ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΠΕΫΖΙΑΝΗ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΩΝ ΓΙΑ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ

ΕΝΤΒΙΝ ΚΑΤΣΜΟΛΙ

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστήμιου Αθηνών στο πλαίσιο του

Προπτυχιακού Προγράμματος Σπουδών

Αθήνα

Φεβρουάριος 2024

ΑΦΙΕΡΩΣΗ

Στην οικογένεια μου,

για την συνεχή υποστήριξη και ενθάρρυνση τους

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τους παρακάτω ανθρώπους για την ανεπιφύλακτη υποστήριξή τους και πολύτιμη βοήθεια τους στην διάρκεια συγγραφής της διπλωματικής εργασίας μου.

Καταρχάς, θα ήθελα να ευχαριστήσω τον επιβλέποντα της εργασίας μου, Καθηγητή Ιωάννη Ντζούφρα για την πολύτιμη γνώση που παρείχα καθώς και την καθοδήγηση, εμπιστοσύνη που μου έδειξε καθ' όλη την διάρκεια της διπλωματικής εργασίας.

Θα ήθελα επίσης να εκφράσω την ειλικρινή μου ευγνωμοσύνη για τον Καθηγητή Παναγιώτη Παπασταμούλη για τα σημαντικά εφόδια και τις πολύτιμες γνώσεις Μπεϋζιανής θεωρίας που απέκτησα κατά την διάρκεια του εξαμήνου καθώς και για τον χρόνο που αφιέρωσε σε εμένα για την κατανόηση δυσκολότερων εννοιών. Πρακτικά, αυτή η εργασία δεν θα ήταν ίδιας ποιότητας χωρίς την βοήθεια του.

Θερμές ευχαριστίες θα ήθελα να απευθύνω στον Καθηγητή Δημήτριο Καρλή για τις χρήσιμες συμβουλές του σχετικά με τα Μεγάλα δεδομένα. Οι έξυπνες παρατηρήσεις και προτάσεις του έχουν αλλάξει σημαντικά τον τρόπο που αντιμετωπίζω κάθε πρόβλημα που σχετίζεται γενικότερα με την επιστήμη της στατιστικής.

Τέλος, θα ήθελα να ευχαριστώ πολύ την οικογένειά μου και τους φίλους μου που είχαν και ακόμα έχουν πίστη σε μένα.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

ΠΕΡΙΛΗΨΗ

Έντβιν Κατσμόλι

ΜΠΕΫΖΙΑΝΗ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΩΝ ΓΙΑ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ

Φεβρουάριος 2024

Παρουσίαση και περιγραφή Μπεϋζιανών αλγορίθμων επιλογής μοντέλων στο πλαίσιο των μεγάλων δεδομένων. Ειδικότερα, θα εισάγουμε βασικές έννοιες της Μπεϋζιανής στατιστικής με σκοπό την καλύτερη κατανόηση των παραπάνω αλγορίθμων. Θα εφαρμόσουμε τους παραπάνω αλγορίθμους σε διάφορες περιπτώσεις προσομοιωμένων δεδομένων και θα προσπαθήσουμε να αναδείξουμε μια πιθανή μέθοδο για την διαχείριση των μεγάλων δεδομένων σε τέτοιες εφαρμογές.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο της διατριβής

Τα μοντέλα αποτελούν ένα από τα σημαντικότερα εργαλεία της επιστήμης της στατιστικής για την αποτύπωση και κατανόηση των δεδομένων. Μέσα από αυτά, μπορούμε να εξηγήσουμε τον μηχανισμό ή φύση των δεδομένων δίνοντας μας την δυνατότητα να λάβουμε αποφάσεις και να κάνουμε προβλέψεις. Σε αυτό το σημείο αξίζει να αναφέρουμε τον γνωστό αφορισμό *"Όλα τα μοντέλα είναι λάθος αλλά κάποια είναι χρήσιμα."* του George Box το 1976. Ο παραπάνω αφορισμός αναγνωρίζει πως η πολυπλοκότητα της πραγματικότητας είναι πάντα αδύνατη να εντοπιστεί από τα στατιστικά μοντέλα, ωστόσο, κάποια από αυτά επαρκούν για να την προσεγγίσουν. Αυτό ισχύει ειδικότερα σε μια εποχή όπου ο όγκος των δεδομένων αυξάνεται με ραγδαίο ρυθμό και η ανάγκη για τα μοντέλα γίνεται όλο και πιο έντονη. Το παραπάνω έχει ως συνέπεια, η αναζήτηση του "πραγματικού" μοντέλου που γέννησε τα δεδομένα να γίνεται όλο και πιο απαιτητική.

Το πρόβλημα επιλογής μοντέλου είναι ένα από τα πιο δύσκολα προβλήματα όπου η επιστήμη της στατιστικής καλείται να αντιμετωπίσει. Γενικότερα, υπάρχει μια μεγάλη ποικιλία μεθόδων που έχει αναπτυχθεί για την επιλογή μοντέλου, αλλά στην παρούσα έρευνα θα παρουσιάσουμε και θα αναλύσουμε ένα υποσύνολο αυτών. Οι περισσότερες επιστήμες αντιμετωπίζουν το παραπάνω πρόβλημα μέσω της κλασικής προσέγγισης, η οποία κάνει χρήση διάφορων αλγορίθμων που θα δούμε εν συνέχεια. Ωστόσο, πέρα από το πρόβλημα επιλογής μεταβλητών, ιδιαίτερο θέμα τίθεται στον μεγάλο όγκο παρατηρήσεων. Επειδή οι περισσότερες μεθοδολογίες επιλογής μοντέλου είχαν αναπτυχθεί σε μια περίοδο σχετικά μικρού συνόλου παρατηρήσεων, στις σημερινές ημέρες διατρέχουμε τον κίνδυνο μη-δυνατής εκτέλεσης του αλγορίθμου λόγου μη-ρεαλιστικού χρόνου αναμονής. Οπότε, ένας από τους σκοπούς αυτής της εργασίας είναι να δούμε κάποιον πιθανό τρόπο αντιμετώπισης.

Η παραπάνω έρευνα επικεντρώνεται κυρίως στην ανάδειξη Μπεϋζιανών τεχνικών επιλογής μοντέλου. Η Μπεϋζιανή στατιστική, γεννήθηκε στα μέσα του 20ού αιώνα και πιο συγκεκριμένα την δεκαετία του 1950. Αξίζει να αναφέρουμε, ότι τα θεμέλια για την γέννηση αυτής της περιοχής οφείλεται στον στατιστικό Thomas Bayes (1701-1761), τον 18ο αιώνα. Οι επιστήμονες που ήταν υπέρ στον τρόπο σκέψης του Thomas Bayes διαχωρίστηκαν από τους υπόλοιπους κλασικούς στατιστικούς, όπου ήταν η επικρατέστερη προσέγγιση στατιστικής αναπτυγμένη από τον Ronald Fisher (1890-1962) και την ομάδα Jerzy Neyman (1894-1981) και Egon Pearson (1895-1980), και ονομάστηκαν Μπεϋζιανοί στατιστικοί. Η αρχή της Μπεϋζιανής θεωρίας είχε σημαντικό αντίκτυπο την δεκαετία του 1950, αλλά η εφαρμογή τους ήταν δύσκολη έως και αδύνατη. Αυτό πήγαζε από το γεγονός της αδύνατης υπολογιστικής τους εκτίμησης εκείνης της εποχής, με την αναλυτική λύση να είναι ανέφικτη σε πρακτικά προβλήματα. Από τις αρχές του 1980 έως και σήμερα, η υπολογιστική δύναμη και οι υπολογιστικές τεχνικές αναπτύσσονται ραγδαία, με αποτέλεσμα την πραγματοποίηση αυτών των υπολογισμών. Το παραπάνω, επιτρέπει τους ερευνητές να εξερευνήσουν όλες τις δυνατότητες που θα μπορούσε να προσφέρει η Μπεϋζιανη στατιστική.

1.2 Δομή της διατριβής

Το υπόλοιπο αυτής της έρευνας θα προχωρήσει ως εξής: Στο Κεφάλαιο 2 αναλύονται οι βασικές Μπεϋζιανές έννοιες. Θα μελετήσουμε αρχικά τις θεμελιώδες διαφορές της έννοια της πιθανότητας και συμπερασματολογίας μεταξύ των δύο προσεγγίσεων και έπειτα θα αναπτύξουμε τη Μπεϋζιανή συμπερασματολογία από το θεώρημα του Bayes §2.4.1 μέχρι και τους Μπεϋζιανούς ελέγχους υποθέσεων §2.4.7. Στην συνέχεια, στο Κεφάλαιο 3, θα εστιάσουμε το ενδιαφέρον μας στο κλασικό κανονικό γραμμικό μοντέλο και ιδιαίτερα στο Μπεϋζιανό αντίστοιχο. Έπειτα, στο Κεφάλαιο 4, θα εισάγουμε τις βασικές έννοιες για την κατανόηση των Markov chain Monte Carlo μεθοδολογιών όπως τον δειγματολήπτη Gibbs και αλγόριθμο Metropolis Hastings. Αυτά θα αφορούν τις Μαρκοβιανές αλυσίδες §4.1 και μεθοδολογίες Monte-Carlo §4.2. Στο Κεφάλαιο 5, θα αναλύσουμε την Μπεϋζιανη επιλογή μοντέλου. Συγκεκριμένα, στο Κεφάλαιο §5.1 θα αναπτύξουμε την θεωρία του συζυγούς μοντέλου. Στο Κεφάλαιο §5.2 θα εξερευνήσουμε κάποιους αλγορίθμους επιλογής μοντέλου όπως Reversible Jump Markov Chain Monte Carlo §5.2.1, Stochastic Search Variable Selection §5.2.2, Gibbs Variable Selection §5.2.3 και Bayesian Adaptive Sampling §5.2.4. Έπειτα, στο Κεφάλαιο 6, και συγκεκριμένα στο Κεφάλαιο §6.1 θα εφαρμόσουμε τα παραπάνω με διαφορετικές περιπτώσεις προσομοιωμένων δεδομένων και θα τις συγκρίνουμε με τις κλασικές μεθόδους. Εν συνέχεια, στο Κεφάλαιο §6.2, θα προσπαθήσουμε να βρούμε τρόπο για μια ταχύτερη εκτέλεση των μεθόδων στην περίπτωση μεγάλου όγκου παρατη-

ρήσεων. Τέλος, στο Κεφάλαιο 7, θα συνοψίσουμε τα συμπεράσματα της παρούσας έρευνας και θα δώσουμε κάποιες περαιτέρω ιδέες και προσεγγίσεις πάνω στο αντικείμενο της Μπεύζιανής επιλογής μοντέλου.

Κεφάλαιο 2

Ερμηνεία Πιθανότητας

Στο κόσμο της στατιστικής υπάρχουν δύο διαφορετικές φιλοσοφίες για το πως γίνεται αντιληπτή η έννοια της πιθανότητας. Αυτές οι δύο φιλοσοφίες είναι η Μπεϋζιανή και κλασική προσέγγιση. Σε αυτό το κεφάλαιο θα εξετάσουμε τις διαφορές τους στην θεμελιώδη έννοια της πιθανότητας καθώς και την συμπερασματολογία τους.

2.1 Φιλοσοφία Κλασικής Στατιστικής

Στον πυρήνα της, η κλασική Στατιστική θεωρεί την πιθανότητα ως μια ιδιότητα ακολουθίας γεγονότων που συμβαίνει επαναλαμβανόμενα κάτω από τις ίδιες συνθήκες. Μετράμε την συχνότητα των φορών μέσα σε αυτήν τη ακολουθία όπου παρατηρούμε ένα συγκεκριμένο αποτέλεσμα και το διαιρούμε με το συνολικό αριθμό επαναλήψεων. Το ευκολότερο παράδειγμα είναι η ρίψη ενός νομίσματος. Αν μας ενδιέφερε να μάθουμε την πιθανότητα για κάποιο από τα δύο ενδεχόμενα, η κλασική προσέγγιση, επειδή δεν γνωρίζει τίποτα για αυτό το νόμισμα, θα πραγματοποιούσε το πείραμα ρίψης νομίσματος και θα μέτραγε το πόσες φορές έτυχε το κάθε ενδεχόμενο προς το συνολικό αριθμό ρίψεων. Θεωρητικά, αν ο αριθμός επαναλήψεων τείνει το άπειρο τόσο πιο κοντά φτάνουμε στην αλήθεια (Νόμος μεγάλων αριθμών).

2.2 Συμπερασματολογία Κλασικής Στατιστικής

Η κλασική Στατιστική συμπερασματολογία, έχει ως σκοπό να εξάγει συμπεράσματα για το υπο-μελέτη πληθυσμό μέσα από ένα δείγμα. Αν μας ενδιέφερε να εξάγουμε συμπεράσματα για το μέσο του πληθυσμού, θα μπορούσε να γίνει με την βοήθεια των ελέγχων υποθέσεων. Οι έλεγχοι υποθέσεων είναι απλά ιδέες που θέλουμε να ελέγξουμε κατά το πόσο ισχύουν. Παρ' όλο αυτά, από το δείγμα μας έχουμε την δυνατότητα να υπολογίσουμε μόνο έναν δειγματικό μέσο όποτε η υπόθεση μας για το μέσο

θα είναι είτε αληθής με πιθανότητα 1 είτε ψευδής με πιθανότητα 0. Αυτό συμβαίνει διότι, δεν είναι πρακτικό ή σχεδόν αδύνατο να πάρουμε πολλά δείγματα από το υπό-μελέτη πληθυσμό και να υπολογίσουμε τους αντίστοιχους δειγματικούς μέσους για να μπορούμε να πούμε κατά πόσο πιθανό είναι να ισχύει η υπόθεση μας ή όχι. Οπότε, παρατηρούμε πως ο ορισμός της πιθανότητας μέσω της κλασικής προσέγγισης των «άπειρων» δειγμάτων είναι θεμελιώδης στην κλασική Στατιστική συμπερασματολογία. Ο τρόπος με τον οποίο αντιμετωπίζεται αυτό το πρόβλημα είναι μέσω της δειγματοληπτικής κατανομής των εκτιμητριών. Για να καταλάβουμε το γιατί, μπορούμε να το σκεφτούμε απλώς ως ένα πρόβλημα προσομοίωσης. Προσομοιώνουμε δειγματοληπτικούς μέσους από την κανονική κατανομή κάτω από την μηδενική υπόθεση, δηλαδή, όταν θεωρούμε ότι η υπόθεση μας είναι ψευδής. Έτσι, συγκρίνουμε τις προσομοιωμένες εκτιμήσεις με αυτό που έχουμε πραγματικά παρατηρήσει. Τέλος, μπορούμε να βγάλουμε μια πιθανότητα του κατά πόσο ισχύει η υπόθεση μας, δηλαδή, την πιθανότητα αυτό που έχουμε παρατηρήσει να είναι πιο ακραίο σε σύγκριση με τις προσομοιωμένες εκτιμήσεις (p-value). Αν έχουμε μικρή πιθανότητα η υπόθεση μας να είναι ψευδής (H_1) τότε απορρίπτουμε την μηδενική υπόθεση (H_0). Ο Eric-Jan Wagenmakers (2007) συζητά εκτενώς το θέμα των p-value που βασίζονται πρακτικά σε απαρατήρητα δεδομένα.

Την ίδια λογική έχουν και τα διαστήματα εμπιστοσύνης. Πολλές φορές, μπορεί να φαίνεται πως έρχονται σε αντίφαση με αυτά που έχουμε πει μέχρι στιγμής καθώς φαίνεται να έχει περισσότερο νόημα να ερμηνευτεί ως «με 95% πιθανότητα η παράμετρος βρίσκεται στο διάστημα εμπιστοσύνης». Ωστόσο, από την άποψη της κλασικής Στατιστικής αυτή η ερμηνεία είναι λάθος αφού δεν ορίζει με αυτόν τον τρόπο την πιθανότητα και κατ' επέκταση τα διαστήματα εμπιστοσύνης. Όπως πριν, επειδή έχουμε ένα δείγμα μπορούμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης που θα θέλαμε ιδανικά να περιέχει τον μέσο. Αν το ερμηνεύσουμε με την κλασική φιλοσοφία τότε το διάστημα είτε θα περιέχει την παράμετρο με πιθανότητα 1 είτε όχι με πιθανότητα 0. Το παραπάνω, συμβαίνει για τους ίδιους λόγους που αναφέραμε προηγουμένως. Ο τρόπος με τον οποίο κατασκευάζεται το διάστημα εμπιστοσύνης είναι πάλι με την βοήθεια της δειγματοληπτικής κατανομής. Αν με παρόμοια λογική, προσομοιώσουμε διαστήματα εμπιστοσύνης τότε μπορούμε να δώσουμε πιθανότητα, αλλά η ερμηνεία που θα δοθεί δεν είναι η ίδια με την αρχική, καθώς εδώ λέμε πως «Όσο ο αριθμός διαστημάτων εμπιστοσύνης τείνουν στο άπειρο τότε τα 95% από αυτά θα περιέχουν τη παράμετρο». Αυτή η ερμηνεία συμφωνεί με τον ορισμό πιθανότητας της κλασικής Στατιστικής καθώς υπάρχει η δειγματοληπτική κατανομή που μας βοηθάει στις «άπειρες» επαναλήψεις πειράματος. Αυτή η ερμηνεία είναι τελείως διαφορετική από αυτή που δόθηκε αρχικά. Στην πρώτη ερμηνεία πρακτικά δίναμε πιθανότητα στην ίδια την παράμετρο κάτι που η κλασική Στατιστική δεν κάνει (πιο αναλυτικά αργότερα). Στην δεύτερη ερμηνεία δίνουμε πιθανότητα στα άκρα του διαστήματος, εκεί που υπάρχει το δειγματοληπτικό σφάλμα, διότι

εκεί μπλέκεται η δειγματοληπτική κατανομή της εκτιμήτριας. Συνοψίζοντας, στα διαστήματα εμπιστοσύνης εννοούμε το διάστημα εμπιστοσύνης να περιέχει την παράμετρο παρά η παράμετρος να εμπεριέχεται στο διάστημα εμπιστοσύνης καθώς δίνουμε αβεβαιότητα στο δειγματοληπτικό σφάλμα μόνο.

Από τα παραπάνω, και ειδικά στην συζήτηση για τα διαστήματα εμπιστοσύνης, παρατηρούμε πως δίνετε ιδιαίτερη έμφαση στο γεγονός της μηδενικής αβεβαιότητας στην παράμετρο. Αυτό το γεγονός, έχει ειπωθεί από την αρχή αλλά δύσκολο να εντοπιστεί. Αν πάμε στο παράδειγμα του νομίσματος της Ενότητας §2.1, ειπώθηκε το γεγονός ότι όσο αυξάνεται ο αριθμός των πειραμάτων τόσο θα τείνουμε στην αλήθεια. Σε αυτό το παράδειγμα, η αλήθεια είναι η πραγματική πιθανότητα νομίσματος το οποίο μπορεί να θεωρηθεί παράμετρος. Ένα άλλο παράδειγμα αφορά το δειγματικό μέσο το οποίο τείνει όλο και περισσότερο στην αλήθεια (πληθυσμιακό μέσο - παράμετρο) όσο το μέγεθος του δείγματος αυξάνεται. Οπότε στην κλασική προσέγγιση, οποιαδήποτε αβεβαιότητα έχει να κάνει με το δειγματοληπτικό σφάλμα και όχι την ίδια την παράμετρο. Όποτε, η παράμετρος στην κλασική Στατιστική θεωρείτε σταθερή αλλά άγνωστη, δηλαδή, υπάρχει μια σωστή απάντηση.

2.3 Φιλοσοφία Μπεϋζιανής Στατιστικής

Από την Ενότητα §2.1, το πρόβλημα με το ορισμό της πιθανότητας της κλασικής προσέγγισης είναι ότι δεν μπορούμε να ορίσουμε ”πραγματικά” την πιθανότητα καθώς στην πραγματική ζωή δεν έχουμε τη δυνατότητα ή δεν μπορούμε να επαναλάβουμε το πείραμα.

Ένα απλό παράδειγμα αφορά τους αγώνες ποδοσφαίρου. Σε κάθε αγώνα, βγαίνουν πιθανότητες νίκης της αντίστοιχης ομάδας. Αυτή η πιθανότητα δεν μπορεί να εξηγηθεί από την κλασική προσέγγιση καθώς είναι αδύνατον ο αγώνας να επαναληφθεί κάτω από τις ίδιες συνθήκες (ίδιοι παίκτες κ.λ.π). Στην πραγματικότητα, η κλασική προσέγγιση θα έλεγε πως η αντίστοιχη ομάδα είτε θα νικήσει με πιθανότητα 1 είτε θα χάσει με πιθανότητα 0. Αυτό συμβαίνει διότι στο τέλος του αγώνα θα έχουμε ένα τελικό αποτέλεσμα οπότε θα έχουμε ένα αποτέλεσμα του πειράματος. Τέλος, η συμπερασματολογία της κλασικής προσέγγισης βασίζεται σε δειγματοληπτικές κατανομές όπου στην ουσία είναι δεδομένα που δεν έχουν παρατηρηθεί ποτέ. Αυτό το γεγονός φέρει αρκετά μειονεκτήματα όπως οι αυστηρές υποθέσεις για τον πληθυσμό, η ευαισθησία στις ακραίες παρατηρήσεις κ.ο.κ.

Από την άλλη πλευρά, η Μπεϋζιανή προσέγγιση χρησιμοποιεί την πιθανότητα ακριβώς όπως ένας άνθρωπος θα τη χρησιμοποιούσε, δηλαδή, η ποσοτικοποίηση της αβεβαιότητας μιας αντίστοιχης πρότασης ή κατάστασης μέσω των πιθανοτήτων. Ο όρος «αβεβαιότητα» μπορεί να έχει πολλές ερμηνείες. Ένα συμβάν μπορεί να είναι αβέβαιο λόγου της εκ φύσεως απροβλεψιμότητας, υπόκειται σε τυχαία μεταβλητότητα. Για παράδειγμα, η θέση ηλεκτρονίων στην κβατική φυσική περιγράφεται με πιθανότητες καθώς είναι εκ φύσεως τυχαίο. Θα μπορούσε να είναι επίσης αβέβαιο απλώς και μόνο επειδή έχουμε ατελή γνώση του, για παράδειγμα, η νίκη σε ποδοσφαιρικό αγώνα. Η Μπεϋζιανή στατιστική βασίζεται στην ιδέα του ότι η μόνη ικανοποιητική περιγραφή της αβεβαιότητας μας επιτυγχάνεται μέσω της πιθανότητας. Γιατί όμως δεν υπάρχει άλλος τρόπος για να περιγράψουμε την αβεβαιότητα; Διότι ο μόνος ικανοποιητικός τρόπος να συνδέσουμε διάφορες γνώμες που αφορούν την αβεβαιότητα είναι να το κάνουμε με τον ίδιο ακριβώς τρόπο όπως χειριζόμαστε τις πιθανότητες.

Εφόσον περιγράφουμε την αβεβαιότητα μας μέσω των πιθανοτήτων, τότε η μαθηματικοποίηση του θα πρέπει να υπακούει την βασική υπόθεση της συνέπειας (coherence) το οποίο αναφέρει πως η υποκειμενική πιθανότητα ακολουθεί τα αξιώματα του λογισμού πιθανοτήτων, έτσι συνεπείς αποφάσεις μπορούν να παρθούν από αυτές τις πιθανότητες.

Αν Ω ο δειγματικός χώρος σε ένα πείραμα. Ας θεωρήσουμε σε ένα ενδεχόμενο A του Ω αντιστοιχείται ένας πραγματικός αριθμός $\mathbb{P}(A)$, αν

- $\mathbb{P}(A) \geq 0$ για κάθε ενδεχόμενο A του Ω.
- Όλα τα δυνατά αποτελέσματα πιθανοτήτων να αθροίζουν στην μονάδα, δηλαδή, $\mathbb{P}(\Omega) = 1$.
- Αν $A_1, A_2, ...$ ακολουθία ξένων ανά δύο ενδεχομένων του Ω τότε $\mathbb{P}(A_1 \cup A_2 \cup ...) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + ...$

Τότε η συνάρτηση $\mathbb{P}(\cdot)$ είναι συνάρτηση πιθανότητας στον δειγματικό χώρο Ω και ο αριθμός $\mathbb{P}(A)$ λέγεται πιθανότητα του ενδεχομένου A. Αξίζει να σημειώσουμε πως οι παραπάνω ιδιότητες είναι γνωστές ως αξιώματα του Kolmogorov. Αυτά είναι τα μόνα εργαλεία μας καθώς η πιθανότητα είναι το μόνο εργαλείο που χρειαζόμαστε για την μελέτη της αβεβαιότητας.

Με βάση αυτά, μπορούμε πλέον να διακρίνουμε τις διαφορές μεταξύ της Μπεϋζιανής προσέγγισης και των Ενοτήτων §2.1 και §2.2. Από την Μπεϋζιανή προσέγγιση, οι άγνωστες παράμετροι αντιμετωπίζονται ως αβέβαιες ποσότητες οι οποίες πρέπει να περιγράφονται δια μέσου πιθανοτήτων (εδώ έχουμε την αβεβαιότητα ίδιας της παραμέτρου). Για παράδειγμα, η κλασική προσέγγιση θα πει πως η πιθανότητα νίκης σε αγώνα θα είναι 0 ή 1 (επειδή δεν μπορούμε να επαναλάβουμε το πείραμα) ενώ η

Μπεϋζιανή προσέγγιση θα μπορούσε να πει πως είναι πιθανότερη η υψηλή πιθανότητα νίκης και λιγότερο πιθανή η χαμηλή πιθανότητα νίκης. Αυτή η αρχική μας γνώμη για μια αβέβαιη ποσότητα είναι η «καρδιά» της Μπεϋζιανής θεωρίας διότι έτσι ακριβώς γίνεται αντιληπτή η έννοια της πιθανότητας σε αυτήν την προσέγγιση (Υποκειμενική πιθανότητα - Subjective Probability).

Τώρα, η πιο βασική και σημαντική ιδιότητα στην Μπεϋζιανη στατιστική είναι η ανανέωση της αρχικής μας γνώμης όσο παρατηρούμε δεδομένα. Οπότε, με βάση όσα έχουν ειπωθεί, στην πραγματικότητα δεν υπάρχει αληθινή πιθανότητα αλλά απλά μια έκφραση της σχέσης μας με τον κόσμο. Στο παράδειγμα με το αγώνα, θα μπορούσε η ομάδα να χάνει στο πρώτο ημίχρονο (παρατηρήθηκε δεδομένο) οπότε ανανεώνουμε την αρχική μας αβεβαιότητα για την πιθανότητα νίκης στο να γίνει ίσως πιθανότερη η χαμηλή πιθανότητα νίκης.

Φυσικά, θα μπορούσε κανείς να πει πως η Μπεϋζιανη προσέγγιση θα έχει πολλές σωστές απαντήσεις καθώς ο καθένας δίνει την δική του υποκειμενική πιθανότητα και έτσι δεν έχει νόημα καν να βρούμε κάποια απάντηση. Πράγματι, η βασική αντίρρηση για αυτή την θεωρία εντοπίζεται στο γεγονός ότι τα συμπεράσματα εξαρτώνται από την προσωπική μας άποψη. Όμως, όπως τονίσαμε και προηγουμένως, κάποιοι υποστηρίζουν πως ακριβώς αυτό το σημείο κρύβει όλη την «ομορφιά» της Μπεϋζιανής θεωρίας καθώς έρχεται σε παράθεση με την λογική του ανθρώπου, εφόσον έχουμε εμπειρίες από τον κόσμο που παρατηρούμε. Άλλοι υποστηρίζουν, πως αυτή η αντίρρηση είναι άστοχη, καθώς η ίδια επιστήμη δεν είναι πραγματικά αντικειμενική. Δυστυχώς, η αντιπαράθεση αυτή θα μπορούσε να οδηγήσει σε μια ατελείωτη συζήτηση πάνω σε αυτό το ζήτημα.

2.4 Συμπερασματολογία Μπεϋζιανής Στατιστικής

Σε αυτό το Κεφάλαιο θα αναλύσουμε τις βασικότερες έννοιες της Μπεϋζιανής στατιστικής συμπερασματολογίας. Πιο συγκεκριμένα, θα αναλύσουμε τον Κανόνα του Bayes, την σύγχρονη θεωρία του, την θεμελιώδη έννοια της συνάρτησης πιθανοφάνειας, την εκ των προτέρων κατανομή με τις αντίστοιχες κατηγορίες και τέλος τους Μπεϋζιανούς ελέγχους υποθέσεων.

2.4.1 Το Θεώρημα του Bayes

Το ερώτημα που καλούμαστε να απαντήσουμε σε αυτήν την ενότητα είναι το πως γίνεται ακριβώς αυτή η ανανέωση της εκ των προτέρων γνώμης όταν ”έρχονται” δεδομένα. Η απάντηση σε αυτό είναι το θεώρημα του Bayes, το οποίο για να γίνει διαισθητικά κατανοητό, η περιγραφή του θα γίνει αρχικά με την βοήθεια ενδεχομένων. Αυτό που υπολογίζουμε στο θεώρημα στην πραγματικότητα είναι μια

δεσμευμένη πιθανότητα, δηλαδή, την πιθανότητα ενός ενδεχομένου δοθέντος ότι έχει συμβεί κάποιο άλλο ενδεχόμενο (ή δοθέντος ότι παρατηρήσαμε τα δεδομένα).

Η δεσμευμένη πιθανότητα ορίζεται ως

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \ , \ \mathbb{P}(B) > 0 \tag{2.4.1}$$

Ο τρόπος με τον οποίο σκεφτόμαστε αυτόν τον τύπο είναι να μην κοιτάμε το σύνολο όλων των δυνατών ενδεχομένων (δειγματικό χώρο Ω) αλλά κοιτάμε έναν «περιορισμένο» δειγματικό χώρο, εφόσον έχει ήδη συμβεί το Β ενδεχόμενο. Το άθροισμα όλων των δυνατών περιπτώσεων που αφορούν και το Β, είναι το $\mathbb{P}(B)$ (ολική πιθανότητα). Διαισθητικά, όταν έχουμε να υπολογίσουμε την απλή πιθανότητα $\mathbb{P}(\mathrm{A})$, τυπικά διαιρούμε με το 1, δηλαδή, $\mathbb{P}(\mathrm{A}) = \frac{\mathbb{P}(A)}{1}$, όπου $\mathbb{P}(\Omega) = 1$ αθροίζουμε δηλαδή όλα τα δυνατά ενδεχόμενα που μπορούν να συμβούν και από όλες αυτές ρωτάμε: Πόσο πιθανό είναι να μου συμβεί το ενδεχόμενο Α (όπου ανήκει σε όλα τα δυνατά ενδεχόμενα). Παρόμοια λογική έχει και η δεσμευμένη πιθανότητα. Αν μας ενδιέφερε μόνο η $\mathbb{P}(\mathrm{A})$ τότε θα έπρεπε να λάβουμε υπόψη όλα τα δυνατά ενδεχόμενα. Τυχαίνει όμως να γνωρίζουμε ότι έχει συμβεί ένα γεγονός Β, τότε ο δειγματικός χώρος περιορίζεται στα ενδεχόμενα που έχει συμβεί και το γεγονός Β. Από όλα τα δυνατά ενδεχόμενα που αφορούν το Β θέλουμε να βρούμε το πόσο πιθανό είναι να συμβεί και το Α.

Παρ' όλο αυτά, η εξίσωση (2.4.1) δεν είναι ακόμα ακριβώς το θεώρημα του Bayes. Από την εξίσωση (2.4.1) μπορούμε να γράψουμε ισοδύναμα την δεσμευμένη πιθανότητα ως

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \tag{2.4.2}$$

Η νέα σχέση (2.4.2) είναι το θεώρημα του Bayes!

Απόδειξη. Παρατηρούμε ότι η μόνη διαφορά της σχέσης (2.4.1) και (2.4.2) είναι ο αριθμητής

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \xrightarrow{(1)+(2)} \mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

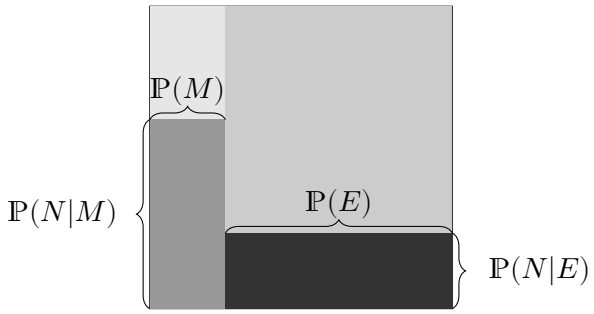
$$\mathbb{P}(\mathrm{B}|\mathrm{A}) = \frac{\mathbb{P}(\mathrm{B} \cap \mathrm{A})}{\mathbb{P}(\mathrm{A})} \Rightarrow \mathbb{P}(\mathrm{B} \cap \mathrm{A}) = \mathbb{P}(\mathrm{B}|\mathrm{A}) \cdot \mathbb{P}(\mathrm{A}) \qquad (1)$$

$$\mathbb{P}(\mathrm{B} \cap \mathrm{A}) = \mathbb{P}(A \cap B) \qquad (2)$$

Πως ακριβώς σχετίζεται η σχέση (2.4.2) με όλα όσα έχουμε πει;

Έστω ότι είμαστε σε ένα πανεπιστήμιο και συναντάμε έναν φοιτητή. Ενώ του μιλάμε, παρατηρούμε ότι είναι ντροπαλός (N). Αν ήταν να μαντέψουμε: Είναι πιο πιθανό ο φοιτητής να είναι στο τμήμα μαθηματικών (M) ή τμήμα επιχειρήσεων (E); Οι περισσότεροι θα βάζανε υψηλότερη πιθανότητα στο τμήμα μαθηματικών διότι η ντροπαλότητα συνδέεται περισσότερο με τμήμα μαθηματικών παρά το τμήμα επιχειρήσεων [1]. Παρ' όλο αυτά, υπάρχει και άλλη σχετική πληροφορία όπου ο κόσμος ξεχνά όταν απαντά σε τέτοια ερωτήματα: Η αναλογία των ατόμων στο τμήμα μαθηματικών σε σύγκριση με το τμήμα επιχειρήσεων. Προφανώς δεν είναι ανάγκη να ξέρουμε ακριβώς αυτή την αναλογία αλλά η εμπειρία δείχνει πως έχουμε λιγότερους φοιτητές μαθηματικών παρά επιχειρήσεων. Ας υποθέσουμε πως η αναλογία είναι $1:10$. Πλέον, έχουμε δύο ειδών πληροφορίες, την αναλογία και την ντροπαλότητα και θέλουμε να τα συνδυάσουμε για να έχουμε μια πιο ολοκληρωμένη εικόνα για τον φοιτητή.

Με όρους πιθανότητας, αυτό που μας ενδιαφέρει στην πραγματικότητα είναι η δεσμευμένη πιθανότητα $\mathbb{P}(\mathrm{M}|\mathrm{N})$.



Διάγραμμα 2.4.1: Η εκ των προτέρων γνώση σε συνδυασμό με το δεδομένο.

Ξεκινώντας με την εκ των προτέρων γνώση για την αναλογία μπορούμε να φανταστούμε ένα δείγμα 10 ατόμων από το τμήμα M και αντίστοιχα ένα δείγμα 100 ατόμων από το τμήμα E. Στο τετράγωνο το τμήμα M έχει ένα μικρό κομμάτι του και το τμήμα E έχει το μεγάλο. Μετά έχουμε την πληροφορία που παρατηρούμε ότι ο φοιτητής είναι ντροπαλός. Μπορούμε να κάνουμε μια εκτίμηση ότι το 70% των μαθηματικών ταιριάζουν με αυτό το χαρακτηριστικό. Για το τμήμα E εκτιμούμε πως το 10% ταιριάζουν με αυτό το χαρακτηριστικό. Από το δείγμα μας, θα είχαμε 7 άτομα από το τμήμα M και 10 από το τμήμα E.

[1] Σκέψη κλασικής προσέγγισης, η πιθανοφάνεια. Θα μπορούσε κανείς να πει πως αυτό είναι ίσως παράδοξο καθώς χρησιμοποιείται προηγούμενη γνώση. Πράγματι, ακόμα και η κλασική συμπερασματολογία χρησιμοποιεί κάποιες προηγούμενες γνώσεις, στην κατασκευή κατάλληλου μοντέλου πιθανοφάνειας

Οπότε:

$$\mathbb{P}(\mathrm{M}|\mathrm{N}) = \frac{\mathbb{P}(\mathrm{M}\cap\mathrm{N})}{\mathbb{P}(\mathrm{N})} = \frac{\mathbb{P}(\mathrm{N}|\mathrm{M})\cdot\mathbb{P}(\mathrm{M})}{\mathbb{P}(\mathrm{N})} = \frac{\frac{7}{10}\cdot\frac{1}{11}}{\frac{7}{110}+\frac{10}{110}} = \frac{7}{17} \approx 41\%$$

Άρα, ακόμα και αν θεωρούμε ότι τα άτομα στο τμήμα Μ ταιριάζουν πολύ περισσότερο στην περιγραφή του φοιτητή, δεν είναι αρκετό να ξεπεράσει την αναλογία που ξέρουμε ότι ισχύει. Αυτή είναι και η βασική ιδέα του θεωρήματος, δηλαδή, νέα στοιχεία δεν μας προκαθορίζουν από μόνα τους το αποτέλεσμα, πρέπει να μας ανανεώσει την εκ των προτέρων γνώμη μας!

2.4.2 Σύγχρονη Θεωρία κατά Bayes

Σε πραγματικά προβλήματα, αυτό που μας ενδιαφέρει από την Μπεϋζιανή προσέγγιση είναι η αντιμετώπιση της άγνωστης παραμέτρου θ ως τυχαία μεταβλητή, ο καθορισμός της εκ των προτέρων κατανομής για το θ και η χρήση του θεωρήματος για τον «εκσυγχρονισμό» της εκ των προτέρων πεποίθησης μας σε εκ των υστέρων πιθανότητες. Το παραπάνω, εκπροσωπεί τις νέες μας απόψεις για την κατανομή του θ (την νέα αβεβαιότητα μιας άγνωστης ποσότητας) για την εξαγωγή κατάλληλης συμπερασματολογίας.

