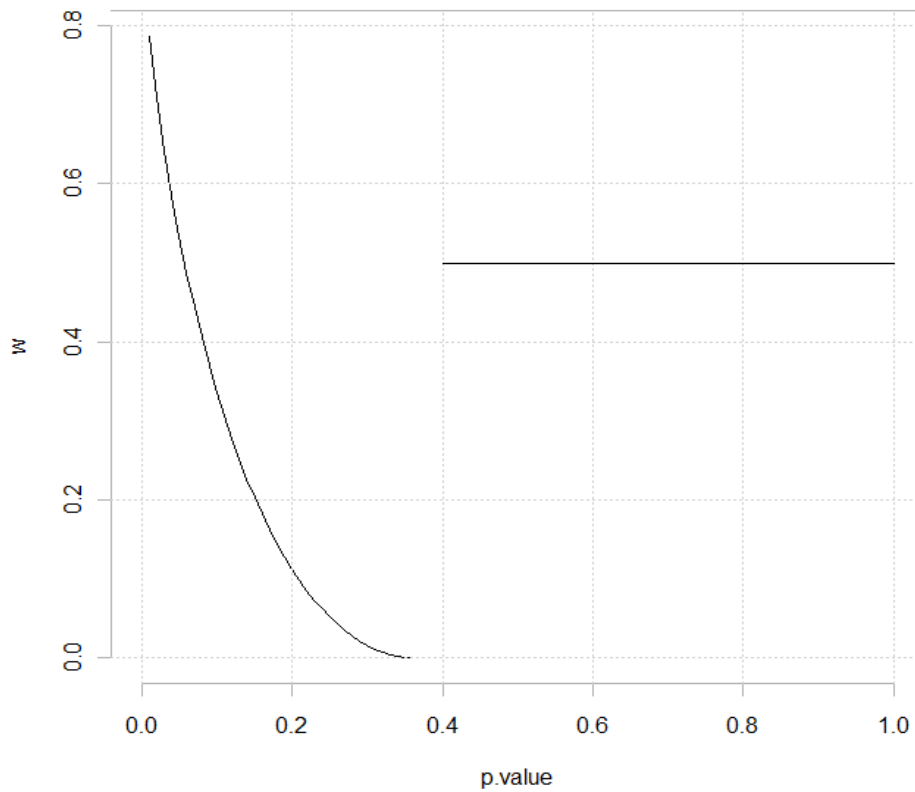
Όπως έχουμε αναφέρει στις εφαρμογές σπάνια έχουν χρησιμοποιηθεί επιλογές πέραν της ομοιόμορφης. Παρ' όλα αυτά θα ήταν χρήσιμο - έστω και περιγραφικά - να έχουμε μια πληροφόρηση για τη σημαντικότητα των ανεξάρτητων μεταβλητών.

Οι Sellke *et al.* (2001) περιγράφουν μια διαδικασία μέσω της οποίας τα p-values των ελέγχων της κλασσικής Στατιστικής για το γραμμικό μοντέλο, δηλαδή των ελέγχων οι οποίοι βασίζονται στη στατιστική συνάρτηση $T = \sqrt{n}\hat{\beta}_i/s$ για την υπόθεση $\beta_i = 0$, μπορούν να χρησιμοποιηθούν ώστε να βρεθεί ένα κάτω φράγμα για τον αντίστοιχο μπεϋζιανό έλεγχο. Αυτές οι τιμές μπορούν να αποτελέσουν τις εκ των προτέρων πιθανότητες για την συμπερίληψη των X_i σε ένα μοντέλο (Clyde, 2006).

Οι τιμές των w_i θα δίνονται από την

$$w_i = \begin{cases} \frac{1}{1 - e \Pr(|T| \geq t) \log \Pr(|T| \geq t)}, & \Pr(|T| \geq t) \leq e^{-1} \\ w^*, & \Pr(|T| \geq t) > e^{-1} \end{cases},$$

όπου e η σταθερά του Euler.



Σχήμα 4-1 : Προσδιορισμός των εκ των προτέρων πιθανοτήτων μέσω των p-values

Η περιγραφόμενη προσέγγιση ισχύει στο διάστημα $[0, e^{-1}]$. Για τιμές των p-values μεγαλύτερες του e^{-1} θεωρείται μια προκαθορισμένη τιμή, w^* .

Στο Σχήμα 4-1 αναπαρίσταται η παραπάνω διαδικασία για $w^* = 0.5$. Επίσης, δίνονται τα w_i για τα δεδομένα του παραδείγματος της 3.9.

	w_i
M	0.99
So	0.51
Ed	0.75
Po1	0.50
Po2	0.50
LF	0.62
M.F	0.67
Pop	0.50
NW	0.52
U1	0.50
U2	0.70
GDP	0.68
Ineq	0.99
Prob	0.99
Time	0.98

Πίνακας 4-1 : Προσδιορισμός των εκ των προτέρων πιθανοτήτων μέσω των p-values

4.2 Το w ως τυχαία μεταβλητή

Μέχρι τώρα θεωρούσαμε ότι η εκ των πρότερων πιθανότητα συμπερίληψης μιας ανεξάρτητης μεταβλητής (κεφ. 2) μπορεί να προσδιοριστεί από τον ερευνητή. Ωστόσο, υπάρχει πάντα η δυνατότητα να υιοθετήσουμε μια ιεραρχική τακτική θεωρώντας και την παράμετρο w ως μια τυχαία μεταβλητή με μία εκ των προτέρων κατανομή.

Θεωρώντας ως εκ των προτέρων την συζυγή οικογένεια κατανομών της διωνυμικής, η οποία είναι η βήτα, $Be(\tau, \lambda)$, $\tau > 0$, $\lambda > 0$ η εκ των προτέρων κατανομή για ένα μοντέλο γ θα δίνεται από την

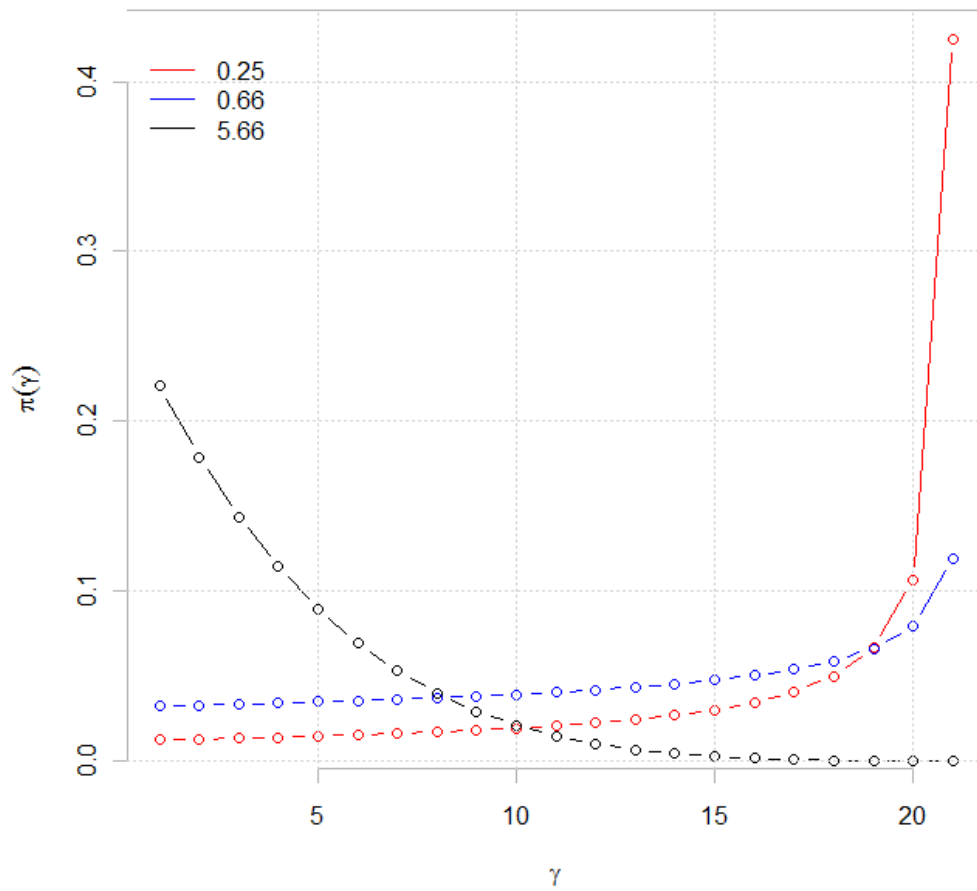
$$\begin{aligned}
 \pi(\gamma) &= \int_0^1 \pi(\gamma | w) \pi(w) dw \\
 &= \binom{p}{q_\gamma} \frac{1}{B(\tau, \lambda)} \int_0^1 \pi^{q_\gamma + \tau - 1} (1 - \pi)^{p - q_\gamma + \lambda - 1} dw \\
 &= \binom{p}{q_\gamma} \frac{B(q_\gamma + \tau, p - q_\gamma + \lambda)}{B(\tau, \lambda)}, \quad q_\gamma = 0, 1, \dots, p,
 \end{aligned}$$

η οποία καλείται Βήτα-Διωνυμική, $Bb(p, \tau, \lambda)$.