Οπότε, αντικαθιστώντας τώρα το ενδεχόμενο Β με τα δεδομένα (παρατηρήσεις) $\boldsymbol{x}$, το ενδεχόμενο Α με ένα σύνολο παραμέτρων $\boldsymbol{\theta}$ και τις πιθανότητες με κατανομές, προκύπτει το θεώρημα του Bayes όπως χρησιμοποείται στην σύγχρονη Μπεϋζιανή στατιστική

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{m(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\ , \tag{2.4.3}$$

όπου $p(\boldsymbol{\theta})$ είναι η κατανομή των παραμέτρων $\boldsymbol{\theta}$ πριν παρατηρήσουμε τα δεδομένα $\boldsymbol{x}$ (εκ των προτέρων ή a-priori κατανομή), $f(\boldsymbol{x}|\boldsymbol{\theta})$ η συνάρτηση πιθανοφάνειας των δεδομένων $\boldsymbol{x}$, $p(\boldsymbol{\theta}|\boldsymbol{x})$ η από κοινού εκ των υστέρων κατανομή των παραμέτρων $\boldsymbol{\theta}$ (a-posteriori) δοθέντων των δεδομένων $\boldsymbol{x}$ και $m(\boldsymbol{x})$ η περιθώρια πιθανοφάνεια των δεδομένων $\boldsymbol{x}$. Η αντίστροφη ποσότητα της $m(\boldsymbol{x})$ ονομάζεται και σταθερά κανονικοποίησης της εκ των υστέρων κατανομής, $p(\boldsymbol{\theta}|\boldsymbol{x})$. Ο υπολογισμός της είναι αρκετά δύσκολος, ακόμα και αδύνατος όταν η διάσταση του προβλήματος είναι σημαντικά αυξημένη. Αφαιρώντας την περιθώρια κατανομή των δεδομένων από την σχέση (2.4.3), η σχέση της ισότητας αντικαθίσταται από τη σχέση αναλογίας.

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.4.4}$$

Έτσι, μπορούμε να πούμε ότι η εκ των υστέρων κατανομή του $\boldsymbol{\theta}$ είναι ανάλογη του γινομένου της συνάρτησης πιθανοφάνειας με την εκ των προτέρων κατανομή. Προκύπτει, συνδυάζοντας τα δεδομένα $\boldsymbol{x}$ που παρατηρήσαμε και την προσωπική μας άποψη. Η ανάλογη εκ των υστέρων κατανομή θα έχει

την ίδια συναρτησιακή μορφή με την εκ των υστέρων κατανομή απλώς το ολοκλήρωμα της δεν θα αθροίζει στην μονάδα διότι λείπει η σταθερά κανονικοποίησης $m(\boldsymbol{x})$. Ένας άλλος τρόπος για να δούμε την σχέση (2.4.3) ή (2.4.4) είναι να σκεφτούμε την a-priori ως «βάρη» που βάζουμε σε κάθε σημείο της πιθανοφάνειας.

Είναι σημαντικό να σημειωθεί πως η σχέση (2.4.3) αναφέρεται σε συνεχή παράμετρο $\boldsymbol{\theta}$. Η αντίστοιχη περίπτωση της διακριτής περίπτωσης αλλάζει μόνο την ποσότητα

$$m(\boldsymbol{x}) = \sum_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

2.4.3 Συνάρτηση Πιθανοφάνειας

Για να μπορούμε να καταλάβουμε την φιλοσοφία της, θα πρέπει αρχικά να καταλάβουμε τη συνάρτηση πυκνότητας $f(x;\theta)$. Θεωρούμε X μια τυχαία μεταβλητή (τ.μ). Γνωρίζουμε πως οι τιμές μιας τ.μ προέρχονται από ένα μηχανισμό ή νόμο. Αυτός ο μηχανισμός, περιγράφεται μέσω μιας μαθηματικής συνάρτησης η οποία περιγράφει την κατανομή της τ.μ και ονομάζετε συνάρτηση πυκνότητας. Η συνάρτηση πυκνότητας όμως εξαρτάται από την παράμετρο θ όπου «καθορίζει» τις τιμές που γεννάει ο μηχανισμός. Στην συνάρτηση πυκνότητας έχουμε τις τιμές της τ.μ X να «τρέχουν» (δηλαδή, μεταβλητές) και το θ είναι σταθερό.

Τώρα, σύμφωνα με την κλασική προσέγγιση, η φύση μας έχει δώσει μια παράμετρο θ, όπου με βάση κάποιο νόμο ή μηχανισμό «γεννιούνται» δεδομένα. Άμα γνωρίζαμε το θ τότε θα είχαμε την κατανομή της X (έτσι δεν θα είχαμε ανάγκη από μοντέλα). Στην Στατική αντιμετωπίζουμε το ανάποδο πρόβλημα. Έχουμε στα χέρια μας ένα δείγμα και προσπαθούμε μέσα από αυτό να υποθέσουμε τη φύση των δεδομένων και με βάση αυτό να πούμε κάτι για το θ.

Αυτό θα γίνει με την βοήθεια της συνάρτησης πιθανοφάνειας όπου στην ουσία είναι η συνάρτηση πυκνότητας διαβασμένη ανάποδα. Θεωρώντας το θ μεταβλητή μπορούμε με την βοήθεια των παρατηρήσεων (σταθερό) να αποτιμήσουμε στον μηχανισμό $f(x;\theta)$ (που υποθέτουμε για την φύση των δεδομένων) διάφορες τιμές του θ και βάσει των δεδομένων η $f(x;\theta)$ θα δίνει διάφορες τιμές πιθανοφάνειας. Αυτές οι τιμές πιθανοφάνειας μας λένε την πυκνότητα ή πόσο πιθανό είναι το κάθε θ να μου γέννησε τις παρατηρήσεις.

Θα μπορούσε κανείς να πει πως αυτό είναι εργαλείο κλασικής προσέγγισης και εκεί θεωρούν το θ σταθερό από την φύση όπως ειπώθηκε και στον ορισμό της συνάρτησης πυκνότητας. Ο τρόπος με

τον οποίο ορίζεται η συνάρτηση πιθανοφάνειας είναι με μεταβλητό θ οπότε αυτή η λογική συνδέεται με την Μπεϋζιανη προσέγγιση. Η κλασική συμπερασματολογία χρησιμοποιεί την συνάρτηση πιθανοφάνειας με σκοπό την μεγιστοποίηση του μόνο για να έχουμε την καλύτερη δυνατή εκτίμηση της σταθερής παραμέτρου από τα δεδομένα.

Όποτε, η πιθανοφάνεια $f(\boldsymbol{x}|\theta)$ εμπεριέχει όλη τη διαθέσιμη πληροφορία που μας δίνεται από το δείγμα μας. Συνήθως συμβολίζεται ως $\mathcal{L}(\theta|\boldsymbol{x})$ και δίνεται από την παρακάτω σχέση

$$\mathcal{L}(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2.4.5}$$

Όπου n το πλήθος παρατηρήσεων, x_i η $i-$οστή παρατήρηση, με τις παρατηρήσεις να είναι ένα τυχαίο δείγμα, δηλαδή, ανεξάρτητες και ισόνομες παρατηρήσεις (i.i.d).

Στην Στατιστική, εργαζόμαστε γενικά με τον λογάριθμο της πιθανοφάνειας καθώς είναι πολύ ευκολότερο στις πράξεις, στην υπολογιστική δύναμη και δίνει ακριβώς τα ίδια αποτελέσματα (ως προς το θ) επειδή ο φυσικός λογάριθμος είναι αύξουσα συνάρτηση.

$$l(\theta|x) = logf(x|\theta) = \sum_{i=1}^{n} logf(x_i|\theta) \tag{2.4.6}$$

Μια τελευταία και βασική έννοια σε αυτή την ενότητα αφορά την αρχή της πιθανοφάνειας Berger, James O et al (1988).

Έστω δύο διαφορετικά μοντέλα που χρησιμοποιούμε για τη μοντελοποίηση ενός συνόλου παρατηρήσεων $\boldsymbol{x} = (x_1, ..., x_n)^\top$.

Μοντέλο 1: με πιθανοφάνεια $\mathcal{L}_1(\theta|\boldsymbol{x})$

Μοντέλο 2: με πιθανοφάνεια $\mathcal{L}_2(\theta|\boldsymbol{x})$

Αν $\mathcal{L}_1(\theta|\boldsymbol{x}) = c\mathcal{L}_2(\theta|\boldsymbol{x})$ τότε τα δύο μοντέλα θα πρέπει να οδηγούν στα ίδια συμπεράσματα για το θ. Στον ορισμό θα μπορούσαμε να είχαμε και δύο διαφορετικά δείγματα $\boldsymbol{x}$ και $\boldsymbol{y}$ και αν ικανοποιείται η συνθήκη $\mathcal{L}_1(\theta|\boldsymbol{x}) = c\mathcal{L}_2(\theta|\boldsymbol{y})$ τότε θα πρέπει να έχουμε την ίδια συμπερασματολογία για το θ.

2.4.4 Εκ των προτέρων κατανομή

Σε πολλά προβλήματα που μας ενδιαφέρει η συμπερασματολογία για μια άγνωστη παράμετρο θ μπορεί να υπάρχει διαθέσιμη εκ των προτέρων πληροφορία, $p(\theta)$, αντιπροσωπεύει «τις πεποιθήσεις» μας για την κατανομή του θ (την αβεβαιότητα μιας άγνωστης ποσότητας) προτού αποκτήσουμε οποιαδήποτε πληροφορία για τα δεδομένα μας, ίσως από προηγούμενη εμπειρία ή έρευνα. Το θεώρημα του

Bayes «ακούει» και τις δύο πηγές πληροφορίας και τις συνθέτει μαζί. Η ισχύς κάθε πληροφορίας υποδεικνύεται από την στενότητα κάθε συνάρτησης, δηλαδή όσο πιο στενή, η συνάρτηση αποκλείει περισσότερες παραμέτρους (περισσότερη βεβαιότητα για συγκεκριμένες παραμέτρους) και έτσι εκπροσωπεί μια ισχυρή εκ των προτέρων πληροφορία. Πρέπει να σημειωθεί πως στην πράξη είναι σημαντική η επιλογή της εκ των προτέρων κατανομής να γίνει βάσει ειλικρινείς γνήσιας πληροφορίας, όχι προσωπικής προκατάληψης. Συνεπώς, η επιλογή της χρήζει ιδιαίτερης προσοχής[2]. Η απόφαση μιας πραγματικά πληροφοριακής a-priori δεν γίνεται με στατιστικούς αλλά με ειδικός στον τομέα εφαρμογής. Η διαδικασία αυτή ονομάζεται «elicitation» και είναι η διαμόρφωση μιας εκ των προτέρων κατανομής όπου περιλαμβάνει έναν διάλογο μεταξύ του ειδικού και ενός στατιστικού. Ο Anthony O'Hagan (2004) συζητά για την αξία μιας πληροφοριακής εκ των προτέρων κατανομής.

2.4.5 Μη-πληροφοριακές εκ των προτέρων κατανομές

Σε πολλούς, αυτή η ιδέα του να εκφράσουμε τις παραμέτρους με πιθανότητες ήταν κομψή. Η μόνη αντίρρηση, ήταν η εκ των προτέρων κατανομή καθώς τα αποτελέσματα μιας έρευνας θα εξαρτυόντουσαν σημαντικά από την επιλογή της εκ των προτέρων κατανομής. Με σκοπό την ικανοποίηση αυτών που ήταν ενάντιας της εκ των προτέρων κατανομής, επινοήθηκαν οι μη-πληροφοριακές εκ των προτέρων κατανομές. Οι μη-πληροφοριακές κατανομές, δημιουργήθηκαν για εκείνους τους ερευνητές που ήθελαν τα αποτελέσματα τους να επηρεάζονται σε σχεδόν μηδενικό βαθμό από τις πεποιθήσεις τους και κυρίως από τα δεδομένα, κρατώντας όμως την Μπεϋζιανή φιλοσοφία για τις παραμέτρους. Ωστόσο, θα μπορούσαν να χρησιμοποιηθούν ακόμα και σε περιπτώσεις που δεν έχουμε καμία εκ των προτέρων πεποίθηση για την παράμετρο ή όταν απλώς δεν υπάρχει εκ των προτέρων πληροφορία. Συνοψίζοντας, υπάρχουν περιπτώσεις όπου είτε δεν υπάρχει εκ των πληροφορία είτε δεν επιθυμούμε να την ενσωματώσουμε.

Η ιδέα των μη-πληροφορικών κατανομών είναι ακριβώς η αντίθετη των πληροφοριακών κατανομών που συζητήθηκε στην Ενότητα §2.4.4. Στην περίπτωση μας όσο πιο απλωμένη η συνάρτηση, αποκλείονται λιγότερες παράμετροι (μικρότερη βεβαιότητα για τις παραμέτρους) έτσι εκπροσωπείται μια άγνοια στην εκ των προτέρων πληροφορία. Με άλλα λόγια, προσπαθούν να τοποθετήσουν το ίδιο βάρος σε όλες τις πιθανές τιμές παραμέτρων. Για παράδειγμα, μια μη-πληροφοριακή ή ασαφής εκ των προτέρων κατανομή θα μπορούσε να θεωρηθεί η κανονική κατανομή με πολύ μεγάλη διακύμανση (το τι θεωρείται μεγάλη διακύμανση εξαρτάται προφανώς από το πρόβλημα) ή την $\mathcal{U}(0,1)$. Η ομοιόμορφη κατανομή είναι η ιδανική περίπτωση μιας «πραγματικά» μη-πληροφοριακής κατανομής καθώς

[2] Γενικά, η επιρροή της a-priori γίνεται ολοένα και μικρότερη καθώς προστίθενται νέα δεδομένα. Σε τέτοιες περιπτώσεις ο «λάθος» προσδιορισμός της έχει μικρή σημασία. Ωστόσο αυτό δεν ισχύει πάντα καθώς θα μπορούσαμε εκ των προτέρων να αποκλείσουμε υποσύνολα του Θ.

δίνει το ίδιο βάρος σε όλες τις παραμέτρους αλλά είναι περιορισμένη σε ένα φραγμένο σύνολο Θ.

Ωστόσο, είναι δυνατόν στη θέση της εκ των προτέρων κατανομής να χρησιμοποιήσουμε οποιαδήποτε μη αρνητική συνάρτηση $p(\theta)$ για την οποία ισχύει ότι $\int_\Theta p(\theta)\, d\theta = \infty$. Για παράδειγμα, σε πρόβλημα όπου $\theta \in \Theta = \mathbb{R}$ μπορούμε (όχι πάντα) να χρησιμοποιήσουμε την $p(\theta) = 1$ (σε όλο το $\mathbb{R}$ δίνουμε το ίδιο βάρος σε όλα τα θ) της ομοιόμορφης αλλά δεν θα είναι πυκνότητα καθώς το ολοκλήρωμα αποκλίνει. Τέτοιες εκ των προτέρων κατανομές ονομάζονται καταχρηστικές κατανομές (Improper Priors) και αφορά μη αρνητικές συναρτήσεις p ορισμένες στο σύνολο Θ για τις οποίες το ολοκλήρωμα τους στο Θ αποκλίνει. Η χρήση καταχρηστικών εκ των προτέρων κατανομών μπορεί να δικαιολογηθεί μόνο όταν η εκ των υστέρων κατανομή μπορεί να οριστεί καλά (κάτι που δεν ισχύει πάντα).

Η αντίστοιχη εκδοχή της "μη-πληροφορικής εκ των προτέρων κατανομής" με $p(\theta) = 1$ είναι η λεγόμενη Fiducial συμπερασματολογία Pedersen, J. G. (1978). Αυτή η συμπερασματολογία ήταν η προσπάθεια του Ronald Fisher να κανονικοποιήσει την συνάρτηση πιθανοφάνειας για να πετύχει την ίδια φιλοσοφία με την Μπεϋζιανή ανάλυση αλλά μόνο βάσει των δεδομένων. Αυτή η προσέγγιση από τότε δεν έγινε ποτέ ευρέως αποδεκτή καθώς ούτε ο ίδιος ο Fisher δεν την είχε κατανοήσει πλήρως.

2.4.6 Συζυγείς Οικογένειες

Η σταθερά κανονικοποίησης $m(\boldsymbol{x})$ είναι από τις σημαντικότερες ποσότητες για την a-posteriori κατανομή, καθώς χωρίς αυτή δεν είναι δυνατή η συμπερασματολογία. Μάλιστα, η ποσότητα $m(\boldsymbol{x})$ ήταν και ένας από τους κυρίους λόγους που η Μπεϋζιανη Στατιστική δεν είχε πρακτική αξία, καθώς ο υπολογισμός της ακόμα και στην απλούστερη περίπτωση της (την μονοδιάστατη περίπτωση) ήταν επίπονη ή ακόμα και αδύνατη. Θα πούμε περισσότερα για την $m(\boldsymbol{x})$ σε επόμενα κεφάλαια αναλυτικότερα αλλά η βασική ιδέα που πρέπει να πάρουμε από εδώ είναι ότι ο υπολογισμός της σταθεράς κανονικοποίησης σε κλειστή μορφή είναι πολύ δύσκολη έως και αδύνατη.

Για παράδειγμα: Αν θεωρήσουμε ένα τυχαίο δείγμα $x_1, ..., x_n \sim \mathcal{P}oisson(\theta)$ και a-priori κατανομή $\theta \sim \mathcal{U}(0, 1)$ τότε η a-posteriori κατανομή $p(\theta|x)$.

$$p(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)p(\theta) \propto e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}$$

Χρειαζόμαστε την σταθερά κανονικοποίησης $m(\boldsymbol{x})$ για να έχουμε την a-posteriori κατανομή.

$$m(\boldsymbol{x}) = \int_0^1 f(\boldsymbol{x}|\theta)p(\theta)\,d\theta = \int_0^1 e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\,d\theta$$

Το ολοκλήρωμα αυτό είναι μια ημιτελής $\mathcal{G}amma$ συνάρτηση και δεν υπάρχει λύση κλειστής μορφής αλλά μπορεί να λυθεί αριθμητικά.

Ο σκοπός μας εδώ είναι να αναρωτηθούμε αν υπάρχει κάποιος τρόπος που μπορούμε να αποφύγουμε τον άμεσο υπολογισμό του ολοκληρώματος. Σε κάθε στατιστικό πρόβλημα, ο ορισμός του ίδιου του προβλήματος είναι αυτό που καθορίζει τον τύπο δεδομένων. Η κατανόηση του ερευνητή για τον μηχανισμό που γέννησε τα δεδομένα είναι αυτό που καθορίζει την συνάρτηση πιθανοφάνειας των δεδομένων. Αυτό σημαίνει ότι ο μόνος τρόπος για να αποφύγουμε τον υπολογισμού του $m(\boldsymbol{x})$ εξαρτάται από την επιλογή της εκ των προτέρων κατανομής. Η ιδέα είναι η εξής: Αν μπορέσουμε να επιλέξουμε μια κατάλληλη εκ των προτέρων κατανομή τέτοια ώστε η πράξη της (2.4.4) να μας δώσει μια ανάλογη εκ των υστέρων κατανομή όπου είναι δυνατή η αναγνώριση της συναρτησιακής μορφής της με μια γνωστή κατανομή τότε έχουμε αποφύγει τον άμεσο υπολογισμό της $m(\boldsymbol{x})$.

Τέτοιες εκ των προτέρων κατανομές υπάρχουν και ονομάζονται συζυγής εκ των προτέρων κατανομές και συζητήθηκαν εκτενώς από τους Raiffa. H, Schlaifer. R (1961). Σύμφωνα με τον Daniel Fink (1997) είναι σημαντικό η επιλογή μιας συζυγής εκ των προτέρων κατανομής να μην γίνει μόνο βάσει μαθηματικής διευκόλυνσης αλλά να γίνει και βάσει των πραγματικών μας πεποιθήσεων πριν δούμε τα δεδομένα. Ωστόσο, ένα μάλλον φιλοσοφικό πρόβλημα που θα μπορούσε να διαπιστώσει κανείς είναι ότι η επιλογή μιας συζυγής εκ των προτέρων κατανομής φαίνεται να γίνεται αφού έχουμε δει τα δεδομένα. Μια πιθανή αιτία για αυτή την διαπίστωση θα μπορούσε να οφείλεται στην πιθανοφάνεια, καθώς η επιλογή γίνεται δοθείσης της πιθανοφάνειας και αυτή έχει άμεση σχέση με τα δεδομένα. Στην πραγματικότητα, οι συζυγείς εκ των προτέρων κατανομές δεν έρχονται σε αντιπαράθεση με την φιλοσοφία των εκ των προτέρων κατανομών. Όπως είχε αναφερθεί και προηγουμένως, δεν είναι ότι βλέπουμε τα δεδομένα αλλά υποθέτουμε την μορφή των δεδομένων μέσω της πιθανοφάνειας. Με άλλα λόγια, αυτή απλά αντλεί πληροφορία από την οικογένεια κατανομών που υποθέτουμε για τον πληθυσμό, χωρίς όμως αυτό να σημαίνει ότι κοιτάμε τα δεδομένα για να την καθορίσουμε. Χρησιμοποιούμε απλώς, τη μορφή της πιθανοφάνειας (ως συνάρτηση).

Οπότε, μπορούμε να δώσουμε πλέον τον επίσημο ορισμό τους:
Μια οικογένεια κατανομών $\mathcal{F}$ με στήριγμα Θ λέμε ότι είναι συζυγής για τη συνάρτηση πιθανοφάνειας $f(x|\theta)$ αν για κάθε εκ των προτέρων κατανομή $p \in \mathcal{F}$, η αντίστοιχη εκ των υστέρων κατανομή ανήκει επίσης στην $\mathcal{F}$.

Παρατηρούμε ότι ο ορισμός αυτός είναι πολύ γενικός καθώς κάτι τέτοιο ισχύει πάντα για την οικογένεια όλων των κατανομών. Οι περιπτώσεις που παρουσιάζουν το περισσότερο ενδιαφέρον επικεντρώνονται σε μια συγκεκριμένη οικογένεια κατανομών, την εκθετική οικογένεια.

2.4.7 Μπεϋζιανοί έλεγχοι υποθέσεων

Στον επιστημονικό κόσμο, ο κάθε ισχυρισμός για ένα φαινόμενο ή μια κατάσταση, θεωρία ή υποψία θα πρέπει να υποστηρίζεται από τα δεδομένα. Αυτή η διαδικασία είναι θεμελιώδης για την επιστήμη καθώς υπάρχει διαφορά μεταξύ θεωρίας και εφαρμογής. Μια θεωρία, προσπαθεί να εξηγήσει το «πως» και «γιατί» μέσα από δομές ιδεών που εξηγούν και ερμηνεύουν γεγονότα. Στην προσπάθεια ερμηνείας ενός γεγονότος, γίνεται πολύ συχνά η χρήση υποθέσεων όπου, συνήθως μέσω της μαθηματικής γλώσσας, χτίζουμε πάνω σε αυτές τις υποθέσεις για να καταλήξουμε σε κάποιο συμπέρασμα. Από την άλλη, η εφαρμογή είναι απλώς οι παρατηρήσεις. Αυτό που κάνει την θεωρία, μια «επιστημονική θεωρία» είναι η δυνατότητα πρακτικής απόδειξης της θεωρίας μέσω δεδομένων (εφαρμογής της θεωρίας στις εμπειρικές επιστήμες), όπου θεωρείται αποδεκτή μέχρι να διαψευστεί.

Οι περισσότερες επιστήμες ελέγχουν τις υποθέσεις τους μέσω στατιστικών τεστ, μέθοδος κλασικής προσέγγισης. Το πιο δημοφιλές μέτρο για τα στατιστικά τεστ είναι το p-value το οποίο είναι εύκολο να χρησιμοποιηθεί, αλλά ελάχιστοι το κατανοούν και το χρησιμοποιούν σωστά. Το μέτρο αυτό έχει υποστεί αυστηρή κριτική από πολλούς ερευνητές και εδώ θα συζητήσουμε δύο από αυτούς τους λόγους. Ένας από τους λόγους είναι η διχοτομική απόφαση, δηλαδή, είτε θα απορριφθεί η μηδενική υπόθεση είτε όχι με βάση κάποιο επίπεδο σημαντικότητας α. Όμως, αυτή η προσέγγιση δεν μας δίνει ένα μέτρο για να αξιολογήσουμε τον βαθμό ενδείξεων, δηλαδή, αν δεν έχουμε ενδείξεις να είμαστε υπέρ (ή κατά), λίγες, μέτριες ή πολλές ενδείξεις. Ο άλλος λόγος αφορά την αρχή της πιθανοφάνειας όπου θα συζητηθεί αργότερα στο τέλος αυτής της ενότητας.

Ο τρόπος με τον οποίο θα καταφέρουμε να έχουμε μια τέτοια ερμηνεία είναι με την βοήθεια του παράγοντα Bayes. Ο παράγοντας Bayes, είχε αρχικά αναπτυχθεί από τον Harold Sir Jeffreys (1935) όπου στην συνέχεια απέκτησε περισσότερη προσοχή μετά την δημοσίευση των Kass. Robert E, Raftery. Adrian E (1995). Προτού τον ορίσουμε, θα στήσουμε πρώτα την σκηνή όπου θα μας οδηγήσει με φυ-

σικό τρόπο σε αυτόν.

Στην επιστημονική κοινότητα, η ανάπτυξη μιας υπόθεσης για ένα γεγονός θα πρέπει να στηρίζεται σε όρους λογικής και φύσης κατά κάποιον τρόπο. Τέτοιες υποθέσεις περιέχουν εμπειρικό ή και θεωρητικό χαρακτήρα και έχουν στόχο να εκπροσωπούν τον κόσμο και να προβλέψουν τα δεδομένα. Για παράδειγμα, σε απλά προβλήματα για έλεγχο υποθέσεων, θα μπορούσε να μας ενδιαφέρει αν ένα νόμισμα είναι δίκαιο, δηλαδή, την υπόθεση $H_0 : \theta = 0.5$ (απλή υπόθεση). Αυτή η υπόθεση είναι ακριβής επειδή γνωρίζουμε και θεωρητικά αλλά και εμπειρικά πως λειτουργούν τα δίκαια νομίσματα. Παρ'όλο αυτά, στα περισσότερα προβλήματα δεν έχουμε την δυνατότητα να γνωρίζουμε ακριβώς μέσω θεωρίας τι να εξετάσουμε για μια παράμετρο. Οπότε, ένας πιο φυσικός τρόπος για να ορίσει κανείς μια υπόθεση είναι να δώσει ένα διάστημα τιμών της παραμέτρου (σύνθετη υπόθεση), βάση εμπειρικού χαρακτήρα.

Η αξιολόγηση των εμπειρικών υποθέσεων γίνεται με τα δεδομένα. Αλλά τι ακριβώς μπορεί να θεωρηθεί αποδεικτικό στοιχείο για μια υπόθεση; Αρχικά όπως έχουμε πει και από την Ενότητα §2.4 τα δεδομένα από μόνα τους δεν εκπροσωπούν αποδεικτικά στοιχεία αλλά θα πρέπει να επιδρούν στις πεποιθήσεις μας. Επειδή η κατανόηση για την φύση μας είναι ατελής μπορούμε να εκφράσουμε την αβεβαιότητα μας με την γλώσσα πιθανοτήτων. Με αυτόν τον τρόπο οι εμπειρικές μας υποθέσεις αποκτούν και έναν βαθμό βεβαιότητας ή αβεβαιότητας για κάθε τιμή που μπορεί να πάρει η παράμετρος.

Πως ακριβώς χρησιμοποιούνται τα δεδομένα προς την επιρροή των υποθέσεων μας; Μια λογική ιδέα είναι να θεωρήσουμε μόνο την σχέση ανάμεσα της υπόθεσης και των δεδομένων. Στην ουσία εδώ λέμε πως μας ενδιαφέρει απλώς το πόσο καλά η υπόθεση μας προσαρμόζεται στα δεδομένα ανεξαρτήτως το πόσο καλά θα μπορούσαν να προσαρμοστούν άλλες υποθέσεις. Για παράδειγμα: Έστω πως έχουμε έναν υποψήφιο για θέση εργασίας. Με βάση αυτή την προσέγγιση θα αξιολογήσουμε τον υποψήφιο ξεχωριστά, χωρίς να μας ενδιέφεραν οι προηγούμενοι υποψήφιοι μας. Θα είχαμε δηλαδή μια εμπειρική υπόθεση για αυτόν και με βάση το βιογραφικό του, θα κρίναμε ξεχωριστά αν ταιριάζει στην δουλειά. Διαπιστώνουμε όμως, πως στην πραγματικότητα δεν μας ενδιαφέρει μια απόλυτη αξιολόγηση καθώς δεν έχουμε κάποιο τρόπο να κρίνουμε την ικανότητα του. Καταλήγουμε λοιπόν, σε μια άλλη ιδέα όπου μας ενδιαφέρει το πόσο καλά προσαρμόζονται δύο ή περισσότερες ανταγωνιστικές υποθέσεις στα δεδομένα και έτσι αντί να πούμε: "Αυτή η υπόθεση υποστηρίζεται", λέμε: "Ποια υπόθεση κάνει την καλύτερη δουλειά σε σχέση με τις άλλες; Αυτή η ιδέα είναι πιο ευέλικτη καθώς σχετίζεται με τον τρόπο σκέψης μας. Συνοπτικά, υποστηρίξαμε πως τα δεδομένα είναι ενδείξεις όταν επηρεάζουν τις πεποιθήσεις μας για μια υπόθεση και ότι αυτή η επιρροή είναι ευκολότερα κατανοητή

στο πλαίσιο σύγκρισης με άλλες υποθέσεις.

Ας δούμε τώρα πως θα εκφράσουμε αυτές τις έννοιες με μαθηματικά. Ξεκινάμε με την ιδέα του λόγου των εκ των προτέρων κατανομών (prior odds) που περιγράφουν τον βαθμό στον οποίο ευνοούμε μια υπόθεση έναντι μιας άλλης πριν δούμε τα δεδομένα

$$\textit{Prior Odds} = \frac{p(H_1)}{p(H_0)} \tag{2.4.7}$$

Μπορούμε επίσης να περιγράψουμε την ιδέα του λόγου των εκ των υστέρων κατανομών (posterior odds) που περιγράφουν τον βαθμό στον οποίο ευνοούμε μια υπόθεση έναντι μιας άλλης μετά την παρατήρηση των δεδομένων

$$\textit{Posterior Odds} = \frac{p(H_1|x)}{p(H_0|x)} \tag{2.4.8}$$

Το ερώτημα τώρα είναι πως θα μετακινηθούμε από τα prior odds στα posterior odds με τον σωστό τρόπο. Το κλειδί προφανώς είναι το θεώρημα του Bayes. Ο κανόνας λέει πως ο βαθμός ευνόησης για τις υποθέσεις θα πρέπει να αλλάξει από τα δεδομένα με έναν συγκεκριμένο τρόπο.

$$\frac{p(H_1|\boldsymbol{x})}{p(H_0|\boldsymbol{x})} = \underbrace{\frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)}}_{\text{Bayes Factor}} \cdot \frac{p(H_1)}{p(H_0)} \tag{2.4.9}$$

Ο μεσαίος όρος της σχέσης (2.4.9) ονομάζεται ο παράγοντας Bayes (Bayes Factor) και είναι ο όρος με τον οποίο πολλαπλασιάζουμε τα prior odds για να γίνουν posterior odds. Εκπροσωπεί ακριβώς αυτό που αναζητάμε, δηλαδή, το πως πρέπει να αλλάξουν οι σχετικές πεποιθήσεις μας υπό το φως των δεδομένων (ενδείξεις), ο παράγοντας με τον οποίο θα πρέπει να αλλάξουν οι πεποιθήσεις.

Πως ακριβώς υπολογίζουμε τον παράγοντα Bayes; Πρέπει να προσδιορίσουμε το πόσο πιθανό είναι να παρατηρήσουμε τα δεδομένα δοθείσης μιας υπόθεσης. Γνωρίζουμε πως εδώ εννοούμε να κάνουμε χρήση της πιθανοφάνειας, αλλά δεν την ορίζουμε με τον ίδιο τρόπο όπως στους έλεγχους υποθέσεων της κλασικής προσέγγισης, καθώς εκεί την μεγιστοποιούσαμε. Στη δική μας περίπτωση, η μεγιστοποίηση δεν έχει νόημα καθώς το θ είναι τ.μ. Θέλουμε να λάβουμε υπόψη όλη την πληροφορία της πιθανοφάνειας για κάθε $\theta \in \Theta$. Επειδή οι πεποιθήσεις μας παίζουν σημαντικό ρόλο στην πιθανότητα να παρατηρήσουμε τα δεδομένα δοθείσης μιας υπόθεσης για το θ τότε κάθε σημείο της πιθανοφάνειας θα σταθμιστεί με την εκ των προτέρων κατανομή της θ μας με στήριγμα τον παραμετρικό χώρο

Θ. Τέλος, αθροίζουμε όλες τις δυνατές περιπτώσεις του Θ με την βοήθεια του ολοκληρώματος και παίρνουμε την απάντηση που αναζητάμε. Αναλυτικά ο υπολογισμός γίνεται ως εξής

$$f(\boldsymbol{x}|H_i) = \int_{\Theta} f(\boldsymbol{x}|\Theta, \mathrm{H}_i) p(\Theta|H_i)\ d\Theta \tag{2.4.10}$$

Γενικά συμβολίζουμε τον παράγοντα Bayes με $B_{ij} = \dfrac{f(\boldsymbol{x}|H_i)}{f(\boldsymbol{x}|H_j)}$, υπέρ της H_i και κατά της H_j. Για παράδειγμα, μπορούμε να έχουμε B_{10} υπέρ της H_1 και κατά H_0 ή B_{01} υπέρ της H_0 και κατά της H_1. Για λόγους οικειότητας με την κλασική προσέγγιση η ερμηνεία θα γίνει βάσει το B_{10}.
Η ερμηνεία του μέτρου είναι το επόμενο βασικό βήμα. Ο Harold Sir Jeffreys (1961) πρότεινε να ερμηνεύσουμε το B_{10} ως εξής

Πίνακας 2.1: Ερμηνεία B_{10} κατά τον Jeffrey.

$log_{10}(B_{10})$	B_{10}	Στοιχεία ενάντια H_0
$0 - 0.5$	$1 - 3.2$	Δεν αξίζει περισσότερο από μια απλή αναφορά
$0.5 - 1$	$3.2 - 10$	Επαρκείς
$1 - 2$	$10 - 100$	Ισχυρά
> 2	> 100	Καθοριστικά

Στο άρθρο τους, ο Kass. Robert E, Raftery. Adrian E (1995) προτείνουν έναν εναλλακτικό τρόπο, ο οποίος συνηθίζεται περισσότερο, για την ερμηνεία του B_{10} ως εξής

Πίνακας 2.2: Ερμηνεία B_{10} κατά τον Kass και Raftery.

$2log_e(B_{10})$	B_{10}	Στοιχεία ενάντια H_0
$0 - 2$	$1 - 3$	Δεν αξίζει περισσότερο από μια απλή αναφορά
$2 - 6$	$3 - 20$	Επαρκείς
$6 - 10$	$20 - 150$	Ισχυρά
> 10	> 150	Καθοριστικά

Τέλος, θα γίνει αναφορά σε κάποια θετικά και αρνητικά του παράγοντα του Bayes.

- Αρνητικά:
  - Τα αποτελέσματα του B_{10} εξαρτώνται σημαντικά από την επιλογή της εκ των προτέρων κατανομής.
  - Υπολογιστικά ακριβός υπολογισμός καθώς η αναλυτική του λύση είναι πολύ δύσκολη έως και αδύνατη στις περισσότερες περιπτώσεις.
- Θετικά:
  - Εκφράζουμε τα στοιχεία με την φυσική γλώσσα της πιθανότητας.

- Σε σύγκριση με την κλασσική προσέγγιση, οι υποθέσεις δεν χρειάζονται να είναι εμφωλευμένες.
- Ακολουθεί την αρχή πιθανοφάνειας σε σύγκριση με τα p-values.

Από την Ενότητα §2.4.3 είδαμε τον ορισμό της αρχής της πιθανοφάνειας. Τώρα θα δούμε με ένα παράδειγμα το λόγο που το p-value δεν ακολουθεί την αρχή της πιθανοφάνειας.

Έστω ένα πείραμα Bernoulli με πιθανότητα επιτυχίας $\theta \in \Theta = (0,1)$. Μας ενδιαφέρει ο έλεγχος υπόθεσης $H_0 : \theta = \frac{1}{2}$ και $H_1 : \theta > \frac{1}{2}$ με βάση ενός συνόλου ανεξάρτητων παρατηρήσεων που αποτελούνται από 9 επιτυχίες και 3 αποτυχίες.

Παρατηρούμε πως η εκφώνηση αυτού του παραδείγματος είναι αρκετά γενική και δεν μας καθορίζει από πριν τι συμβαίνει ακριβώς. Άρα, ας αντιμετωπίσουμε αυτό το πρόβλημα με δύο λογικές προσεγγίσεις:

- Θεωρούμε ότι τα δεδομένα είναι παρατηρήσεις από μια διωνυμική κατανομή με ένα προκαθορισμένο σύνολο 12 ανεξαρτήτων δοκιμών (δικιά μας υπόθεση καθώς η εκφώνηση δεν μας το καθορίζει).
- Ένας άλλος τρόπος είναι να θεωρήσουμε ότι τα δεδομένα του πειράματος εκτελέστηκαν μέχρι να έχω τρεις αποτυχίες (θα μπορούσε δηλαδή ο συνολικός αριθμός δοκιμών να είναι μεγαλύτερο του 12)

Αντιμετωπίζουμε αρχικά το πρόβλημα βάση της πρώτης προσέγγισης:

$$X \sim \mathcal{B}in(12, \theta)$$

$$\mathcal{L}_1(\theta|\mathbf{x}) = \binom{12}{9} \theta^9 (1-\theta)^3 \ , \ \theta \in \Theta = (0,1)$$

$$\mathbb{P}(X > 9|H_0) = \sum_{x=10}^{12} \mathcal{L}_1(\theta|\mathbf{x}) = \sum_{x=10}^{12} \binom{12}{x} 0.5^x 0.5^{12-x} = 0.073$$

Σε αυτή την περίπτωση αν $\alpha = 0.05$ τότε δεν απορρίπτουμε την H_0.

Τέλος, αντιμετωπίζουμε αυτό το πρόβλημα βάση της δεύτερης προσέγγισης:

$$X \sim \mathcal{NB}(3, 1-\theta)$$

$$\mathcal{L}_2(\theta|\mathbf{x}) = \binom{11}{9} \theta^9(1-\theta)^3 \ , \ \theta \in \Theta = (0,1)$$

$$\mathbb{P}(X > 9|H_0) = \sum_{x=10}^{\infty} \mathcal{L}_2(\theta|\mathbf{x}) = \sum_{x=10}^{\infty} \binom{x+2}{x} 0.5^x 0.5^3 \approx 0.0327$$

Σε αυτή την περίπτωση όμως με $\alpha = 0.05$ απορρίπτουμε την H_0

Οι δύο πιθανοφάνειες είναι ανάλογες όποτε θα έπρεπε να βγάζουν το ίδιο συμπέρασμα αλλά δεν συμβαίνει. Άρα, το p-value δεν συνάδει με την αρχή αυτή. Ο λόγος που συμβαίνει κάτι τέτοιο είναι επειδή, όπως είχε αναφερθεί στην Ενότητα §2.2, βασιζόμαστε σε μη παρατηρηθέντα δεδομένα.

2.5 Συμπεράσματα Κεφαλαίου

Σε αυτό το Κεφάλαιο συζητήσαμε εκτενώς τις διαφορές μεταξύ Μπεϋζιανής και κλασικής προσέγγισης. Πιο συγκεκριμένα, είδαμε τις διαφορές τους στον ορισμό πιθανότητας καθώς και στην αντίστοιχη συμπερασματολογία τους. Εν συνέχεια είδαμε λεπτομερώς την φιλοσοφία του κανόνα του Bayes και τα επιμέρους κομμάτια της, από την πιθανοφάνεια μέχρι και την έννοια της εκ των προτέρων κατανομής μάζι με τις διαφορετικές εκδοχές της. Τέλος, αναλύσαμε τους Μπεϋζιανούς ελέγχους υποθέσεων και καταλήξαμε με φυσικό τρόπο στο παράγοντα Bayes μέσω ενός παραδείγματος.

Κεφάλαιο 3

Συμπερασματολογία κανονικών γραμμικών μοντέλων

Η βασική ιδέα των στατιστικών μοντέλων είναι να παρέχει ένα μαθηματικό δομημένο πλαίσιο με σκοπό την κατανόηση της φύσης μας. Αυτό συνήθως επιτυγχάνεται με την βοήθεια σχέσεων και μοτίβων που υπάρχουν στον κόσμο μας. Η πιο γνωστή μεθοδολογία είναι αυτή της κλασικής προσέγγισης η οποία θα συζητηθεί περιληπτικά παρακάτω. Από την άλλη, υπάρχει και η Μπεϋζιανή προσέγγιση στατιστικής μοντελοποίησης όπου θα συζητήσουμε εκτενώς στην συνέχεια.

3.1 Το κλασσικό κανονικό γραμμικό μοντέλο

Έστω $\{Y_i\}_{i=1}^n$ ένα σύνολο n ανεξάρτητων και ισόνομων τ.μ, $\{y_i\}_{i=1}^n$ οι ανεξάρτητες παρατηρήσεις, όπου θα την ονομάσουμε απαντητική μεταβλητή και $\{x_{ij}\}_{j=1}^q$ η $j-$οστή επεξηγηματική μεταβλητή για το i. Θεωρούμε $\boldsymbol{y} = (y_1, ..., y_n)^\top$ ένα $n \times 1$ διάνυσμα και $\mathbf{X} = (1, x_{i2}, ..., x_{iq})_{i=1}^n$ ένας $n \times q$ πίνακας σχεδιασμού με $q = p + 1$ το οποίο περιέχει όλη την πληροφορία που θα μπορούσε να σχετίζεται με την απαντητική μεταβλητή $\boldsymbol{y}$. Το κλασικό γραμμικό μοντέλο μπορεί να γραφεί ως εξής

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{όπου } \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \tag{3.1.1}$$

Το διάνυσμα $\boldsymbol{\beta} = (\beta_0, ..., \beta_q)^\top$, $q \times 1$ διαστάσεων είναι οι παράμετροι του μοντέλου. Μια παράμετρος, περιγράφει και ποσοτικοποιεί την σχέση μιας επεξηγηματικής με την απαντητική μεταβλητή, με άλλα λόγια μας λέει πόσο η επεξηγηματική μεταβλητή επηρεάζει την απαντητική. Για παράδειγμα, θα μπορούσε να μας ενδιαφέρει πόσο τα κέρδη μιας εταιρείας επηρεάζονται από τις πωλήσεις της. Ο όρος $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$ είναι ένα διάνυσμα διαστάσεων $n \times 1$ και είναι ο στοχαστικός όρος του μοντέλου

ή διαφορετικά το σφάλμα. Το σφάλμα, περιγράφει την τυχαιότητα που υπάρχει στην φύση μας και θεωρούμε την κατανομή της μια κανονική κατανομή με μέσο μηδέν και σταθερή διακύμανση.

Επειδή έχουμε υποθέσει κανονικά σφάλματα, αυτή η ιδιότητα περνάει και στα $\boldsymbol{y}$ οπότε υποθέτουμε ότι τα δεδομένα μας προέρχονται από $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ όπου στην ουσία μας λέει πως το γραμμικό μας υπόδειγμα μοντελοποιεί τον μέσο μιας δεσμευμένης κανονικής κατανομής. Από αυτή την ιδιότητα προκύπτει άμεσα ο στόχος της συμπερασματολογίας για τον πληθυσμό μέσω ενός δείγματος. Θέλουμε να βρούμε την πραγματική σχέση που συνδέει γραμμικά τις μεταβλητές μας όταν έχουμε ενδείξεις ότι το μοντέλο μας δεν απορρίπτει τις υποθέσεις που έχουμε κάνει.

Γιατί ακριβώς κάναμε αυτές τις υποθέσεις; Επειδή ο σκοπός μας είναι η συμπερασματολογία και επειδή υποθέσαμε την γραμμική σχέση μεταξύ των μεταβλητών τότε προκύπτει άμεσα πως η κατανομή που θα μας επιτρέψει κάτι τέτοιο είναι η δεσμευμένη κανονική. Η γραμμικότητα και η ομοσκεδαστικότητα (σταθερή διακύμανση) προκύπτει άμεσα από τις ιδιότητες της δεσμευμένης κανονικής κατανομής. Ο λόγος που υποθέσαμε την αναμενόμενη τιμή των σφαλμάτων με μηδέν είναι καθαρά για διαισθητικούς λόγους καθώς αναμένουμε γενικά η τυχαιότητα να μην επηρεάζει κάποια δομή που θα μπορούσε να υπήρχε στα δεδομένα. Αν παρατηρηθεί κάποια συστηματική συμπεριφορά των καταλοίπων του μοντέλου τότε μπορούμε απλά να την αφαιρέσουμε. Τέλος, υποθέσαμε επίσης και ανεξαρτησία των δεδομένων. Παρ' όλο που ακούγεται λογικό υπάρχει και μια εξήγηση πίσω από αυτό. Πριν το κανονικό γραμμικό μοντέλο, υπάρχει και το στατιστικό γραμμικό μοντέλο το οποίο δεν υποθέτει κανονικότητα και έτσι δεν μπορεί να γίνει συμπερασματολογία. Ένα από τα πράγματα που υποθέτει είναι ασυσχέτιστα σφάλματα καθώς υποτίθεται την γραμμική δομή την συλλαμβάνει το γραμμικό μοντέλο. Στην δική μας περίπτωση τα ασυσχέτιστα σφάλματα θα γίνουν τελείως ανεξάρτητα μεταξύ τους λόγου της ιδιότητας της κανονικής κατανομής. Οπότε αν παρατηρηθεί οποιαδήποτε μορφή εξάρτησης στα κατάλοιπα θα πρέπει να ληφθεί υπόψη και να διορθωθεί. Συνοψίζοντας, οι υποθέσεις του κανονικού γραμμικού μοντέλου είναι συνέπειες της δεσμευμένης κανονικής κατανομής και οποιαδήποτε παραβίαση τους καθιστά το μοντέλο αναξιόπιστο.

Προηγουμένως, αναφέρθηκε ο στόχος του γραμμικού μοντέλου και γενικά οποιουδήποτε μοντέλου. Προφανώς, επειδή στην διάθεση μας έχουμε δεδομένα δεν θα μπορέσουμε να ξέρουμε ποτέ την αλήθεια για την φύση των δεδομένων αλλά τουλάχιστον μας βοηθάνε στην καλύτερη κατανόηση της φύσης μας. Ο George Box το 1976 συνοψίζει τα μοντέλα ως: *"Όλα τα μοντέλα είναι λάθος, αλλά κάποια είναι χρήσιμα"*.

Αυτό που μπορούμε να κάνουμε για τις παραμέτρους του μοντέλου μας είναι να τα εκτιμήσουμε βάση των δεδομένων για να προσεγγίσουμε όσο μπορούμε την αλήθεια. Ο πιο συνηθισμένος τρόπος για το πρόβλημα της εκτίμησης είναι με την μεγιστοποίηση της πιθανοφάνειας, συνάρτηση του $\boldsymbol{y}$ δοθέντος τα β και σ^2 το οποίο ορίζεται ως εξής:

$$\prod_{i=1}^{n} f(Y_i = y_i|\mathbf{X}, \beta, \sigma^2) = f_Y(\boldsymbol{y}|\mathbf{X}, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\Big\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\Big\} \quad (3.1.2)$$

Η μεγιστοποίηση του (3.1.2) για αναλυτική ή και υπολογιστική ευκολία γίνεται με την βοήθεια του φυσικού λογαρίθμου το οποίο δεν επηρεάζει τα αποτελέσματα επειδή είναι αύξουσα συνάρτηση. Έτσι, η πιθανοφάνεια (3.1.2) γίνεται

$$log f_Y(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \quad (3.1.3)$$

Αποδεικνύεται ότι η εξίσωση (3.1.3) μεγιστοποιείται στα σημεία

Μερική παραγώγιση ως προς β

$$\frac{\partial ln f_Y(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{2\sigma^2} 2\mathbf{X}^\top \boldsymbol{y} - 2\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

Μερική παραγώγιση ως προς σ^2

$$\frac{\partial ln f_Y(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{2(\hat{\sigma}^2)^2} = 0$$

$$= \hat{\sigma}^2 n + (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})}{n}$$

Επειδή έχουμε εκτιμήσει το β μπορούμε να βρούμε την δειγματοληπτική κατανομή του $\hat{\beta}$.

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})]$$

$$= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}] = \boldsymbol{\beta}$$

$$Var[\hat{\boldsymbol{\beta}}] = Var[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Var[\boldsymbol{y}]((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top$$
$$= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\underbrace{\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}}_{\mathbf{I}} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

Με βάση αυτά η δειγματοληπτική κατανομή του $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_q(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$.

3.2 Το Μπεϋζιανό κανονικό γραμμικό μοντέλο

Από την προηγούμενη Ενότητα §3.1 καταλήξαμε πως γενικότερα τα μοντέλα και συγκεκριμένα για τους σκοπούς μας, τα γραμμικά μοντέλα είναι στατιστικές μέθοδοι τα οποία τα χρησιμοποιούμε για την κατανόηση σχέσεων μεταξύ μεταβλητών. Σε αυτή την Ενότητα θα στρέψουμε την προσοχή μας στην μοντελοποίηση γραμμικού υποδείγματος αλλά από την σκοπιά της Μπεϋζιανής φιλοσοφίας.

Για λόγους απλότητας θα θεωρήσουμε ένα απλό γραμμικό υπόδειγμα. Για την καλύτερη κατανόηση της έννοιας, θα πάρουμε και το παράδειγμα της Ενότητας §3.1 όπου θεωρούμε απαντητική μεταβλητή ($\mathbf{Y}$) τα κέρδη επιχειρήσεων και επεξηγηματική μεταβλητή ($\boldsymbol{x}$) τις πωλήσεις επιχειρήσεων. Ο τρόπος με τον οποίο εκφράστηκε η σχέση μεταξύ μεταβλητών ή επιρροή της $\boldsymbol{x}$ στην $\mathbf{Y}$ ήταν μέσου των παραμέτρων του μοντέλου. Η κλασική προσέγγιση, όπως γνωρίζουμε, απλά θα εκτιμούσε αυτές τις παραμέτρους βάσει των δεδομένων. Από την Μπεϋζιανή σκοπιά προφανώς θα θεωρήσουμε τις παραμέτρους του μοντέλου αβέβαιες ποσότητες. Από το παράδειγμα μας, μια λογική σκέψη είναι να περιμένουμε τα κέρδη να επηρεάζονται από τις πωλήσεις πριν καν δούμε τα δεδομένα με την αντίστοιχη αβεβαιότητα μας γύρω από αυτή την επιρροή. Όπως γνωρίζουμε, αυτήν την αβεβαιότητα για την παράμετρο, μπορούμε να την μαθηματικοποιήσουμε ορίζοντας μια εκ των προτέρων κατανομή για τα $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. Όταν παρατηρήσουμε πλέον τα δεδομένα $\boldsymbol{x}$ και $\boldsymbol{y}$ απλά θα αλλάξουμε τις πεποιθήσεις μας για την επιρροή της $\boldsymbol{x}$ στο $\boldsymbol{y}$[1] δίνοντας μας την εκ των υστέρων κατανομή των παραμέτρων $\boldsymbol{\beta}$.

Χρησιμοποιώντας την θεωρία πιθανοτήτων για τις παραμέτρους του μοντέλου έχουμε ως άμεση συνέπεια την αβεβαιότητα στα μοντέλα μας. Οπότε, αντί για μια εκτίμηση της $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x]$ που θα παίρναμε από την κλασσική προσέγγιση, πλέον θα έχουμε πολλές εκτιμήσεις της $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x]$ από τα πολλά μοντέλα με τις αντίστοιχες αβεβαιότητες που υπάρχουν. Έτσι, αυτό που παίρνουμε στην πραγματικότητα είναι απλώς μια εκ των υστέρων κατανομή της $\mathbf{Y}|\mathbf{X} = x$.

[1] Όπως και στην κλασική προσέγγιση, σε ένα γραμμικό υπόδειγμα, θεωρούμε τα $\mathbf{Y}$ τ.μ και τα $\boldsymbol{x}$ είναι σταθερά πριν δούμε τα δεδομένα. Τα $\mathbf{Y}$ θα γίνουν $\boldsymbol{y}$ όταν θα έχουμε πραγματοποίηση τ.μ.

3.2.1 Συζυγής εκ των προτέρων κατανομή

Κατ' αρχήν, γνωρίζουμε πως μπορούμε να χρησιμοποιήσουμε οποιαδήποτε εκ των προτέρων κατανομή της επιλογής μας. Παρ' όλο αυτά, όπως θα συζητηθεί και αργότερα, τα αποτελέσματα της εκ των υστέρων κατανομής δεν θα είναι σε κλειστή μορφή. Οπότε, θα χρησιμοποιήσουμε την έννοια της συζυγής εκ των προτέρων κατανομής που συζητήσαμε στην Ενότητα §2.4.6 λόγου των μαθηματικών ιδιοτήτων που προσφέρει.

Από τη μορφή της πιθανοφάνειας (3.1.2) υποδεικνύεται η επιλογή της οικογένειας κανονικής - αντίστροφης Γάμμα ως εκ των προτέρων κατανομή των παραμέτρων $(\boldsymbol{\beta}, \sigma^2)$.

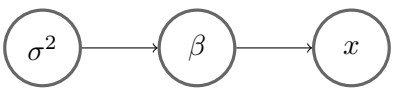
$$(\boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}_q(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}\sigma^2)\mathcal{IG}(a, b) \tag{3.2.1}$$

Προφανώς, η επιλογή της κανονικής κατανομής ως εκ των προτέρων κατανομή για τα $\boldsymbol{\beta}$ είναι λογική. Από την άλλη, η επιλογή της αντίστροφης Γάμμα για το σ^2 μπορεί να μην είναι τόσο. Ο λόγος για μια τέτοια επιλογή κρύβεται στην συνάρτηση πυκνότητας της αντίστροφης Γάμμα. Η συνάρτηση πυκνότητας της αντίστροφης Γάμμα φέρει την μορφή της συνάρτησης πυκνότητας της κανονικής επειδή έχει το σ^2 στον παρανομαστή.

Η συνάρτηση πυκνότητας πιθανότητας της (3.2.1) γράφεται ως

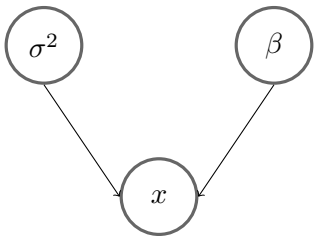
$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) \tag{3.2.2}$$

Παρατηρούμε από την εξίσωση (3.2.2) ότι μαθαίνουμε τα $\boldsymbol{\beta}$ και την διακύμανση σ^2 δεσμεύοντας την εκ των προτέρων κατανομή στο $\boldsymbol{\beta}$ ως προς το σ^2. Τέτοιου είδους εκ των προτέρων κατανομών ονομάζονται ιεραρχικά και περισσότερες πληροφορίες σχετικά με αυτά μπορούν να βρεθούν στο βιβλίο των Gelman. A, Carlin. J.B, et al. (2013). Συνοπτικά όμως, σε προβλήματα όπου εμπλέκονται πολλές παράμετροι είναι λογικό να υποθέσει κανείς πως μπορούν να θεωρηθούν ως συνδεδεμένες μεταξύ τους ανάλογα με την φύση του προβλήματος. Εδώ, υπονοείται ότι η από κοινού κατανομή των παραμέτρων θα πρέπει να εκπροσωπεί την εξάρτηση που υπάρχει. Στην περίπτωση μας, αυτό που λέμε είναι ότι το σ^2 είναι άγνωστη, οπότε θα έχει την δική της εκ των προτέρων κατανομή $p(\sigma^2)$. Έτσι, αυτό που θα πάρουμε για τα $\boldsymbol{\beta}$ θα εξαρτηθεί αρχικά από αυτό που θα πάρουμε από την $p(\sigma^2)$.



Διάγραμμα 3.2.1: Το ιεραρχικό μοντέλο.

Ανάλογα με το πρόβλημα, θα μπορούσε κανείς να μην υποθέσει τέτοια ιεραρχία αλλά να υποθέσει ανεξαρτησία δηλαδή, $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2)$. Ωστόσο, υπάρχει πρόβλημα καθώς αποδεικνύεται πως δεν έχουμε συζυγή εκ των προτέρων κατανομή. Το ολοκλήρωμα της εκ των υστέρων κατανομής δεν μπορεί να υπολογιστεί αναλυτικά οπότε αναγκαζόμαστε σε αυτή την περίπτωση να καταφύγουμε σε μεθόδους Markov Chain Monte Carlo.



Διάγραμμα 3.2.2: Το ανεξάρτητο μοντέλο.

Η εξίσωση (3.2.2) γράφεται ως

$$p(\boldsymbol{\beta}, \sigma^2) = \frac{b^a}{(2\pi)^{q/2}\Gamma(a)|V|^{1/2}(\sigma^2)^{a+1+q/2}} exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^\top V^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta) - \frac{b}{\sigma^2}\right\} \mathrm{I}_{\mathbb{R}^q \times (0,\infty)}(\boldsymbol{\beta}, \sigma^2) \tag{3.2.3}$$

Η εξίσωση (3.2.3) ονομάζεται η εκ των προτέρων κατανομή του Zellner με τον πίνακα $V = g(\mathrm{X}^\top X)^{-1}$ και $g > 0$ όπου συνήθως $g = n$. Η g-prior του Zellner παίρνει στην ουσία πληροφορία από πίνακα πληροφορίας $(\mathrm{X}^\top X)^{-1}$ και εκπροσωπεί μια αντικειμενική εκ των προτέρων κατανομή. Επιπλέον, βοηθά στο πρόβλημα πολυσυγγραμμικότητας καθώς στα διαγώνια στοιχεία του πίνακα V προστίθενται κάποιοι επιπλέον όροι που βοηθούν στη πιο εύκολη αντιστροφή του πίνακα $(\mathrm{X}^\top X)^{-1}$. Είναι σημαντικό να παρατηρήσουμε, πως ο πίνακας V δεν χρησιμοποιεί πραγματικά τα δεδομένα. Αυτό συμβαίνει, από την υπόθεση του σταθερού πίνακα σχεδιασμού X από το γραμμικό μοντέλο.

3.2.2 Εκ των υστέρων κατανομή

Από τις εξισώσεις (3.1.2) και (3.2.3) η εκ των υστέρων κατανομή των παραμέτρων γράφεται ως

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \mathbf{X}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})p(\boldsymbol{\beta}, \sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{1+(2a+q+n)/2}} exp\left\{-\frac{b}{\sigma^2}\right\} \\ &\times exp\left\{-\frac{1}{2\sigma^2}[(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top V^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})]\right\} \end{aligned}$$

Κάνουμε πράξεις και συμπληρώνουμε την τετραγωνική μορφή στον όρο του δεύτερου εκθετικού

$$\begin{aligned} &(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}) \\ &\qquad = \boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{y}^\top\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{X}^\top\boldsymbol{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

$$
\begin{aligned}
&+ \boldsymbol{\beta}^\top \mathbf{V}^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\beta} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta \\
&= \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{V}^{-1}\boldsymbol{\beta} - 2\boldsymbol{y}^\top \mathbf{X}\boldsymbol{\beta} \\
&- 2\boldsymbol{\beta}^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta \\
&= \boldsymbol{\beta}^\top \underbrace{(\mathbf{X}^\top \mathbf{X} + \mathbf{V}^{-1})}_{\boldsymbol{\Psi}^{-1}} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \underbrace{(\mathbf{X}^\top \boldsymbol{y} + \mathbf{V}^{-1}\boldsymbol{\mu}_\beta)}_{\mathbf{w}} \\
&- \underbrace{(\mathbf{X}^\top \boldsymbol{y} + \mathbf{V}^{-1}\boldsymbol{\mu}_\beta)^\top}_{\mathbf{w}^\top} \underbrace{(\mathbf{X}^\top \mathbf{X} + \mathbf{V}^{-1})^{-1}}_{\boldsymbol{\Psi}} \underbrace{(\mathbf{X}^\top \boldsymbol{y} + \mathbf{V}^{-1}\boldsymbol{\mu}_\beta)}_{\mathbf{w}} \\
&+ \boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta \\
&= \boldsymbol{\beta}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{w} - \mathbf{w}^\top \boldsymbol{\Psi}\mathbf{w} + \boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta \\
&= (\boldsymbol{\beta} - \boldsymbol{\Psi}\mathbf{w})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta} - \boldsymbol{\Psi}\mathbf{w}) - \mathbf{w}^\top \boldsymbol{\Psi}\mathbf{w} + \boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta
\end{aligned}
$$

Όπου $\boldsymbol{\Psi} = (\mathbf{X}^\top \mathbf{X} + \mathbf{V}^{-1})^{-1}$ και $\mathbf{w} = \mathbf{X}^\top \boldsymbol{y} + \mathbf{V}^{-1}\boldsymbol{\mu}_\beta$. Αντικαθιστώντας αυτά που βρήκαμε και χωρίζοντας τους όρους με το $\boldsymbol{\beta}$ και τους όρους χωρίς το $\boldsymbol{\beta}$, έχουμε

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \mathbf{X}) &\propto \frac{1}{(\sigma^2)^{1+(2a+q+n)/2}} \times exp\Bigg\{ -\frac{1}{2\sigma^2}\Bigg[b + \frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta - \mathbf{w}^\top \boldsymbol{\Psi}\mathbf{w}) \Bigg] \Bigg\} \\
&\times exp\Bigg\{ -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\Psi}\mathbf{w})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta} - \boldsymbol{\Psi}\mathbf{w}) \Bigg\} \\
&\propto \frac{1}{(\sigma^2)^{1+(2a+n)/2}} exp\Bigg\{ -\frac{1}{\sigma^2}\Bigg[b + \frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta - \mathbf{w}^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}\mathbf{w}) \Bigg] \Bigg\} \\
&\times \frac{1}{(\sigma^2)^{q/2}} exp\Bigg\{ -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\Psi}\mathbf{w})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta} - \boldsymbol{\Psi}\mathbf{w}) \Bigg\} \\
&\propto \frac{1}{(\sigma^2)^{\tilde{a}+1}} exp\Bigg\{ -\frac{\tilde{b}}{\sigma^2} \Bigg\} \times \frac{1}{(\sigma^2)^{q/2}} exp\Bigg\{ -\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \Bigg\} \qquad (3.2.4)
\end{aligned}
$$

Όπου $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Psi}\mathbf{w}$, $\tilde{b} = b + \frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{y} + \boldsymbol{\mu}_\beta^\top \mathbf{V}^{-1}\boldsymbol{\mu}_\beta - \tilde{\boldsymbol{\beta}}^\top \boldsymbol{\Psi}^{-1}\tilde{\boldsymbol{\beta}})$ και $\tilde{a} = a + n/2$. Επιπλέον, παρατηρούμε πως οι δύο παράγοντες της εξίσωσης (3.2.4) αναγνωρίζονται ως οι ανάλογες πυκνότητες της αντίστροφης Γάμμα και κανονικής κατανομής αντίστοιχα. Οπότε η από κοινού εκ των υστέρων κατανομή των $(\boldsymbol{\beta}, \sigma^2)$ είναι

$$
\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \mathbf{X} \sim \mathcal{N}_q(\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{\Psi}) \mathcal{IG}(\tilde{a}, \tilde{b}) \qquad (3.2.5)
$$

Ως προς τις αντίστοιχες περιθωριακές κατανομές έχουμε τα εξής αποτελέσματα

Για την περιθώρια κατανομή του σ^2 προκύπτει άμεσα από την εξίσωση (3.2.5) ότι

$$
\sigma^2 | \boldsymbol{y}, \mathbf{X} \sim \mathcal{IG}(\tilde{a}, \tilde{b})
$$

Αυτό το αποτέλεσμα είναι αναμενόμενο καθώς και η εκ των προτέρων κατανομή ήταν ορισμένη ανεξάρτητα.

Για την περιθώρια κατανομή του $\boldsymbol{\beta}$ δεν είναι τόσο προφανές καθώς υπάρχει εξάρτηση από το σ^2. Οπότε, για τη εύρεση της θα πρέπει να ολοκληρώσουμε την από κοινού εκ των υστέρων κατανομή ως προς το σ^2

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X}) &= \int_0^\infty p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\mathbf{X})\,d\sigma^2 \\
&\propto \int_0^\infty \frac{1}{(\sigma^2)^{1+(2\tilde{a}+q)/2}} exp\Big\{-\frac{\tilde{b}}{\sigma^2}\Big\} exp\Big\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big\}\,d\sigma^2 \\
&\propto \int_0^\infty \frac{1}{(\sigma^2)^{1+(2\tilde{a}+q)/2}} exp\Big\{-\frac{1}{\sigma^2}\Big[\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big]\Big\}\,d\sigma^2 \\
&\propto \frac{\Gamma\Big(\frac{2\tilde{a}+q}{2}\Big)}{\Big[\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big]^{(2\tilde{a}+q)/2}} \\
&\times \underbrace{\int_0^\infty \frac{\Big[\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big]^{(2\tilde{a}+q)/2}}{\Gamma\Big(\frac{2\tilde{a}+q}{2}\Big)} exp\Big\{-\frac{1}{\sigma^2}\Big[\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big]\Big\}\,d\sigma^2}_{\text{σ.π.π } \mathcal{IG}(\frac{2\tilde{a}+q}{2},\,\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}))\,=\,1} \\
&\propto \Gamma\Big(\frac{2\tilde{a}+q}{2}\Big)\Big[\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big]^{-(2\tilde{a}+q)/2} \\
&\propto \Big[1+\frac{1}{2\tilde{a}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \Big(\frac{\tilde{b}}{\tilde{a}}\boldsymbol{\Psi}\Big)^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big]^{-(2\tilde{a}+q)/2} \qquad (3.2.6)
\end{aligned}
$$

Η τελευταία έκφραση (3.2.6) είναι ανάλογη της πολυμεταβλητής $\mathcal{T}$ κατανομής διάστασης q με $2\tilde{a}$ βαθμούς ελευθερίας. Οπότε, η περιθωριακή εκ των υστέρων κατανομή των $\boldsymbol{\beta}$ είναι:

$$
\boldsymbol{\beta}|\boldsymbol{y},\mathbf{X} \sim \mathcal{T}_q(\tilde{\boldsymbol{\beta}}, \frac{\tilde{b}}{\tilde{a}}\boldsymbol{\Psi}, 2\tilde{a}) \qquad (3.2.7)
$$

Το παραπάνω αποτέλεσμα είναι λογικό καθώς ακόμα και στην κλασική προσέγγιση του γραμμικού μοντέλου όταν δεν γνωρίζαμε το σ^2 κάναμε συμπερασματολογία με την κατανομή $t-$Student.

3.3 Συμπεράσματα Κεφαλαίου

Σε αυτό το Κεφάλαιο καταφέραμε να εισάγουμε την έννοια του Μπεϋζιανού γραμμικού μοντέλου. Αρχικά, υπενθυμίσαμε την θεωρία του κλασικού κανονικού γραμμικού μοντέλου για να είναι πιο ευδιάκριτες οι διαφορές μεταξύ των δύο προσεγγίσεων. Στην συνέχεια, είδαμε τι σημαίνει διαισθητικά να προσεγγίζουμε ένα γραμμικό μοντέλο μέσω της Μπεϋζιανής οπτικής και στην συνέχεια αναπτύξαμε την συζυγή ανάλυση του. Εκμεταλλευτήκαμε τις ιδιότητες της συζυγής ανάλυσης με σκοπό να αντλήσουμε την εκ των υστέρων κατανομή και τις αντίστοιχες περιθωριακές εκ των υστέρων κατανομές σε κλειστή μορφή .

Κεφάλαιο 4

Markov Chain Monte Carlo

Σε αυτό το κεφάλαιο θα συζητήσουμε για τις μεθόδους Markov chain Monte Carlo (MCMC). Οι μέθοδοι MCMC έχουν υπάρξει σχεδόν για όσο καιρό υπήρξαν και οι τεχνικές Monte-Carlo. Παρ' όλο αυτά, η επιρροή του στην Στατιστική δεν είχε αναδειχτεί ακόμα μέχρι τις αρχές της δεκαετίας του 90. Καταρχήν, ένα από τα βασικότερα ερωτήματα που καλούμαστε να απαντήσουμε είναι να κατανοήσουμε την έννοια «MCMC», καθώς περιέχει έναν συνδυασμό Μαρκοβιανών αλυσίδων και Monte-Carlo. Για αυτόν τον σκοπό, θα δούμε τις βασικές έννοιες Μαρκοβιανών αλυσίδων και Monte-Carlo, και στην συνέχεια θα προχωρήσουμε σε μεθόδους που χρησιμοποιούν MCMC. Οι μέθοδοί MCMC, στην πραγματικότητα είναι απλά αλγόριθμοι οι οποίοι εκμεταλλεύονται τις μαθηματικές ιδιότητες αυτών των δύο εννοιών. Αυτοί οι αλγόριθμοι πλέον, χρησιμοποιούνται ευρέως τεχνικά σε όλες τις περιοχές της Στατιστικής, ειδικότερα στη Μπεϋζιανή, καθώς και σε άλλες επιστήμες όπως φυσική, πληροφορική κλπ. Όπως έχει αναφερθεί σε προηγούμενα κεφάλαια, ο λόγος που χρησιμοποιείται στην Μπεϋζιανή Στατιστική οφείλεται στον αδύνατο αναλυτικό υπολογισμό της εκ των υστέρων κατανομής, ειδικά σε πολυμεταβλητά προβλήματα, λόγου της σταθεράς κανονικοποίησης. Ως εκ τούτου, οι μέθοδοί MCMC μας παρέχουν την δυνατότητα να προσομοιώσουμε τουλάχιστον δείγμα από αυτή, με αποτέλεσμα να μπορούμε να κάνουμε την συμπερασματολογία.

4.1 Έννοιες Μαρκοβιανών αλυσίδων

Οι Μαρκοβιανές αλυσίδες πήραν το όνομα τους από τον μαθηματικό Andrey Markov, ο οποίος μελέτησε εκτενώς της Μαρκοβιανές διαδικασίες στις αρχές του 20ου αιώνα και δημοσίευσε το πρώτου του άρθρο πάνω σε αυτό το θέμα το 1906.

Μία Μαρκοβιανή αλυσίδα είναι μια ακολουθία τυχαίων μεταβλητών που μπορεί να θεωρηθεί ότι εξελίσσεται χρονικά. Συγκεκριμένα, ας θεωρήσουμε μια διακριτή τ.μ X_0 την χρονική στιγμή $t = 0$. Αυτή η τ.μ προφανώς έχει μια κατανομή και για λόγους απλότητας ας θεωρήσουμε ότι έχει 4 πιθανές καταστάσεις (χώρος καταστάσεων $\mathcal{S}$[1]) με τις αντίστοιχες πιθανότητες τους. Την επόμενη χρονική στιγμή $t = 1$ η τ.μ γίνεται X_1 και μπορούμε να φανταστούμε ότι η κατανομή της τ.μ για τις 4 πιθανές καταστάσεις έχει αλλάξει. Με την ίδια λογική, για $t = 2$ έχουμε X_2 κ.ο.κ. Αυτό που πρέπει να παρατηρήσουμε εδώ είναι ότι η τ.μ X ακόμα έχει 4 καταστάσεις, δηλαδή, εκπροσωπεί το ίδιο φυσικό φαινόμενο, το μόνο που αλλάζει είναι η κατανομή του φαινομένου στον χρόνο.

Μια λογική ερώτηση εδώ είναι: *"Τι προκαλεί τέτοια συμπεριφορά;"*. Ένα παράδειγμα θα μπορούσε να είχε να κάνει με τον καιρό. Παρ' όλο που υπάρχουν πολλοί εξωτερικοί παράγοντες που τελικά επηρεάζουν την πιθανότητα βροχής, μπορούμε να κατανοήσουμε πως μια άμεση επίδραση στο αν θα βρέξει σήμερα έχει να κάνει με το αν είχε βρέξει χθες. Παρατηρούμε εδώ, πως έχουμε δύο πιθανές καταστάσεις κάθε φορά και η κατανομή αυτών των καταστάσεων αλλάζει με τον χρόνο ανάλογα με την έκβαση της προηγούμενης τ.μ. Τώρα, αν τελικά χθες είχε βρέξει, τότε η επιρροή αυτής της έκβασης στον καιρό σήμερα θα περιέχει πάντα αβεβαιότητα. Προφανώς, αυτήν την αβεβαιότητα μπορούμε να την εκφράσουμε με πιθανότητες όπου πλέον η τ.μ θα έχει για παράδειγμα, $1/2$ πιθανότητα να μην βρέξει και $1/2$ πιθανότητα να βρέξει σήμερα. Από την άλλη, αν δεν είχε βρέξει τότε θα μπορούσε να είχαμε $1/3$ πιθανότητα να βρέξει και $2/3$ να μην βρέξει. Αυτές τις πιθανότητες τις ονομάζουμε πιθανότητες μετάβασης από την μια κατάσταση σε άλλες. Αυτή η ιδέα μιας τ.μ σε μια συγκεκριμένη χρονική στιγμή να εξαρτάται μόνο από την προηγούμενη χρονική στιγμή δημιουργεί μια Μαρκοβιανή αλυσίδα. Η αλυσίδα λέγεται Μαρκοβιανή λόγου της Μαρκοβιανής ιδιότητας η οποία ορίζεται ως εξής

$$\mathbb{P}(\underbrace{X_{t+1} = j}_{\text{Μέλλον}} \mid \underbrace{X_t = i}_{\text{Παρόν}}, \underbrace{X_{t-1} = i_{t-1}, ..., X_0 = i_0}_{\text{Παρελθόν}}) = \mathbb{P}(X_{t+1} = j | X_t = i) \tag{4.1.1}$$

[1] Ο χώρος καταστάσεων $\mathcal{S}$ είναι το σύνολο των δυνατών τιμών που μπορεί να βρεθεί η αλυσίδα και μπορεί να είναι είτε διακριτό είτε συνεχές σύνολο

Η εξίσωση (4.1.1) μας λέει ότι η επόμενη κατάσταση της αλυσίδας εξαρτάται από την κατάσταση που βρισκόταν η αλυσίδα την προηγούμενη χρονική στιγμή αλλά όχι από τις προηγούμενες καταστάσεις αυτής. Μια τέτοια ιδιότητα θα μπορούσε να μην θεωρηθεί τόσο λογική αλλά για παράδειγμα σε μια επιχείρηση παλιά έσοδα και τα χρέη μπορούν να συνοψιστούν στο παρόν, δηλαδή, όλα αυτά μας οδηγούν στο αποτέλεσμα του παρόντος. Ως εκ τούτου, αυτή η ιδιότητα ίσως τελικά να μην είναι τόσο παράλογη. Επιπλέον, είναι πολύ βολική από την πλευρά υπολογιστικού κόστους αλλά και μαθηματικής άποψης[2]. Τέλος, θα υποθέσουμε ότι οι πιθανότητες μετάβασης είναι τα ίδια για κάθε $t = 0, 1, 2, ...$, δηλαδή, δεν αλλάζουν στον χρόνο. Μια τέτοια Μαρκοβιανή αλυσίδα καλείται ομογενής

$$\mathbb{P}(X_{t+1} = j|X_t = i) = \mathbb{P}(X_1 = j|X_0 = i) = p_{ij}$$

Με βάση τα παραπάνω έχουμε πλέον την δυνατότητα να εκφράσουμε τις πιθανότητες μετάβασης στο αρχικό μας παράδειγμα (διακριτό σύνολο καταστάσεων $\mathcal{S}$) με την βοήθεια πινάκων (πίνακας πιθανοτήτων μετάβασης). Αυτός ο πίνακας θα είναι πάντα τετραγωνικός με τις εξής ιδιότητες

$$\mathbb{P}(X_{t+1} = j|X_t = i) = p_{ij} \geq 0 \ , \ \sum_{j=1}^{k} p_{ij} = 1$$

Ένας τέτοιος πίνακας όπου ικανοποιεί αυτές τις ιδιότητες ονομάζεται στοχαστικός.