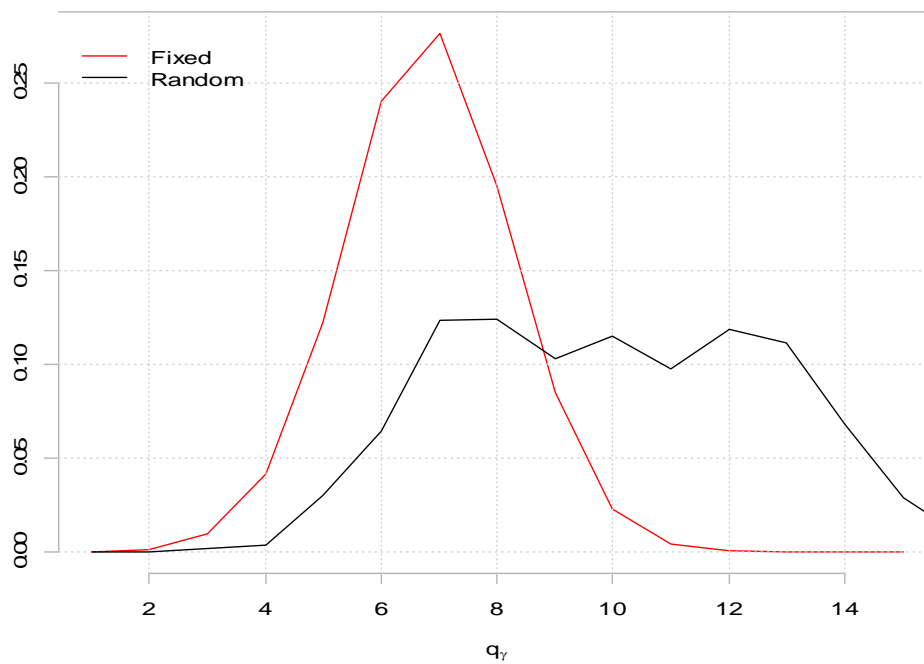
Το Σχήμα 4-2 αναπαριστά τις συναρτήσεις πιθανότητας της Βήτα-Διωνυμικής για $p = 20$ και $\lambda = 0.25, 0.66, 5.66$ και $\tau = 1$.

Ας θεωρήσουμε $\tau = 1$. Τότε, το ζητούμενο είναι να προσδιοριστεί η υπερπαράμετρος λ . Χρησιμοποιώντας το γεγονός ότι η αναμενόμενη τιμή μιας τυχαίας μεταβλητής η οποία έχει συνάρτηση πιθανότητας $Bb(p, \tau, \lambda)$ είναι $\tau p / (\tau + \lambda)$, μπορούμε να γράψουμε $\lambda = (p - \bar{q}) / \bar{q}$, όπου \bar{q} ο εκ των προτέρων αναμενόμενος αριθμός μεταβλητών που περιλαμβάνεται στο μοντέλο.

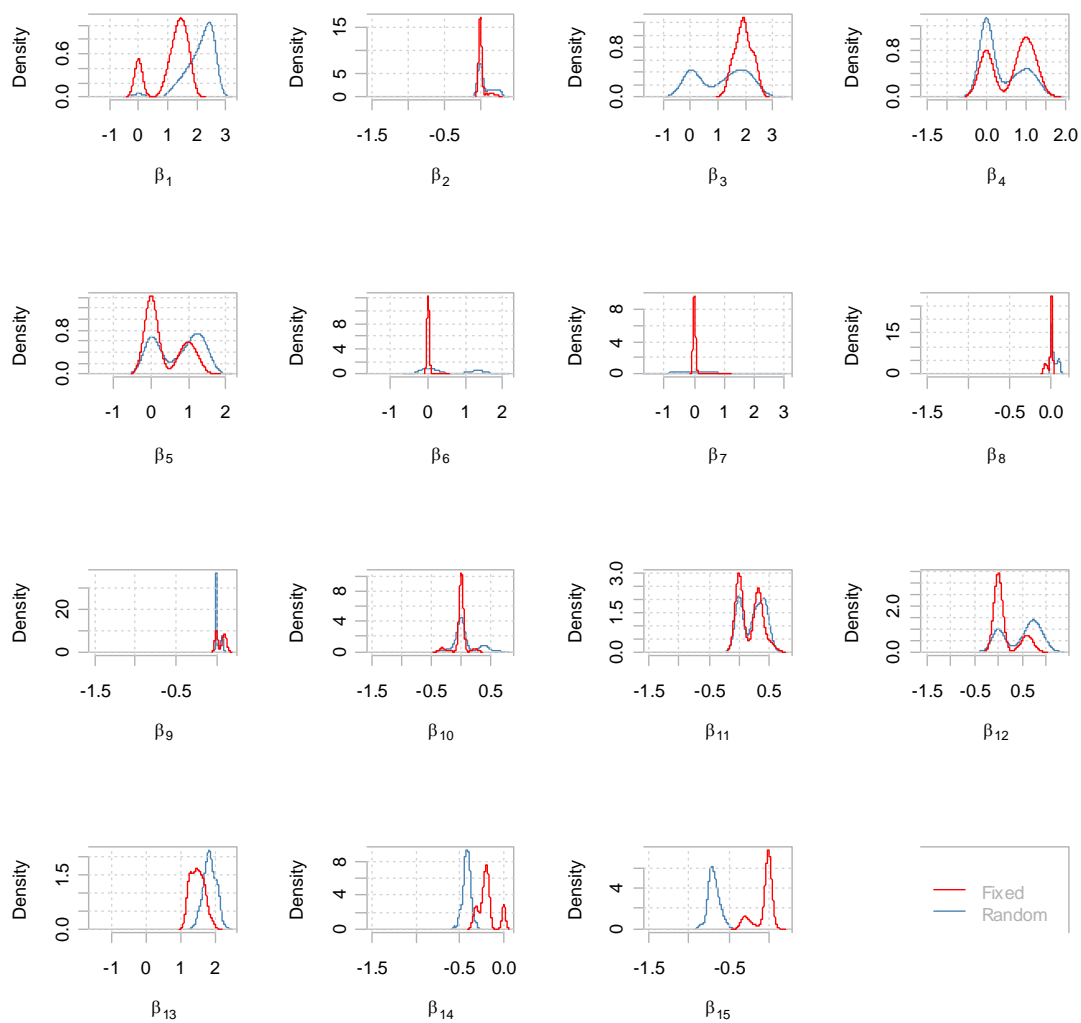
Για να ελέγξουμε τη διαφορά μεταξύ των δύο επιλογών επαναλάβουμε την ανάλυση της Ενότητας 3.9 χρησιμοποιώντας τον αλγόριθμο MC³. Το Σχήμα 4-2 καταδεικνύει τη διαφοροποίηση. Η σημειακή εκτίμηση για το μέγεθος των μοντέλων που επιλέχθηκαν συχνότερα είναι $E^w(q_\gamma | Y) = 9.05$ (έναντι περίπου 6.8 για τους υπόλοιπους στοχαστικούς αλγόριθμους και 6.5 για το Παράθυρο του Οκκάμ). Επίσης, η επιλογή της αντιστοίχισης μιας εκ των προτέρων κατανομής στην παράμετρο χειροτερεύει και το Μέσο Τετραγωνικό Σφάλμα (0.2704). Τέλος, οι κατανομές των συντελεστών των δύο κύριων μεταβλητών (Prob, Time) εμφανίζονται σημαντικά διαφοροποιημένες, όπως δείχνει το Σχήμα 4-4.



Σχήμα 4-2 : Η Βήτα-Διωνυμική κατανομή



Σχήμα 4-3 : Εκ των υστέρων κατανομή του μεγέθους των μοντέλων για σταθερή/τυχαία επιλογή



Σχήμα 4-4 : Εκ των υστέρων κατανομές των συντελεστών για σταθερή/τυχαία επιλογή

4.3 Εμπειρική μπεϋζιανή εκτίμηση της υπερπαραμέτρου ϕ

Εκτός από τις «αυτόματες» επιλογές για την τιμή της υπερπαραμέτρου ϕ , έχουν προταθεί και εκτιμητές βασισμένοι σε εμπειρικές μπεϋζιανές μεθόδους (*Empirical Bayes methods*) (Hansen & Yu, 1999). Οι μέθοδοι αυτές κινούνται μεταξύ της επιλογής μιας προκαθορισμένης τιμής της υπερπαραμέτρου και της χρήσης μιας ιεραρχικής λογικής.

4.2.1 Τοπικοί Εμπειρικοί εκτιμητές Bayes (*Local Empirical Bayes Estimates*)

Οι τοπικοί εμπειρικοί εκτιμητές Bayes (EB(L)) εκτιμούν την παράμετρο ϕ για κάθε μοντέλο ξεχωριστά.

Ο EB(L) – εκτιμητής είναι αυτός που μεγιστοποιεί την πιθανοφάνεια (3.1.4) υπό τον περιορισμό να είναι μη αρνητικός.

Ο εκτιμητής είναι ο

$$\hat{\phi}_\gamma^{EB(L)} = (F_\gamma - 1) \vee 0, \quad (4.2.1)$$

όπου F_γ είναι η συνήθης στατιστική συνάρτηση για τον έλεγχο της υπόθεσης $\beta_\gamma = 0$, δηλαδή

$$F_\gamma = \frac{\frac{R_\gamma^2}{q_\gamma}}{\frac{(1 - R_\gamma^2)}{(n - q_\gamma - 1)}}.$$

Εάν $\hat{\phi}_\gamma^{EB(L)} = 0$, τότε $F_\gamma \leq 1$. Η ερμηνεία είναι ευθεία καθώς ο έλεγχος της κλασσικής στατιστικής για τιμές της στατιστικής συνάρτησης F_γ μικρότερες του 1 δεν απορρίπτει την υπόθεση ότι το μηδενικό μοντέλο είναι ισοδύναμο με το μοντέλο γ . Συνεπώς, μπορούμε να αντιστοιχίσουμε την εκτίμηση $\hat{\phi}_\gamma^{EB(L)} = 0$ στο μηδενικό μοντέλο.

Τέλος, από την παρατηρούμενη πληροφορία μπορούμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης χρησιμοποιώντας το γεγονός ότι η εκτίμηση της (ασυμπτωτικής) διακύμανσης του εκτιμητή (για $\hat{\phi}_\gamma^{EB(L)} > 0$) είναι

$$\frac{1}{2} \left[\frac{q_\gamma + 1 - n}{(q_\gamma + 1)^2} - \left(\frac{\frac{R_\gamma^2 - 1}{\hat{\phi}_\gamma^{EB(L)} (R_\gamma^2 - 1) - 1}}{\sqrt{(n - 1)}} \right)^2 \right].$$

4.2.2 Καθολικοί Εμπειρικοί εκτιμητές Bayes (*Global Empirical Bayes Estimates*)

Επίσης, μπορούμε να υποθέσουμε ότι υπάρχει μία παράμετρος, ϕ , κοινή για όλα τα μοντέλα.

Σε αυτήν την περίπτωση ο εκτιμητής μεγιστοποιεί την (συνδυασμένη) περιθωριακή πιθανοφάνεια όλων των $\gamma \in \Gamma$

$$\sum_{\gamma} \left[\pi(\gamma) \frac{(1 + \phi)^{\binom{n-1-q_{\gamma}}{2}}}{[\phi(R_{\gamma}^2 - 1) + 1]^{\binom{n-1}{2}}} \right],$$

υπό τον περιορισμό να είναι μη αρνητικός, δηλαδή

$$\hat{\phi}^{EB(G)} = \arg \max_{\phi > 0} \sum_{\gamma} \left[\pi(\gamma) \frac{(1 + \phi)^{\binom{n-1-q_{\gamma}}{2}}}{[\phi(R_{\gamma}^2 - 1) + 1]^{\binom{n-1}{2}}} \right]. \quad (4.2.2)$$

Δυστυχώς, ο εκτιμητής δεν δίνεται σε κλειστή μορφή. Οι George & Foster (2000) χρησιμοποίησαν αριθμητικές μεθόδους για την εύρεση του εκτιμητή. Εναλλακτικά, μπορούμε να εφαρμόσουμε ένα αλγόριθμο EM¹ ακολουθώντας τους Liang *et al.* (2008).

Για να υλοποιήσουμε τον EM αλγόριθμο αντιμετωπίζουμε το μοντέλο γ και την παράμετρο σ^2 ως λανθάνοντα (*latent*) δεδομένα.

¹ Για τον αλγόριθμο EM δες Dempster *et al.* (1977), McLachlan & Krishnan (2008), Ghosh *et al.* (2008).

Αλγόριθμος 3: EM αλγόριθμος για EB(G) εκτιμητές (EM-EB)

Στην $i + 1$ επανάληψη του αλγορίθμου

E – βήμα : Υπολόγισε τις επόμενες αναμενόμενες τιμές για κάθε $\gamma \in \Gamma$

$$E(\sigma^2 | y, \gamma, \hat{\phi}^{(i)}) = \frac{n - 1}{\left(SSY - \frac{\hat{\phi}^{(i)}}{1 + \hat{\phi}^{(i)}} SSR \right)}$$

$$E(\gamma | y, \hat{\phi}^{(i)}) = \frac{p(Y | \gamma, \hat{\phi}^{(i)})}{\sum_{\gamma} p(Y | \gamma, \hat{\phi}^{(i)})}$$

M – βήμα : Ο (τρέχων) εκτιμητής είναι

$$\hat{\phi}^{(i+1)} = \left[\frac{\sum_{\gamma} E(\gamma | y, \hat{\phi}^{(i)}) E(\sigma^2 | y, \gamma, \hat{\phi}^{(i)}) SSR}{\sum_{\gamma} E(\gamma | y, \hat{\phi}^{(i)}) q_{\gamma}} - 1 \right] \vee 0$$

,όπου $SSR = y^T H_{\gamma} y$ το γνωστό άθροισμα τετραγώνων λόγω της παλινδρόμησης υπό το μοντέλο γ και $SSY = (y - \bar{y})^T (y - \bar{y})$ το ολικό άθροισμα τετραγώνων.

Βήμα ελέγχου : Εάν $|\hat{\phi}^{(i+1)} - \hat{\phi}^{(i)}| < \varepsilon$ τότε $\hat{\phi}^{EB(G)} = \hat{\phi}^{(i+1)}$.

Για αρχική τιμή του αλγορίθμου, $\hat{\phi}^{(0)}$, μπορούμε να θεωρήσουμε την εκτίμηση με βάση κάποιο «παραδοσιακό» εκτιμητή όπως αυτοί του κεφαλαίου 3.

Η δεύτερη αναμενόμενη τιμή του *E* – βήματος είναι η εκ των υστέρων πιθανότητα του μοντέλου σε κάθε επανάληψη του αλγορίθμου, $\pi^{(i)}(\gamma | y)$. Συνεπώς όταν ο αλγόριθμος έχει συγκλίνει, θα έχουν υπολογιστεί και οι εκ των υστέρων πιθανότητες των μοντέλων (οι οποίες είναι το ζητούμενο!).

Συνεχίζοντας θα μπορούσαμε να κάνουμε μια, διαισθητική περισσότερο, παρατήρηση. Ο εκτιμητής στο M – βήμα μπορεί να γραφεί ως μία (μπεϋζιανή) σταθμισμένη στατιστική συνάρτηση F :

$$\hat{\phi}^{(i+1)} = \left[\sum_{\gamma} \pi^{(i)}(\gamma | y) \frac{SSR / \sum_{\gamma} \pi^{(i)}(\gamma | y) q_{\gamma}}{\left(SSY - \frac{\hat{\phi}^{(i)}}{1 + \hat{\phi}^{(i)}} SSR \right) / (n - 1)} - 1 \right] \vee 0.$$

Μέχρι αυτό το σημείο πρέπει να έχει γίνει εμφανές ότι για κάθε εκτιμητή ο οποίος δίνεται σε μορφή αθροίσματος ως προς το σύνολο Γ , ο υπολογισμός είναι εξαιρετικά επίπονος. Ακριβώς αυτό ισχύει και για τον $EB(G)$ εκτιμητή. Για την ακρίβεια, στην περίπτωση του $EB(G)$ εκτιμητή το πρόβλημα είναι ακόμα εντονότερο καθώς σε κάθε επανάληψη του αλγόριθμου θα επαναλαμβάνεται η άθροιση στο σύνολο Γ . Η λύση είναι και πάλι η χρήση ενός ελαττωμένου συνόλου μοντέλων. Εξ άλλου υπάρχει πάντα και η εναλλακτική λύση των στοχαστικών αλγόριθμων. Για παράδειγμα, μπορούμε να εφαρμόσουμε μια στοχαστική αναζήτηση μεταξύ των μοντέλων (όπως με τον RJMCMC) και να εκτιμήσουμε την κατανομή του $EB(L)$ -εκτιμητή. Χρησιμοποιώντας την κατανομή του $EB(L)$ -εκτιμητή μπορούμε να εκτιμήσουμε τον $EB(G)$.

Τέλος, αναφέρουμε με τη μορφή λήμματος το επόμενο αποτέλεσμα (Liang *et al.*, 2008).

Λήμμα 4.1 Οι EB – εκτιμητές είναι της ίδιας τάξης, $\hat{\phi}^{EB(L)} = \hat{\phi}^{EB(G)} = O_p(n)$.

Επίσης, ασυμπτωτικά έχουν την ίδια τάξη με το BIC.

4.4 Ένα ακόμα παράδοξο σχετικά με την υπερπαράμετρο ϕ (ή γιατί χρειαζόμαστε νέους εκτιμητές)

Στο δεύτερο κεφάλαιο συνδέσαμε τη λανθασμένη επιλογή της υπερπαραμέτρου ϕ με το παράδοξο του Bartlett. Δυστυχώς, δεν αρκεί μια λανθασμένη (μεγάλη) τιμή της υπερπαραμέτρου για να οδηγηθούμε σε ένα παράδοξο. Ακόμα, και η χρήση μιας σταθερής, προκαθορισμένης λογικής τιμής μάς «εξασφαλίζει» την παρουσία ενός παράδοξου.