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

Οι γραμμές του πίνακα υποδηλώνουν το X_t και οι στήλες του πίνακα το X_{t+1}.

Εφόσον ορίσαμε τις πιθανότητες μετάβασης, θα μπορούσε κανείς να αναρωτηθεί για την πιθανότητα μετάβασης από μια κατάσταση στην άλλη μετά από δύο βήματα ή την δεύτερη χρονική στιγμή. Η απάντηση πίσω από αυτό το ερώτημα βρίσκεται στις εξισώσεις Chapman-Kolmogorov οι οποίες μπορούν να γενικεύσουν το πρόβλημα σε m βήματα.

$$\mathbb{P}(X_t = j|X_0 = i) = \sum_{k \in \mathcal{S}} \mathbb{P}(X_t = j|X_{t-1} = k)\mathbb{P}(X_{t-1} = k|X_0 = i) \tag{4.1.2}$$

Η έκφραση (4.1.2) κάνει χρήση της ολικής πιθανότητας, δηλαδή, λαμβάνονται υπόψη όλες οι δυνατές καταστάσεις ανάμεσα στην αρχική μας κατάσταση και την τελική μας κατάσταση. Με ισοδύναμο

[2]Υπάρχουν και Μαρκοβιανές αλυσίδες όπου λαμβάνουν υπόψη και προηγούμενες χρονικές περιόδους και ονομάζονται Μαρκοβιανές αλυσίδες κ-τάξης δηλαδή δεσμεύουμε κ-χρονικές περιόδους. Αλλά στην βιβλιογραφία όταν λέμε Μαρκοβιανές αλυσίδες εννοούμε αυτές της πρώτης τάξης.

τρόπο, μπορούμε να γράψουμε την εξίσωση (4.1.2) ως

$$p_{ij}^{(t)} = \sum_{k \in \mathcal{S}} p_{kj} p_{ik}^{(t-1)}$$

Διαπιστώνουμε πως το άθροισμα εκπροσωπεί γενικότερα απλώς γινόμενο πινάκων για όλες τις δυνατές καταστάσεις. Έτσι, αν θέλουμε το $p_{ij}^{(t)}$ μπορούμε να δούμε το (i, j) στοιχείο του πίνακα P^t.

Ένα σημαντικό ερώτημα που καλούμαστε να απαντήσουμε είναι η κατανομή των τ.μ στις αντίστοιχες χρονικές τους στιγμές. Η κατανομή μιας τ.μ σε κάποια χρονική t αρχικά εξαρτάται από την προηγούμενη τ.μ την χρονική στιγμή $t-1$, άρα εξαρτάται από τις πιθανότητες μετάβασης για κάθε κατάσταση που ανήκει στο $\mathcal{S}$. Οι πιθανότητες μετάβασης όμως από μόνες τους δεν αρκούν για να περιγράψουν την κατανομή της X_t καθώς πρέπει να λάβουμε υπόψη και την κατανομή της X_0 η οποία ξεκινά την αλυσίδα. Οπότε, με βάση τα παραπάνω θα πρέπει πάλι να υπολογίσουμε μια ολική πιθανότητα για να λάβουμε υπόψη όλες τις δυνατές περιπτώσεις λόγου της X_0. Συμβολίζουμε λοιπόν, την κατανομή της X_0 με $\pi^{(0)} = [\pi_1^{(0)}, ..., \pi_{\mathcal{S}}^{(0)}]^\top$ και με βάση την εξίσωση (4.1.2) έχουμε

$$\begin{aligned}\mathbb{P}(X_t = j) &= \sum_{k \in \mathcal{S}} \mathbb{P}(X_t = j | X_0 = k) \mathbb{P}(X_0 = k) \\ &= \sum_{k \in \mathcal{S}} p_{kj}^{(t)} \pi_k^{(0)} = P^t \pi^{(0)}\end{aligned}$$

Από το παραπάνω αποτέλεσμα μπορούμε να κάνουμε την σημαντικότερη ερώτηση αυτή της ενότητας: *Υπάρχει κάποια χρονική στιγμή στην αλυσίδα όπου η κατανομή της X_t να μην αλλάζει πλέον;*

$$\pi = P\pi \ \text{, με} \ \sum_{k \in \mathcal{S}} \pi_k = 1 \ \text{ και } \ \pi_k \geq 0 \tag{4.1.3}$$

Η μετάφρασης αυτής της ερώτησης με μαθηματικά είναι η εξίσωση (4.1.3). Άρα, ψάχνουμε μια κατανομή της αλυσίδας τέτοια ώστε όταν πολλαπλασιαστεί με τον πίνακα μετάβασης P η κατανομή της θα παραμείνει αμετάβλητη. Αν τελικά, υπάρχει μια τέτοια αλυσίδα και αν φτάσει σε αυτή την κατανομή τότε λέμε πως φτάσαμε την στάσιμη κατανομή.

Η ανάγκη για την έννοια της στάσιμης κατανομής ανακύπτει εξαιτίας της κατά κόρον χρήση της στις MCMC μεθόδους. Αξιοποιούμε αυτήν την ιδιότητα για να επιτύχουμε, εν τέλει, τη λήψη δείγματος από την κατανομή στόχο. Ωστόσο, σε αυτό το σημείο πρέπει να σημειώσουμε πως η Μαρκοβιανή αλυσίδα δεν θα φτάσει στην στάσιμη κατανομή όσο την αφήνουμε να «τρέχει» αλλά αν φτάσει τότε θα παραμείνει εκεί. Οπότε, δεν μπορούμε να εγγυηθούμε κάποια σύγκλιση ακόμα καθώς η αλυσίδα μπορεί να φτάσει κάποια στιγμή (σε μη ρεαλιστικό χρόνο) ή να μην φτάσει ποτέ.

Αν όμως μια Μαρκοβιανή αλυσίδα είναι αδιαχώριστη (irreducible) και απεριοδική (aperiodic) και αν μια στάσιμη κατανομη $\boldsymbol{\pi}$ υπάρχει, τότε η αλυσίδα θα συγκλίνει σε αυτή την κατανομή όσο $t \to \infty$.

Μια Μαρκοβιανή αλυσίδα καλείται αδιαχώριστη αν και μόνο αν όλες οι καταστάσεις της αλυσίδας επικοινωνούν μεταξύ τους, δηλαδή, η Μαρκοβιανή αλυσίδα περιέχει μόνο μια κλάση. Μια αδιαχώριστη Μαρκοβιανή αλυσίδα είναι σημαντική για τη σύγκλιση καθώς θέλουμε η σύγκλιση να συμβαίνει ανεξαρτήτως από την αρχική κατάσταση και να μην κολλήσει σε συγκεκριμένες καταστάσεις.

Μια Μαρκοβιανή αλυσίδα καλείται απεριοδική όταν για κάθε κατάσταση δεν υπάρχει πολλαπλάσια περιοδικότητα για να γυρίσουμε στην ίδια κατάσταση. Με άλλα λόγια, δεν υπάρχει κάποιο μοτίβο στο χρόνο που θα κάνει να γυρίσει σε κάποια κατάσταση.

Στην πράξη μια επαρκή συνθήκη η οποία χρησιμοποιείται συχνά για λόγους ευκολίας στους MCMC αλγορίθμους για την επιβεβαίωση στασιμότητας είναι η ιδιότητα της χρονικής αντιστρεψιμότητας (Reversibility).

$$\pi_i P_{ij} = \pi_j P_{ji} \text{ για κάθε } i, j \in \mathcal{S} \tag{4.1.4}$$

Η παραπάνω εξίσωση λέει πως αν ξεκινήσουμε από την κατάσταση i η οποία είναι κατανέμεται με βάση την π_i και στην συνέχεια μεταβαίνουμε στην j κατάσταση τότε θα έχουμε την πιθανότητα μετάβασης από $i \to j$. Το δεξί μέλος της εξίσωσης λέει τότε την πιθανότητα μετάβασης $j \to i$. Ποια η σχέση όμως της χρονικής αντιστρεψιμότητας με την στασιμότητα; Εάν λάβουμε υπόψη την σχέση (4.1.4) για όλα τα j, τότε έχουμε την πιθανότητα να ξεκινήσουμε από την κατάσταση i $(X_t = i)$ να είναι η ίδια με την πιθανότητα να καταλήξουμε στην κατάσταση i $(X_{t+1} = i)$. Με αυτόν τον τρόπο η κατανομή του X_{t+1} είναι η ίδια με αυτή της X_t, κάνοντας την π μια στάσιμη κατανομή.

$$\sum_j \pi_i P_{ij} = \sum_j \pi_j P_{ji} \Rightarrow \pi = \pi P$$

Το παραπάνω δίνει και μια διαίσθηση στο γιατί αυτή η ιδιότητα ονομάζεται χρονική αντιστρεψιμότητα. Η προς τα εμπρός κίνηση στην αλυσίδα θα πρέπει να μοιάζει η ίδια με την προς τα πίσω κίνηση, δηλαδή, οι πιθανότητες είναι ισορροπημένες.

Έως τώρα, έχουμε αναλύσει της Μαρκοβιανές αλυσίδες σε διακριτό χώρο καταστάσεων $\mathcal{S}$. Οι αντιστοιχίες σε συνεχή χώρο καταστάσεων Θ είναι πρακτικά ίδιας λογικής. Σε αυτή περίπτωση ο πίνακας μετάβασης πλέον θεωρείται μια δεσμευμένη συνάρτηση πυκνότητας $F(\theta_t, \theta_{t+1})$ την οποία την ονο-

μάζουμε πυρήνα μετάβασης.

$$\theta_{t+1} \sim F(\theta_t, \cdot)$$

Παρομοίως, μας ενδιαφέρει η αλυσίδα να συγκλίνει επίσης στην στάσιμη κατανομή π.

$$\theta_t \sim \pi \Rightarrow \theta_{t+1} \sim \pi$$

και η χρονική αντιστρεψιμότητα γίνεται ως

$$\pi(x)F(x,y) = \pi(y)F(y,x)$$

$$\int \pi(x)F(x,y)\,dx = \int \pi(y)F(y,x)\,dx \Rightarrow \pi(y)\int F(y,x)\,dx = \pi(y)$$

4.2 Monte-Carlo

Monte-Carlo είναι η τεχνική της εκτίμησης ιδιοτήτων μιας κατανομής μέσω τυχαίων δειγμάτων από την συγκεκριμένη κατανομή. Με άλλα λόγια, γίνεται η εκμετάλλευση της στενής σχέσης που υπάρχει μεταξύ δείγματος και συνάρτηση πυκνότητας για τον υπολογισμό ποσοτήτων. Για παράδειγμα, αντί να υπολογίσουμε αναλυτικά τον μέσο μιας γνωστής κατανομής, μπορούμε να προσεγγίσουμε το πρόβλημα μέσω Monte-Carlo παίρνοντας ένα μεγάλο δείγμα από εκείνη την κατανομή και υπολογίζοντας τον δειγματικό μέσο. Η αιτία αυτής της χρησιμότητας πηγάζει από το γεγονός ότι ο υπολογισμός ενός δειγματικού μέσου είναι πολύ πιο απλός από τον αναλυτικό υπολογισμό του ολοκληρώματος. Τα πλεονεκτήματα του γίνονται πιο έντονα όταν τα τυχαία δείγματα είναι εύκολα να παραχθούν και όταν το πρόβλημα είναι δύσκολο να λυθεί αναλυτικά. Οι μέθοδοι Monte-Carlo μπορούν να θεωρηθούν ως μια συλλογή από υπολογιστικές τεχνικές για την λύση ενός μαθηματικού προβλήματος. Οι μέθοδοι αυτοί στοχεύουν συνήθως την λύση δύο προβλημάτων.

- Monte-Carlo προσομοίωση: Την προσομοίωση ανεξάρτητων δειγμάτων από κατανομή στόχο.
- Monte-Carlo ολοκλήρωση: Τον υπολογισμό αναμενόμενων τιμών υπό αυτή την κατανομή στόχο.

$$\Phi = \mathbb{E}_f[\varphi(\theta)] = \int_\Theta \varphi(\theta) f(\theta)\,d\theta \tag{4.2.1}$$

4.2.1 Monte-Carlo ολοκλήρωση

Αν υποθέσουμε αρχικά ότι το πρόβλημα προσομοίωσης δειγμάτων είναι εύκολο, τότε μπορούμε να λύσουμε το δεύτερο πρόβλημα με μια δειγματική μέση τιμή.

$$\bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n} \varphi(\theta_i) \quad , \quad \theta_i \sim f \tag{4.2.2}$$

Η ποσότητα (4.2.2) είναι η κλασική Monte-Carlo ολοκλήρωση. Μας λέει να προσομοιώσουμε n ανεξάρτητες τιμές του θ από την συνάρτηση πυκνότητας $f(\theta)$, αντικαθιστούμε εκείνες τις τιμές στην συνάρτηση φ και παίρνουμε τον μέσο όρο εκείνων των τιμών.

$$\mathbb{E}[\bar{\varphi}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_f[\varphi(\theta_i)] = \frac{1}{n}n\mathbb{E}_f[\varphi(\theta)] = \Phi \ , \ n \to \infty \tag{4.2.3}$$

$$Var[\bar{\varphi}] = \frac{1}{n}Var_f[\varphi(\theta)] = \frac{1}{n}\underbrace{\int_\Theta (\varphi(\theta)-\Phi)^2 f(\theta)\,d\theta}_{\sigma^2} = \frac{\sigma^2}{n} \ , \ n \to \infty \tag{4.2.4}$$

Όπου σ^2 είναι σταθερά. Άρα, η εκτιμήτρια (4.2.2) είναι αμερόληπτη, με τυπικό σφάλμα ανάλογη της $\frac{1}{\sqrt{n}}$ και συνεπής. Ιδιαίτερο ενδιαφέρον παρουσιάζει η εξίσωση (4.2.4). Συγκεκριμένα, παρατηρούμε ότι όσο $n \to \infty$ το $Var[\bar{\varphi}] \to 0$. Παρ' όλο που θα φτάσουμε στην αλήθεια αν $n \to \infty$, από την πλευρά υπολογιστικού κόστους ένα τεράστιο n μπορεί να γίνει πολύ χρονοβόρο στην εκτίμηση του Φ. Παρατηρούμε επίσης, ότι το σφάλμα του εκτιμητή θα τείνει στο μηδέν στην τάξη του $\sqrt{n}$. Αυτό υποδεικνύει ότι θα χρειαστούμε ένα ακόμα μεγαλύτερο n καθώς σε συνάρτηση του n ο ρυθμός αύξησης θα είναι πιο αργός από όσο ήταν προηγουμένως. Για αυτόν τον λόγο, ιδιαίτερη σημασία παρουσιάζει ο όρος σ^2 της εξίσωσης (4.2.4). Αν το σ^2 είναι γενικά αρκετά μικρό το σφάλμα μας θα χρειαστεί μικρότερο δείγμα για να τείνει στο μηδέν σε σύγκριση με κάποιο άλλο σ^2 το οποίο είναι μεγαλύτερο, και άρα θα χρειαστεί μεγαλύτερο δείγμα για να κάνει την ίδια δουλεία με το μικρό σ^2. Οπότε, θα χρειαστεί να επιλέξουμε κατάλληλη f με σκοπό να ελαχιστοποιήσουμε το σ^2 της (4.2.4).

Δύο βασικές τεχνικές που θα μας βοηθήσουν να καταφέρουμε τα παραπάνω είναι το Importance Sampling και Control Variates.

Το Importance Sampling, στην ουσία κάνει ένα μαθηματικό τέχνασμα με σκοπό την περαιτέρω μείωση της διακύμανσης σε σύγκριση με το κλασικό Monte-Carlo. Η ιδέα ξεκινά δοθέντος μια άλλη συνάρτηση πυκνότητας, έστω $q(\theta)$, την οποία επιλέγουμε με κάποιον συγκεκριμένο τρόπο. Ο σκοπός μας, όπως πριν, είναι η εκτίμηση της ποσότητας (4.2.1) όπου μπορούμε ισοδύναμα να την γράψουμε ως

$$\Phi = \mathbb{E}_f[\varphi(\theta)] = \int_\Theta \varphi(\theta) f(\theta)\,d\theta$$

$$= \int_\Theta \left[\frac{f(\theta)}{q(\theta)}\varphi(\theta)\right] q(\theta)\,d\theta = \mathbb{E}_q\left[\frac{f(\theta)}{q(\theta)}\varphi(\theta)\right] \tag{4.2.5}$$

Αν θυμηθούμε την ιδέα του κλασικού Monte-Carlo τότε μπορούμε να εκτιμήσουμε την ποσότητα (4.2.5) αλλά με δείγματα από την συνάρτηση πυκνότητας q.

$$\bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(\theta_i)}{q(\theta_i)}\varphi(\theta_i) \quad , \quad \theta_i \sim q \tag{4.2.6}$$

Από πλευρά μαθηματικών πράξεων στην πραγματικότητα δεν έχει αλλάξει τίποτα καθώς οι ποσότητες (4.2.1) και (4.2.5) είναι ισοδύναμες.

$$\mathbb{E}[\bar{\varphi}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_q\left[\frac{f(\theta_i)}{q(\theta_i)}\varphi(\theta_i)\right] = \Phi \quad , \quad n \to \infty \tag{4.2.7}$$

$$Var[\bar{\varphi}] = \frac{1}{n}\int_{\Theta}\left[\frac{f(\theta)}{q(\theta)}\varphi(\theta) - \Phi\right]^2 q(\theta)\, d\theta = \frac{1}{n}Var_q\left[\frac{f(\theta)}{q(\theta)}\varphi(\theta)\right] \quad , \quad n \to \infty \tag{4.2.8}$$

Όπως και με το κλασικό Monte-Carlo, η νέα μας εκτιμήτρια με βάση την (4.2.7) είναι αμερόληπτη αλλά έχουμε διαφορετικό και μάλλον καλύτερο σφάλμα της εκτίμησης, την εξίσωση (4.2.8). Η ελπίδα με αυτή την τεχνική είναι να πετύχουμε το παρακάτω

$$Var_q\left[\frac{f(\theta)}{q(\theta)}\varphi(\theta)\right] < Var_f[\varphi(\theta)].$$

Το παραπάνω εξαρτάται από την επιλογή της συνάρτησης πυκνότητας της q. Για να πετύχουμε την μείωση της διακύμανσης θέλουμε η επιλογή της q να φέρει όσο το περισσότερο δυνατόν, την μορφή της συνάρτησης $\varphi(\theta)f(\theta)$. Δεδομένου ότι προσπαθούμε να εκτιμήσουμε το ολοκλήρωμα της συνάρτησης, το δείγμα θα πρέπει να παρθεί με αποτελεσματικό τρόπο ως προς την συνάρτηση $\varphi(\theta)f(\theta)$. Άρα, η ιδέα του Importance Sampling στηρίζεται στην ίδια ιδέα με του κλασικού Monte-Carlo μόνο που η επιλογή της f για το δείγμα μπορεί να μην είναι τόσο αποτελεσματική τέτοια ώστε να «καλύπτει» την $\varphi(\theta)f(\theta)$ κατάλληλα. Για περισσότερες πληροφορίες σχετικά με αυτή την τεχνική, ανατρέξτε στο άρθρο του Eric C. Anderson (1999).

Η τεχνική Control Variates είναι επίσης μια μέθοδο που εκμεταλλεύεται μαθηματικό τέχνασμα με σκοπό την ελάττωση διασποράς. Αν $\Phi = \mathbb{E}_f[\varphi(\theta)]$ είναι η ποσότητα που μας ενδιαφέρει τότε υποθέτουμε πως γνωρίζουμε την ποσότητα $\mu = \mathbb{E}_f[h(\theta)]$ όπου $\varphi(\theta) \approx h(\theta)$. Η συνάρτηση $h(\theta)$ ονομάζεται Control Variate. Θέτοντας από το κλασικό Monte Carlo $\bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n}\varphi(\theta_i)$ και $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}h(\theta_i)$ τότε μπορούμε να εκτιμήσουμε το Φ με τον εκτιμητή διαφοράς.

$$\bar{\varphi}_{diff} = \frac{1}{n}\sum_{i=1}^{n}(\varphi(\theta_i) - h(\theta_i)) + \mu = \bar{\varphi} - \bar{\mu} + \mu \tag{4.2.9}$$

Παρ' όλο που η ποσότητα (4.2.9) δεν είναι μαθηματικά ισοδύναμη με την (4.2.2) ή την (4.2.6) έχει τις εξής στατιστικές ιδιότητες

$$\mathbb{E}(\bar{\varphi}_{diff}) = \Phi - \mu + \mu = \Phi \ \ , \ \ n \to \infty \tag{4.2.10}$$

$$Var[\bar{\varphi}_{diff}] = \frac{1}{n} Var[\varphi(\theta) - h(\theta)] \ \ , \ \ n \to \infty \tag{4.2.11}$$

Άρα, ο εκτιμητής διαφοράς είναι αμερόληπτος του Φ. Το παραπάνω μας λέει ότι αντιμετωπίζουμε το ίδιο πρόβλημα με όλα τα προηγούμενα αλλά όπως και με το Importance Sampling έχουμε μια διαφορετική διακύμανση εκτιμητή, την ποσότητα (4.2.11). Ο λόγος που θέλουμε $\varphi(\theta) \approx h(\theta)$ είναι ακριβώς για την (4.2.11) διότι το $Var[\bar{\varphi}_{diff}] \to 0$. Αξίζει να σημειωθεί, πως η μέθοδο αυτή έχει άμεση σχέση με την παλινδρόμηση. Για περαιτέρω πληροφορίες σχετικά με αυτή την τεχνική ο αναγνώστης μπορεί να ανακαλέσει το βιβλίο των Art B. Owen και Peter W. Glynn (2016).

Τέλος, οι Monte-Carlo εκτιμήσεις μπορούν να γενικευτούν ανεξάρτητα από την διάσταση του προβλήματος. Αν υποθέσουμε, δηλαδή, ότι είναι εύκολη η προσομοίωση δειγμάτων από την συνάρτηση πυκνότητας τότε μπορούμε να χρησιμοποιήσουμε αυτές τις τεχνικές Monte-Carlo και σε μεγαλύτερες διαστάσεις. Παρ' όλο αυτά, πρέπει να σημειωθεί πως σε τέτοιες περιπτώσεις υπάρχουν και μειονεκτήματα που ο Mackay, D.J.C. (1998) αναλύει στο άρθρο του.

4.2.2 Monte-Carlo προσομοίωση

Μέχρι τώρα, υποθέταμε πως η προσομοίωση δειγμάτων για τις μεθόδους ολοκλήρωσης Monte-Carlo ήταν εύκολη. Ωστόσο, στην πραγματικότητα κάτι τέτοιο δεν ισχύει. Σε γενικότερες περιπτώσεις, όταν θέλουμε να προσομοιώσουμε τιμές από την $f(x)$, στην διάθεση μας θα έχουμε μόνο μια ανάλογη συνάρτηση της, την $f^*(x)$.

$$f(x) = \frac{f^*(x)}{C}$$

Η άγνοια της σταθεράς κανονικοποίησης είναι ένας από τους κύριους λόγους η προσομοίωση δειγμάτων είναι δύσκολη. Παρ' όλο αυτά, ακόμα και με τη γνώση της σταθεράς κανονικοποίησης, μια τέτοια ενέργεια αποδεικνύεται απαιτητική, ειδικά σε περισσότερες διαστάσεις.

Σε αυτή την ενότητα θα δούμε πως προσομοιώνουμε δείγματα στην απλή μονοδιάστατη περίπτωση. Συγκεκριμένα, θα μας απασχολήσουν δύο μέθοδοι, αυτό της αντιστροφής και απόρριψης.

Ενσωματωμένα μες στους επεξεργαστές μας υπάρχει ένας αλγόριθμος ο οποίος παράγει "τυχαίους" αριθμούς από την ομοιόμορφη κατανομή στο διάστημα $(0, 1)$. Οι αλγόριθμοι που παράγουν αυτούς

τους αριθμούς ονομάζονται αναγωγικές γεννήτριες και είναι της μορφής

$$x_{n+1} = (ax_n + b)\ mod\ m\ ,\ n = 0, 1, 2, ...$$

Με x_0 να είναι ένας «σπόρος» (αρχική τιμή), m μέτρο, b το βήμα (αύξηση) και a ο πολλαπλασιαστής. Όλες αυτές οι ποσότητες είναι θετικοί ακέραιοι αριθμοί και αν ζητούμε τιμές στο $(0, 1)$, τότε

$$u_n = \frac{x_n}{m} \Rightarrow u_n \in (0, 1)$$

Με βάση τα παραπάνω, παρατηρούμε πως στην πραγματικότητα αν γνωρίζουμε αυτές τις παραμέτρους τότε μπορούμε να προβλέψουμε ακριβώς τις τιμές που θα παράγει ο αλγόριθμος. Πρακτικά, αυτές οι τιμές δεν είναι γνωστές αλλά επειδή μιλάμε για ντετερμινιστικό αλγόριθμο τότε στην πραγματικότητα οι τιμές που παράγει είναι ψεύδο-τυχαίοι. Ο λόγος που τις χρησιμοποιούμε είναι επειδή συμπεριφέρονται αποδοτικά ως ”τυχαίοι” ανεξάρτητοι αριθμοί. Συνεπώς, ο μόνος «έτοιμος» αλγόριθμος είναι η αναγωγική γεννήτρια και θα πρέπει να εκμεταλλευτούμε την ομοιόμορφη κατανομή για να παράγουμε ”τυχαίους” αριθμούς από άλλες, άγνωστες, κατανομές.

Η πρώτη μας μέθοδο είναι αυτό της αντιστροφής. Η ιδέα του είναι η χρήση της αθροιστικής συνάρτησης κατανομής $F(x)$. Αν παρατηρήσουμε, το πεδίο τιμών της $F(x) \in [0, 1]$. Έτσι, αν υποθέσουμε μια καλός ορισμένη αντίστροφη αθροιστική συνάρτηση $F^{-1}(u)$ με $u \in [0, 1]$ τότε θα μπορέσουμε να πάρουμε δείγμα από την συνάρτηση πυκνότητας f.

Όλη αυτή η διαδικασία, στηρίζεται στο γεγονός ότι αν $U \sim \mathcal{U}nif(0, 1)$, τότε η $X = F^{-1}(U)$ έχει την ζητούμενη κατανομή

$$X = F^{-1}(U) \Rightarrow \mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(X) \leq u) = \mathbb{P}(U \leq F(x)) = F(x)$$

Ο αλγόριθμος ξεκινά προσομοιώνοντας δείγμα από την $U \sim \mathcal{U}nif(0, 1)$. Στην συνέχεια θέτουμε το $X = F^{-1}(U)$ με αποτέλεσμα το $X \sim f$.

Τα θετικά με αυτή την μέθοδο είναι η ευκολία της και ότι είναι 100% αποδοτικό, με την έννοια ότι δεν χάνετε κανένα δεδομένο. Από την άλλη πλευρά, για την εφαρμογή αυτής της μεθόδου χρειαζόμαστε μια καλώς ορισμένη σ.π.π, ωστόσο τις περισσότερες φορές έχουμε μια ανάλογη συνάρτηση (ειδικά σε περισσότερες διαστάσεις). Επιπλέον, η αναλυτική συνάρτηση της αντίστροφης αθροιστικής, στην πράξη δεν βρίσκεται εύκολα ή μπορεί να μην είναι εφικτό καθώς προϋποθέτει να υπάρχει η

γενικευμένη αντίστροφη συνάρτηση της $F_X(x)$.

Η επόμενη μέθοδος είναι αυτό της απόρριψης. Η ιδέα είναι να πάρουμε ένα δείγμα από μια άλλη, πιο εύκολη, συνάρτηση πυκνότητας q. Τα δείγματα που θα παράγει η q θα πρέπει με κάποιον τρόπο να τα δεχθούμε ή να τα απορρίψουμε ως δείγματα της f με κάποια αντίστοιχη πιθανότητα. Η q θα πρέπει αρχικά να έχει το ίδιο πεδίο ορισμού με την f για να έχουν νόημα οι τιμές. Επιπλέον, θα θέλαμε ιδανικά να είναι «κοντά» στην f. Με αυτόν τον τρόπο, η πιθανότητα αποδοχής του δείγματος θα είναι συχνότερη με αποτέλεσμα ο χρόνος εκτέλεσης του αλγόριθμου να είναι σημαντικά μικρότερος.

Ο τρόπος με τον οποίο θα βρούμε αυτή την πιθανότητα, είναι να βρούμε κάποιο $M > 0$, και συγκεκριμένα $M > 1$ τέτοιο ώστε να καλύπτουμε την f, δηλαδή, $\mathrm{M}q(x) > f(x)$ για κάθε x. Επειδή η $Mq(x)$ (Συνάρτηση Φάκελος) θα είναι πάντα πάνω από την $f(x)$, μπορούμε να ορίσουμε μια πιθανότητα όπου οι τιμές που γεννήθηκαν από την q θα είναι περισσότερο πιθανές να γίνουν αποδεκτές όταν η πυκνότητα της f σε εκείνο το σημείο είναι μεγάλη. Λόγου αυτού, ορίζουμε την πιθανότητα αποδοχής ως $\frac{f(x)}{\mathrm{M}q(x)}$.

Για να έχουμε πιθανότητα αποδοχής θα πρέπει προφανώς να βρούμε το κατάλληλο M. Υπάρχουν διαφορετικού τρόποι για την εύρεση του $\mathrm{M} = sup\frac{f(x)}{q(x)}$, το οποίο μπορεί να βρεθεί και αριθμητικά, ή $f(x) = \mathrm{M}q(x)$.

Αλγόριθμος 4.2.1 : Μέθοδος Απόρριψης.

---

1. Προσομοιώνουμε $X \sim q$ και $U \sim \mathcal{U}nif(0,1)$.

2. Θέτουμε $X = x$ και $U = u$.

3. **Αν** $u \leq \dfrac{f(x)}{\mathrm{M}q(x)}$ **τότε**
$y = x$
Αλλιώς
Επιστροφή στο βήμα **1**

Τα θετικά με αυτή την μέθοδο είναι ότι δεν προϋποθέτει τα πράγματα που χρειάζεται η μέθοδος αντιστροφής, δηλαδή, μπορούμε να εφαρμόσουμε αυτή την μέθοδο ακόμα και όταν έχουμε μια ανάλογη συνάρτηση $f^*(x)$[3]. Ένα από τα αρνητικά της, είναι ότι η μέθοδος αυτή βασίζεται πολύ στην συνάρτηση φάκελο με την έννοια ότι αν δεν είναι «κοντά» στην $f(x)$ θα πάρει πολύ χρόνο να εκτελεστεί ο αλγόριθμος. Το πόσο χρόνο θα πάρει το ποσοτικοποιούμε με την πιθανότητα αποδοχής όπου είναι $\frac{\int_D f(x)\,dx}{M}$. Τέλος επειδή αυτός ο αλγόριθμος απορρίπτει δείγματα τότε είναι πολύ πιθανό να καταλήξουμε με λιγότερο δείγμα από όσο θα θέλαμε στην αρχή.

[3]Σε αυτή τη περίπτωση ενδέχεται τότε $0 < M < 1$ καθώς λείπει η σταθερά κανονικοποίησης.

Ωστόσο, αυτές οι μέθοδοι δεν είναι πρακτικά χρήσιμες σε μεγαλύτερες διαστάσεις. Για την μέθοδο αντιστροφής δεν είναι ρεαλιστικό να υποθέσουμε πως έχουμε στην διάθεση μας την σταθερά κανονικοποίησης. Ακόμα και να την είχαμε στην διάθεση μας, ο υπολογισμός της αντίστροφης αθροιστικής συνάρτηση δεν θα είναι εφικτή. Η μέθοδος απόρριψης είναι αρκετά γενική και μπορεί να εφαρμοστεί θεωρητικά σε μεγαλύτερες διαστάσεις. Σε τέτοια προβλήματα ωστόσο, το άνω φράγμα M θα γίνει υπερβολικά μεγάλο, καθώς προσπαθούμε να καλύψουμε μια άλλη συνάρτηση στον πολυδιάστατο χώρο. Η συνέπεια είναι πως η αποδοχή δειγμάτων θα είναι πολύ σπάνια, καθιστώντας αυτήν τη μέθοδο μη-πρακτική σε πολυδιάστατα προβλήματα.

4.3 Αλγόριθμος του δειγματολήπτη Gibbs

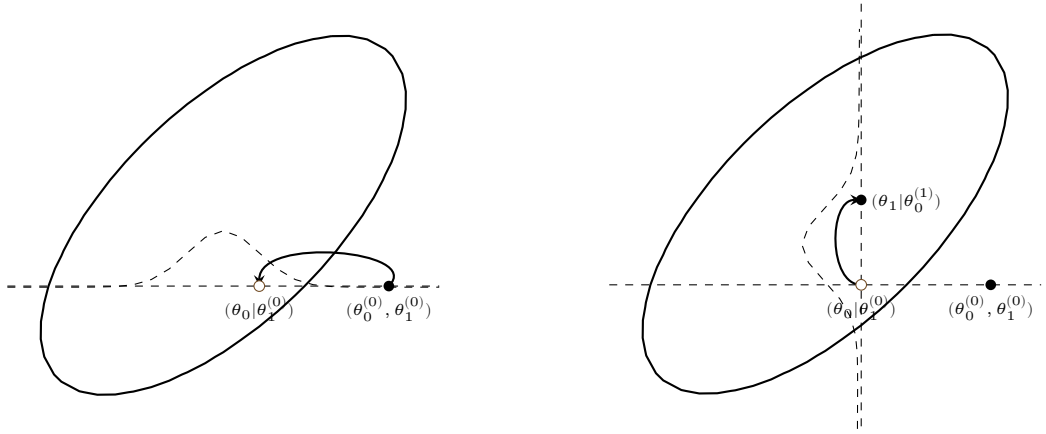
Από τις προηγούμενες ενότητες ένα από τα προβλήματα που διαπιστώσαμε είναι ότι η προσομοίωση δειγμάτων από πολυμεταβλητές κατανομές είναι πολύ δύσκολη έως και αδύνατη. Εξαιτίας αυτού, θα πρέπει να σκεφτούμε μια εναλλακτική προσέγγιση για την λήψη δειγμάτων από πολυμεταβλητές κατανομές. Το κλειδί σε αυτή την απάντηση, βρίσκεται στις μεθόδους MCMC που είχαμε αναφέρει και στην αρχή αυτού του κεφαλαίου. Σε αυτό το κεφάλαιο θα συζητήσουμε για μια από τις πιο γνωστές τεχνικές MCMC, τον αλγόριθμο του δειγματολήπτη Gibbs.

Ο δειγματολήπτης Gibbs πήρε το όνομα του από τον φυσικό Josiah Willard Gibbs $(1839–1903)$, λόγου μιας σχέσης μεταξύ του αλγορίθμου δειγματοληψίας και της Στατιστικής Φυσικής. Παρ' όλο που αρχικά ήταν μια τεχνική στην Στατιστική Φυσική, το άρθρο των Geman.S και Geman.D (1984) ήταν εκείνο που κατάφερε να το φέρει στο πεδίο της Στατιστικής. Στο άρθρο τους, εφάρμοσαν την ιδέα αυτή στη μελέτη τυχαίων πεδίων Gibbs με εφαρμογή σε προβλήματα επεξεργασίας εικόνων. Οι Tanner, M.A. και Wong, W.H. (1987) χρησιμοποίησαν τις ίδιες ιδέες για να προσομοιώσουν τιμές από μία εκ των υστέρων κατανομή $p(\theta|x)$. Ωστόσο, το άρθρο των Gelfand, A.E και Adrian F.M. Smith (1990) ήταν εκείνο το οποίο εδραίωσε τον αλγόριθμο ως μια γενικότερη μεθοδολογία προσομοίωσης τιμών από κατανομές. Μια τέτοια καθιέρωση, ήταν τελικά ο κυριότερος λόγος της μεγάλης ανάπτυξης των μεθόδων MCMC μέχρι και την σημερινή εποχή.

Αυτό που κάνει ο δειγματολήπτης Gibbs είναι να «σπάσει» την δυσκολία ενός πολυδιάστατου δειγματοληπτικού προβλήματος σε πιο απλοϊκά προβλήματα, δηλαδή, σε προβλήματα προσομοίωσης μικρότερες διάστασης όπου μπορούμε να διαχειριστούμε. Για να καταλάβουμε καλύτερα τι εννοούμε από το παραπάνω, θα θεωρήσουμε πρόβλημα προσομοίωσης σε δύο διαστάσεις και στην συνέχεια θα γενικεύσουμε. Η κεντρική ιδέα του αλγορίθμου είναι η διαδοχική προσομοίωση τιμών από τις

δεσμευμένες κατανομές $(\theta_0|\theta_1)$ και $(\theta_1|\theta_0)$ το οποίο οδηγεί (για μεγάλο πλήθος επαναλήψεων) στην προσομοίωση ενός δείγματος από την από κοινού κατανομή του (θ_0, θ_1). Στην ουσία αυτές οι δεσμευμένες κατανομές «σαρώνουν» την από κοινού κατανομή.

Ο τρόπος με τον οποίο ξεκινάμε είναι βάζοντας μια αρχική τιμή $\theta^{(0)} = (\theta_0^{(0)}, \theta_1^{(0)}) \in \Theta$. Η επόμενη τιμή της θ_0 θα είναι μια προσομοιωμένη τιμή από την δεσμευμένη κατανομή $\theta_0^{(1)} \sim f(\theta_0|\theta_1^{(0)})$. Για την επόμενη τιμή του θ_1, θα προσομοιώσουμε από την δεσμευμένη κατανομή $\theta_1^{(1)} \sim f(\theta_1|\theta_1^{(1)})$, δίνοντας μας την πρώτη παρατήρηση του αλγορίθμου $(\theta_0^{(1)}, \theta_1^{(1)})$.



Διάγραμμα 4.3.1: Αλγόριθμος Gibbs σε δύο διαστάσεις .