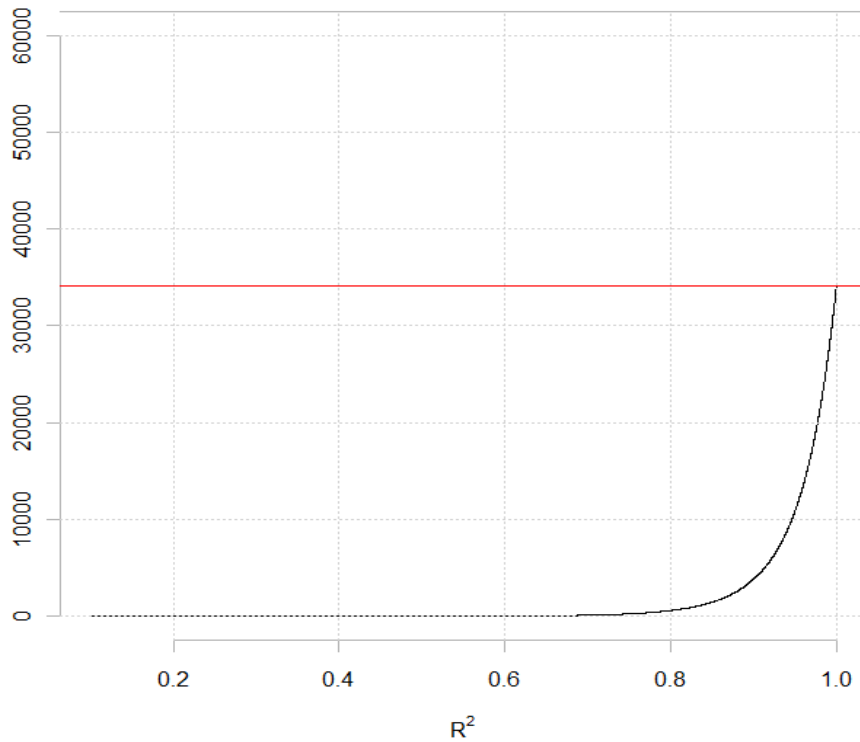
Ας δούμε το ζήτημα υπό την οπτική της επιλογής ενός μοντέλου. Έστω ότι το μοντέλο γ είναι πραγματικά αξιόλογο, δηλαδή έχει «μεγάλη» τιμή της στατιστικής συνάρτησης F , «μεγάλη» τιμή R_γ^2 και «μεγάλους» συντελεστές β_i . Σε αυτήν την περίπτωση, είναι χρήσιμο οι συγκρίσεις των μοντέλων να γίνονται με βάση κάποιο μοντέλο αναφοράς. Ως τέτοιο λαμβάνεται συνήθως το μηδενικό, $\gamma = \mathbf{0}_p$, ή το πλήρες, $\gamma = \mathbf{1}_p$.

Στα επόμενα θεωρούμε το μηδενικό ως μοντέλο αναφοράς, ακολουθώντας τους Berger & Pericchi (2001). Σε αυτήν την περίπτωση η σύγκριση των δύο μοντέλων (θεωρώντας ίσες εκ των προτέρων πιθανότητες) θα βασιστεί στην $\pi(\gamma | y) / \pi(\gamma = \mathbf{0}_p | y)$. Καθώς $R_\gamma^2 \rightarrow 1$, θα αναμενόταν η εκ των υστέρων πιθανότητα του μοντέλου γ να συμπεριφέρεται παρόμοια, δηλαδή $\pi(\gamma | y) \rightarrow 1$. Συνεπώς, η $\pi(\gamma | y) / \pi(\gamma = \mathbf{0}_p | y)$ θα αυξάνει συνεχώς για σταθερές τιμές των n, q_γ καθώς η «πληροφορία» που προσφέρουν τα δεδομένα είναι εις βάρος του μηδενικού ($\pi(\gamma | y) \uparrow, \pi(\gamma = \mathbf{0}_p | y) \downarrow$).

Παραδόξως, η πραγματική οριακή συμπεριφορά της στατιστικής σύγκρισης είναι διαφορετική, καθώς

$$\frac{\pi(\gamma | y)}{\pi(\gamma = \mathbf{0}_p | y)} \rightarrow (1 + \phi)^{(n - q_\gamma - 1)/2},$$

όπως φαίνεται και στο επόμενο γράφημα για $n = 15, q_\gamma = 5, \phi = 2$



Σχήμα 4-5 : Το «παράδοξο της πληροφορίας»

Αντίθετα, με την επιλογή μιας σταθερής, προκαθορισμένης τιμής της υπερπαραμέτρου ϕ , οι EB – εκτιμητές δεν επιβαρύνονται από το παραπάνω παράδοξο, το οποίο οι Liang *et al.* (2008) ονομάζουν «παράδοξο της πληροφορίας» (*information paradox*).

Θεώρημα 4.1 Υπό τον EB(L) ή EB(G)-εκτιμητή δεν εμφανίζεται το «παράδοξο της πληροφορίας».

Απόδειξη. Είναι προφανές από την (4.2.2) ότι καθώς $R_\gamma^2 \rightarrow 1$ ο $\hat{\phi}_\gamma^{EB(L)} \rightarrow \infty$, οπότε

$$\pi(\gamma | y) / \pi(\gamma = \mathbf{0}_p | y) \rightarrow \infty.$$

Επίσης, το μέγιστο της (4.2.3) τείνει στο ∞ όταν υπάρχει ένα μοντέλο γ για το οποίο $R_\gamma^2 \rightarrow 1$. Αντίστοιχα, $\pi(\gamma | y) / \pi(\gamma = \mathbf{0}_p | y) \rightarrow \infty$ για $\phi = \hat{\phi}^{EB(G)}$.

4.5 Χρήση της εκ των πρότερων κατανομής των Zellner – Siow

Τα προαναφερθέντα στην 4.4 είναι γνωστά ήδη από την μονογραφία του Jeffreys (1961). Ο Jeffreys βασισμένος στο παράδοξο της πληροφορίας απέρριψε τη χρήση κατανομών που πάσχουν από αυτό στα πλαίσια των ελέγχων για τον μέσο μιας κανονικής κατανομής. Η λύση που πρότεινε, ήταν η χρήση της κατανομής Cauchy ως εκ των προτέρων κατανομής. Στα 1980 οι Zellner και Siow επεξεύειναν την λογική του Jeffreys για τους συντελεστές του γραμμικού μοντέλου.

Ορισμός 4.5.1 Ένα τυχαίο διάνυσμα Y μήκους n ακολουθεί την n -διάστατη κατανομή Cauchy, C_n , όταν η συνάρτηση πυκνότητας πιθανότητάς του δίνεται από την

$$p(y) = \frac{\Gamma[(n+1)/2]}{\pi^{(n+1)/2}} \left(1 + \frac{y^T y^T}{\sigma^2} \right)^{-(n+1)/2}, \quad y \in \mathbb{R}^n.$$

Για διαδικασίες επιλογής μοντέλου και τη BMA, η εκ των προτέρων κατανομή των Zellner – Siow που χρησιμοποιήσαν οι Liang *et al.* (2008) είναι η

$$\pi(\beta_\gamma | \sigma^2, \gamma) \propto \frac{\Gamma(q_\gamma / 2)}{\pi^{q_\gamma/2}} \left| \frac{X_\gamma^T X_\gamma^T}{n\sigma^2} \right|^{1/2} \left(1 + \beta_\gamma^T \frac{X_\gamma^T X_\gamma^T}{n\sigma^2} \beta_\gamma \right)^{-q_\gamma/2}, \quad (4.5.1)$$

η οποία είναι μια q_γ -διάστατη κατανομή Cauchy, $C_{q_\gamma}(0, n\sigma^2(X_\gamma^T X_\gamma)^{-1})$, με κέντρο 0 (δηλ. το μηδενικό μοντέλο) και πίνακα scale ο οποίος καθορίζεται από τον πίνακα πληροφορίας $X_\gamma^T X_\gamma$.

Για τις παραμέτρους οι οποίες είναι κοινές μεταξύ των μοντέλων (με βάση το επιχείρημα της ορθογωνιότητας) χρησιμοποιείται και πάλι η καταχρηστική εκ των προτέρων κατανομή

$$\pi(\alpha, \sigma^2) \propto \frac{1}{\sigma^2}, \quad \alpha \in \mathbb{R}, \sigma > 0.$$

Οι υπολογισμοί με την (4.5.1) δεν είναι απλοί. Οι Zellner και Siow ανέπτυξαν προσεγγίσεις για την περιθωριακή πιθανοφάνεια όμως η ακρίβεια της προσέγγισης ελαττώνεται με την διάσταση του παραμετρικού χώρου. Μια άλλη λύση είναι να παρατηρήσουμε ότι η (4.5.1) μπορεί να γραφεί ως μείξη κατανομών Zellner με εκ των πρότερων κατανομή μείξης για την ϕ την $IG(1/2, n/2)$

$$\pi(\beta_\gamma | \sigma^2, \gamma) \propto \int N\left(\beta_\gamma | 0, \sigma^2 \phi (X_\gamma^T X_\gamma)^{-1}\right) \pi(\phi) d\phi \quad (4.5.2)$$

και να χρησιμοποιήσουμε προσεγγιστικές μεθόδους.

Επίσης, η πιθανοφάνεια μπορεί να γραφεί ως

$$\begin{aligned} p(y | \gamma) &\propto \int (1 + \phi)^{\frac{n-1-q_\gamma}{2}} \left[\phi (R_\gamma^2 - 1) + 1 \right]^{-\binom{n-1}{2}} \pi(\phi) d\phi \\ &= \int (1 + \phi)^{\frac{n-1-q_\gamma}{2}} \left[\phi (R_\gamma^2 - 1) + 1 \right]^{-\binom{n-1}{2}} \phi^{-3/2} e^{-n/2\phi} d\phi, \end{aligned} \quad (4.5.3)$$

η οποία πρέπει να προσεγγιστεί.