Το Διάγραμμα 4.3.1 απεικονίζει μια ισοϋψή καμπύλη μιας διμεταβλητής κανονικής κατανομής, με τις μαύρες ευθείες να εκπροσωπούν, τις δεσμευμένες κατανομές των μεταβλητών. Η διαδικασία του αλγορίθμου, όπως αναφέρθηκε, έχει να κάνει με τις διαδοχικές προσομοιώσεις τιμών. Αυτό που αξίζει να αναφερθεί σε αυτό το σημείο είναι πως πλέον δεν έχουμε στην διάθεση μας ένα ανεξάρτητο δείγμα, αλλά ένα εξαρτημένο καθώς η κάθε διαδοχική επανάληψη εξαρτάται από την προηγούμενη (Μαρκοβιανές αλυσίδες).

Στην Ενότητα §4.1 είχε γίνει η αναφορά στις στάσιμες κατανομές ως μια κατανομή όπου η αλυσίδα συγκλίνει όσο $t \to \infty$ κάτω από κάποιες υποθέσεις. Αποδεικνύεται, πως ο δειγματολήπτης Gibbs συγκλίνει στην στάσιμη κατανομή με αποτέλεσμα να έχουμε δείγμα από την κατανομή στόχο.

Η σπουδαιότητα του δειγματολήπτη Gibbs δεν σταματά στα δισδιάστατα προβλήματα καθώς μπορεί να γενικευτεί για οποιαδήποτε διάσταση.

Υποθέτουμε πως μας ενδιαφέρει να προσομοιώσουμε ένα δείγμα από μια πολυμεταβλητή κατανομή $f(\boldsymbol{\theta})$ με $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)$. Μπορούμε, να προσομοιώσουμε δείγμα για κάθε συνιστώσα θ_j από

την δεσμευμένη κατανομή δοθέντος όλες τις υπόλοιπες συνιστώσες $\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_d$. Αυτή η κατανομή καλείται η πλήρη δεσμευμένη κατανομή της θ_j (full conditional) και θα είναι της μορφής $\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_d$ για κάθε j. Η επαναληπτική προσομοίωση διαδοχικών δειγμάτων μας δίνει αυτόν τον γενικό αλγόριθμο

Αλγόριθμος 4.3.1 : Δειγματολήπτης Gibbs.

1. Αρχικοποίηση $\theta^{(0)} = (\theta_0^{(0)}, ..., \theta_d^{(0)})$.

2. Για $t = 1, 2, ...$ **επανάλαβε**
Προσομοίωσε $\theta_0^{(t)}$ από την δεσμευμένη $\theta_0|\theta_1^{(t-1)}, ..., \theta_d^{(t-1)}$.
Προσομοίωσε $\theta_1^{(t)}$ από την δεσμευμένη $\theta_1|\theta_0^{(t)}, \theta_2^{(t-1)}, ..., \theta_d^{(t-1)}$.
...
Προσομοίωσε $\theta_d^{(t)}$ από την δεσμευμένη $\theta_d|\theta_0^{(t)}, ..., \theta_{d-1}^{(t)}$.
Αποθήκευση t-οστού δείγματος.

Τέλος επανάληψης

Σε αυτό το σημείο θα πρέπει να αναφερθούν και κάποια αρνητικά σχετικά με αυτή τη τεχνική. Πρώτον και κυριότερο, αυτός ο αλγόριθμος προϋποθέτει τις πλήρες δεσμευμένες εκ των υστέρων κατανομές (full conditional) το οποίο προφανώς μπορεί και να μην είναι δυνατό. Επιπλέον, ο ρυθμός σύγκλισης του αλγορίθμου εξαρτάται σοβαρά από την αρχική τιμή, ειδικά όταν πρόκειται για δείγμα συσχετισμένων μεταβλητών. Σε περίπτωση ανεξάρτητων μεταβλητών ο ρυθμός σύγκλισης είναι πιο γρήγορος σε σύγκριση με εξαρτημένες μεταβλητές. Παρ' όλο αυτά, το δείγμα και στις δύο περιπτώσεις είναι εξαρτημένο, με αποτέλεσμα η διασπορά των εκτιμητών είναι μεγαλύτερη καθώς η εκτίμηση μπορεί να γίνεται σε μη αντιπροσωπευτική περιοχή της κατανομή. Για αυτόν τον λόγο πρέπει να περιμένουμε για δείγμα από όλο τον χώρο που καλύπτει η κατανομή.

4.4 Αλγόριθμος Metropolis-Hastings

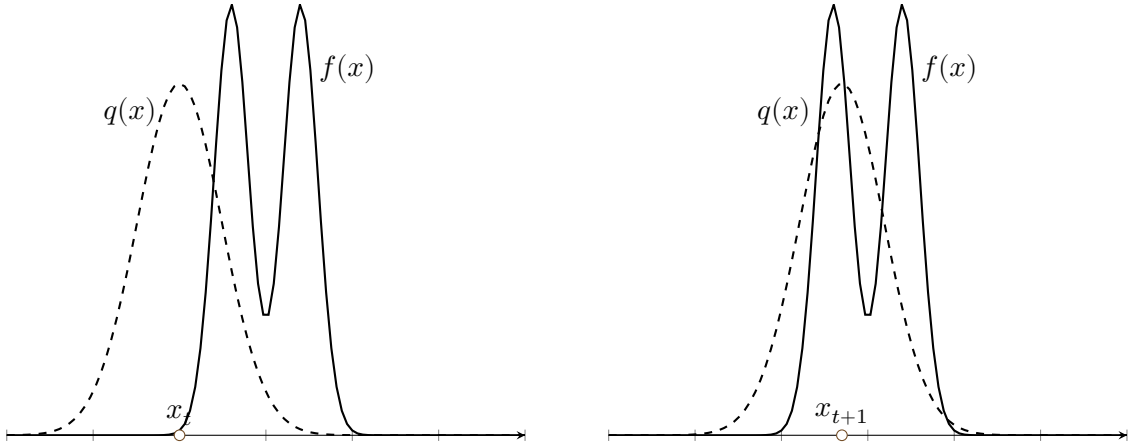
Ο αλγόριθμος Metropolis-Hastings ήταν ένας από τους πρώτος αλγορίθμους MCMC. Αναπτύχθηκε την δεκαετία του 50 από τον Μητρόπουλο Νικόλαο, et al (1953), όπου υιοθετήθηκε πιο κοντά στη Στατιστική πρακτικά από το έργο του Hastings, W. K. (1970). Ο Peter J. Green (1995) προσάρμοσε την ιδέα του αλγορίθμου Metropolis-Hastings στην περίπτωση που η κατανομή-στόχος έχει στήριγμα χώρους μεταβλητής μεγάλης διάστασης. Το 2000, είχε πάρει την πρώτη θέση των πιο επιδραστικών αλγορίθμων του προηγούμενου αιώνα. Ο αλγόριθμος πλέον χρησιμοποιείται εκτενώς από πολλούς επιστήμονες και είναι ένα ιδιαίτερα πολύτιμο εργαλείο στην Μπεϋζιανή Στατιστική.

Ο αλγόριθμος Metropolis κάνει χρήση μιας κατανομής πρότασης q το οποίο θα δίνει προτάσεις δειγμάτων, δηλαδή, θα γεννά τιμές. Η πυκνότητα q θα εξαρτάται από την τρέχουσα χρονική στιγμή x_t και την επόμενη χρονική στιγμή θα έχουμε δείγμα από την $q(x_{t+1}|x_t)$. Αφού προτείνει δείγμα, τότε βάση κάποιου κριτηρίου θα πρέπει να δεχθούμε ή όχι το δείγμα ως δείγμα της f. Το κριτήριο αποδοχής του δείγματος x_{t+1} θα δίνετε από την ποσότητα

$$\alpha = \frac{f(x_{t+1})}{f(x_t)} \frac{q(x_t|x_{t+1})}{q(x_{t+1}|x_t)} \tag{4.4.1}$$

Η ποσότητα α δεν είναι πάντα πιθανότητα καθώς ο παρανομαστής δεν θα είναι πάντα μεγαλύτερος από τον αριθμητή. Συγκεκριμένα, αγνοώντας προς το παρόν τον όρο $\frac{q(x_t|x_{t+1})}{q(x_{t+1}|x_t)}$ στην εξίσωση (4.4.1), παρατηρούμε πως το υπόλοιπο μέρος του α θα είναι μεγάλο αν η πρόταση νέου δείγματος βρίσκεται σε περιοχή με μεγαλύτερη πυκνότητα από την προηγούμενη πρόταση. Προφανώς αυτό σημαίνει πως η αποδοχή του νέου δείγματος είναι βέβαιη καθώς παίρνουμε τιμές σε πυκνότερες περιοχές της f, όπως φαίνεται και από το Σχήμα 4.4.1. Από την άλλη αν το νέο δείγμα βρίσκεται σε περιοχή με μικρότερη πυκνότητα από την προηγούμενη, τότε το $\alpha \in (0, 1)$ και εδώ πλέον μιλάμε για πιθανότητα αποδοχής. Μια τέτοια συμπεριφορά είναι λογική καθώς, το παραπάνω μας «εξασφαλίζει» λιγότερο δείγμα, σε λιγότερες πυκνές περιοχές της f. Τώρα ο όρος $\frac{q(x_t|x_{t+1})}{q(x_{t+1}|x_t)}$ παίζει τον ρόλο διόρθωσης του αλγορίθμου. Όταν η κατανομή πρότασης q δεν είναι συμμετρική τότε ο υπολογισμός του α είναι λανθασμένος. Έτσι, ο γενικότερος υπολογισμός του α ανάγεται στην εξίσωση (4.4.1) όπου λαμβάνει υπόψη κάθε είδους κατανομή πρότασης.

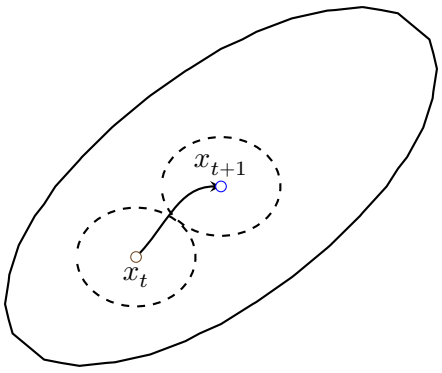
Παρατηρούμε ωστόσο ότι αυτή η φιλοσοφία προσομοίωσης μοιάζει αρκετά με την μέθοδο απόρριψης. Η μέθοδος απόρριψης όμως προϋποθέτει η κατανομή πρότασης να μοιάζει με την κατανομή στόχο για να λειτουργεί αποδοτικά. Σε μεγάλα και πολύπλοκα προβλήματα όμως, η δημιουργία μιας τέτοιας q είναι πολύ δύσκολη. Στον Metropolis αλγόριθμο η κατανομή πρότασης δεν είναι αναγκασμένη να μοιάζει με την κατανομή στόχο. Μια συνηθισμένη επιλογή της q θα μπορούσε να είναι μια απλή κατανομή όπως την κανονική κεντραρισμένη στο x_t και η κατανομή στόχος να είναι τελείως διαφορετική.



Διάγραμμα 4.4.1: Metropolis αλγόριθμος σε μια διάσταση.

Επειδή πρόκειται για μια Μαρκοβιανή αλυσίδα, αν παίρνουμε για αρκετό χρόνο δείγμα, θα καταλήξουμε πλέον να παίρνουμε δείγμα από την στάσιμη κατανομή, την $f(x)$.

Αν η κατανομή στόχος είναι μια διμεταβλητή κατανομή τότε, θα πρέπει αντίστοιχα η κατανομή πρότασης να είναι επίσης διμεταβλητή κατανομή. Η ιδέα του αλγορίθμου δεν αλλάζει, συνεχίζουμε την προσομοίωση τιμών από την κατανομή πρόταση και αποδεχόμαστε ή όχι αναλόγως το α.



Διάγραμμα 4.4.2: Metropolis αλγόριθμος σε δύο διαστάσεις.

Γενικεύουμε το αλγόριθμο Metropolis-Hastings ως εξής

Αλγόριθμος 4.4.1 : Metropolis-Hastings Αλγόριθμος.

1. Αρχικοποίηση $\theta^{(0)}$.

2. Για $t = 1, 2, ...$ **επανάλαβε**
Προσομοίωσε $x_{t+1} \sim q(x_{t+1}|x_t)$ και $u \sim \mathcal{U}nif(0, 1)$.

3. Υπολόγισε $\alpha(x_t, x_{t+1}) = min\left\{1, \frac{f(x_{t+1})q(x_t|x_{t+1})}{f(x_t)q(x_{t+1}|x_t)}\right\}$

4. Αν $u < a(x_t, x_{t+1})$ **τοτε**
Αποθήκευσε x_{t+1}

Αλλιώς
Θέσε $x_{t+1} = x_t$

Τέλος αν
Τέλος επανάληψης

Αυτός ο αλγόριθμος έχει όλα τα θετικά της μέθοδου απόρριψης, με το επιπλέον σημαντικό πλεονέκτημα της γενίκευσης.

Παρ' όλο αυτά, επειδή είναι μέθοδος MCMC, θα έχει το ίδιο πρόβλημα διασποράς εκτιμητών όπως τον Gibbs. Επιπλέον, πρέπει να σημειωθεί πως η επιλογή της q είναι αρκετά σημαντική σε προβλήματα όπου η κατανομή στόχο έχει παχές ουρές. Σε τέτοιες περιπτώσεις αν η q είναι για παράδειγμα κανονική τότε δεν θα έχουμε σχεδόν ποτέ δείγμα από τις ουρές σε λογικό χρόνο. Θα πρέπει δηλαδή να ληφθούν υπόψη οι παχές ουρές μέσω της q. Τέλος, σε συνδυασμό με το παραπάνω πρόβλημα έχουμε και την επιλογή της σ^2 στις κατανομές πρότασης (το βήμα). Πολύ μεγάλα ή μικρά σ^2 επηρεάζουν αρνητικά την ορθή κάλυψη της κατανομής και τον ρυθμό σύγκλισης. Οπότε, ο σκοπός είναι η επιλογή μιας σ^2 όπου προσπαθεί να καλύψει την κατανομή σε λογικό χρόνο.

4.5 Σύγκλιση και διαχείριση δείγματος

Μετά την εκτέλεση των αλγορίθμων, τίθεται το ερώτημα της σύγκλισης του. Ο πιο κλασικός τρόπος αντιμετώπισης αυτού του προβλήματος είναι η επιλογή διαφορετικών αρχικών τιμών. Αν η κάθε αλυσίδα συγκλίνει στην ίδια περιοχή τιμών (όχι απαραίτητα με την ίδια ταχύτητα) τότε ξέρουμε πως παίρνουμε σωστά δείγμα.

Στο πλαίσιο δείγματος πρέπει να λάβουμε υπόψη το γεγονός πως τα αρχικά δείγματα της αλυσίδας δεν είναι από την κατανομή στόχο. Είναι απλώς τιμές που είναι απαραίτητες για να φτάσει η αλυσίδα

στην στάσιμη κατανομή. Οπότε μπορούμε να διώξουμε κάποια αρχικά δείγματα (το ποσό δειγμάτων εξαρτάται από ρυθμό σύγκλισης). Αυτή η περίοδος καλείται burn-in. Τέλος, επειδή αναφέραμε αυτοσυσχέτιση μεταξύ δειγμάτων, μπορούμε να διώξουμε δείγματα που έχουν υψηλή αυτοσυσχέτιση σε κάποιο συγκεκριμένη χρονική υστέρηση (lag). Το παραπάνω μας βοηθά στην μείωση διασποράς των εκτιμητών.

4.6 Συμπεράσματα Κεφαλαίου

Ο σκοπός αυτού το Κεφαλαίου ήταν η βασική εισαγωγή στις μεθόδους Markov chain Monte Carlo (MCMC). Για να πετύχουμε τον παραπάνω στόχο, έγινε αρχικά η εισαγωγή των Μαρκοβιανών αλυσίδων. Έπειτα, δώσαμε μια περιληπτική εισαγωγή στις μεθόδους ολοκλήρωσης Monte Carlo. Ο λόγος είναι διότι οι παραπάνω μέθοδοι μπορούν να χρησιμοποιηθούν και σε πολυμεταβλητά προβλήματα. Σε τέτοια προβλήματα, ο χώρος αυξάνεται δραματικά με αποτέλεσμα οι εκτιμητές μας να έχουν μεγάλο σφάλμα. Έτσι, αυτές οι μεθοδολογίες μας παρέχουν αποδοτικούς τρόπους να εκτιμήσουμε ποσότητες ακόμα και σε πολυδιάστατες κατανομές. Στην συνέχεια, αναλύσαμε την Monte Carlo προσομοίωση. Εντοπίσαμε τον λόγο που αποτυγχάνουν σε πολυμεταβλητά προβλήματα και δώσαμε γενικεύσεις της παραπάνω διαδικασίας με την βοήθεια δύο βασικών αλγορίθμων για προσομοίωση δειγμάτων από πολυμεταβλητές κατανομές, τον δειγματολήπτη Gibbs και Metropolis-Hastings αλγόριθμο.

Κεφάλαιο 5

Επιλογή Μοντέλου

Ένα από τα προβλήματα που καλείται να αντιμετωπίσει ένας ερευνητής στην μοντελοποίηση των δεδομένων, είναι η επιλογή του «βέλτιστου» μοντέλου, του μοντέλου το οποίο προσεγγίζει πιο αποτελεσματικά την αλήθεια. Συγκεκριμένα, το πρόβλημα αφορά την επιλογή συγκεκριμένων μεταβλητών από ένα ευρύτερο σύνολο επεξηγηματικών μεταβλητών, με σκοπό την βέλτιστη εξήγηση της απαντητικής μεταβλητής. Η διαδικασία εύρεσης ενός τέτοιου μοντέλου ανάγεται στην σύγκριση μοντέλων. Οι πιο συνηθισμένες στρατηγικές έχουν να κάνουν με ελέγχους σημαντικότητας τα οποία είναι βασισμένες σε ασυμπτωματικές κατανομές. Παρ' όλο αυτά, υπάρχουν αρκετά προβλήματα με αυτή την προσέγγιση. Δύο βασικά προβλήματα αφορούν τα μη-φωλιασμένα μοντέλα και τα είδη λαθών (τύπου I και II). Για την αποφυγή τέτοιων προβλημάτων, συνήθως καταφεύγουμε στην χρήση αλγοριθμικών μεθόδων, όπως Stepwise με κριτήρια πληροφορίας AIC (Akaike Information Criterion) ή και BIC (Bayesian Information Criterion). Ωστόσο, οι Stepwise αλγόριθμοι υποκύπτουν σε αυστηρή κρητική καθώς πρόκειται για διαδικασίες οι οποίες θεωρούν ότι όλες οι υποθέσεις του μοντέλου ισχύουν και δεν λαμβάνουν υπόψη μεταβλητές οι οποίες προηγουμένως δεν θεωρούνταν σημαντικές (Backward και Forward selection). Στην πραγματικότητα, δεν υπάρχει κανένα ευρέως αποδεκτό κριτήριο για την τελική απόφαση του «βέλτιστου» μοντέλου. Σε γενικότερες γραμμές, αυτό που προτείνετε είναι η χρήση αυτών των "βρώμικων" αλλά γρήγορων διαδικασιών για να μας ξεκαθαρίσουν και να μας δώσουν μια ιδέα για το τι προτείνεται και να αξιολογήσουμε. Υπογραμμίζουμε, πως δεν τις εμπιστευόμαστε από μόνες τους και η τελική απόφαση θα πρέπει να παρθεί βάσει πολλών και διαφορετικών κριτηρίων που στην ουσία κοιτάνε το ίδιο πρόβλημα από διαφορετικές οπτικές.

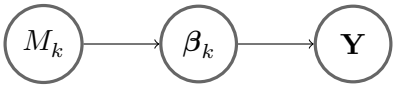
Το βασικότερο μειονέκτημα με τις κλασικές μεθόδους είναι η επιλογή ενός μοναδικού μοντέλου. Ας υποθέσουμε την περίπτωση που έχουμε στην διάθεση μας πολλά δυνατά μοντέλα τα οποία προσαρμόζονται επαρκώς στα δεδομένα αλλά έχουν εντελώς διαφορετική ερμηνεία. Η επιλογή ενός μοναδικού μοντέλου με κλασικές μεθόδους σε αυτή την περίπτωση δεν θα λαμβάνει υπόψη την αντίστοιχη

αβεβαιότητα των μοντέλων, δηλαδή, μοντέλων που μπορούν να παρέχουν ισοδύναμη και χρήσιμη πληροφορία με αποτέλεσμα η συμπερασματολογία να είναι μάλλον μεροληπτική. Με άλλα λόγια, επειδή η κατανόηση μας για ένα φυσικό φαινόμενο είναι ατελές, ο εντοπισμός ενός μοναδικού βέλτιστου μοντέλου είναι αδύνατη, ειδικά όταν έχουμε πολλά και σχεδόν ισοπίθανα δυνατά μοντέλα που περιγράφουν το φαινόμενο. Οπότε, γενικότερα υπάρχει αβεβαιότητα στην επιλογή του βέλτιστου μοντέλου από ένα σύνολο μοντέλων. Η μεροληπτική συμπερασματολογία από τις κλασικές προσεγγίσεις θα μπορούσε να οφείλετε στο γεγονός πως για ένα άλλο σύνολο δεδομένων, το μοντέλο το οποίο υποστηρίζετε να είναι διαφορετικό από το υπάρχων. Άρα, θα θέλαμε να ποσοτικοποιήσουμε την αβεβαιότητα γύρω από το μοντέλο και να γνωρίζουμε πόσο σίγουροι νιώθουμε για το αντίστοιχο μοντέλο.

5.1 Το πιθανοθεωρητικό πλαίσιο αβεβαιότητας μοντέλων

Η ποσοτικοποίηση της αβεβαιότητας θα γίνει με την πιθανότητα για το εκάστοτε μοντέλο από ένα σύνολο μοντέλων. Η πιθανότητα ενός μοντέλου ορίζεται ως ο βαθμός βεβαιότητας του μοντέλου να είναι το πραγματικό δοθέντος ότι το πραγματικό μοντέλο υπάρχει στο σύνολο των δυνατών μοντέλων. Πρακτικά όμως, επειδή όλα τα μοντέλα είναι λάθος, είναι πιο λογικό να θεωρήσουμε την πιθανότητα του μοντέλου ως βαθμό βεβαιότητας ότι εκείνο το μοντέλο είναι το «βέλτιστο» στο σύνολο μοντέλων.

Έτσι λοιπόν, θεωρούμε το πρόβλημα επιλογής μοντέλου ένα πεπερασμένο σύνολο από K μοντέλα $\mathcal{M} = \{M_1, ..., M_K\}$, με $\mathcal{M}$ να συμβολίζει το σύνολο των πιθανών μοντέλων. Αυτά τα K δυνατά μοντέλα είναι υποψήφια ως προς την μοντελοποίηση των δεδομένων $\mathbf{Y}$ και είναι $2^{p+1} - 1$ σε πλήθος. Η αντίστοιχη πιθανοφάνεια του M_k είναι η $f(\mathbf{Y}|\boldsymbol{\beta}_k, M_k)$ (καθώς το μοντέλο ορίζεται και από τις επεξηγηματικές του μεταβλητές οπότε, και από τις παραμέτρους του) με $\boldsymbol{\beta}_k$ να είναι το διάνυσμα άγνωστων παραμέτρων του μοντέλου M_k. Η Μπεϋζιανή προσέγγιση φυσικά αναθέτει μια εκ των προτέρων κατανομή $p(\boldsymbol{\beta}_k|M_k)$ στις παραμέτρους του κάθε μοντέλου, και μια εκ των προτέρων κατανομή $p(M_k)$ για τα ίδια τα μοντέλα. Τα παραπάνω είναι η ιεραρχική δομή του μοντέλου με την έννοια ότι πρώτα παράγεται το μοντέλο M_k από την $p(M_k)$, στην συνέχεια το διάνυσμα $\boldsymbol{\beta}_k$ παράγετε από το $p(\boldsymbol{\beta}_k|M_k)$ και τέλος, τα δεδομένα παράγονται από το $f(\mathbf{Y}|\boldsymbol{\beta}_k, \mathrm{M}_k)$.



Διάγραμμα 5.1.1: Δομή ιεραρχικού μοντέλου.

Έτσι ο σκοπός μας είναι το να βρούμε εκείνο το μοντέλο (από όλα τα πιθανά) στο $\mathcal{M}$ το οποίο πραγματικά γέννησε τα δεδομένα μας, δηλαδή, εκείνο το μοντέλο που γεννήθηκε από την $p(M_k)$. Με άλλα λόγια, θέλουμε να αξιολογήσουμε την πιθανότητα ότι το M_k ήταν εκείνο το μοντέλο, δοθέντος τα $\mathbf{Y}$. Όποτε, με την βοήθεια της εκ των υστέρων κατανομής (5.1.1) θα αξιολογήσουμε τα πολλά μοντέλα και η επιλογή ενός μοντέλου ως το "πραγματικό" θα γίνει βάσει κάποιων κριτηρίων.

$$p(M_k|\boldsymbol{y}) = \frac{p(M_k)f(\boldsymbol{y}|M_k)}{\sum_{k\in\mathcal{M}} p(M_k)f(\boldsymbol{y}|M_k)} \tag{5.1.1}$$

όπου

$$f(\boldsymbol{y}|M_k) = \int f(\boldsymbol{y}|\boldsymbol{\beta}_k, M_k)p(\boldsymbol{\beta}_k|M_k)\, d\boldsymbol{\beta}_k \tag{5.1.2}$$

Στην εξίσωση (5.1.2) ολοκληρώνουμε ως προς το β_k, δηλαδή, λαμβάνουμε υπόψη όλα τα δυνατά β_k του μοντέλου M_k για να έχουμε την πιθανοφάνεια των δεδομένων δοθέντος του M_k.

Αξίζει να σημειωθεί, πως η διαδικασία επιλογής μοντέλου, είναι η άμεση γενίκευση των Μπεϋζιανών ελέγχων υποθέσεων για τα μοντέλα, καθώς πρόκειται για την άμεση σύγκριση πολλών μοντέλων ταυτόχρονα. Παρ' όλο που η επιλογή μοντέλου μπορεί να γίνει με την χρήση του παράγοντα Bayes, το παραπάνω αφορά μόνο την επιλογή μεταξύ δύο ανταγωνιστικών μοντέλων το οποίο δεν είναι τόσο άμεσο όταν γίνεται σύγκριση παραπάνω από δύο ταυτόχρονα.

Αυτό που κάνουμε δηλαδή, είναι οι ανά δύο συγκρίσεις των μοντέλων ας πούμε M_0 και M_1

$$\frac{p(M_0|\boldsymbol{y})}{p(M_1|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|M_0)}{f(\boldsymbol{y}|M_1)}\frac{p(M_0)}{p(M_1)}$$

Η παραπάνω έκφραση, λέει πως τα δεδομένα μέσα από τον παράγοντα Bayes $\frac{f(\boldsymbol{y}|M_0)}{f(\boldsymbol{y}|M_1)}$ αναβαθμίζουν τα prior odds $\frac{p(M_0)}{p(M_1)}$ για να μας δώσουν τα posterior odds.

Για να υπολογίσουμε την εκ των υστέρων κατανομή (5.1.1), στο πρόβλημα επιλογής μεταβλητών για το γραμμικό μοντέλο, χρειαζόμαστε να υπολογίσουμε την ποσότητα (5.1.2) για κάθε δυνατό μοντέλο. Από το συζυγές μας γραμμικό μοντέλο, μπορούμε να γράψουμε με ισοδύναμο τρόπο την εξίσωση (5.1.2) ως

$$f(\boldsymbol{y}|M) = \int\int f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})p(\boldsymbol{\beta}, \sigma^2)\, d\sigma^2\, d\boldsymbol{\beta} = \int\int p(\beta, \sigma^2|\boldsymbol{y}, \mathbf{X})\, d\sigma^2\, d\boldsymbol{\beta}$$

Ωστόσο, οι παραπάνω πράξεις θα είναι επίπονες, οπότε θα πρέπει να προσεγγίσουμε το πρόβλημα με διαφορετικό τρόπο. Θα εκμεταλλευτούμε τις ιδιότητες της κανονικής κατανομής από το συζυγές

γραμμικό μοντέλο

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \;\text{ με }\; \boldsymbol{\varepsilon}_1 \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$$

$$\boldsymbol{\beta} = \boldsymbol{\mu}_{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_2 \;\text{ με }\; \boldsymbol{\varepsilon}_2 \sim \mathcal{N}_q(\mathbf{0}_q, \sigma^2\mathbf{V})$$

Τα τυχαία διανύσματα $\boldsymbol{\varepsilon}_1$ και $\boldsymbol{\varepsilon}_2$ είναι ανεξάρτητα. Αντικαθιστώντας ως προς $\boldsymbol{\beta}$ έχουμε

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \mathbf{X}\boldsymbol{\varepsilon}_2 + \boldsymbol{\varepsilon}_1$$

Η παραπάνω ποσότητα είναι η δεσμευμένη κατανομή του $\boldsymbol{y}|\sigma^2 \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top))$ διότι

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}] &= \mathbb{E}[\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}] + \mathbb{E}[\mathbf{X}\boldsymbol{\varepsilon}_2] + \mathbb{E}[\boldsymbol{\varepsilon}_1] = \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\
Var(\boldsymbol{y}) &= Var(\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \mathbf{X}\boldsymbol{\varepsilon}_2 + \boldsymbol{\varepsilon}_1) = Var(\mathbf{X}\boldsymbol{\varepsilon}_2) + Var(\boldsymbol{\varepsilon}_1) \\
&= \mathbf{X}Var(\boldsymbol{\varepsilon}_2)\mathbf{X}^\top + \sigma^2\mathbf{I}_n = \mathbf{X}\sigma^2\mathbf{V}\mathbf{X}^\top + \sigma^2\mathbf{I}_n = \sigma^2(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)
\end{aligned}$$

Για να μπορούμε να πάρουμε την περιθωριακή κατανομή των δεδομένων $f(\boldsymbol{y}|M)$ πρέπει να ολοκληρώσουμε την από κοινού κατανομή $(\boldsymbol{y}, \sigma^2)$ ως προς το σ^2.

$$\begin{aligned}
f(\boldsymbol{y}|\mathrm{M}) &= \int_0^\infty f(\boldsymbol{y}|\sigma^2, M)p(\sigma^2)\,d\sigma^2 \\
&\propto \int_0^\infty \frac{1}{(\sigma^2)^{n/2+a+1}} exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \right\} exp\left\{ -\frac{b}{\sigma^2} \right\} d\sigma^2 \\
&\propto \int_0^\infty \frac{1}{(\sigma^2)^{n/2+a+1}} exp\left\{ -\frac{1}{\sigma^2}\left[b + \frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \right] \right\} d\sigma^2 \\
&\propto \left[b + \frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \right]^{-\frac{2a+n}{2}} \\
&\propto \left[1 + \frac{1}{2b}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \right]^{-\frac{2a+n}{2}} \\
&\propto \left[1 + \frac{a}{2ab}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \right]^{-\frac{2a+n}{2}} \\
&\propto \left[1 + \frac{1}{2a}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top \left\{ \frac{b}{a}(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top) \right\}^{-1} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \right]^{-\frac{2a+n}{2}}
\end{aligned}$$

Οπότε η πιθανοφάνεια των δεδομένων δοθέντος κάποιο μοντέλο[1] θα ακολουθεί την πολυμεταβλητή $\mathcal{T}$ κατανομή για το συζυγές γραμμικό μοντέλο.

$$\mathbf{Y}|M \sim \mathcal{T}_n(\mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}}, \frac{b}{a}(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top), 2a) \tag{5.1.3}$$

[1] Η περιθωριακή κατανομή των δεδομένων το οποίο εξαρτάτε από το δοθέν μοντέλο, δηλαδή, τον πίνακα σχεδιασμού $\mathbf{X}$.

Με περιθωριακή πυκνότητα ίση με

$$f(\boldsymbol{y}|M) = \frac{\Gamma(\frac{2a+n}{2})}{|\frac{b}{a}(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)|^{1/2}(2a\pi)^{n/2}\Gamma(\frac{2a}{2})} \times \left[1 + \frac{1}{2a}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top \left\{\frac{b}{a}(\mathbf{I}_n + \mathbf{X}\mathbf{V}\mathbf{X}^\top)\right\}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}_{\boldsymbol{\beta}})\right]^{-\frac{2a+n}{2}} \tag{5.1.4}$$

Οπότε, μπορούμε να υπολογίσουμε την περιθωριακή πυκνότητα (5.1.4) για τα διαφορετικά k μοντέλα, αλλάζοντας κάθε φορά τον πίνακα σχεδιασμού για το αντίστοιχο μοντέλο.

Για τον τελικό υπολογισμό της εκ των υστέρων κατανομής (5.1.1) χρειαζόμαστε μια εκ των προτέρων κατανομή για το σύνολο μοντέλων $\mathcal{M}$. Σύμφωνα με τους Hugh Chipman et.al (2001) η επιλογή μιας υποκειμενικής εκ των προτέρων κατανομή για τα μοντέλα έχει αρκετούς περιορισμούς. Σε πρακτικά προβλήματα ο τεράστιος αριθμός καθώς και η πολυπλοκότητα των μοντέλων καθιστά μη-ρεαλιστική την προσπάθεια λογικής περιγραφής της αβεβαιότητας. Έτσι, η πιο συνηθισμένη επιλογή είναι μια αντικειμενική εκ των προτέρων κατανομή

$$p(M_k) = \frac{1}{K} \tag{5.1.5}$$

Η κατανομή της (5.1.5) είναι η διακριτή ομοιόμορφη η οποία είναι μια αντικειμενική, με την έννοια ότι όλα τα μοντέλα είναι εκ των προτέρων ισοπίθανα. Μια χρήσιμη παρατήρηση είναι πως η κατανομή (5.1.5) δεν είναι πραγματικά αντικειμενική. Αυτό συμβαίνει διότι δίνουμε το ίδιο βάρος ακόμα και σε πιο περίπλοκα μοντέλα, δηλαδή, είμαστε πιο πληροφοριακοί για εκείνα τα μοντέλα. Μια πιο δίκαιη εκ των προτέρων κατανομή είναι να εισάγουμε διαφορετικό ομοιόμορφο βάρος το οποίο θα εξαρτάται από τον αριθμό μεταβλητών για το αντίστοιχο μοντέλο. Ωστόσο, για λόγους απλότητας θα χρησιμοποιήσουμε την διακριτή ομοιόμορφη κατανομή (5.1.5).

Αφού έχουν επιλεχθεί και οι εκ των προτέρων κατανομές, όλη η απαραίτητη πληροφορία για την αβεβαιότητα των μοντέλων εμπεριέχεται στην εκ των υστέρων κατανομή. Παρ' όλο αυτά, για την επιλογή ενός μοντέλου θα πρέπει να υιοθετήσουμε κάποιο κριτήριο το οποίο θα βασίζεται στην εκ των υστέρων κατανομή. Ένα λογικό κριτήριο είναι η επιλογή του μοντέλου με την μέγιστη εκ των υστέρων πιθανότητα $\max_k p(M_k|\boldsymbol{y})$. Το παραπάνω κριτήριο ονομάζεται Maximum A Posteriori (MAP) και είναι απλώς η επιλογή της κορυφής της εκ των υστέρων κατανομής του μοντέλου. Ωστόσο, το παραπάνω κριτήριο στην πράξη έχει ένα σημαντικό μειονέκτημα. Σε περιπτώσεις όπου το πλήθος των μοντέλων είναι μεγάλο, η μέγιστη πιθανότητα θα είναι σχεδόν πάντα μικρή. Για παράδειγμα, η απόφαση επιλογής μοντέλου με πιθανότητα του MAP 10% με κάποια μοντέλα να έχουν 9% εκ των υστέρων πιθανότητα

τότε μάλλον δεν είναι σωστή, καθώς υπάρχει υπερβολική αβεβαιότητα γύρω από αυτά τα μοντέλα και θα έχουμε γενικότερα 90% πιθανότητα να κάνουμε λάθος σε αυτή την απόφαση.

Ένας εναλλακτικός τρόπος απόφασης είναι με την πιθανότητα ένταξης (Posterior Inclusion Probability - PIP). Προτού όμως πούμε για την πιθανότητα ένταξης, θα πρέπει να εισάγουμε μια γενικότερη έννοια, το Bayesian Model Averaging.

Το Bayesian Model Averaging (BMA), είναι μια φυσική εξέλιξη της αβεβαιότητας των μοντέλων. Συνοπτικά, είναι ένας τρόπος ο οποίος συνδυάζει πληροφορία μεταξύ στατιστικών μοντέλων και λαμβάνει υπόψη την αβεβαιότητα που υπάρχει στο καθένα. Τυπικά, αυτό που κάνουμε σε μια ανάλυση, είναι να επιλέξουμε ένα μοναδικό μοντέλο για την περιγραφή του μηχανισμού των δεδομένων. Για παράδειγμα, η επιλογή μπορεί να γίνει με το κριτήριο του MAP. Ωστόσο, οποιαδήποτε αβεβαιότητα που υπάρχει στην ίδια την διαδικασία επιλογής μοντέλου αγνοείται. Στην συνέχεια, αυτό που συνήθως μας ενδιαφέρει είναι απλώς ο υπολογισμός της εκ των υστέρων κατανομής για τις παραμέτρους του δοθέν μοντέλου $p(\boldsymbol{\beta}|\boldsymbol{y}, M_k)$. Οι συνέπειες μιας τέτοιας διαδικασίας είχαν αναφερθεί και προηγουμένως, αλλά με λίγα λόγια ο πραγματικός κόσμος είναι συνήθως πιο περίπλοκος από όσο θα μπορούσε να αποτυπωθεί από ένα μοντέλο. Εν συνέχεια, αναφέρεται η ποσοτικοποίηση αβεβαιότητας για το κάθε μοντέλο και το πρακτικό πρόβλημα αβεβαιότητας στην τελική επιλογή ενός μοντέλου. Ως τώρα, έχουμε συζητήσει μόνο για την ποσοτικοποίηση αβεβαιότητας του εκάστοτε μοντέλου. Τώρα, θα αναλύσουμε και το δεύτερο πρόβλημα.

Με πολλά δυνατά μοντέλα, μπορούμε να πάρουμε έναν σταθμισμένο μέσο των εκ των υστέρων κατανομών $p(\boldsymbol{\beta}|\boldsymbol{y})$ κάτω από το εκάστοτε μοντέλο.

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{k \in \mathcal{M}} p(\boldsymbol{\beta}|M_k, \boldsymbol{y}) p(M_k|\boldsymbol{y}) \tag{5.1.6}$$

Εδώ, η $p(\boldsymbol{\beta}|M_k, \boldsymbol{y})$ είναι η εκ των υστέρων κατανομή του $\boldsymbol{\beta}$ για το συγκεκριμένο μοντέλο M_k και η $p(M_k|\boldsymbol{y})$ είναι η εκ των υστέρων πιθανότητα του μοντέλου (5.1.1). Το BMA παρέχει μια αξιολόγηση της εκ των υστέρων κατανομής του $\boldsymbol{\beta}$ που λαμβάνει υπόψη όλα τα δυνατά μοντέλα ταυτόχρονα. Τα βάρη στην εξίσωση (5.1.6)[2] είναι το $p(M_k|\boldsymbol{y})$ καθώς αν η αντίστοιχη πιθανότητα του μοντέλου είναι μικρή τότε η $p(\boldsymbol{\beta}|M_k, \boldsymbol{y})$ δεν θα ληφθεί ιδιαίτερα υπόψη.

Το αξιοσημείωτο με την εξίσωση (5.1.6) είναι ότι μπορεί να γενικευτεί για μια οποιαδήποτε καλώς

[2]Τονίζουμε πως η παραπάνω σχέση έχει νόημα μόνο σε προβλήματα που η παράμετρος έχει το ίδιο νόημα για όλα τα επιμέρους μοντέλα (δηλαδή ισχύει στο πρόβλημα επιλογής μοντέλου αλλά όχι γενικά).

ορισμένη παράμετρο ενδιαφέροντος Δ για κάθε μοντέλο. Το Δ μπορεί να είναι η εκ των υστέρων των $\boldsymbol{\beta}$ ή κάποιο β_j της μεταβλητής X_j κ.ο.κ. Γενικώς, έχουμε

$$p(\Delta|\boldsymbol{y}) = \sum_{k\in\mathcal{M}} p(\Delta|M_k, \boldsymbol{y})p(M_k|\boldsymbol{y}) \tag{5.1.7}$$

Το παραπάνω λαμβάνει υπόψη την αβεβαιότητα της μορφής του μοντέλο βάζοντας βάρη στην δεσμευμένη εκ των υστέρων κατανομή βάση της εκ των υστέρων πιθανότητες του κάθε μοντέλου.
Ο εκ των υστέρων μέσος και διακύμανση της Δ είναι οι εξής

$$\begin{aligned}
\mathbb{E}[\Delta|\boldsymbol{y}] &= \sum_{k\in\mathcal{M}} \mathbb{E}[\Delta|M_k, \boldsymbol{y}]p(M_k|\boldsymbol{y}) \\
Var(\Delta|\boldsymbol{y}) &= \sum_{k\in\mathcal{M}} (Var(\Delta|M_k, \boldsymbol{y}) + \mathbb{E}[\Delta|M_k, \boldsymbol{y}]^2)p(M_k|\boldsymbol{y}) - \mathbb{E}[\Delta|\boldsymbol{y}]^2
\end{aligned}$$

Αυτό που θέλουμε τώρα είναι να βρούμε έναν τρόπο να αξιολογήσουμε την σημαντικότητα μιας μεταβλητής από τα δεδομένα. Στο BMA η απάντηση είναι να υπολογίσουμε για κάθε μεταβλητή την πιθανότητα ένταξης του. Η πιθανότητα ένταξης είναι το άθροισμα των εκ των υστέρων πιθανοτήτων των μοντέλων με τις εκ των υστέρων πιθανότητες των μοντέλων που εμπεριέχουν μια συγκεκριμένη επεξηγηματική. Την πιθανότητα ένταξης μιας μεταβλητής j την ορίζουμε ως το μοντέλο k να περιέχει την επεξηγηματική μεταβλητή j. Έτσι λοιπόν ορίζουμε τις δίτιμες μεταβλητές.

$$\gamma_{kj} = \begin{cases} 1, & \text{Αν το } M_k \text{ μοντέλο περιέχει την μεταβλητή } j \\ 0, & \text{Αλλιώς} \end{cases}$$

Η ποσότητα γ_{kj} είναι μια δίτιμη μεταβλητή η οποία λαμβάνει τιμή ίση με μονάδα αν η μεταβλητή j εμπεριέχεται στο μοντέλο k και μηδέν διαφορετικά. Τότε η πιθανότητα ένταξης της μεταβλητής j είναι ίση με

$$\gamma_{\cdot j} = \sum_{k\in\mathcal{M}} \gamma_{kj} p(M_k|\boldsymbol{y}) \tag{5.1.8}$$

Η εξίσωση (5.1.8) μας λέει στην ουσία ότι η πιθανότητα ένταξης λαμβάνει υπόψη όλα τα δυνατά μοντέλα και μετράμε με συγκεκριμένα βάρη εκείνα τα μοντέλα τα οποία περιέχουν την μεταβλητή j. Αν τελικά η μεταβλητή j είναι πληροφοριακή, τότε αναμένουμε η πιθανότητα ένταξης του θα είναι μεγάλη καθώς εμπεριέχεται σε πολλά μοντέλα τα οποία έχουν και μεγάλη βαρύτητα. Σε αντίθετη περίπτωση αναμένουμε μικρή πιθανότητα καθώς εμπεριέχεται από μοντέλα μικρής βαρύτητας. Τιμές της πιθανότητας ένταξης μεγαλύτερες του 0.5 μας δείχνει πως η επεξηγηματική μεταβλητή j θεωρείται σημαντική.

5.2 Αλγόριθμοι για επιλογή μοντέλου

Οι παραδοσιακές μέθοδοι MCMC που παρουσιάστηκαν στο Κεφάλαιο 4 ήταν θεμελιώδης για την δειγματοληψία από την εκ των υστέρων κατανομή ενός μοντέλου, ειδικά σε πολυδιάστατα προβλήματα. Ωστόσο, αυτές οι μέθοδοι αντιμετωπίζουν σημαντικές δυσκολίες στην Μπεϋζιανή επιλογή μοντέλου καθώς η διάσταση των διαφορετικών μοντέλων δεν είναι σταθερή. Στο κεφάλαιο αυτό, θα δούμε τέσσερις αλγορίθμους οι οποίοι αντιμετωπίζουν το παραπάνω πρόβλημα με διαφορετικούς τρόπους.

5.2.1 Reversible Jump Markov Chain Monte Carlo

Το έργο του Peter J. Green (1995) επεκτείνει την ιδέα του Metropolis-Hastings αλγορίθμου εφαρμόζοντας την σε προβλήματα μεταβαλλόμενης διάστασης. Με αυτόν τον τρόπο το Reversible Jump Markov Chain Monte Carlo (RJMCMC) δίνει λύση στην σύγκριση μοντέλων διαφορετικής διάστασης. Αυτή η μεθοδολογία δεν δίνει μόνο περαιτέρω αξία στην Μπεϋζιανή επιλογή μοντέλου αλλά επεκτείνει και σε γενικότερο βαθμό τις δυνατότητες της Metropolis-Hastings μεθοδολογίας.

Στην επιλογή μοντέλου η έκφραση *"Ο αριθμός των πραγμάτων που δεν γνωρίζουμε είναι ένα από τα πράγματα που δεν γνωρίζουμε."*[3] είναι ένα από τα "παράδοξα" ζητήματα που καλούμαστε να αντιμετωπίσουμε. Συγκεκριμένα, στην Μπεϋζιανή στατιστική δεν υπάρχει μόνο η αβεβαιότητα στις παραμέτρους του μοντέλου αλλά και για το ίδιο το μοντέλο. Ο λόγος πηγάζει από το γεγονός που στην πράξη θέλουμε να επιλέξουμε το πραγματικό ή το πιο κατάλληλο μοντέλο με αποτέλεσμα να υπάρχει το στοιχείο της αβεβαιότητας και στον αριθμό παραμέτρων. Η ιδέα του Reversible Jump είναι να επιτρέπει στην διάσταση του παραμετρικού χώρου του μοντέλου να μεταβάλλεται και στο εσωτερικό κάθε διάστασης μοντέλου να λαμβάνονται οι παράμετροι του μοντέλου με κλασικό Metropolis-Hastings. Πηδήματα μεταξύ μοντέλων από έναν υπόχωρο σε έναν άλλο, τίθεται θέμα προς συζήτηση.

Υποθέτουμε ένα σύνολο από υποψήφια μοντέλα $\{M_1, M_2, ...\}$ τα οποία περιγράφουν τα δεδομένα $\boldsymbol{y}$. Κάθε μοντέλο M_k συνδέεται με ένα τυχαίο διάνυσμα παραμέτρων $\boldsymbol{\beta}_k$ με εκ των προτέρων πυκνότητα $p(\boldsymbol{\beta}_k|M_k)$ στον παραμετρικό χώρο $\mathcal{B}_k$, την συνάρτηση πιθανοφάνειας $f(\boldsymbol{y}|\boldsymbol{\beta}_k, M_k)$ και την εκ των προτέρων πιθανότητα $p(M_k)$ για το ίδιο το μοντέλο. Αυτό που μας ενδιαφέρει εμάς είναι να κάνουμε συμπερασματολογία για το μοντέλο αλλά και για τις παραμέτρους του μοντέλου. Το παραπάνω βασίζεται στον υπολογισμό της από κοινού εκ των υστέρων κατανομής

$$p(M_k, \boldsymbol{\beta}_k|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}_k, M_k)p(\boldsymbol{\beta}_k|M_k)p(M_k)}{\sum_j (\int_{\mathcal{B}_j} f(\boldsymbol{y}|\boldsymbol{\beta}_j, M_j)p(\boldsymbol{\beta}_j|M_j)\, d\boldsymbol{\beta}_j)p(M_j)} \tag{5.2.1}$$

[3] Peter J. Green 1995

Η εκ των υστέρων κατανομή (5.2.1) ορίζεται στο σύνολο $\mathcal{B} = \bigcup_k \{k\} \times \mathcal{B}_k$. Η ερώτηση η οποία καλούμαστε να απαντήσουμε είναι το πως θα γίνει το βήμα από $k \times \mathcal{B}_k$ στο $k' \times \mathcal{B}_{k'}$. Αυτό που θέλουμε είναι να προτείνουμε έναν πυρήνα μετάβασης $q(k \to k')$. Ο παραπάνω πυρήνας μας δίνει την πιθανότητα να μετακινηθούμε από το μοντέλο M_k στο $M_{k'}$. Στην συνέχεια θα πρέπει να προτείνουμε μετακίνηση σημείου από $\boldsymbol{\beta}_k$ στο $\boldsymbol{\beta}_{k'}$, διάστασης d_k και $d_{k'}$ αντίστοιχα. Οπότε, υποθέτοντας για μια στιγμή πως το $dim(x) < dim(x')$, για να μπορούμε να πηδήξουμε από το $x = (k, \boldsymbol{\beta}_k)$ στο $x' = (k', \boldsymbol{\beta}_{k'})$ πρέπει να κάνουμε αντιστοίχηση διαστάσεων μεταξύ των δύο συνόλων. Για αυτόν τον λόγο, εισάγουμε την βοηθητική μεταβλητή $u_{k\to k'} \sim \varphi_{k\to k'}$ όπου πλέον το $\boldsymbol{\beta}_{k'}$ θα είναι συνάρτηση του $\boldsymbol{\beta}_k$ και $u_{k\to k'}$, δηλαδή, $\boldsymbol{\beta}_{k'} = G_{k\to k'}(\boldsymbol{\beta}_k, u_{k\to k'})$. Για την περίπτωση $dim(x) > dim(x')$, δηλαδή, να πηδήξουμε από μεγαλύτερη διάσταση σε μικρότερη, θα πρέπει να γεννήσουμε τυχαίο διάνυσμα $u' \sim \varphi' \Rightarrow u_{k'\to k} \sim \varphi_{k\to k'}$ με τέτοιον τρόπο ώστε να ικανοποιείται η σχέση (5.2.2). Τότε ορίζουμε έναν αντίστροφο μετασχηματισμό $\boldsymbol{\beta}_k = G_{k'\to k}(\boldsymbol{\beta}_{k'}, u_{k'\to k})$.

Οπότε για να έχουμε μεταβάσεις σε διαφορετικές διαστάσεις θα πρέπει για δύο μοντέλα $(\boldsymbol{\beta}_k, u_{k\to k'})$ και $(\boldsymbol{\beta}_{k'}, u_{k'\to k})$ να ισχύει

$$d_k + dim(u_{k\to k'}) = d_{k'} + dim(u_{k'\to k}) \tag{5.2.2}$$

Τώρα που έχουμε προτάσεις από $\mathcal{B}_k$ στο $\mathcal{B}_{k'}$ ποια είναι η πιθανότητα αποδοχής; Εκμεταλλευόμαστε την ιδιότητα χρονικής αντιστρεψιμότητας από το Κεφάλαιο §4.1 για να αντλήσουμε την πιθανότητα αποδοχής $\alpha(x \to x')$ όπως και το Metropolis-Hastings. Η πιθανότητα αποδοχής εισάγεται στην χρονική αντιστρεψιμότητα, όπου η μετακίνηση στο x' γίνεται αποδεκτή με πιθανότητα $\alpha(x \to x')$. Εάν απορριφθεί τότε η αλυσίδα παραμένει στην κατάσταση x. Υπό αυτές τις συνθήκες έχουμε

$$\int_{(x,x')\in A\times B} p(x)q(x \to x')\alpha(x \to x')\, dxdx' = \int_{(x,x')\in A\times B} p(x')q(x' \to x)a(x' \to x)\, dx'dx \tag{5.2.3}$$

Με αποτέλεσμα από την εξίσωση (5.2.3) να προκύπτει η πιθανότητα αποδοχής

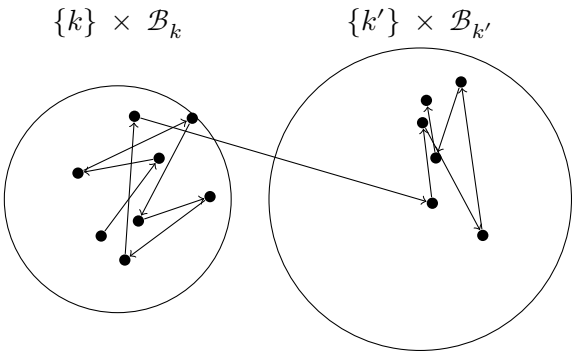
$$\alpha_{k\to k'} = \min\left\{1, \frac{p(\boldsymbol{\beta}_{k'})\varphi_{k'\to k}(u_{k'\to k})q(k' \to k)}{p(\boldsymbol{\beta}_k)\varphi_{k\to k'}(u_{k\to k'})q(k \to k')} J_{k\to k'}(\boldsymbol{\beta}_k, u_{k\to k'})\right\} \tag{5.2.4}$$

όπου

$$J_{k\to k'}(\boldsymbol{\beta}_k, u_{k\to k'}) = \left|\frac{\partial G_{k\to k'}(\boldsymbol{\beta}_k, u_{k\to k'})}{\partial(\boldsymbol{\beta}_k, u_{k\to k'})}\right|$$

Η πιθανότητα αποδοχής (5.2.4) είναι στην ουσία μια πιθανότητα αποδοχής επόμενου μοντέλου και σημείου. Ο όρος $\frac{p(\boldsymbol{\beta}_{k'})}{p(\boldsymbol{\beta}_k)}$ αναφέρεται στο πόσο εύλογη είναι η μετακίνηση από το $\boldsymbol{\beta}_k$ στο $\boldsymbol{\beta}_{k'}$. Ο όρος $\frac{\varphi_{k'\to k}(u_{k'\to k})q(k'\to k)}{\varphi_{k\to k'}(u_{k\to k'})q(k\to k')}$ αναφέρεται στις πιθανές ασύμμετρες κατανομές πρότασης των παραμέτρων (μέσω της φ) και των μοντέλων (μέσω της q). Παρ' όλο αυτά, οι όροι αυτοί δεν είναι συγκρίσιμοι καθώς

δεν έχουμε λάβει υπόψη την γεωμετρία των διαφορετικών διαστάσεων ή τον μετασχηματισμό του $\beta_{k'}$ λόγου της $G_{k\to k'}$. Για αυτόν τον λόγο εισάγουμε και την Ιακωβιανή $J_{k\to k'}(\boldsymbol{\beta}_k, u_{k\to k'})$ το οποίο επιτρέπει πλέον την σωστή σύγκριση.



Διάγραμμα 5.2.1: Απεικόνιση του RJMCMC.

Έχοντας πλέον περιγράψει την διαδικασία των μεταξύ-μοντέλων (”between-models”), μπορούμε να πούμε για τον συνολικό αλγόριθμο ο οποίος επιτρέπει την δειγματοληψία από την εκ των υστέρων κατανομή που ορίζεται $\mathcal{B} = \bigcup_k \{k\} \times \mathcal{B}_k$. Για κάθε μοντέλο M_k εισάγουμε ένα εντός-μοντέλο (”within-model”) το οποίο αξιοποιεί τον κλασικό Metropolis-Hastings αλγόριθμο με πυρήνα μετάβασης F_k. Με πιθανότητα p θα κάνουμε ”within-model” κινήσεις, το οποίο θα μπορούσαμε να το θέσουμε κοντά στην μονάδα, με σκοπό οι περισσότερες κινήσεις να έχουν να κάνουν με τα ”within-models” και κάποιες περιστασιακές ”between-models” κινήσεις με πιθανότητα $1-p$.

Ξεκινάμε από το μοντέλο $k^{(0)}$ και παράμετρο $\boldsymbol{\beta}_k^{(0)} \in \mathcal{B}_{k^{(0)}}$.

Αλγόριθμος 5.2.1 : Reversible Jump Markov Chain Monte Carlo.

---

1. Αρχικοποίηση $(k^{(0)}, \boldsymbol{\beta}_k^{(0)})$ και **επανάλαβε** για $t = 1, 2, 3, ...$

2. Με πιθανότητα p, θέσε $k^{(t)} = k^{(t-1)}$ και εκτέλεσε ένα βήμα $F_{k^{(t)}}$.

3. Με πιθανότητα $1-p$, δώσε πρόταση για μοντέλο $k' \sim q(k'|k^{(t-1)})$ και προσομοίωσε μια τυχαία μεταβλητή $u_{k^{(t-1)}\to k'} \sim \varphi_{k^{(t-1)}\to k'}$ και από την $G_{k^{(t-1)}\to k'}$ λάβε την πρόταση $\boldsymbol{\beta}' \in \mathcal{B}_{k'}$.
Με πιθανότητα

$$\alpha(\boldsymbol{\beta}^{(t-1)} \to \boldsymbol{\beta}') = \min\left\{1, \frac{p(\boldsymbol{\beta}')\varphi_{k'\to k^{(t-1)}}(u_{k'\to k^{(t-1)}})q(k^{(t-1)}|k')}{p(\boldsymbol{\beta}^{(t-1)})\varphi_{k^{(t-1)}\to k'}(u_{k^{(t-1)}\to k'})q(k'|k^{(t-1)})} J_{k^{(t-1)}\to k'}(\boldsymbol{\beta}', u_{k^{(t-1)}\to k'})\right\}$$

Δέξου, δηλαδή, όρισε $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}'$, $k^{(t)} = k'$. Αλλιώς απόρριψε, δηλαδή, θέσε $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$, $k^{(t)} = k^{(t-1)}$.

Συνοψίζοντας, ο Reversible Jump αλγόριθμος θα χρησιμοποιεί επαναληπτικά “within-model” και “between-models” πυρήνες μετάβασης. Στην “within-model” περίπτωση χρησιμοποιούμε το κλασικό Metropolis-

Hastings αλγόριθμο και στο “between-models” κατασκευάσαμε την μετάθεση από έναν υπόχωρο σε έναν άλλο.

Αξίζει να σημειώσουμε πως το RJMCMC είναι μια γενική μεθοδολογία για οποιοδήποτε πρόβλημα επιλογής μοντέλων, σε αντίθεση με άλλους αλγορίθμους όπως το SSVS (που θα δούμε στην συνέχεια) που είναι τεχνική ειδικά σχεδιασμένη για πρόβλημα επιλογής μεταβλητών σε μοντέλα παλινδρόμησης.

5.2.2 Stochastic Search Variable Selection

Ο George, Edward I και Robert E. McCulloch (1993) παρουσίασαν στο έργο τους μια διαφορετική μεθοδολογία για την Μπεϋζιανή επιλογή μοντέλου, το Stochastic Search Variable Selection (SSVS). Ο παραπάνω αλγόριθμος είναι βασισμένος σε μια γενικότερη κατηγορία Gibbs μεθοδολογίας για την επιλογή μοντέλου. Με λίγα λόγια, ο αλγόριθμος ψάχνει με έναν στοχαστικό τρόπο τα πιο ”πιθανά” υποσύνολα μεταβλητών. Εκείνα τα μοντέλα μπορούν να εντοπιστούν από την συχνότητα που έχουν εμφανιστεί κατά την δειγματοληψία.

Η μεθοδολογία SSVS θεωρεί ένα κανονικό γραμμικό μοντέλο $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ με p πιθανές επεξηγηματικές μεταβλητές (χωρίς την σταθερά). Όπου $\mathbf{Y}$ ένα $n \times 1$ διάνυσμα, $\mathbf{X}$ ο πίνακας σχεδιασμού διάστασης $n \times p$, $\boldsymbol{\beta}$ το διάνυσμα με τις παραμέτρους του μοντέλου διάστασης $p \times 1$. Γνωρίζουμε πως, η επιλογή ενός μοντέλου είναι ισοδύναμη με το να θεωρήσουμε πως οι αντίστοιχες παράμετροι $\boldsymbol{\beta}_j$ δεν συμπεριλαμβάνονται στο μοντέλο, δηλαδή, εκείνα τα $\boldsymbol{\beta}_j$ ισούνται με μηδέν. Το παραπάνω σημαίνει πως ο αντίστοιχος πίνακας σχεδιασμού για κάθε μοντέλου θα είναι διαφορετικός ανάλογα με ποιες μεταβλητές συμπεριλαμβάνονται στο μοντέλο (άρα θα αλλάζει η διάσταση του μοντέλου). Αυτό που κάνει το SSVS είναι να θεωρήσει πως το $\mathbf{X}$ περιέχει όλες τις μεταβλητές οπότε και το $\boldsymbol{\beta}$ θα είναι σταθερής διάστασης,δηλαδή, δεν θα εξαρτάται από το μοντέλο. Βάσει αυτής της προσέγγισης οι παράμετροι που δεν περιέχονται στο μοντέλο δεν θα είναι ακριβώς μηδέν αλλά θα θεωρούνται απών όταν θα είναι σε μια γειτονιά γύρω από το μηδέν, πρακτικά αμελητέα. Έτσι, λόγου της σταθερής διάστασης των $\boldsymbol{\beta}$, το παραπάνω μας επιτρέπει να εφαρμόσουμε τον δειγματολήπτη Gibbs.

Αυτό μπορούμε να το καταφέρουμε θεωρώντας μια εκ των προτέρων κατανομή για το όταν συμπεριλαμβάνεται η j μεταβλητή στο μοντέλο $\boldsymbol{\beta}_j|\gamma_j = 1 \sim \mathcal{N}(0, \tau_j^2 c_j^2)$ και όταν δεν συμπεριλαμβάνεται έχουμε $\boldsymbol{\beta}_j|\gamma_j = 0 \sim \mathcal{N}(0, \tau_j^2)$. Οπότε μπορούμε να κάνουμε μια μείξη κανονικών κατανομών ως μια εκ των προτέρων κατανομή των $\boldsymbol{\beta}_j$.

$$\boldsymbol{\beta}_j|\gamma_j \sim (1-\gamma_j)\mathcal{N}(0, \tau_j^2) + \gamma_j\mathcal{N}(0, \tau_j^2 c_j^2) \tag{5.2.5}$$

Η γ_j είναι μια δίτιμη μεταβλητή η οποία μετρά το αν υπάρχει ή όχι η j μεταβλητή. Οι υπερ-παράμετροι της (5.2.5) παίζουν τον ρόλο της πληροφοριακής ή μη-πληροφοριακής εκ των προτέρων κατανομής. Συγκεκριμένα, επιλέγουμε το τ_j^2 να είναι μικρό και το $\tau_j^2 c_j^2$ να είναι μεγάλο. Αν $\gamma_j = 1$, δηλαδή, υπάρχει η παράμετρος β_j στο μοντέλο τότε έχουμε μια μη-πληροφοριακή εκ των προτέρων κατανομή λόγου της μεγάλης διακύμανσης με αποτέλεσμα, η εκ των υστέρων κατανομή να καθορίζεται σε μεγάλο βαθμό από τα δεδομένα. Στην αντίθετη περίπτωση όπου $\gamma_j = 0$, δηλαδή, η παράμετρος β_j δεν υπάρχει στο μοντέλο τότε έχουμε μια ισχυρά πληροφοριακή εκ των προτέρων κατανομή καθώς η κατανομή της β_j αναγκάζεται να συγκεντρώνεται γύρω από το μηδέν.

Μπορούμε να γενικεύσουμε την εξίσωση (5.2.5) σε μορφή πίνακα ως

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}\mathbf{D}_{\boldsymbol{\gamma}}) \tag{5.2.6}$$

Όπου $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)^\top$, $\mathbf{D}_\gamma$ ένας διαγώνιος πίνακας με στοιχεία $(\alpha_1\tau_1, ..., \alpha_p\tau_p)$ όπου αν $\gamma_j = 1$ τότε $\alpha_j = c_j$ ή αν $\gamma_j = 0$ τότε $\alpha_j = 1$. Η επιλογή των υπερ-παραμέτρων c_j^2 και τ_j^2 πρέπει να είναι τέτοια ώστε ο αλγόριθμος να μπορεί εξερευνήσει αποδοτικά τον χώρο. Παρατηρήστε πως για $\tau_j^2 \to 0$ η αποδοτικότητα του αλγορίθμου θα πέσει σημαντικά καθώς θα κάνει πολύ μικρές κινήσεις. Τέλος, ο πίνακας $\mathbf{R}$ ο οποίος είναι ο εκ των προτέρων πίνακας συσχετίσεων. Δύο συνηθισμένες επιλογές για το $\mathbf{R}$ είναι $\mathbf{R} = \mathbf{I}$, δηλαδή, τα β είναι ανεξάρτητα εκ των προτέρων ή $\mathbf{R} = n(\mathbf{X}^\top\mathbf{X})^{-1}$ η g-prior του Zellner που συζητήθηκε και στην Ενότητα §3.2.1.

Στην συνέχεια θα θέλαμε και μια εκ των προτέρων κατανομή για το σ^2. Μια βολική και λογική, επιλογή είναι η αντίστροφη-Γάμμα που είχαμε συζητήσει και στην Ενότητα §3.2.1.

$$\sigma^2|\boldsymbol{\gamma} \sim \mathcal{IG}(\nu_{\boldsymbol{\gamma}}/2, \nu_{\boldsymbol{\gamma}}\lambda_{\boldsymbol{\gamma}}/2) \tag{5.2.7}$$

Μπορούμε να κάνουμε την εκ των προτέρων κατανομή (5.2.7) μη-πληροφοριακή θέτοντας τις υπερ-παραμέτρους $v_{\boldsymbol{\gamma}} \to 0$ και $\lambda_{\boldsymbol{\gamma}}$ μια οποιαδήποτε τιμή καθώς δεν συνεισφέρει στην εκ των υστέρων κατανομή.

Τέλος, θα πρέπει να έχουμε μια εκ των προτέρων κατανομή για τα ίδια τα μοντέλα. Ο μηχανισμός πιθανοτήτων που θα χρησιμοποιήσουμε θα πρέπει να μας δώσει δείγματα της μορφής $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)^\top$. Μια τέτοια κατανομή θα μπορούσε να ήταν

$$p(\boldsymbol{\gamma}) = \prod_j p_j^{\gamma_j}(1 - p_j)^{(1-\gamma_j)} \tag{5.2.8}$$

Η $p(\gamma)$ είναι η από κοινού εκ των προτέρων συνάρτηση μάζας ανεξαρτήτων $Bernoulli$ δοκιμών. Στην ουσία η εξίσωση (5.2.8), θα μας δώσει την εκ των προτέρων πιθανότητα του μοντέλου γ. Παρ' όλο που η υπόθεση ανεξαρτησίας μπορεί να μην φαίνεται τόσο λογική, οι συγγραφείς υποστηρίζουν πως η παραπάνω υπόθεση λειτουργεί αρκετά καλά σε διάφορες περιπτώσεις. Αξίζει να σημειωθεί πως η "αντικειμενική" εκ των προτέρων κατανομή $p(\gamma) = 1/2^p$ είναι μια ειδική περίπτωση της (5.2.8), όταν δηλαδή, το κάθε $\mathbf{X}_j$ έχει ίση πιθανότητα ένταξης $(p_j = 1/2)$. Εφόσον η (5.2.8) είναι μια γενική περίπτωση, τότε θα μπορούσαμε να βάλουμε και αντίστοιχα βάρη για να έχουμε μια πιο δίκαια ή αντικειμενική εκ των προτέρων κατανομή.

Όλη αυτή η διαδικασία οδηγεί σε υψηλή εκ των υστέρων πιθανότητα να συγκεντρώνετε στα μοντέλα εκ των οποίων οι παράμετροι τους διαφέρουν σημαντικά από το μηδέν, δηλαδή, στα πιο πιθανά μοντέλα. Ωστόσο, η από κοινού εκ των υστέρων κατανομή $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\gamma})p(\sigma^2|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ (γινόμενο κανονικών, αντίστροφης Γάμμα και Bernoulli) είναι μια κατανομή που δεν μπορούμε να αναγνωρίσουμε. Σε αυτό το σημείο, μπορούμε να χρησιμοποιήσουμε τον δειγματολήπτη Gibbs καθώς οι πλήρες δεσμευμένες κατανομές είναι γνωστές. Ο κύριος λόγος για την χρήση της ιεραρχικής δομής στις παραμέτρους είναι για να μπορούμε να πάρουμε την περιθωριακή εκ των υστέρων κατανομή $p(\boldsymbol{\gamma}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ η οποία περιέχει την πληροφορία για την επιλογή μεταβλητών. Η εκ των υστέρων κατανομή $p(\boldsymbol{\gamma}|\mathbf{Y})$ αναβαθμίζει τις εκ των προτέρων πιθανότητες για κάθε ένα από τις 2^p δυνατές τιμές του διανύσματος $\boldsymbol{\gamma}$.

Ωστόσο, από τις σχέσεις (5.2.6) - (5.2.8) προκύπτει μια ιεραρχική μείξη εκ των προτέρων κατανομών $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}|\boldsymbol{\gamma})p(\sigma^2|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$. Παρατηρούμε πως αυτή η κατανομή δεν είναι συζυγής καθώς $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2)$, δηλαδή, έχουμε εκ των προτέρων ανεξαρτησία μεταξύ των $\boldsymbol{\beta}$ και σ^2. Προφανώς, θα μπορούσαμε να υποθέσουμε τελικά μια συζυγής ιεραρχική μίξη εκ των προτέρων κατανομών, δηλαδή, $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma})p(\sigma^2|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ το οποίο μας μας δίνει αναλυτικά την $p(\boldsymbol{\gamma}|\mathbf{Y})$. Ωστόσο, το κόστος στους αναλυτικούς υπολογισμούς της $p(\boldsymbol{\gamma}|\mathbf{Y})$ μπορούν να γίνουν για σχετικά μικρό p και με MCMC για μεγαλύτερο.

Παρ' όλο, που ο αναλυτικός υπολογισμός της $p(\boldsymbol{\gamma}|\mathbf{Y})$ δεν είναι εφικτός στη δική μας περίπτωση, μπορούμε να χρησιμοποιήσουμε την μορφή της $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}|\mathbf{Y})$ σε συνδυασμό με τον δειγματολήπτη Gibbs για να εξερευνήσουμε την εκ των υστέρων κατανομή $p(\boldsymbol{\gamma}|\mathbf{Y})$.

Μπορούμε να κατασκευάσουμε μια Μαρκοβιανή αλυσίδα με αρχικές τιμές $\boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}, \boldsymbol{\gamma}^{(0)}$ η οποία

γεννά μια ακολουθία τιμών

$$\boldsymbol{\beta}^{(1)}, \sigma^{2^{(1)}}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\beta}^{(2)}, \sigma^{2^{(2)}}, \boldsymbol{\gamma}^{(2)}, ... \tag{5.2.9}$$

θα συγκλίνουν στην εκ των υστέρων κατανομή $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}|\mathbf{Y})$. Με την ίδια λογική η υποακολουθία τιμών

$$\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, ... \tag{5.2.10}$$

θα συγκλίνει επίσης στην περιθώρια εκ των υστέρων κατανομή $p(\boldsymbol{\gamma}|\mathbf{Y})$. Σε σχετικά μικρό p ο αλγόριθμος θα εξερευνήσει όλη την $p(\boldsymbol{\gamma}|\mathbf{Y})$, σε διαφορετική περίπτωση πάλι θα μπορούσε να μας παρέχει χρήσιμη πληροφορία.

Η ακολουθία τιμών (5.2.9) και κατ επέκταση η υποακολουθία (5.2.10) αποκτώνται από διαδοχική προσομοίωση από τις πλήρες δεσμευμένες κατανομές

$$\begin{gathered} p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}, \mathbf{Y}) \\ p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{Y}) = p(\sigma^2|\boldsymbol{\beta}, \mathbf{Y}) \\ p(\gamma_j|\boldsymbol{\beta}, \sigma^2, \gamma_{-j}, \mathbf{Y}) = p(\gamma_j|\boldsymbol{\beta}, \gamma_{-j}) \end{gathered} \tag{5.2.11}$$

Παρατηρούμε πως η δεύτερη εξίσωση στην πραγματικότητα δεσμεύεται μόνο από τα $\boldsymbol{\beta}$ το οποίο έρχεται σε αντιπαράθεση με την ιεραρχική δομή που υποθέσαμε. Η παραπάνω συμπεριφορά οφείλεται στον δειγματολήπτη Gibbs ο οποίος χρειάζεται τα $\boldsymbol{\beta}$ για να κάνει τις διαδοχικές επαναλήψεις. Η τρίτη εξίσωση, επίσης δεσμεύεται στην πραγματικότητα μόνο από τα $\boldsymbol{\beta}$. Αυτό συμβαίνει επειδή η πληροφορία των δεδομένων και για το αν η j μεταβλητή είναι σημαντική, εμπεριέχεται στα $\boldsymbol{\beta}$. Επιπλέον το διάνυσμα $\boldsymbol{\gamma}$ προσομοιώνεται ανά συνιστώσα γ_j δοθέντος τα $\boldsymbol{\beta}$ και τα $\boldsymbol{\gamma}_{-j} = (\gamma_1, ..., \gamma_{j-1}, \gamma_{j+1}, ..., \gamma_p)^\top$ όπου $p(\gamma_j|\boldsymbol{\beta}, \gamma_{-j}) \propto p(\boldsymbol{\gamma}|\boldsymbol{\beta})$. Η διαδικασία με την οποία λαμβάνουμε το $\boldsymbol{\gamma}^{(1)}$ είναι

Η επιτυχία του παραπάνω αλγορίθμου οφείλεται στο γεγονός πως η δειγματοληψία γίνεται από γνωστές κατανομές.

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}, \mathbf{Y}) \sim \mathcal{N}_p((\mathbf{X}^\top\mathbf{X} + \sigma^2(\mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}})^{-1})^{-1}\mathbf{X}^\top\mathbf{Y}, \sigma^2(\mathbf{X}^\top\mathbf{X} + \sigma^2(\mathbf{D}_{\boldsymbol{\gamma}}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{D}_{\boldsymbol{\gamma}})^{-1})^{-1})$$

$$p(\sigma^2|\boldsymbol{\beta}, \mathbf{Y}) \sim \mathcal{IG}\left(\frac{n + \nu_{\boldsymbol{\gamma}}}{2}, \frac{|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}|^2 + \nu_{\boldsymbol{\gamma}}\lambda_{\boldsymbol{\gamma}}}{2}\right)$$

Αλγόριθμος 5.2.2 : Stochastic Search Variable Selection.

1. Αρχικοποίηση $(\boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}, \gamma_1^{(0)}, ..., \gamma_p^{(0)})$ και **επανάλαβε** για $t = 1, 2, 3, ...$

2. Προσομοίωσε $p(\boldsymbol{\beta}^{(t)}|\sigma^{2^{(t-1)}}, \boldsymbol{\gamma}^{(t-1)})$

3. Προσομοίωσε $p(\sigma^{2^{(t)}}|\boldsymbol{\beta}^{(t)}, \mathbf{Y})$

4. Δοθέν $(\gamma_1^{(t-1)}, ..., \gamma_p^{(t-1)})$
$\gamma_1^{(t)} \sim p(\gamma_1|\boldsymbol{\beta}^{(t)}, \gamma_2^{(t-1)}, ..., \gamma_p^{(t-1)})$

$\gamma_2^{(t)} \sim p(\gamma_2|\boldsymbol{\beta}^{(t)}, \gamma_1^{(t)}, \gamma_3^{(t-1)}, ..., \gamma_p^{(t-1)})$

$\vdots$

$\gamma_p^{(t)} \sim p(\gamma_p|\boldsymbol{\beta}^{(t)}, \gamma_1^{(t)}, \gamma_2^{(t)}, ..., \gamma_{p-1}^{(t)})$

Τέλος, κάθε γ_j προσομοιώνεται με τυχαία σειρά από την κατανομή $\mathcal{B}ernoulli$ με πιθανότητα

$$p(\gamma_j = 1|\boldsymbol{\beta}, \gamma_{-j}) = \frac{a}{a+b}$$

Όπου

$$a = p(\boldsymbol{\beta}|\gamma_{-j}, \gamma_j = 1)p(\gamma_{-j}, \gamma_j = 1)$$
$$b = p(\boldsymbol{\beta}|\gamma_{-j}, \gamma_j = 0)p(\gamma_{-j}, \gamma_j = 0)$$

Κάτι που έχουμε αγνοήσει μέχρι στιγμής είναι η σταθερά β_0 του μοντέλου. Μπορούμε να αντιμετωπίσουμε την σταθερά β_0, βάζοντας την απλώς στον πίνακα σχεδιασμού την στήλη $\mathbf{X}_0 = (1, ..., 1)^\top$ και εκ των προτέρων πιθανότητα ίση με μονάδα, δηλαδή, $\mathbb{P}(\gamma_0 = 1) = 1 = p_0$.