Τέλος, ακόμα μια ποσότητα θα απαιτήσει προσεγγιστική λύση. Η ιεραρχική δομή της εκ των προτέρων κατανομής των Zellner – Siow όπως δηλώνεται από την (4.5.2), περιπλέκει τον υπολογισμό των αναμενόμενων τιμών οι οποίες θα χρησιμοποιηθούν για την πρόβλεψη. Για την εκ των προτέρων κατανομή του Zellner ήταν εύκολος ο υπολογισμός της πρόβλεψης $E(Y^* | \gamma, y)$ καθώς δινόταν από την

$$E(Y^* | \gamma, y) = \hat{\alpha} + \frac{\phi}{1 + \phi} X_\gamma \hat{\beta}_\gamma,$$

όπου $\hat{\alpha}, \hat{\beta}_\gamma$ οι συνήθεις εκτιμητές μέγιστης πιθανοφάνειας του γραμμικού μοντέλου.

Αντίστοιχα, για την διεξαγωγή προβλέψεων υπό την εκ των προτέρων κατανομή των Zellner – Siow απαιτείται ο υπολογισμός της

$$E(Y^* | \gamma, y) = \hat{\alpha} + E\left(\frac{\phi}{1 + \phi} \middle| \gamma, y\right) X_\gamma \hat{\beta}_\gamma.$$

Γενικά, ο υπολογισμός της δεν είναι εύκολος, λόγω της ύπαρξης της αναμενόμενης τιμής $E\left(\frac{\phi}{1 + \phi} \middle| \gamma, y\right)$. Οι Liang *et al.* (2008) περιγράφουν μια διαδικασία η οποία

βασίζεται στην προσέγγιση του ολοκληρώματος της αναμενόμενης τιμής μέσω της μεθόδου Laplace (Tierney & Kadane, 1986, Ghosh, 2008, Παράρτημα Α).

Έστω οι ακόλουθες συναρτήσεις, $h_1(\phi) = \log\left(\frac{\phi}{1+\phi}\right) + \log p(y|\phi) + \log \pi(\phi)$ και

$$h_2(\phi) = \log p(y|\phi) + \log \pi(\phi).$$

Τότε,

$$\begin{aligned} E\left(\frac{\phi}{1+\phi} \middle| \gamma, y\right) &= \frac{\int \frac{\phi}{1+\phi} p(y|\phi)\pi(\phi)d\phi}{\int p(y|\phi)\pi(\phi)d\phi} \\ &\approx \frac{\tilde{\sigma}_{h_1}}{\tilde{\sigma}_{h_2}} \exp[h_1(\tilde{\phi}_{h_1}) - h_2(\tilde{\phi}_{h_2})], \end{aligned} \quad (4.5.4)$$

όπου $\tilde{\phi}_{h_i}$ είναι η εκ των υστέρων κορυφή της ϕ και $\tilde{\sigma}_{h_i} = \left(\frac{d^2}{d\phi^2} h_i(\phi) \Big|_{\phi=\tilde{\phi}_{h_i}}\right)^{-1/2}$,

$i = 1, 2$.

Ο παρονομαστής στην (4.5.4) είναι η περιθωριακή πιθανοφάνεια, $p(y|\gamma)$, συνεπώς μπορούμε να προσεγγίσουμε και την περιθωριακή πιθανοφάνεια μέσω της προσέγγισης Laplace.

Η εφαρμογή της προσέγγισης βασίζεται στην ύπαρξη και την εύρεση της εκ των υστέρων κορυφής της ϕ . Οι Liang *et al.* (2008) προσέγγισαν την αναμενόμενη τιμή

$$\text{της } g(\phi) = \phi^a (1+\phi)^b$$

$$\int_0^{\infty} \exp[h(\phi)]d\phi = \int_0^{\infty} (1+\phi)^{\frac{n-1-q_\gamma+b}{2}} \left[\phi(R_\gamma^2 - 1) + 1\right]^{-\left(\frac{n-1}{2}\right)} \phi^{a-3/2} e^{-n/2\phi} d\phi,$$

η οποία για $a = b = 0$ ανάγεται στην περιθωριακή πιθανοφάνεια και για $a = 1, b = -1$ στον αριθμητή της (4.5.4). Η εκ των υστέρων κορυφή είναι η λύση της επόμενης κυβικής εξίσωσης,

$$-(1 - R_\gamma^2)[q_\gamma + 3 - 2(a - b)]\phi^3 + \{n - q_\gamma + 2b - 4 + 2[a + b - (1 - a)(1 - R_\gamma^2)]\}\phi^2 + [n(2 - R_\gamma^2) + 2a - 3]\phi + n = 0$$

για την οποία υπάρχει μοναδική, θετική ρίζα (δες το Παράρτημα των Liang *et al.*, 2008).

Ολοκληρώνοντας, αναφέρουμε μερικά ασυμπτωτικά αποτελέσματα για την εκ των προτέρων κατανομή των Zellner – Siow.

Θεώρημα 4.2 Για την εκ των προτέρων κατανομή των Zellner – Siow ισχύουν τα εξής :

- (i) Είναι απαλλαγμένη από το «παράδοξο της πληροφορίας»
- (ii) Είναι συνεπής στην επιλογή του πραγματικού μοντέλου
- (iii) Ο BMA-εκτιμητής της πρόβλεψης υπό την εκ των προτέρων κατανομή των Zellner – Siow είναι συνεπής ως προς την πρόβλεψη.

Απόδειξη. Δες Liang *et al.* (2008)

Ιδιαίτερο ενδιαφέρον έχει το (iii), την απόδειξη του οποίου σκιαγραφούμε.

Αρχικά, ας ορίσουμε την συνέπεια ως προς την πρόβλεψη.

Ορισμός 4.1 Ένας εκτιμητής, \hat{Y}^* , μιας μελλοντικής παρατήρησης ονομάζεται συνεπής όταν $\hat{Y}^* \xrightarrow{p} EY^*$, όπου $EY^* = \alpha + X_\gamma^T \beta_\gamma$ και $\alpha, \beta_\gamma, \sigma^2$ είναι οι πραγματικές τιμές των παραμέτρων.

Ο BMA – εκτιμητής για την πρόβλεψη μιας μελλοντική τιμής θα είναι

$$\begin{aligned} \hat{Y}^* &= \hat{\alpha} + \sum_\gamma \pi(\gamma | Y) E\left(\frac{\phi}{1 + \phi} | \gamma, Y\right) X_\gamma \hat{\beta}_\gamma \\ &= \hat{\alpha} + \sum_\gamma \pi(\gamma | Y) X_\gamma \hat{\beta}_\gamma \int_0^\infty \frac{\phi}{1 + \phi} p(\phi | \gamma, Y) d\phi. \end{aligned} \quad (4.5.5)$$

Για να αποδείξουμε το (iii) αρκεί να δείξουμε ότι καθώς $n \rightarrow \infty$

$$\alpha) \int_0^{\infty} \frac{\phi}{1+\phi} p(\phi | \gamma, y) d\phi \xrightarrow{p} 1$$

$$\beta) \hat{\beta}_\gamma \xrightarrow{p} \beta_\gamma$$

$$\gamma) \pi(\gamma | y) \xrightarrow{p} 1$$

Το β) προκύπτει από τη συνέπεια των εκτιμητών μέγιστης πιθανοφάνειας. Το γ) από τη συνέπεια στην επιλογή μοντέλου. Για να δείξουμε το α) γραφούμε την

$$\int_0^{\infty} \frac{\phi}{1+\phi} p(\phi | \gamma, y) d\phi = \frac{\int_0^{\infty} \frac{\phi}{1+\phi} L(\phi) p(\phi) d\phi}{\int_0^{\infty} L(\phi) p(\phi) d\phi}, \quad (4.5.6)$$

όπου $L(\phi) = (1+\phi)^{-a/2} \left(1 - R_\gamma^2 \frac{\phi}{1+\phi}\right)^{-\binom{n-1}{2}}$. Το ζητούμενο πλέον είναι να

προσεγγιστεί η (4.5.6). Οι Liang *et al.* (2008) χρησιμοποίησαν μια παραλλαγή της προσέγγισης Laplace για την (4.5.6). Ακολουθώντας τους Kass & Raftery (1995) χρησιμοποίησαν στην θέση της εκ των υστέρων κορυφής τον εκτιμητή μέγιστης πιθανοφάνειας και στην θέση της $\hat{\sigma}_{h_i}$ την τετραγωνική ρίζα της αντιστρόφου της παρατηρούμενης πληροφορίας. Οι Kass & Raftery (1995) έδειξαν ότι η προσέγγιση είναι της ίδιας τάξης με την αρχική προσέγγιση Laplace, δηλαδή $O(n^{-1})$. Το επόμενο θεώρημα συνοψίζει αρκετά βήματα της απόδειξης των Liang *et al.* (2008).

Θεώρημα 4.3 Η προσέγγιση Laplace η οποία βασίζεται στην πιθανοφάνεια σε συνδυασμό με το γεγονός ότι η $L(\phi)$ παρουσιάζει μέγιστο στο $\hat{\phi}^{EB(L)}$ συνεπάγεται ότι

$$\int_0^{\infty} \frac{\phi}{1+\phi} p(\phi | \gamma, Y) d\phi = \frac{\hat{\phi}^{EB(L)}}{1+\hat{\phi}^{EB(L)}} (1 + O(n^{-1})),$$

όπου η $p(\phi | \gamma, Y)$ είναι είτε η εκ των πρότερων κατανομή των Zellner – Siow είτε η (σημειακή στο $\hat{\phi}^{EB(L)}$) κατανομή του EB(L) – εκτιμητή.

Από τα παραπάνω συμπεραίνεται ότι

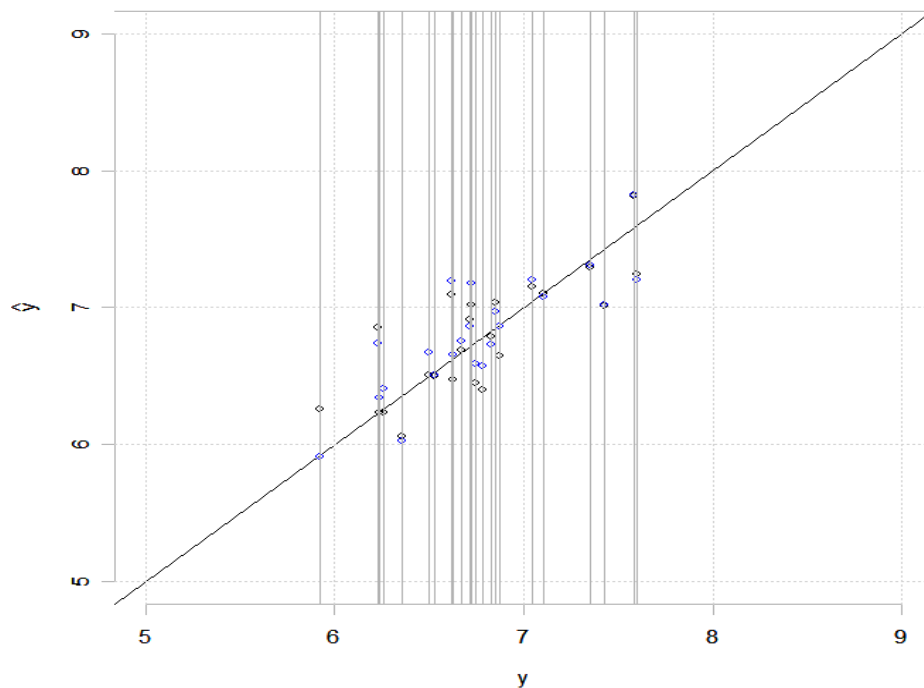
$$\int_0^{\infty} \frac{\phi}{1+\phi} p(\phi | \gamma, Y) d\phi \xrightarrow{p} 1$$

καθώς $\hat{\phi}^{EB(L)} \rightarrow \infty$. Επομένως, ο εκτιμητής (4.5.5) είναι συνεπής ως προς την πρόβλεψη τόσο για τον EB(L) – εκτιμητή όσο και για αυτόν που βασίζεται στην εκ των πρότερων κατανομή των Zellner – Siow. Τέλος, από το Λήμμα 4.1 το ίδιο ισχύει και για τον EB(G) – εκτιμητή.

Για την εφαρμογή της 3.9 έχουμε τα εξής αποτελέσματα

	Pr($\beta_i \neq 0 Y$)					Stepwise
	HRM		Zellner		Zellner – Siow	
	MC3	Occam's Window	MC3	RJMCMC	MC3	
M	0.65	0.92	0.65	0.64	0.96	*
So	0.17	0.04	0.13	0.12	0.49	
Ed	0.85	0.99	0.80	0.79	0.70	*
Po1	0.64	0.53	0.56	0.55	0.57	
Po2	0.59	0.47	0.47	0.48	0.66	
LF	0.10	0.03	0.11	0.11	0.52	
M.F	0.10	0.07	0.11	0.11	0.51	
Pop	0.22	0.36	0.19	0.19	0.52	*
NW	0.36	0.59	0.25	0.25	0.43	*
U1	0.10	0.12	0.10	0.11	0.47	
U2	0.25	0.64	0.24	0.24	0.67	*
GDP	0.22	0.18	0.12	0.12	0.66	*
Ineq	0.97	1.00	0.96	0.96	0.98	*
Prob	0.61	0.91	0.56	0.56	0.96	*
Time	0.14	0.28	0.12	0.12	0.95	

Πίνακας 4-2: BMA χρησιμοποιώντας την εκ των προτέρων κατανομή των Zellner – Siow



Σχήμα 4-6 : Πραγματικές και προβλεπόμενες τιμές υπό την εκ των προτέρων κατανομή των Zellner-Siow

Μια πρώτη παρατήρηση προέρχεται από τον αριθμό των ανεξάρτητων μεταβλητών με υψηλή εκ των υστέρων πιθανότητα. Για παράδειγμα, εάν ορίζαμε το μοντέλο διάμεσης πιθανότητας (Median Probability Model) το οποίο αποτελείται από τις ανεξάρτητες μεταβλητές οι οποίες έχουν εκ των υστέρων πιθανότητα άνω του 0.5, προκύπτει από τα αποτελέσματα του αλγορίθμου MC³ για την Zellner – Siow ότι αυτό περιλαμβάνει 12 ανεξάρτητες μεταβλητές!

Παραδόξως, τα αποτελέσματα των προβλέψεων είναι χειρότερα τόσο από τις υπόλοιπες διαδικασίες BMA όσο και από τη stepwise!

	HRM	Zellner	Stepwise	Zellner – Siow
RMSE	0.1875916	0.2312784	0.2537807	0.2665655

Πίνακας 4-2 : Σύγκριση του RMSE χρησιμοποιώντας την εκ των προτέρων κατανομή των Zellner – Siow

5.

Προεκτάσεις και σχολιασμός

Το παρόν κεφάλαιο είναι κατά μια έννοια «αυτόνομο». Αποτελεί μια βιβλιογραφική αναφορά σε εφαρμογές πέραν του γραμμικού μοντέλου, σε ανταγωνιστικές μορφές *model averaging* καθώς και σε εναλλακτικές υπολογιστικές μεθόδους.

5.1 Πέρα από το γραμμικό μοντέλο

Το γραμμικό μοντέλο είναι αναμφισβήτητα το δημοφιλέστερο εργαλείο ανάλυσης. Παρ' όλ' αυτά είναι γνωστό ότι είναι ακατάλληλο για συγκεκριμένα δεδομένα (Dobson, 2002). Επομένως, για να αποτελέσει η BMA μια ολοκληρωμένη στατιστική διαδικασία είναι απαραίτητο να συμπεριλάβει εκτεταμένες κλάσεις μοντέλων, όπως γενικευμένα γραμμικά μοντέλα καθώς και μοντέλα ανάλυσης επιβίωσης.

Ο Raftery (1996) ανέπτυξε προσεγγίσεις για τις (περιθωριακές) πιθανοφάνειες γενικευμένων γραμμικών μοντέλων. Με την αβεβαιότητα ως προς την επιλογή της συνάρτησης σύνδεσης ασχολήθηκαν οι Ntzoufras *et al.* (2003). Εξ άλλου, οι Volinsky *et al.* (1997), Raftery *et al.* (1996), μελέτησαν την εφαρμογή της BMA στο μοντέλο των ανάλογων κινδύνων του Cox (CPH). Στο πλαίσιο της διδακτορικής του διατριβής ο Nobble (2001) εξέτασε τη BMA σε μεθόδους της Πολυμεταβλητής Ανάλυσης, όπως PCA, CCA και CVA.

Η BMA αναπτύσσεται δυναμικά τόσο στην οικονομετρική κοινότητα καθώς και σε χρηματοοικονομικές εφαρμογές. Για παράδειγμα οι Koop & Potter (2008) μελέτησαν την αποτελεσματικότητα της BMA σε μοντέλα δυναμικών παραγόντων ενώ οι Vrontos *et al.* (2000) χρησιμοποίησαν τη BMA για να μελετήσουν την αβεβαιότητα μεταξύ δύο κλάσεων μοντέλων, των γενικευμένων αυτοπαλίνδρομων υπό συνθήκη ετεροσκεδαστικών υποδείγματων (GARCH) και των εκθετικά γενικευμένων αυτοπαλίνδρομων υπό συνθήκη ετεροσκεδαστικών υποδειγμάτων (EGARCH). Με τη

τιμολόγηση χρηματοοικονομικών παραγώγων ασχολήθηκαν οι Bunnin *et al.* (2002). Οι Pesaran *et al.* (2006) χρησιμοποίησαν τεχνικές BMA για πρόβλεψη με μοντέλα χρονολογικών σειρών. Τέλος οι Rachev *et al.* (2008) στη μονογραφία τους σχετικά με μπεϋζιανές μεθόδους στη Χρηματοοικονομική Ανάλυση, αφιερώνουν ένα κεφάλαιο στην αντιμετώπιση της αβεβαιότητας την οποία παρέχει η επιλογή ενός θεωρητικού μοντέλου.

Οι Chirpman *et al.* (2008) ανέπτυξαν μια συγγενή μέθοδο για τη μπεϋζιανή ανάλυση Δένδρων Ταξινόμησης και Παλινδρόμησης την οποία ονομάζουν BART (Bayesian Additive Regression Trees). Χρησιμοποίησαν δένδρα τα οποία χαρακτηρίζονται ως «ασθενείς μαθητές» (*weak learners*) μέσω της εκ των προτέρων κατανομής ώστε εκ των υστέρων να συνεισφέρουν ελάχιστα στο σύνολο.