5.2.3 Gibbs Variable Selection

Σε αυτή την Ενότητα θα εξερευνήσουμε τον αλγόριθμο Gibbs Variable Selection (GVS) ο οποίος εισήχθηκε από τον Ioannis Ntzoufras (1999) και Dellaportas Petros et al (2002). Συγκεκριμένα, αυτός ο αλγόριθμος είναι ένας φυσικός συνδυασμός του SSVS που συζητήσαμε στην προηγούμενη ενότητα, και του Kuo Lynn και Bani Mallick (1998), KM αλγορίθμου. Ο στόχος του GVS αλγορίθμου είναι να συνδυάσει τα θετικά του SSVS και του KM δειγματολήπτη.

Ορίζουμε για τα στατιστικά μοντέλα την γνωστή δίτιμη μεταβλητή γ και p τον αριθμό των μεταβλητών στο μοντέλο. Επιπλέον, μπορούμε να γράψουμε τον γραμμικό όρο του μοντέλου ως

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \gamma_j \mathbf{X}_j \boldsymbol{\beta}_j \tag{5.2.12}$$

Παρατηρούμε πως από την εξίσωση (5.2.12) ο πίνακας σχεδιασμού $\mathbf{X}$ και άρα η διάσταση του μοντέλου είναι σταθερή. Αυτό που αλλάζει είναι η πιθανοφάνεια καθώς εξαρτάται από το γ. Στην συνέχεια, θεωρούμε την εκ των προτέρων κατανομή $p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$ όπου η $p(\boldsymbol{\gamma})$ μας λέει για την δομή των μοντέλων και το $p(\boldsymbol{\beta}|\boldsymbol{\gamma})$ λέει για τα $\boldsymbol{\beta}$ δοθέντος του μοντέλου από την $p(\boldsymbol{\gamma})$.

Λόγου της (5.2.12) θα σπάσουμε το διάνυσμα $\boldsymbol{\beta}$ σε δύο διανύσματα $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ και $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$. Η $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ περιέχει τις παραμέτρους για τις μεταβλητές που εμπεριέχονται στο μοντέλο και η $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ εκείνες που δεν συμπεριλαμβάνονται. Το παραπάνω έχει ως συνέπεια η πιθανοφάνεια να γράφεται στην πραγματικότητα ως $f(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ και η εκ των προτέρων κατανομή ως

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})$$

Επειδή οι όροι του διανύσματος $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ δεν θα υπάρχουν στο μοντέλο, η αντίστοιχη πιθανοφάνεια δεν θα επηρεαστεί από αυτά. Με αποτέλεσμα, η εκ των υστέρων κατανομή για κάθε μοντέλο γ να είναι της μορφής

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{y}) = p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y})p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \tag{5.2.13}$$

Η εξίσωση (5.2.13) επιβεβαιώνει πως η διάσταση του μοντέλου έχει παραμείνει σταθερή, σπάζοντας την εκ των υστέρων κατανομή των $\boldsymbol{\beta}$, στην πραγματική εκ των υστέρων κατανομή $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{y})$ για το μοντέλο γ και την εκ των προτέρων κατανομή $p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ των $\boldsymbol{\beta}$ που δεν υπάρχουν στο μοντέλο. Η $p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ ονομάζεται ψευδό εκ των προτέρων κατανομή και κατά αναλογία η $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})$ είναι η πραγματική εκ των προτέρων κατανομή των μοντέλων. Η ψευδό εκ των προτέρων κατανομή δεν επηρεάζεται από τα δεδομένα (αφού δεν υπάρχει στο μοντέλο) και έτσι δεν επηρεάζει την εκ των υστέρων

κατανομή.

Η ψευδό εκ των προτέρων κατανομή είναι χρήσιμη καθώς λειτουργεί ως μια βοηθητική μεταβλητή που κρατά την διάσταση του μοντέλου σταθερή για να είναι εφικτός ο δειγματολήπτης Gibbs[4]. Ωστόσο, η παραπάνω ποσότητα παρ' όλο που δεν παίζει ρόλο στην εκ των υστέρων κατανομή, επηρεάζει σε σημαντικό βαθμό τον MCMC αλγόριθμο μας. Οι συγγραφείς τονίζουν πως η επιλογή της $p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ πρέπει να βοηθά στην αποδοτικότητα του αλγορίθμου.

Αλγόριθμος 5.2.3 : Gibbs Variable Selection.

---

1. Προσομοίωσε τις παραμέτρους που υπάρχουν στο μοντέλο από την εκ των υστέρων κατανομή

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\beta}_{-\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})\, p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\, p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \tag{5.2.14}$$

2. Προσομοίωσε τις παραμέτρους που δεν υπάρχουν στο μοντέλο από την ψευδό εκ των προτέρων κατανομή

$$p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$$

3. Προσομοίωσε ξεχωριστά κάθε γ_j από μια $\mathcal{B}ernoulli$ κατανομή με πιθανότητα επιτυχίας $p = O_j/(1+O_j)$ με O_j

$$O_j = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma_j = 0, \boldsymbol{\gamma}_{-j})} \frac{p(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{p(\boldsymbol{\beta}|\gamma_j = 0, \boldsymbol{\gamma}_{-j})} \frac{p(\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{p(\gamma_j = 0, \boldsymbol{\gamma}_{-j})}$$

Όπως σημειώθηκε και από Dellaportas Petros et al (2002) η εξάρτηση της πλήρης δεσμευμένη κατανομή (5.2.14) στην $p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ δεν φαίνεται τόσο λογική αλλά θα μπορούσε να ήταν χρήσιμη στην περίπτωση πολυσυγγραμικότητας. Ένας πιο βολικός και πιθανός τρόπος αντιμετώπισης είναι να θεωρήσουμε ανεξαρτησία μεταξύ των $\boldsymbol{\beta}$ (που θα δούμε στην συνέχεια) και έτσι ο όρος $p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ από την εξίσωση (5.2.14) μπορεί να παραλειφθεί.

Η επιλογή εκ των προτέρων κατανομή $p(\boldsymbol{\beta}|\boldsymbol{\gamma})$, με $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{-\boldsymbol{\gamma}})^\top$ θα μπορούσε να ήταν της εξής μορφής

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{j=1}^{p} p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}_j)$$

Θεωρώντας δηλαδή ανεξαρτησία μεταξύ των $\boldsymbol{\beta}$, με την απλή επιλογή $p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}_j) = (1-\gamma_j)\, p(\boldsymbol{\beta}_j|\gamma_j = 0) + \gamma_j\, p(\boldsymbol{\beta}_j|\gamma_j = 1)$. Το παραπάνω οδηγεί την πραγματική εκ των προτέρων κατανομή να έχει την μορφή $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = \prod_{\gamma_j=1} p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}_j)$ και η ψευδό εκ των προτέρων κατανομή $p(\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) = \prod_{\gamma_j=0} p(\beta_j|\gamma_j)$. Όπως και στο SSVS μπορούμε να έχουμε μια μείξη κανονικών κατανομών $p(\boldsymbol{\beta}_j|\gamma_j = 0) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_j)$ και $p(\boldsymbol{\beta}_j|\gamma_j = 1) \sim \mathcal{N}(\bar{\mu}_j, \mathbf{S}_j)$. Μια επαρκή επιλογή των υπερ-παραμέτρων αυτής

[4] Υιοθετεί μια παρόμοια ιδέα όπως το RJMCMC.

της μείξης είναι $\bar{\mu}_j = 0$ και $\mathbf{S}_j = \mathbf{\Sigma}_j/k^2$ με $k = 10$ το οποίο ακολουθεί την ίδια λογική, πληροφοριακής και μη-πληροφοριακής εκ των προτέρων κατανομής που πρότειναν ο George, Edward I και Robert E. McCulloch (1993). Τέλος, η εκ των προτέρων κατανομή $p(\boldsymbol{\gamma})$ μπορεί να είναι όπως και η εξίσωση (5.2.8).

5.2.4 Bayesian Adaptive Sampling

Οι γνωστοί MCMC αλγόριθμοι που έχουμε μελετήσει μέχρι στιγμής παίρνουν δείγμα το οποίο είναι εξαρτημένο, αλλά έχουν και την ιδιότητα της επανατοποθέτησης. Οι Merlise A. Clyde et al (2011) παρουσίασαν τον Bayesian Adaptive Sampling (BAS) αλγόριθμο ο οποίος αντιμετωπίζει το πρόβλημα επιλογής μοντέλο στην γραμμική παλινδρόμηση παίρνοντας ανεξάρτητο δείγμα μοντέλων χωρίς επανατοποθέτηση από το σύνολο μοντέλων.

Το κίνητρο του BAS έχει να κάνει με τον τρόπο που οι MCMC αλγόριθμοι εξερευνούν το σύνολο των δυνατών μοντέλων. Προφανώς, σε προβλήματα με μικρό p η απαρίθμηση του χώρου των μοντέλων είναι δυνατή για κάθε αλγόριθμο. Εφόσον μπορεί να εξερευνήσει όλο τον χώρο ο αλγόριθμος, αυτό που μένει είναι να εκτιμήσει τις αντίστοιχες εκ των υστέρων πιθανότητες βάση της αντίστοιχη συχνότητα επίσκεψης στο μοντέλο. Επειδή, υπάρχει η δυνατότητα απαρίθμησης (καθώς θα επισκεφτούμε όλα τα μοντέλα) το δείγμα από την εκ των υστέρων κατανομή θα είναι αντιπροσωπευτικό. Από την άλλη πλευρά, σε πολύπλοκα προβλήματα με μεγάλο p η απαρίθμηση δεν είναι πλέον δυνατή, δηλαδή, οι αλγόριθμοι δεν μπορούν να εξερευνήσουν όλο τον χώρο σε πεπερασμένο χρόνο. Οπότε, αυτό που κάνουν οι MCMC αλγόριθμοι είναι απλώς η αναζήτηση μοντέλων με τις υψηλότερες πιθανότητες, δηλαδή, με τις υψηλότερες συχνότητες. Με αυτόν τον τρόπο μπορούμε να αγνοήσουμε εκείνα τα μοντέλα που δεν επισκεφτήκαμε ποτέ καθώς δεν είχαν σημαντική πιθανότητα. Ωστόσο, σε αυτό το σημείο οι συγγραφείς εντοπίζουν ένα σημαντικό μειονέκτημα. Παρ' όλο που δεν μπορούμε να έχουμε όλη την εκ των υστέρων κατανομή, θα θέλαμε τουλάχιστον ένα αντιπροσωπευτικό δείγμα από αυτή, δηλαδή, ένα αντιπροσωπευτικό υποσύνολο του $\mathcal{M}$. Στους MCMC αλγορίθμους όμως, η αλυσίδα θα ασχοληθεί σε σημαντικό βαθμό του πεπερασμένου χρόνου με ένα πολύ μικρό υποσύνολο του $\mathcal{M}$, δηλαδή, εκείνα τα μοντέλα με την υψηλότερη πιθανότητα και θα αγνοήσει σε μεγάλο βαθμό τα υπόλοιπα μοντέλα στο $\mathcal{M}$ τα οποία θα μπορούσαν να μας προσφέρουν χρήσιμη πληροφορία στις αποφάσεις. Οπότε, ο σκοπός του αλγόριθμου BAS είναι πάρει δείγμα μια φορά από το κάθε μοντέλο με σκοπό την αποδοτικότητα και ένα αντιπροσωπευτικό δείγμα από την εκ των υστέρων κατανομή.

Ο στόχος είναι η δειγματοληψία χωρίς επανατοποθέτηση από τον πεπερασμένο πληθυσμό $\mathcal{M}$. Η πιο εύκολη μέθοδος δειγματοληψίας που υπάρχει είναι η απλή τυχαία. Με την απλή τυχαία δειγματο-

ληψία, μπορούμε να πάρουμε ένα τυχαίο δείγμα μεγέθους N βάζοντας ίσες πιθανότητες στο κάθε μοντέλο στο $\mathcal{M}$. Η ίδια πιθανότητα μοντέλων προκαλεί σημαντικά προβλήματα καθώς μοντέλα με υψηλή εκ των υστέρων πιθανότητα έχουν κίνδυνο να αγνοηθούν, εκτός αν το δείγμα N είναι κοντά στο $|\mathcal{M}|$. Ένας τρόπος για να αντιμετωπίσουμε το παραπάνω πρόβλημα, είναι να βάλουμε την πιθανότητα δειγματοληψίας του μοντέλου να εξαρτάται από μια βοηθητική μεταβλητή η οποία μετρά την "σημαντικότητα" του μοντέλου. Έτσι, για κάθε δειγματοληψία μοντέλου θα γνωρίζουμε πως πρόκειται για μοντέλο με υψηλή εκ των υστέρων πιθανότητα. Ωστόσο, μετά από κάθε δειγματοληψία θα πρέπει η μάζα εκείνου του μοντέλου να αφαιρεθεί με τέτοιο τρόπο ώστε η επόμενη δειγματοληψία να είναι κανονικοποιημένη. Πρέπει να αναφερθεί πως στο έργο τους, οι συγγραφείς σημειώνουν πως σε περίπλοκα προβλήματα, η λίστα των μοντέλων και οι αντίστοιχες πιθανότητες τους, δεν θα είναι δυνατή καθώς πρόκειται για το ίδιο πρόβλημα μη απαρίθμησης των μοντέλων.

Για να παρακάμψουμε την ανάγκη για την λίστα μοντέλων και τις αντίστοιχες πιθανότητες τους είναι να ορίσουμε έναν μηχανισμό πιθανοτήτων το οποίο θα γεννά τυχαία από το $\mathcal{M}$.

$$f(\boldsymbol{\gamma}) = \prod_{j=1}^{p} p_j^{\gamma_j}(1-p_j)^{1-\gamma_j} \tag{5.2.15}$$

Η $f(\boldsymbol{\gamma})$ είναι η από κοινού συνάρτηση μάζας ανεξάρτητων $\mathcal{B}ernoulli$ δοκιμών. Όπου στην ουσία θα μας δώσει την πιθανότητα μοντέλου και θα μας γεννά $\mathrm{M} \in \mathcal{M}$ της μορφής $(\gamma_1, ..., \gamma_p)^\top$ με την αντίστοιχη πιθανότητα p_j, η μεταβλητή j να εμπεριέχεται στο μοντέλο.

Επειδή μιλάμε για μοντέλα, θα μπορούσαμε να γράψουμε γενικότερα πως οποιαδήποτε από κοινού συνάρτηση μάζας θα μπορούσε να γραφεί ως το γινόμενο δεσμευμένων κατανομών της μορφής

$$f(\boldsymbol{\gamma}) = \prod_{j=1}^{p} f(\gamma_j|\gamma_{<j}) \tag{5.2.16}$$

Όπου με τον συμβολισμό $\gamma_{<j}$ εννοούμε την ένταξη εκείνων των μεταβλητών $\{\gamma_k\}$ για $k < j$. Για $j = 1$, $f(\gamma_1|\gamma_{<1}) = f(\gamma_1)$. Εφόσον τα γ_j είναι δίτιμες, τότε μπορούμε να εκφράσουμε την σχέση (5.2.16) ως

$$f(\boldsymbol{\gamma}|\boldsymbol{p}) = \prod_{j=1}^{p} (p_{j|<j})^{\gamma_j}(1-p_{j|<j})^{1-\gamma_j} \tag{5.2.17}$$

Όπου $p_{j|<j} = f(\gamma_j = 1|\boldsymbol{\gamma}_{<j})$ και $\boldsymbol{p}$ είναι η συλλογή όλων των δεσμευμένων πιθανοτήτων $\{p_{j|<j}\}$ δειγματοληψίας των μεταβλητών. Μετά από την δειγματοληψία ενός μοντέλου από την (5.2.17), η μορφή της κατανομής θα παραμείνει η ίδια, απλώς με νέο $\boldsymbol{p}$ για να είναι καλώς ορισμένη η συνάρτηση μάζας. Η απόδειξη βρίσκεται στο άρθρο του αλγορίθμου BAS.

Αν παρατηρήσουμε τον τρόπο με τον οποίο ορίζουμε την $f(\boldsymbol{\gamma})$, τότε θεωρητικά οποιαδήποτε εκ των υστέρων κατανομή θα μπορούσε να οριστεί βάσει της (5.2.16) επιτρέποντας μας δείγμα χωρίς επανατοποθέτηση από την $\mathcal{M}$. Από την απόδειξη στους υπολογισμούς δεσμευμένων πιθανοτήτων $\boldsymbol{p} = \{p_{j|<j}\}$, δείχθηκε ότι η υπολογιστική πολυπλοκότητα είναι ισοδύναμη με αυτή της απαλοιφής του συνόλου $\mathcal{M}$. Παρ' όλο αυτά, μπορούμε να προχωρήσουμε την δειγματοληψία χωρίς επανατοποθέτηση με την βοήθεια κάποιας ποσότητας η οποία είναι κοντά στην $\boldsymbol{p}$.

Αυτό που προτείνεται είναι οι δεσμευμένες πιθανότητες $p_{j|<j} = f(\gamma_j = 1|\gamma_{<j})$ να προσεγγιστούν από τις πιθανότητες ένταξης $p(\gamma_j = 1|\boldsymbol{y}) = \pi_j$ έτσι ώστε να έχουμε μια προσέγγιση της εκ των υστέρων κατανομής. Όμως, επειδή δεν γνωρίζουμε εκ των προτέρων τις πιθανότητες ένταξης, μπορούμε να ανανεώσουμε επαναληπτικά τις εκ των υστέρων πιθανότητες ένταξης. Υποθέτοντας μια κατάλληλη αρχική τιμή, μπορούμε να χρησιμοποιήσουμε τις περιθώριες πιθανοφάνειες των μοντέλων που έχουν ήδη γεννηθεί για να εκτιμήσουμε ακολουθιακά τα $\pi_j^{(t)}$, $j = 1, ..., p$.

$$\hat{\pi}_j^{(t)} = \frac{\sum_{\boldsymbol{\gamma}\in\mathcal{S}_t} f(\mathbf{Y}|M_{\boldsymbol{\gamma}})\gamma_j}{\sum_{\boldsymbol{\gamma}\in\mathcal{S}_t} f(\mathbf{Y}|M_{\boldsymbol{\gamma}})} \tag{5.2.18}$$

Όπου το σύνολο $\mathcal{S}_t$ είναι το σύνολο των μοντέλων που έχουν γεννηθεί στον χρόνο t και καταφέραμε να ανανεώσουμε το $\boldsymbol{p}$ μέσου των $\hat{\boldsymbol{\pi}}^{(t)}$. Ωστόσο, η σχέση (5.2.18) είναι επίσης υπολογιστικά ακριβή καθώς σε κάθε χρόνο t υπολογίζουμε περιθώριες πιθανοφάνειες για να αθροίζουν οι πιθανότητες στην μονάδα. Αυτό που προτείνεται είναι ο υπολογισμός της πιθανοφάνειας κάθε U επαναλήψεις. Οπότε, αρχικοποιούμε $\boldsymbol{p} = \boldsymbol{p}^{(0)}$, μετά από κάθε U επανάληψη ανανέωσε την τιμή του $\boldsymbol{p}$ μέσο του $\hat{\boldsymbol{\pi}}^{(U)}$ αν υπάρχουν σημαντικές διαφορές μεταξύ τους, δηλαδή, $\|\hat{\boldsymbol{\pi}}^{(t)} - \hat{\boldsymbol{\pi}}^{(t-U)}\|^2/p > \varepsilon, \varepsilon > 0$. Έτσι σε περίπλοκα προβλήματα θα υπάρχει ανανέωση τιμής αν υπάρχει σημαντική αλλαγή στις πιθανότητες ένταξης. Ιδιαίτερη προσοχή πρέπει να δοθεί στην σχέση (5.2.18) καθώς τα $\hat{\pi}_j^{(t)}$ θα είναι 0 ή 1 αν το γ_j είναι πάντα δίτιμη στο δείγμα $\mathcal{S}_t$. Αυτό που κάνουμε είναι να περιορίσουμε τις $p_{j|<j}^{(t)} \in (\delta, 1-\delta)$ έτσι ώστε τα μοντέλα να έχουν θετική πιθανότητα δειγματοληψίας. Οι συγγραφείς προτείνουν $\delta = 0.025$ και $\varepsilon = \sqrt{\delta}$.

Τέλος, οι συγγραφείς συζητάνε τρεις τρόπους για την αρχικοποίηση των πιθανοτήτων δειγματοληψίας των μεταβλητών. Αρχικά, προτείνουν την επιλογή $p_{j|<j}^{(0)} = 1/2$ το οποίο ισοδυναμεί με την απλή τυχαία δειγματοληψία. Η δεύτερη προσέγγιση είναι η εκτίμηση του $p_{j|<j}^{(0)}$ μέσω των p-values, αποτέλεσμα των Selke T. et al (2001). Ο τρίτος τρόπος, προτείνετε ιδιαίτερα σε περίπτωση πολυσυγγραμικότητας, να εκτιμήσει την $p_{j|<j}^{(0)}$ με Monte-Carlo συχνότητες. Για περισσότερες λεπτομέρειες σχετικά με αυτές τις προσεγγίσεις, ανατρέξτε στο Κεφάλαιο 3.2 του άρθρου Merlise A. Clyde et al (2011).

5.3 Συμπεράσματα Κεφαλαίου

Στο Κεφάλαιο αυτό αναλύσαμε σε σημαντικό βαθμό την θεωρία της Μπεϋζιανης επιλογής μοντέλου. Αρχικά, είδαμε την φιλοσοφία σε ένα γενικότερο πλαίσιο και έπειτα μέσω της συζυγής ανάλυσης από προηγούμενα κεφάλαια. Επειδή η συζυγής ανάλυση θα μπορούσε να θεωρηθεί αρκετά περιοριστικός, είδαμε και άλλες προσεγγίσεις όπου στην ουσία αντιμετωπίζουν το ίδιο πρόβλημα από διαφορετικές οπτικές γωνίες.

Κεφάλαιο 6

Πειράματα Προσομοίωσης

Ο σκοπός αυτής της Ενότητας είναι να εφαρμόσουμε τους αλγόριθμους Μπεϋζιανής επιλογής μοντέλου σε μια ποικιλία προσομοιωμένων δεδομένων κάτω από διαφορετικές συνθήκες. Ο σκοπός, είναι η δίκαια σύγκριση και ανάλυση συμπεριφοράς της συζυγής ανάλυσης (CA), του SSVS, GVS και BAS αλγορίθμου. Επιπροσθέτως, οι παραπάνω Μπεϋζιανές προσεγγίσεις στην επιλογή μοντέλου θα συγκριθούν και με μεθόδους κλασικής προσέγγισης, όπως Stepwise και LASSO, για μια πιο ολοκληρωμένη εικόνα. Στην συνέχεια, θα δοκιμάσουμε τις παραπάνω μεθόδους με την βοήθεια μιας πιθανής τεχνικής για τη περίπτωση των μεγάλων δεδομένων. Τέλος, οι αλγόριθμοι CA, BAS, Stepwise και LASSO έχουν εκτελεστεί στην R με την βοήθεια πακέτων και το SSVS με GVS έχουν εκτελεστεί στο WinBUGS με πρωτογενή κώδικα.

6.1 Εφαρμογή αλγορίθμων

Θα ασχοληθούμε με ένα απλό πρόβλημα επιλογής μοντέλων με μόλις $p = 5$ επεξηγηματικές μεταβλητές, δίνοντας μας 63 πιθανά μοντέλα. Όλα τα σύνολα δεδομένων είναι μεγέθους $n = 100$ και έχουν προσομοιωθεί υπό τρεις κύριες περιπτώσεις δυσκολίας.

- Απλής δυσκολίας: Οι επεξηγηματικές μεταβλητές $\mathbf{X} \sim \mathcal{N}_5(\mathbf{0}_5, \mathbf{I})$, με $\mathbf{I}$ τον μοναδιαίο 5×5 πίνακα διακύμανσης. Τα πραγματικά $\boldsymbol{\beta} = (1, -3, 2, 0, 0.5, 0)^\top$ και σφάλμα $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, (0.025)^2)$.

- Μέτριας δυσκολίας: Οι επεξηγηματικές μεταβλητές είναι συσχετισμένες μεταξύ τους. Συγκεκριμένα, $X_2 = -2X_1 + \boldsymbol{\varepsilon}$, $X_3 = 2X_2 + X_1 + \boldsymbol{\varepsilon}$, $X_4 = 0.5X_3 + \boldsymbol{\varepsilon}$ και $X_5 = X_2 + X_4 + \boldsymbol{\varepsilon}$. Τα πραγματικά $\boldsymbol{\beta}$ είναι τα ίδια όπως και της απλής περίπτωσης με σφάλμα $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, (0.025)^2)$.

- Απαιτητικής δυσκολίας:

  – Οι επεξηγηματικές μεταβλητές είναι συσχετισμένες με τον ίδιο τρόπο όπως της μέτριας περίπτωσης. Τα πραγματικά $\boldsymbol{\beta} = (0.5, 0, 0.1, 0, 0, 0)^\top$ και σφάλμα $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, (0.025)^2)$.

– Οι επεξηγηματικές μεταβλητές είναι συσχετισμένες με τον ίδιο τρόπο όπως της μέτριας περίπτωσης. Τα πραγματικά $\boldsymbol{\beta} = (0.5, -0.33, 0.1, 0, 0.8, 0)^\top$ και σφάλμα $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, (0.025)^2)$.

Ο σκοπός των $\boldsymbol{\beta}$ στην απαιτητική περίπτωση είναι για να δούμε πως οι μέθοδοι συμπεριφέρονται κάτω από την πλειοψηφία των μεταβλητών να είναι θόρυβος και αντίστοιχα, αν μπορούν να εντοπίσουν μια σχετικά ασθενή πληροφορία.

Όλες οι μέθοδοι επιλογής μοντέλου χρησιμοποιούν στα $\boldsymbol{\beta}$, την g-prior που συζητήσαμε στο Κεφάλαιο §3.2.1 και την ομοιόμορφη ως εκ των προτέρων κατανομή στα ίδια τα μοντέλα. Πρέπει να σημειωθεί πως οι αλγόριθμοι GVS και SSVS χρησιμοποιούν μια παραλλαγή της g-prior με $g = n$ η οποία καλείται μίξη g-prior. Η μείξη g-prior που χρησιμοποιείται είναι της εξής μορφής

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{T}_{\gamma,\beta}^{-1})$$

Ο πίνακας $\mathbf{T}_{\gamma,\beta}$ είναι ο πίνακας ακρίβειας (αντίστροφο ανάλογο της διακύμανσης) και ορίζεται ως

$$T_{jk} = \frac{\gamma_j \gamma_k}{n\sigma^2}[\mathbf{X}^\top \mathbf{X}]_{jk} + (1 - \gamma_j \gamma_k) I(j = k) S_{\beta_j} \quad \text{με } S_{\beta_j} = 100$$

Ο όρος $n^{-1}\sigma^{-2}\mathbf{X}^\top\mathbf{X}$ είναι στην περίπτωση όπου και οι δύο μεταβλητές X_j και X_k είναι στο μοντέλο, δηλαδή, μικρή ακρίβεια ή μεγάλη διακύμανση. Όταν τουλάχιστον ένα από τα δύο δεν εμπεριέχεται στο μοντέλο, τότε $T_{jk} = 0$. Τα διαγώνια στοιχεία του T είναι όταν η μεταβλητή X_j δεν εμπεριέχεται στο μοντέλο, δηλαδή, μεγάλη ακρίβεια ή μικρή διακύμανση. Τέλος, στον SSVS και GVS χρησιμοποιούμε $\sigma^{-2} \sim \mathcal{G}amma(\alpha, \beta)$ με $\alpha = \beta = 0.01$ (καθώς χρησιμοποιούμε την έννοια της ακρίβειας). Ενώ στην συζυγή ανάλυση $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$ με $\alpha = \beta = 0.01$.

Οι Πίνακες 6.1 μας δείχνουν τη συμπεριφορά των μεθόδων για τα διαφορετικά σύνολα δεδομένων. Συγκεκριμένα, παρουσιάζονται οι πιθανότητες ένταξης γ_j και οι τρεις μεγαλύτερες εκ των υστέρων πιθανότητες μοντέλων και από κάτω τα αντίστοιχα μοντέλα $M_j|\boldsymbol{y}$ κατά φθίνουσα σειρά. Υποδεικνύουμε τα μοντέλα με έναν βολικό συμβολισμό αναπαράστασης των μοντέλων με την χρήση δίτιμων μεταβλητών. Πρέπει να σημειωθεί πως τα αποτελέσματα των MCMC αλγορίθμων έχουν βασιστεί σε μια εκτέλεση του κάθε αλγορίθμου με ένα σύνολο αρχικών τιμών. Ο συνολικός αριθμός MCMC επαναλήψεων ήταν $10,000$ με burn-in τις πρώτες $1,000$ επαναλήψεις.

Τα Διαγράμματα 6.1.1 - 6.1.4, μας δείχνουν τα $\beta_j\gamma_j$ και β_j της αλυσίδας για τον GVS και SSVS αλγόριθμο αντίστοιχα. Τα διαγράμματα έχουν χωριστεί στα πληροφοριακά και μη πληροφοριακά $\boldsymbol{\beta}$ για κάθε περίπτωση προσομοιωμένων δεδομένων. Εν συνέχεια το Διάγραμμα 6.1.5 μας δείχνει τις πιθα-

νότητες ένταξης της αλυσίδας για τον GVS και SSVS αλγόριθμο αντίστοιχα. Το διάγραμμα αυτό είναι ένας τρόπος για να επιβεβαιώσουμε την σύγκλιση των αλγορίθμων.

Πίνακας 6.1: Αποτελέσματα μεθόδων σε κάθε περίπτωση.

	CA	GVS	SSVS	BAS
γ_0	1	1	1	1
γ_1	1	1	1	1
γ_2	1	1	1	1
γ_3	0.320	0.307	0.235	0.318
γ_4	0.995	0.994	0.906	0.995
γ_5	0.092	0.093	0.129	0.092
$M_1\|\boldsymbol{y}$	0.613 **111010**	0.624 **111010**	0.599 **111010**	0.615 **111010**
$M_2\|\boldsymbol{y}$	0.290 **111110**	0.276 **111110**	0.188 **111110**	0.288 **111110**
$M_3\|\boldsymbol{y}$	0.062 **111011**	0.063 **111011**	0.093 **111011**	0.063 **111011**

(α') Απλή δυσκολία
Πραγματικό Μοντέλο: **111010**

	CA	GVS	SSVS	BAS
γ_0	1	1	1	1
γ_1	1	1	1	1
γ_2	1	1	1	1
γ_3	0.096	0.107	0.048	0.097
γ_4	0.976	0.975	0.227	0.975
γ_5	0.112	0.112	0.071	0.112
$M_1\|\boldsymbol{y}$	0.795 **111010**	0.788 **111010**	0.685 **111000**	0.795 **111010**
$M_2\|\boldsymbol{y}$	0.091 **111011**	0.092 **111011**	0.197 **111010**	0.091 **111011**
$M_3\|\boldsymbol{y}$	0.080 **111110**	0.084 **111110**	0.049 **111001**	0.080 **111110**

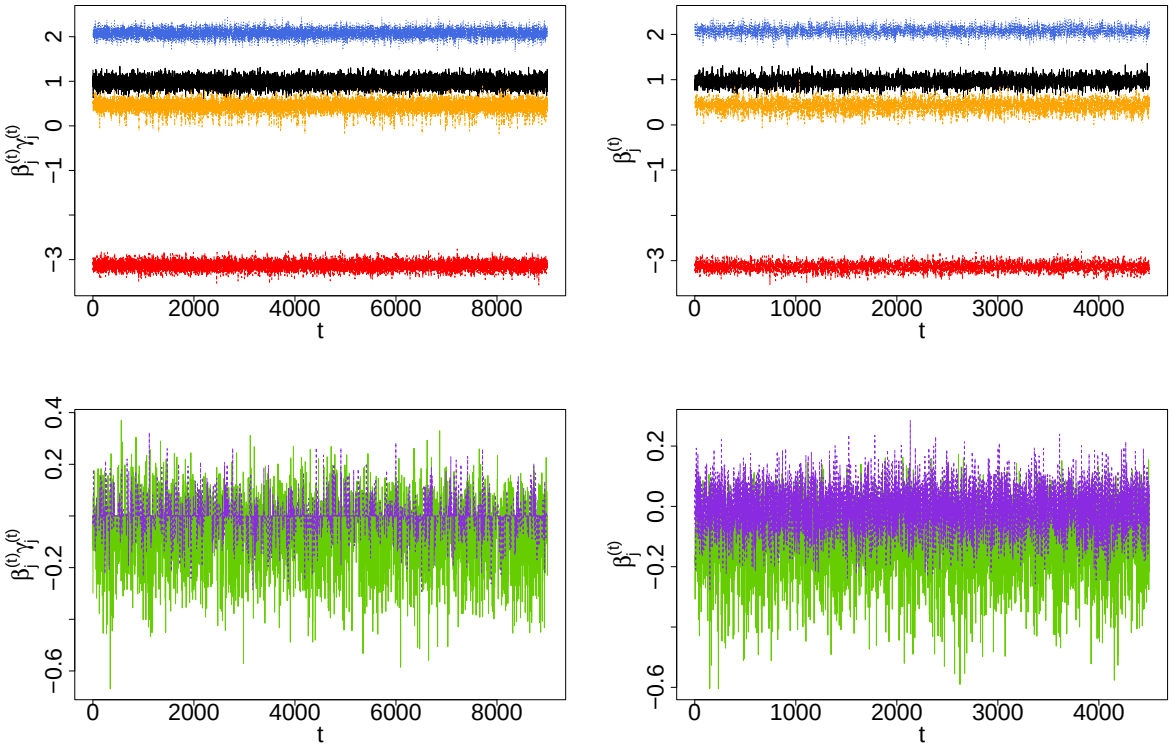
(β') Μέτρια δυσκολία
Πραγματικό Μοντέλο: **111010**

	CA	GVS	SSVS	BAS
γ_0	0.999	0.999	0.980	1
γ_1	0.376	0.384	0.202	0.375
γ_2	0.465	0.472	0.328	0.463
γ_3	0.206	0.198	0.220	0.207
γ_4	0.136	0.128	0.183	0.136
γ_5	0.163	0.174	0.238	0.162
$M_1\|\boldsymbol{y}$	0.254 **101000**	0.247 **101000**	0.194 **100000**	0.253 **101000**
$M_2\|\boldsymbol{y}$	0.237 **110000**	0.244 **110000**	0.157 **101000**	0.237 **110000**
$M_3\|\boldsymbol{y}$	0.093 **100100**	0.082 **100100**	0.094 **100100**	0.094 **100100**

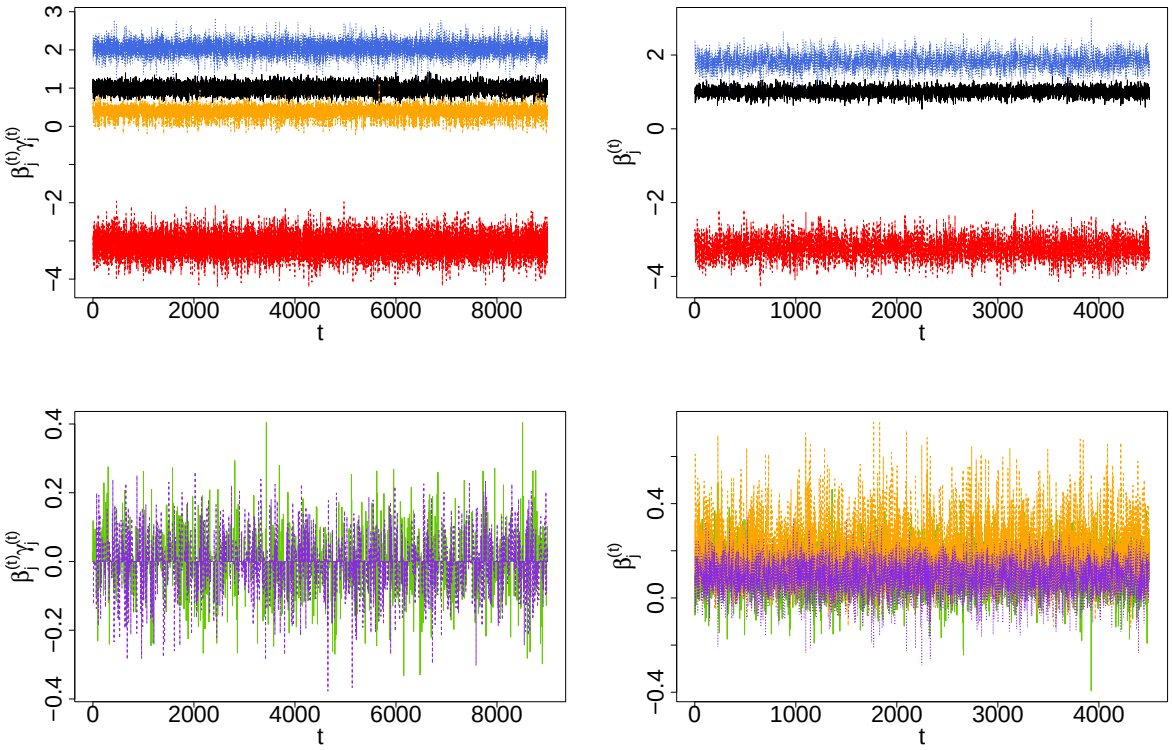
(γ') Απαιτητική δυσκολία 1
Πραγματικό μοντέλο: **101000**

	CA	GVS	SSVS	BAS
γ_0	0.999	1	0.984	1
γ_1	0.790	0.791	0.670	0.788
γ_2	0.338	0.339	0.230	0.338
γ_3	0.140	0.132	0.145	0.140
γ_4	0.999	1	0.999	0.999
γ_5	0.105	0.116	0.170	0.105
$M_1\|\boldsymbol{y}$	0.502 **110010**	0.499 **110010**	0.404 **110010**	0.502 **110010**
$M_2\|\boldsymbol{y}$	0.156 **101010**	0.161 **101010**	0.128 **100010**	0.158 **101010**
$M_3\|\boldsymbol{y}$	0.108 **111010**	0.105 **111010**	0.095 **101010**	0.107 **111010**