5.2 Άλλες διαδικασίες Model Averaging

Παράλληλα με τη BMA αναπτύχθηκαν στα πλαίσια της ίδιας λογικής, και άλλες διαδικασίες. Στις εφαρμογές η γνωστότερη εναλλακτική διαδικασία είναι πρόταση των Burnham & Anderson (2003), χρησιμοποιεί ως βάρη στην (2.1.1) τις ποσότητες $\exp\left(-\frac{1}{2}AIC_{\gamma}\right) / \sum_{\gamma} \exp\left(-\frac{1}{2}AIC_{\gamma}\right)$. Οι Claeskens & Hjort (2008) ανέπτυξαν το «κλασσικό» ισοδύναμο της BMA, τη FMA (Frequentistic Model Averaging) βασισμένοι στο local misspecification model. Μάλιστα, έδειξαν και τη σύνδεση αυτών των δύο προσεγγίζοντας την $p(Y | \gamma)$. Επίσης, μια άλλη λύση θα ήταν η υλοποίηση μια διαδικασία bootstrap η οποία θα περιλαμβάνει και την επιλογή μοντέλου. Τέλος, ο Hansen μελέτησε τη χρήση της ποσότητας του Mallows (MBA, Hansen, 2007) για την επιλογή των σταθμών ενώ ερευνά και την απόδοση της εύρεσης των σταθμών από την ελαχιστοποίηση κριτηρίων cross-validation (Jackknife Model Averaging, Hansen & Racine, 2007).

5.3 Συλλογισμοί σχετικά με τη BMA

Η BMA δεν είναι μια τέλεια διαδικασία. Μέχρι τώρα τα μοντέλα που παρουσιάστηκαν δεν λαμβάνουν υπόψη τη συσχέτιση των ανεξαρτήτων μεταβλητών, με συνέπεια την «άνιση» κατανομή των εκ των πρότερων πιθανοτήτων υπό την ομοιόμορφη κατανομή. Επίσης, η Ενότητα 3.9 κατέδειξε την μέτρια επίδοση της BMA όταν ο λόγος p/n δεν είναι αρκετά μικρός. Εξ άλλου, υπάρχουν και περιπτώσεις όπου θα ισχύει $p > n$ ή ο πίνακας X_γ δεν θα είναι πλήρους τάξης

$$(\text{rank}(X_\gamma) < q_\gamma + 1).$$

Τέλος, αν και η BMA φαίνεται να αντιμετωπίζει την αβεβαιότητα, δεν είναι δυνατό να ισχυριστεί κανείς ότι την εξαλείφει. Για παράδειγμα, ακόμα και εφαρμόζοντας τη BMA ανάμεσα σε διαφορετικές κλάσεις μοντέλων (π.χ κανονικό γραμμικό μοντέλο και t-γραμμικό μοντέλο) θα υπάρχει πάντα η αβεβαιότητα των κλάσεων που δεν συμπεριελήφθησαν. Μάλλον η σημαντικότερη προσφορά της BMA (και των υπόλοιπων διαδικασιών Model Averaging) είναι η συνειδητοποίηση της αβεβαιότητας.

Παράρτημα Α

Προσέγγιση με τη μέθοδο του Laplace

Συχνά απαιτείται ο υπολογισμός ολοκληρωμάτων της μορφής

$$V = \int_{-\infty}^{\infty} g(\theta)e^{nh(\theta)}d\theta,$$

όπου οι h, g είναι ομαλές συναρτήσεις (με την έννοια της ύπαρξης παραγώγων τουλάχιστον δεύτερης τάξης για την g και τρίτης τάξης για την h). Επίσης, η h παρουσιάζει ολικό μέγιστο στο σημείο θ_0 . Τότε, καθώς $n \rightarrow \infty$,

$$V \sim V^* = \int_{\theta_0-\delta}^{\theta_0+\delta} g(\theta)e^{nh(\theta)}d\theta$$

Το παραπάνω προκύπτει από το γεγονός ότι η σημαντικότερη «μάζα» είναι λογικό να συγκεντρώνεται σε μια περιοχή του μεγίστου.

Χρησιμοποιώντας το ανάπτυγμα Taylor των h, g γύρω από το θ_0 έχουμε

$$\begin{aligned} V &\sim \int_{\theta_0-\delta}^{\theta_0+\delta} \left[g(\theta_0) + (\theta - \theta_0)g'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 g''(\theta_0) + \dots \right] \\ &\quad \times \exp \left[nh(\theta_0) + n(\theta - \theta_0)h'(\theta_0) + \frac{n}{2}(\theta - \theta_0)^2 h''(\theta_0) + \dots \right] d\theta \\ &\sim g(\theta_0)e^{nh(\theta_0)} \int_{\theta_0-\delta}^{\theta_0+\delta} \left[1 + (\theta - \theta_0) \frac{g'(\theta_0)}{g(\theta_0)} + \frac{1}{2}(\theta - \theta_0)^2 \frac{g''(\theta_0)}{g(\theta_0)} \right] \exp \left[\frac{n}{2}(\theta - \theta_0)h''(\theta_0) \right] d\theta \end{aligned}$$

Εάν $h''(\theta_0) < 0$ τότε (χρησιμοποιώντας και μια αλλαγή μεταβλητής στο ολοκλήρωμα) η προσέγγιση είναι

$$V \sim g(\theta_0)e^{nh(\theta_0)} \frac{\sqrt{2\pi}}{\sqrt{n[-h''(\theta_0)]}} [1 + O(n^{-1})].$$

Το επόμενο θεώρημα (Severini, 2005) συνοψίζει τα παραπάνω.

Θεώρημα Β Έστω το ολοκλήρωμα V όπου (i) η h είναι τρεις φορές παραγωγίσιμη, (ii) η g είναι δύο φορές παραγωγίσιμη και υπάρχει σταθερά c τέτοια ώστε καθώς $|\theta| \rightarrow \infty$ ισχύει $g(\theta) = O(\exp(c\theta^2))$, (iii) η g παρουσιάζει μέγιστο στο σημείο θ_0 και (δ) $h'(\theta_0) = 0, |h'(\theta_0)| > 0$ για κάθε $\theta \neq \theta_0$ και $h''(\theta_0) < 0$.

Εάν $|h(\theta_0)| > 0$ τότε $V = g(\theta_0)e^{nh(\theta_0)} \frac{\sqrt{2\pi}}{\sqrt{n[-h''(\theta_0)]}} [1 + O(n^{-1})]$, καθώς $n \rightarrow \infty$.

Το αντίστοιχο για ένα παραμετρικό διάνυσμα $\underline{\theta}$ μήκους p είναι

$$V = e^{nh(\theta_0)} \left(\frac{2\pi}{n}\right)^{p/2} |H(\theta_0)|^{-1/2} g(\theta_0) [1 + O(n^{-1})],$$

όπου $H(\theta_0) = \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} h(\theta) \Big|_{\theta=\theta_0} \right)_{p \times p}$ ο Εσσιανός πίνακας της h .

Παράρτημα Β

Λογισμικό

Για την ολοκλήρωση αυτής της εργασίας χρησιμοποιήθηκε η γλώσσα R^1 . Τα πακέτα BMA και BAS παρείχαν τα κυριότερα στοιχεία των μοντέλων, επίσης το πακέτο leaps χρησιμοποιήθηκε για την εφαρμογή του παραθύρου του Οκκάμ.

Το BMA είναι το πακέτο των Hoeting *et al.* (1997) ενώ το BAS υλοποιεί τη μοντελοποίηση κατά Zellner του άρθρου των Ljiang *et al.* (2008). Το τελευταίο

¹ Η R διατίθεται σύμφωνα με την GNU General Public License Version 2 (Ιούνιος 1991).

πακέτο δεν παράγει άμεσα συγκρίσιμα αποτελέσματα καθώς χρησιμοποιεί δειγματοληψία χωρίς επανάθεση στο Γ σε αντίθεση με τους αλγορίθμους που περιγράφηκαν στο Κεφάλαιο 4.

Εναλλακτικά, το πακέτο `Grocer` του `Scilab` προσφέρουν τη δυνατότητα μιας πλήρους ανάλυσης υλοποιώντας το μοντελο του `Zellner`.²

Παράρτημα Γ

Δεδομένα

Η περιγραφή των μεταβλητών του συνόλου δεδομένων είναι η ακόλουθη

M	Ποσοστό ανδρών ηλικίας 14-24
So	Δείκτρια μεταβλητή Νότιας Πολιτείας
Ed	Μέσος χρόνος εκπαίδευσης
Po1	Αστυνομικές δαπάνες (1960)
Po2	Αστυνομικές δαπάνες (1959)
LF	Αναλογία συμμετοχής εργατικού δυναμικού
M.F	Αριθμός ανδρών ανά 1000 γυναίκες
Pop	Πληθυσμός Πολιτείας
NW	Αριθμός μη λευκών ανα 1000 πολίτες
U1	Λόγος ανεργίας για κατοίκους αστικών κέντρων ηλικίας 14-24
U2	Λόγος ανεργίας για κατοίκους αστικών κέντρων ηλικίας 35-39
GDP	Κατά κεφαλήν ΑΕΠ
Ineq	Ανισότητα Εισοδήματος
Prob	Πιθανότητα φυλάκισης
Time	Μέσος χρόνος φυλάκισης σε πολιτειακές φυλακές
y	Αναλογία κατά κεφαλή εγκλημάτων για μια συγκεκριμένη κατηγορία

Η τελευταία μεταβλητή (y) είναι η εξαρτημένη μεταβλητή ενώ οι κύριες μεταβλητές είναι η πιθανότητα φυλάκισης (`Prob`) και ο μέσος χρόνος φυλάκισης σε πολιτειακές φυλακές (`Time`).

² Το `Scilab` καθώς και το `Grocer` διατίθενται σύμφωνα με την άδεια `CeCILL` (`GNU GPL compatible`).

Ο Ehrlich (1973) θεώρησε ότι η παραβατική συμπεριφορά είναι αποτέλεσμα λογικής οικονομικής απόφασης καθώς και ότι η πιθανότητα τιμωρίας αυτής της απόφασης δρα ανασταλτικά ως προς τη λήψη μιας τέτοιας απόφασης. Το συγκεκριμένο σύνολο δεδομένων αποτελείται από τα δεδομένα 47 πολιτειών των Η.Π.Α το έτος 1960. Το σύνολο δεδομένων περιείχε λάθη τα οποία διόρθωσε ο Vandaele το 1978.

Οι μεταβλητές έχουν τυποποιηθεί σε βολικούς αριθμούς από τον Ehrlich ο οποίος ανέλυσε πρώτος τα παραπάνω δεδομένα. Επίσης, όλες οι μεταβλητές (εκτός της δείκτης) έχουν μετασχηματιστεί λογαριθμικά.

Βιβλιογραφία

Altman, D. G. & P. K. Andersen. (1989), Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine* , **8**, pp. 771-783.

Anderson, D. R & Burnham, K. P. (2004), *Model Selection and Multi-Model Inference A Practical Information-Theoretic Approach*, Springer, New York.

Bartlett, M. S. (1957), A comment on D.V.Lindley's statistical paradox, *Biometrika*, **44**, pp. 533-534.

Berger, J. O. & Pericchi, L. R. (2001), Objective Bayesian methods for model selection: introduction and comparison (with discussion). In *Model Selection* (P. Lahiri, ed.) pp. 135-207. IMS, Beachwood, OH.

Bunnin, F., Guo, Y., & Ren, Y. (2002). Option pricing under model and parameter uncertainty using predictive densities. *Statistics and Computing*, **12**, No. 1, pp. 37–44.

Chipman, H., George, E. & McCullough, R. E. (2008), BART: Bayesian Additive Regression Trees, (<http://arxiv.org/abs/0806.3286>)

Chipman, H., George, E. I. and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection (with discussion). In *Model Selection* (P. Lahiri, ed.) pp. 66-134. IMS, Beachwood, OH.

Claeskens & Hjort (2008), *Model Selection and Model Averaging*, Cambridge University Press.

Clyde, M. A. (2006), Bayesian Model Averaging and Stochastic Search using Bayesian Adaptive Sampling. (<http://www.isds.duke.edu/~clyde/BAS/>)

Clyde, M. A. and George, E. I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet-Based Models. Lecture Notes in Statistics*, **141**, pp. 309-322. Springer, New York.

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

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998), Automatic Bayesian Curve Fitting, *Journal of the Royal Statistical Society*, B, **60**, No. 2, pp. 333-350.

Derksen, S. & H. J. Keselman. (1992), Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* **45**, pp. 265-282.

Dobson, A. J. (2002), *An Introduction to Generalized Linear Models* (Second Edition), Chapman & Hall/CRC.

- Draper, D. (1995), Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, B*, **57**: pp. 45–97.
- Ehrlich, I. (1973), Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of Political Economy*, **81**, pp. 521–565.
- Eklund, J. & Karlsson, S. , (2007) Computational Efficiency in Bayesian Model and Variable Selection, *Economics Working Paper 35*, Department of Economics, Central bank of Iceland.
- Fernández, C., Ley, E. & Steel, M. (2001), Benchmark priors for Bayesian model averaging, *Journal of Econometrics*, **100**, Issue 2, pp. 381-427.
- Foster, D. P. & George, E. I. (1994), The Risk Inflation Criterion for Multiple Regression, *The Annals of Statistics*, **22**, No. 4. pp. 1947-1975.
- Furnival, G.M. & Wilson, R.W. (1974), Regression by leaps and bounds, *Technometrics*, **16**, pp. 499-511.
- George, E. & Foster, D. P. (2000), Calibration and empirical Bayes variable selection, *Biometrika*, **87**, No. 4, pp. 731-747(17).
- Ghosh, J. K., Delampady, M. & Samanta, T. (2006), *An Introduction to Bayesian Analysis - Theory and Methods*, Springer, New York.
- Green, P. J. (1995), Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, **82**, No. 4. pp. 711-732.
- Hansen, B. E. & Racine, J. (2007), Jackknife Model Averaging, Working Paper.
- Hansen, B. E. (2007), Least Squares Model Averaging, *Econometrica*, **75**, pp. 1175-1189.
- Hansen, M. & Yu, B. (2001), Model selection and the principle of minimum description, *Journal of the American Statistical Association*, **96**, No. 454, pp. 746-774 (29).
- Hoeting, J.A, Madigan,D., Raftery,A.E & Volinsky C.T (1999), Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors), *Statistical Science*, **14**, No. 4, pp. 382-417.
- Jeffreys, H. (1961). *The Theory of Probability*, The Oxford University Press.
- Kass, R. E. & Wasserman, L. (1995), A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* , **90**, pp. 928-934.
- Kass, R.E. & Raftery, A.E. (1995), Bayes factors, *Journal of the American Statistical Association*, **90**, pp. 773-795.
- Konishi & Kitagawa (2008), *Information Criteria and Statistical Modeling*, Springer Science+Business Media.

Koop, G. & Potter, S. (2008) Forecasting in Dynamic Factor Models using Bayesian Model Averaging, *Econometrics Journal* (forthcoming).

Koop, G. (2003), *Bayesian Econometrics*, John Wiley & Sons.

Leeb, H. & Pötscher, B. M., (2005), Model selection and inference: Facts and fiction, *Econometric Theory*, **21**, pp. 21-59.

Liang, F., Paolo, R., Molina, G., Clyde, M.A. & Berger J.O (2008), Mixtures of g Priors for Bayesian Variable Selection, *Journal of the American Statistical Association*, **103**, pp. 410-423.

Lopes, H. P. (2008), A Note On Reversible Jump Markov Chain Monte Carlo. (<http://faculty.chicagogsb.edu/hedibert.lopes/teaching/BayesianEconometrics/rjmcmc.pdf>)

Madigan, D.M. & Raftery, A.E. (1994), Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association*, **89**, pp. 1335-1346.

Massart, P. (2003), *Concentration Inequalities and Model Selection*, Springer-Verlag Berlin Heidelberg. (http://www.math.u-psud.fr/~massart/stf2003_massart.pdf)

McLachlan, G. J. & Krishnan, T. (2008), *The EM Algorithm and Extensions* (Second Edition), John Wiley & Sons.

Miller, A. (2002), *Subset Selection in Regression* (Second Edition), Chapman & Hall/CRC.

Nishii, R. (1984), Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression, *Annals of Statistics*, **12**, No. 2, pp. 758-765.

Noble, R.B. (2001) Multivariate Applications of Bayesian Model Averaging, PhD Dissertation, Virginia Polytechnic Institute & State University.

Ntzoufras, I., (1999), *Aspects of Bayesian Model and Variable Selection Using MCMC*, PhD Thesis, Dep. of Statistics, Athens University of Economics and Business.

Ntzoufras, I., Dellaportas, P. & Forster, J.J. (2003), Bayesian variable and link determination for generalised linear models, *Journal of Statistical Planning and Inference*, **111**, Vol. 1-2, pp. 165-180.

Pesaran, M. H., Pettenuzzo, D. & Timmermann, A. (2006), Forecasting Time Series Subject To Multiple Structural Breaks, *Review of Economic Studies*, **73**, No. 4, pp. 1057-1084.

Pötscher, B. M. & Novák, A. J. (1998). The distribution of estimators after model selection: Large and small sample results. *Journal of Statistical Computation and Simulation*, **60**, pp. 19-56.

Rachev, S. T., Hsu, J. S. H. & Bagasheva, B. S. (2008), *Bayesian Methods in Finance*, John Wiley & Sons.

- Raftery, A.E. (1996), Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, **83**, pp. 251-266.
- Raftery, A.E., Madigan, D. & Hoeting, J.A. (1997). Bayesian model averaging for regression models. *Journal of the American Statistical Association*, **92**, pp. 179-191.
- Raftery, A.E., Madigan, D., & Volinsky, C.T. (1995), Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance, *Bayesian Statistics 5*, Oxford University Press, pp. 323-349.
- Sellke, T., Bayarri, M.J. & Berger, J.O. (2001), Calibration of P-values for testing precise null hypotheses, *The American Statistician*, **55**, pp. 62-71.
- Severini, T. A. (2005), *Elements of Distribution Theory*, Cambridge University Press.
- Smith, M. & Kohn, R. (1996), Nonparametric regression using Bayesian variable selection, *Journal of Econometrics*, **75**, pp. 317-344.
- Tierney, L. & Kadane, J. (1986), Accurate approximation for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, pp. 82-86.
- Volinsky, C.T., Madigan, D., Raftery, A.E. & Kronmal, R.A. (1997), Bayesian Model Averaging in Proportional Hazard Models: Predicting the Risk of a Stroke, *Applied Statistics*, **46**, pp. 443-448.
- Volinsky, C.T., Madigan, D., Raftery, A.E. & Kronmal, R.A. (1997), Bayesian model averaging in proportional hazard models: Assessing stroke risk. *Journal of the Royal Statistical Society, series C---Applied Statistics*, **46**, pp. 433-448.
- Vrontos, I. Dellaportas, P. & Politis, D. (2000) Full Bayesian inference for GARCH and EGARCH models. *Journal of Business and Economics Statistics*, pp. 187-198.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.), pp. 233-243. North-Holland, Amsterdam.