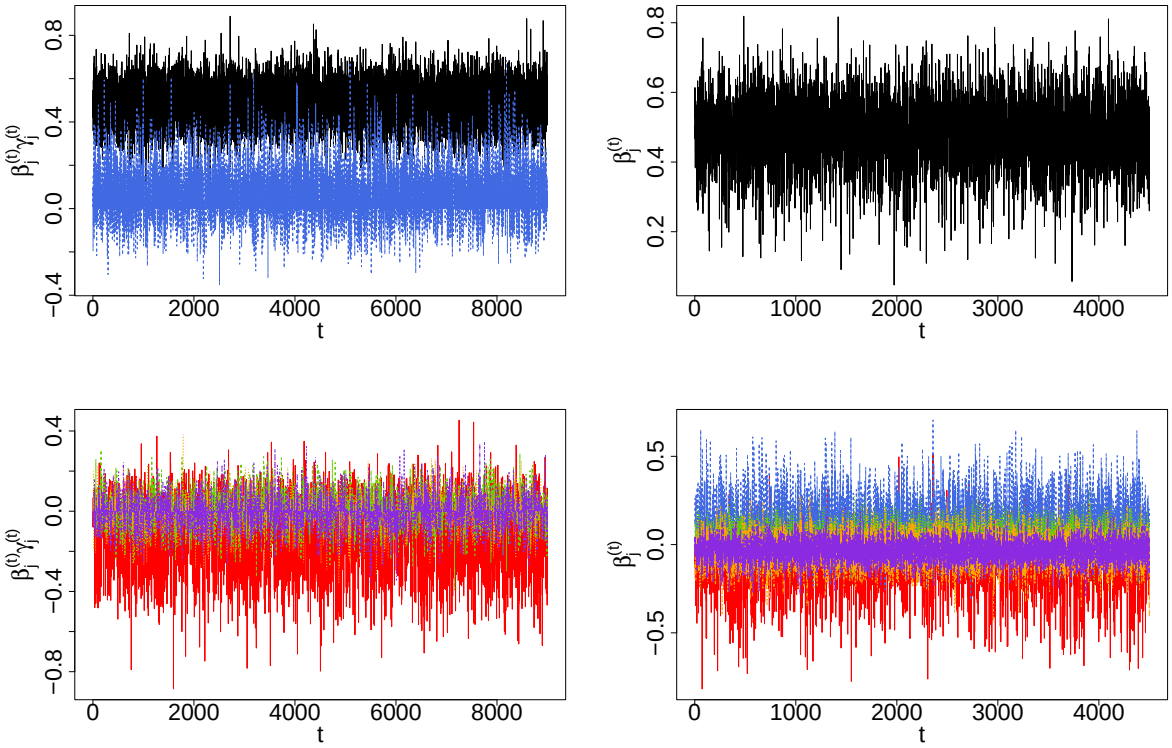
(δ') Απαιτητική δυσκολία 2
Πραγματικό μοντέλο: **111010**



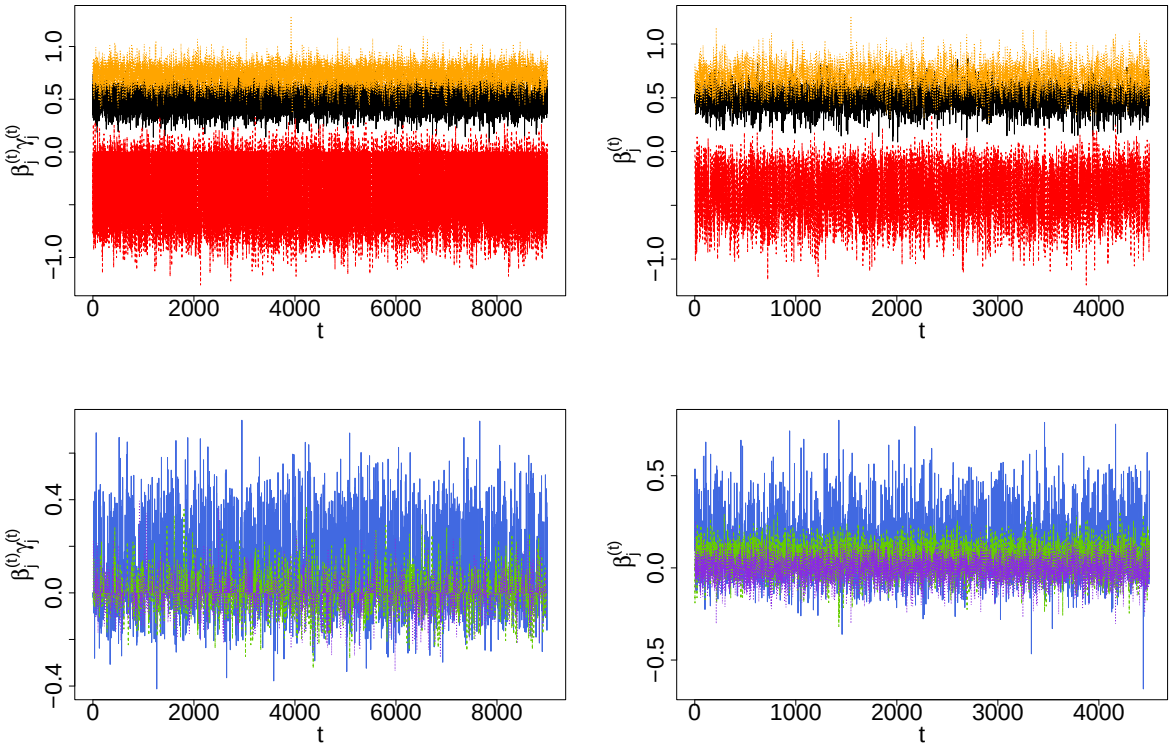
Διάγραμμα 6.1.1: Αποτελέσματα της MCMC αλυσίδας του GVS (αριστερά) και SSVS (δεξιά) αλγορίθμου για τα β στην απλή δυσκολία προσομοιωμένων δεδομένων. Η πρώτη γραμμή αφορά τις σημαντικές μεταβλητές που εντοπίζουν οι αλγόριθμοι και η τελευταία, τις μη-σημαντικές. Τα β_j της ανάλυσης είναι: β_0 (∘), β_1 (∘), β_2 (∘), β_3 (∘) β_4 (∘), β_5 (∘).



Διάγραμμα 6.1.2: Αποτελέσματα της MCMC αλυσίδας του GVS (αριστερά) και SSVS (δεξιά) αλγορίθμου για τα β στην μέτρια δυσκολία προσομοιωμένων δεδομένων. Η πρώτη γραμμή αφορά τις σημαντικές μεταβλητές που εντοπίζουν οι αλγόριθμοι και η τελευταία, τις μη-σημαντικές. Τα β_j της ανάλυσης είναι: β_0 (∘), β_1 (∘), β_2 (∘), β_3 (∘) β_4 (∘), β_5 (∘).



Διάγραμμα 6.1.3: Αποτελέσματα της MCMC αλυσίδας του GVS (αριστερά) και SSVS (δεξιά) αλγορίθμου για τα β στην απαιτητική δυσκολία 1 προσομοιωμένων δεδομένων. Η πρώτη γραμμή αφορά τις σημαντικές μεταβλητές που εντοπίζουν οι αλγόριθμοι και η τελευταία, τις μη-σημαντικές. Τα β_j της ανάλυσης είναι: β_0 (∘), β_1 (∘), β_2 (∘), β_3 (∘) β_4 (∘), β_5 (∘).



Διάγραμμα 6.1.4: Αποτελέσματα της MCMC αλυσίδας του GVS (αριστερά) και SSVS (δεξιά) αλγορίθμου για τα β στην απαιτητική δυσκολία 2 προσομοιωμένων δεδομένων. Η πρώτη γραμμή αφορά τις σημαντικές μεταβλητές που εντοπίζουν οι αλγόριθμοι και η τελευταία, τις μη-σημαντικές. Τα β_j της ανάλυσης είναι: β_0 (∘), β_1 (∘), β_2 (∘), β_3 (∘) β_4 (∘), β_5 (∘).

Από τον Πίνακα 6.1 (α'), παρατηρούμε μια αρκετά ικανοποιητική συμπεριφορά από όλες τις μεθόδους. Συγκεκριμένα, μπορούμε να δούμε πως οι πιθανότητες ένταξης γ_j είναι πολύ υψηλές για εκείνες τις μεταβλητές οι οποίες εμπεριέχονται στο πραγματικό μοντέλο και σχετικά μικρές για εκείνες οι οποίες δεν εμπεριέχονται (ειδικά η γ_5). Το παραπάνω αντανακλάται και στις εκ των υστέρων πιθανότητες το οποίο γίνεται μέγιστο στο πραγματικό μοντέλο για όλες τις μεθόδους.

Στον Πίνακα 6.1 (β') έχουμε τα αποτελέσματα της μέτριας δυσκολίας καθώς εισάγουμε το πρόβλημα πολυσυγγραμμικότητας. Από τις πιθανότητες ένταξης παρατηρούμε πως έχουμε όλες τις μεθόδους εκτός του SSVS να καταλήγουν στο πραγματικό μοντέλο. Ωστόσο, σε σύγκριση με τον Πίνακα 6.1 (α') οι μέθοδοι CA, GVS και BAS φαίνονται να έχουν μια ακόμα καλύτερη εικόνα. Παρατηρήστε ότι η πιθανότητα ένταξης γ_3 μειώνεται σημαντικά σε σύγκριση με του (α'), με την γ_5 να αυξάνεται ελάχιστα. Αυτό σημαίνει ότι μπορούμε να πούμε με περισσότερη βεβαιότητα πως οι μεταβλητές X_3 και X_5 δεν εμπεριέχονται στο μοντέλο. Αυτό φαίνεται και στις εκ των υστέρων πιθανότητες το οποίο γίνεται με μεγαλύτερη πιθανότητα μέγιστο πάλι στο πραγματικό μοντέλο (για το CA, GVS και BAS). Το παραπάνω θα μπορούσε να οφείλεται στην χρήση της g-prior καθώς είναι πιο ανθεκτικό στα προβλήματα πο-

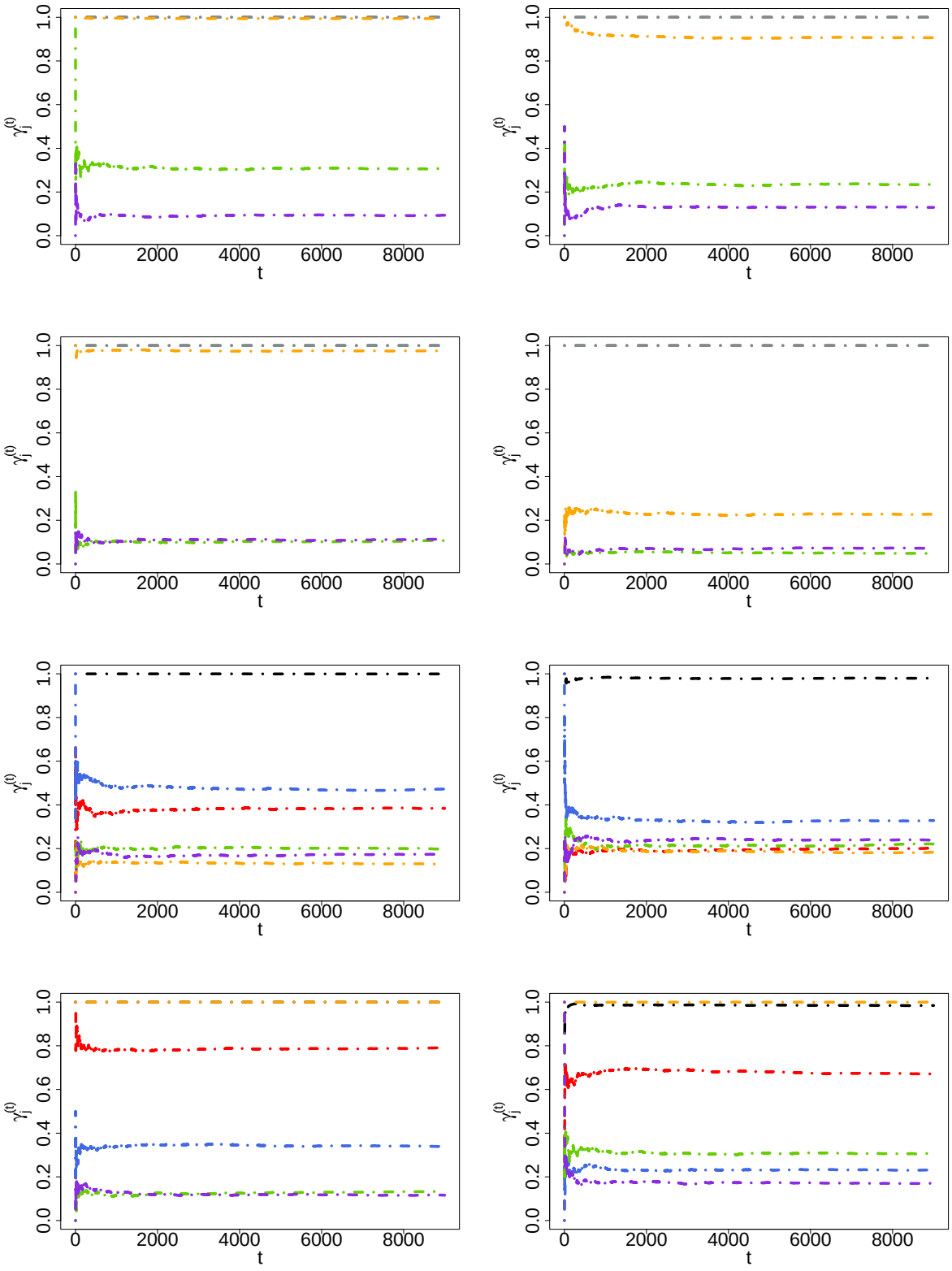
λυσυγγραμμικότητας. Ωστόσο, ο SSVS αλγόριθμος φαίνεται να μην βρίσκει το πραγματικό μοντέλο καθώς η μεταβλητή X_4 που στην πραγματικότητα υπάρχει, έχει πολύ μικρή πιθανότητα ένταξης γ_4. Οπότε, το SSVS δίνει την μέγιστη εκ των υστέρων πιθανότητα στο μη-πραγματικό μοντέλο και την δεύτερη μεγαλύτερη πιθανότητα στο πραγματικό. Αυτό φαίνεται και πιο αναλυτικά στον Πίνακα 6.1 (β').

Για τις δύο δύσκολες περιπτώσεις, όλες οι μέθοδοι δυσκολεύονται στο να βρουν την πραγματικότητα. Αρχικά, από τον Πίνακα 6.1 (γ') φαίνεται πως για όλες τις μεθόδους έχουμε υψηλή αβεβαιότητα λόγου της μικρής εκ των υστέρων πιθανότητας για κάθε μοντέλο. Παρ' όλο που δεν μπορούμε να εμπιστευτούμε το MAP, οι μέθοδοι CA, GVS και BAS, εξακολουθούν να δίνουν την μέγιστη πιθανότητα στο πραγματικό μοντέλο. Από την άλλη, παρ΄ όλο που το SSVS συνεχίζει να δίνει μέγιστη πιθανότητα στο λάθος μοντέλο, μπορούμε να δούμε πως βελτιώνεται στον εντοπισμό θορύβου. Αν συγκρίνουμε τις πιθανότητες του SSVS στον Πίνακα 6.1 (β') με του (γ') θα παρατηρήσουμε πως έχουμε πολύ περισσότερη αβεβαιότητα στην επιλογή του MAP σε σύγκριση με το (β'). Αξίζει να σημειωθεί πως στον Πίνακα 6.1 (γ') το δεύτερο μοντέλο με την μεγαλύτερη εκ των υστέρων πιθανότητα είναι το πραγματικό μοντέλο. Τέλος, από τον Πίνακα 6.1 (δ') παρατηρούμε μια ενδιάμεση περίπτωση του (β') και (γ') όπου δεν έχουμε την αβεβαιότητα του (γ') αλλά δεν έχουμε και την βεβαιότητα του (β'). Όλες οι μέθοδοι σε αυτή την περίπτωση, αποτυγχάνουν να εντοπίσουν το πραγματικό μοντέλο βάσει του MAP. Ωστόσο, για το CA, GVS και BAS το πραγματικό μοντέλο βρίσκεται τουλάχιστον στα τρία μοντέλα με την μεγαλύτερη εκ των υστέρων πιθανότητα. Σε αυτή την περίπτωση, θα ήταν ριψοκίνδυνο να επιλέξουμε εδώ βάση του MAP καθώς θα είχαμε 50% πιθανότητα να κάνουμε λάθος. Από την άλλη πλευρά, το SSVS κατατάσσει το πραγματικό μοντέλο ως το έκτο πιο πιθανό μοντέλο με εκ των υστέρων πιθανότητα ίση με 0.065.

Πίνακας 6.2: Αριθμός μοντέλων που οι αλγόριθμοι επισκέφτηκαν στις 10,000 επαναλήψεις και burn-in τις πρώτες 1,000.

	Απλή δυσκολία	Μέτρια δυσκολία	Απαιτητική δυσκολία 1	Απαιτητική δυσκολία 2
GVS	8	8	34	16
SSVS	8	8	53	31

Από τον Πίνακα 6.2 μπορούμε να δούμε πως στις δύο πρώτες περιπτώσεις και οι δύο αλγόριθμοι επισκέπτονται τον ίδιο αριθμό μοντέλων. Στις δύσκολες περιπτώσεις αυτός ο αριθμος αυξάνεται και αυτό ισχύει ειδικά για τον SSVS αλγόριθμο.



Διάγραμμα 6.1.5: Αποτελέσματα της MCMC αλυσίδας του GVS (αριστερά) και SSVS (δεξιά) αλγορίθμου για τα γ σε κάθε περίπτωση προσομοιωμένων δεδομένων σε αύξουσα σειρά. Τα γ_j της ανάλυσης είναι: γ_0 (—), γ_1 (—), γ_2 (—), γ_3 (—) γ_4 (—), γ_5 (—) και γκρι ευθεία (—) αφορά εκείνα τα γ_j τα οποία έχουν πάντα πιθανότητα ένταξης ίση με μονάδα.

Σε αυτό το σημείο θα γίνει η αναφορά και σε ένα άλλο σημαντικό μέτρο για τα αποτελέσματα του MCMC. Το μέτρο αυτό ονομάζεται Monte Carlo Error (MC error) και μετρά στην ουσία το σφάλμα κάθε εκτίμησης λόγου της προσομοίωσης του MCMC. Οπότε, θα θέλαμε το MC error να είναι μικρό για να έχουμε ακρίβεια στις εκτιμήσεις μας. Ο τρόπος με τον οποίο θα υπολογίσουμε το MC Error είναι με την μέθοδο batch mean. Το batch mean διαμερίζει τις επαναλήψεις της αλυσίδας Markov σε k (συνήθως $30-50$) υποσύνολα (batches), στην συνέχεια εκτιμάμε την ποσότητα θ που μας ενδιαφέρει σε κάθε υποσύνολο και υπολογίζουμε την τυπική απόκλιση αυτών των εκτιμήσεων από όλα τα υποσύνολα ως εξής

$$MCE_b(\hat{\theta}) = \sqrt{\frac{Var(\hat{\theta})}{N_b}}$$

Όπου N_b ο αριθμός των υποσυνόλων (στην δική μας περίπτωση $N_b = 30$) και το $Var(\hat{\theta})$ είναι η διακύμανση εκτιμώμενων τιμών $\hat{\theta_b}$ κάθε υποσυνόλου b, δηλαδή, $Var(\hat{\theta}) = \frac{1}{N_b - 1}\sum_{b=1}^{N_b}(\hat{\theta_b} - \hat{\theta})^2$. Αυτό που θέλουμε να υπολογίσουμε είναι το Monte Carlo error για τα μοντέλα της αλυσίδας. Με την βοήθεια μιας δίτιμης μεταβλητής θα βρούμε για κάθε μοντέλο την πιθανότητα εμφάνισης στα υποσύνολα και θα έχουμε έτσι την δυνατότητα να υπολογίσουμε το MCE_b.

Ο Πίνακας 6.3 μας δείχνει το MC error για τα τρία μοντέλα με την μεγαλύτερη εκ των υστέρων πιθανότητα σε φθίνουσα σειρά.

Πίνακας 6.3: Monte Carlo Error των μοντέλων

	MC error^a	MC error^b	MC error^c	MC error^d
GVS	0.0049 **111010**	0.0045 **111010**	0.0040 **101000**	0.0053 **110010**
	0.0049 **111110**	0.0035 **111011**	0.0044 **110000**	0.0033 **101010**
	0.0024 **111011**	0.0028 **111110**	0.0028 **100100**	0.0034 **111010**
SSVS	0.0053 **111010**	0.0047 **111000**	0.0040 **100000**	0.0051 **110010**
	0.0038 **111110**	0.0036 **111010**	0.0032 **101000**	0.0034 **100010**
	0.0030 **111011**	0.0025 **111001**	0.0030 **100100**	0.0029 **101010**

^{a} Απλή δυσκολία, ^{b} Μέτρια δυσκολία, ^{c} Απαιτητική δυσκολία 1. ^{d} Απαιτητική δυσκολία 2

Τέλος, θα συγκρίνουμε την Μπεϋζιανή προσέγγιση επιλογής μοντέλου με αυτή της κλασικής. Συγκεκριμένα, θα χρησιμοποιήσουμε τα κριτήρια πληροφορίας AIC και BIC για να εφαρμόσουμε τον Stepwise αλγόριθμο (ξεκινώντας από το μηδενικό και πλήρες μοντέλο) και στην συνέχεια θα εφαρμόσουμε το LASSO.

Ο Stepwise αλγόριθμος, ανάλογα με το μοντέλο που ξεκινάμε, θα κάνει προς τα εμπρός και προς τα πίσω κινήσεις σε κάθε βήμα του με σκοπό να δει πιθανές αλλαγές που μπορεί να έχουν συμβεί με την πρόσθεση ή αφαίρεση κάποια μεταβλητής. Η δε απόφαση γίνεται συνήθως με κάποιο κριτήριο πληροφορίας.

Το LASSO εισήχθηκε το 1996 ως μια μέθοδος που μπορεί να κάνει και επιλογή μεταβλητών και «κανονικοποίηση» (Regularization) με σκοπό την αύξηση της προβλεπτικής ικανότητας ενός στατιστικού μοντέλου. Ο όρος «κανονικοποίηση» αναφέρετε στην χρήση μιας τεχνητής πληροφορίας με σκοπό την λύση ενός προβλήματος όπου προηγουμένως δεν μπορούσε να λυθεί. Εμείς θα χρησιμοποιήσουμε το LASSO ως μια μέθοδος επιλογής μεταβλητών. Ο τρόπο με τον οποίο αυτό επιτυγχάνετε είναι βάζοντας έναν επιπλέον περιορισμό στο πρόβλημα ελαχίστων τετραγώνων ως εξής

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\top} + \lambda \sum_{j} |\beta_j| \right\}$$

Το παραπάνω το κάνουμε με την ελπίδα ότι τα β που αντιστοιχούν σε μεταβλητές που δεν είναι ενδιαφέρουσες, να είναι αυτές που θα θυσιαστούν για να ικανοποιηθεί ο περιορισμός. Πρακτικά αυτή η ποινή μας οδηγεί στο να μηδενίσουμε τα ανίσχυρα β. Ωστόσο, μεγάλες τιμές του λ (tuning parameter) οδηγούν σε μεγαλύτερη ποινή. Έτσι, είναι σημαντικό να αναφερθεί πως η επιλογή του λ είναι κρίσιμη καθώς αυτή καθορίζει τα αποτελέσματα μας. Η επιλογή του λ γίνεται πρακτικά με Cross-Validation, εξού και προβλεπτικό μοντέλο. Επειδή σε κάθε διαφορετική εκτέλεση του LASSO τα αποτελέσματα για το λ θα αλλάζουν, θα εκτελέσουμε 100 φορές το LASSO όπου στην εκάστοτε επανάληψη για κάθε λ, θα κρατάμε τα ίδια folds για να έχουμε συγκρίσιμα MSE.

Για αυτόν τον λόγο, λήφθηκε υπόψη ένα σύνολο τιμών $\lambda \in [0.0001, 10]$ καθώς αυτό αποτελεί ένα ικανοποιητικό εύρος για τις ποινές των μεταβλητών. Σε αυτό το σημείο, αξίζει να σημειωθεί πως στην πραγματικότητα το LASSO δεν κάνει τυπικά επιλογή μεταβλητών αλλά κάνει «Screening». Το Screening είναι η ιδέα που αναφέραμε και προηγουμένως, θα διώξουμε τις περιττές μεταβλητές. Ωστόσο, θα έχει την ιδιότητα να υπερεκτιμά τον αριθμό των μεταβλητών για λόγους σιγουριάς. Οι δύο πιο συνηθισμένες επιλογές του λ είναι το λ_{min} (εκείνο το λ που ελαχιστοποεί το MSE) και το λ_{se} (εκείνο το λ που βρίσκεται ένα τυπικό σφάλμα δεξιά από το λ_{min}). Έχει παρατηρηθεί πως το λ_{min} κάνει δυστυχώς πιο έντονο το πρόβλημα υπερεκτίμησης των μεταβλητών. Για να ελαττώσουμε το παραπάνω πρόβλημα θα επιλέξουμε το λ_{se} όπου τείνει να οδηγεί σε λιγότερες επιλεγμένες μεταβλητές.

Πίνακας 6.4: Αποτελέσματα μεθόδων σε κάθε περίπτωση.

	AIC	BIC
Stepwise (Null)	111110	111010
Stepwise (Full)	111110	111010
LASSO (λ_{se})	111110	$(0.142 - 0.271)$

(α') Απλή δυσκολία

	AIC	BIC
Stepwise (Null)	111010	111010
Stepwise (Full)	111110	111010
LASSO (λ_{se})	111010	$(0.236 - 0.411)$

(β') Μέτρια δυσκολία

	AIC	BIC
Stepwise (Null)	101000	101000
Stepwise (Full)	101000	101000
LASSO (λ_{se})	100000	$(10\ ,\ \ sd_{\lambda_{se}} = 0)$

(γ') Απαιτητική δυσκολία 1

	AIC	BIC
Stepwise (Null)	110010	110010
Stepwise (Full)	110010	110010
LASSO (λ_{se})	111010	$(0.236 - 0.411)$

(δ') Απαιτητική δυσκολία 2

Από τον Πίνακα 6.4 βλέπουμε τα μοντέλα όπου επέλεξαν οι μέθοδοι της κλασικής προσέγγισης. Αρχικά, ο Stepwise αλγόριθμος με το κριτήριο πληροφορίας AIC, φαίνεται στην πλειοψηφία των περιπτώσεων να αποτυγχάνει στην επιλογή του πραγματικού είτε ξεκινώντας από το μηδενικό είτε πλήρες μοντέλο. Οι μόνες εξαιρέσεις είναι στην περίπτωση της μέτριας δυσκολίας (όταν ξεκινάμε από το μηδενικό μοντέλο) και απαιτητικής δυσκολίας (γ'). Από την άλλη, παρ' όλο που δεν έχει ιδιαίτερο νόημα η σύγκριση Μπεϋζιανης επιλογής μοντέλου με το κριτήριο πληροφορίας BIC[1], παρατηρούμε πως σε κάθε περίπτωση επιλέγουμε το αντίστοιχο MAP μοντέλο του Πίνακα 6.1.

Ως προς το LASSO, παρατηρούμε εσφαλμένη επιλογή μοντέλου στις μισές περιπτώσεις προσομοιωμένων δεδομένων. Πιο συγκεκριμένα. φαίνεται να βρίσκει το πραγματικό μοντέλο στην Μέτρια δυσκολία και την Απαιτητική δυσκολία 2. Στις 100 επαναλήψεις που εκτελέστηκαν, παρουσίαζε κάθε φορά τις σωστές μεταβλητές πάντα με καμία φορά τις μη-πληροφοριακές μεταβλητές. Από την άλλη πλευρά το LASSO φαίνεται να έχει αποτυγχάνει στην απλή δυσκολία και την απαιτητική δυσκολία 1. Για την απλή δυσκολία στις 100 επαναλήψεις είχαμε και στις 100 τις σωστές μεταβλητές. Ωστόσο, η μεταβλητή X_3 εμφανίστηκε 53. Θα μπορούσε κανείς να υιοθετήσει ένα πολύ αυστηρό κριτήριο και να δεχθεί μόνο αυτές που εμφανίστηκαν και τις 100 φορές καθώς αυτές είναι οι σίγουρες μεταβλητές. Σε αυτή την ανάλυση βάλαμε το κριτήριο πως κάθε μεταβλητή θα πρέπει να εμφανίζεται τουλάχιστον 50 φορές για να συμπεριληφθεί στο μοντέλο. Για την απαιτητική δυσκολία 2, το LASSO σε κάθε επανάληψη βάζει την μεγαλύτερη ποινή με αποτέλεσμα να μην γίνει καμία φορά δεχτεί καμία μεταβλητή εκτός από την σταθερά.

[1] Το BIC είναι μια ασυμπτωματική προσέγγιση του Παράγοντα Bayes. Για περισσότερες πληροφορίες ανατρέξτε στο άρθρο του Gideon Schwarz (1978)

Τέλος, υπενθυμίζουμε πως ένα σημαντικό μειονέκτημα της κλασικής προσέγγισης είναι η ανικανότητα της ποσοτικοποίησης της αβεβαιότητας των ίδιων των μοντέλων. Αυτό είναι ιδιαίτερα κρίσιμο σε περιπτώσεις όπως του Πίνακα 6.4 (β') με το κριτήριο πληροφορίας AIC. Ο Stepwise αλγόριθμος διαφωνεί ανάλογα με την αφετηρία του μοντέλου.

6.2 Εφαρμογή σε μεγάλα δεδομένα

Ο όρος "Μεγάλα δεδομένα" περιγράφει σύνολα δεδομένων τα οποία είναι τόσο σύνθετα που καθιστούν αδύνατη την αποθήκευση τους ή και ανάλυσης τους με παραδοσιακές τεχνικές. Αν θεωρήσουμε τον κλασικό τρόπο αποθήκευσης δεδομένων σε ορθογώνιο τότε τα μεγάλα δεδομένα έχουν να κάνουν με έναν τεράστιο όγκο δείγματος n και τεράστιο αριθμό μεταβλητών p. Σε μια τέτοια εποχή, τίθεται μεγάλο θέμα η διαχείριση τους αλλά και η ανάλυση τους. Πιο συγκεκριμένα, πολλές φορές ο όγκος των δεδομένων είναι τόσο μεγάλος που καθιστά αδύνατη την φόρτωση τους στην μνήμη. Ακόμα και αν είναι τελικά εφικτή η φόρτωση τους, η εφαρμογή διάφορων αλγορίθμων πάνω σε αυτά, προκαλεί προβλήματα τόσο χρονικής εκτέλεσης όσο και υπολογιστικής φύσεως.

Σε αυτή την Ενότητα, θα μας απασχολήσει μόνο το πρόβλημα του μεγάλου δείγματος n. Έτσι, καλούμαστε να ελέγξουμε έναν πιθανό τρόπο εκτέλεσης των Μπεϋζιανών αλγορίθμων επιλογής μοντέλων όταν έχουμε στην διάθεση μας πολλές παρατηρήσεις. Για λόγους απλότητας, θα θεωρήσουμε ένα παιδικό παράδειγμα με σύνολο δεδομένων $n = 5,000$. Αν θεωρήσουμε ότι η εκτέλεση των αλγορίθμων με αυτά τα δεδομένα ή και η φόρτωση τους στην μνήμη είναι αδύνατη, τότε αυτό που θα κάνουμε είναι να σπάσουμε τα δεδομένα με τυχαίο τρόπο σε $k = 10$ κομμάτια. Όποτε, θα έχουμε στην διάθεση μας 10 τυχαία υποσύνολα δεδομένων μεγέθους $n_k = 500$ όπου πλέον είναι δυνατή η εκτέλεση των αλγορίθμων σε κάθε υποσύνολο ξεχωριστά.

Εφόσον πλέον είναι δυνατή η εκτέλεση, θα πρέπει να βρούμε έναν τρόπο να συνθέσουμε αυτά τα 10 αποτελέσματα που μας δίνει ένας αλγόριθμος με τέτοιον τρόπο ώστε να συμφωνεί με τα αποτελέσματα του αλγορίθμου για τα συνολικά δεδομένα. Ένας άλλος παράγοντας που καλούμαστε να ελέγξουμε είναι η σύγκριση χρόνου εκτέλεσης μεταξύ συνολικού δείγματος και των 10 τυχαίων υποσυνόλων (συνολικά). Ελπίζουμε ο χρόνος εκτέλεσης των υποσυνόλων συνολικά να είναι ταχύτερος από το συνολικό δείγμα, εφόσον κάθε φορά βάζουμε έναν πολύ μικρότερο αριθμό δείγματος όπου ξεχωριστά είναι σίγουρα ταχύτερος αλλά όχι απαραίτητα συνολικά. Εφόσον πρακτικά μπορούμε να εκτελέσουμε τον αλγόριθμο με το συνολικό δείγμα, όλοι αυτοί οι έλεγχοι είναι εφικτοί.

Τέλος, θα δούμε δύο πιθανούς τρόπους για να συνθέσουμε τα αποτελέσματα μας. Αυτό θα γίνει με την βοήθεια της πλειοψηφίας και τον μέσο όρο των πιθανοτήτων εντάξεως. Για την πλειοψηφία, θα δούμε πόσες φορές εμφανίστηκε η κάθε μεταβλητή στο MAP μοντέλο που επιλέγει ο αλγόριθμος σε κάθε υποσύνολο δεδομένων και έτσι θα κάνουμε ένα σύστημα ψηφοφορίας για το ποιες μεταβλητές επιλέχθηκαν κάθε φορά. Για τους μέσους όρους πιθανοτήτων ένταξης, προσπαθούμε να λάβουμε υπόψη και τα άλλα μοντέλα τα οποία θα μπορούσαν να προσφέρουν πληροφορία καθώς το MAP δεν την παρέχει. Ελπίζουμε πως οι πραγματικά πληροφοριακές μεταβλητές θα εμφανιστούν στα περισσότερα υποσύνολα και κατά μέσο όρο οι πιθανότητες ένταξης τους να τείνουν σε εκείνα των συνολικών.

Πίνακας 6.5: Αποτελέσματα μεθόδων για $n = 5,000$.

	CA	GVS	SSVS	BAS
γ_0	1	1	1	1
γ_1	1	1	1	1
γ_2	1	1	1	1
γ_3	0.044	0.042	0.092	0.044
γ_4	1	1	0.999	1
γ_5	0.014	0.015	0.090	0.014
$M_1\|\boldsymbol{y}$	0.941 **111010**	0.942 **111010**	0.827 **111010**	0.941 **111010**
$M_2\|\boldsymbol{y}$	0.044 **111110**	0.042 **111110**	0.082 **111110**	0.044 **111110**
$M_3\|\boldsymbol{y}$	0.014 **111011**	0.014 **111011**	0.080 **111011**	0.014 **111011**

(α') Απλή δυσκολία
Πραγματικό Μοντέλο: **111010**

	CA	GVS	SSVS	BAS
γ_0	1	1	1	1
γ_1	1	1	1	1
γ_2	1	1	1	1
γ_3	0.023	0.024	0.108	0.023
γ_4	1	1	0.905	1
γ_5	0.014	0.013	0.078	0.014
$M_1\|\boldsymbol{y}$	0.962 **111010**	0.962 **111010**	0.746 **111010**	0.962 **111010**
$M_2\|\boldsymbol{y}$	0.022 **111110**	0.023 **111110**	0.090 **111110**	0.022 **111110**
$M_3\|\boldsymbol{y}$	0.014 **111011**	0.013 **111011**	0.074 **111000**	0.014 **111011**

(β') Μέτρια δυσκολία
Πραγματικό Μοντέλο: **111010**

	CA	GVS	SSVS	BAS
γ_0	1	1	1	1
γ_1	0.014	0.013	0.070	0.014
γ_2	0.999	1	0.250	0.999
γ_3	0.040	0.050	0.189	0.040
γ_4	0.045	0.041	0.130	0.045
γ_5	0.022	0.020	0.206	0.022
$M_1\|\boldsymbol{y}$	0.881 **101000**	0.877 **101000**	0.320 **100000**	0.881 **101000**
$M_2\|\boldsymbol{y}$	0.043 **101010**	0.047 **101100**	0.171 **101000**	0.043 **101010**
$M_3\|\boldsymbol{y}$	0.038 **101100**	0.039 **101010**	0.134 **100001**	0.038 **101100**

(γ') Απαιτητική δυσκολία 1
Πραγματικό Μοντέλο:**101000**

	CA	GVS	SSVS	BAS
γ_0	1	1	1	1
γ_1	1	1	0.792	1
γ_2	0.999	1	0.110	0.999
γ_3	0.023	0.026	0.133	0.023
γ_4	1	1	1	1
γ_5	0.014	0.010	0.107	0.014
$M_1\|\boldsymbol{y}$	0.962 **111010**	0.963 **111010**	0.557 **110010**	0.962 **111010**
$M_2\|\boldsymbol{y}$	0.022 **111110**	0.026 **111110**	0.124 **100010**	0.022 **111110**
$M_3\|\boldsymbol{y}$	0.014 **111011**	0.010 **111011**	0.087 **110110**	0.014 **111011**

(δ') Απαιτητική δυσκολία 2
Πραγματικό Μοντέλο:**111010**

Από τον Πίνακα 6.5 έχουμε τα αποτελέσματα που είχαμε και στην Ενότητα §6.1. Πιο συγκεκριμένα, έχουμε τις πιθανότητες ένταξης γ_j και τις τρεις μεγαλύτερες εκ των υστέρων πιθανότητες $p(M_j|\boldsymbol{y})$ κατά φθίνουσα σειρά. Προφανώς, επειδή έχουμε αυξήσει σημαντικά το δείγμα n από 100 σε 5000, τα αποτελέσματα των αλγορίθμων σε κάθε περίπτωση είναι πολύ πιο ικανοποιητικά σε σύγκριση με τον Πίνακα 6.1. Ωστόσο ο SSVS αλγόριθμος στις δύο απαιτητικές περιπτώσεις φαίνεται να αποτυγχάνει κάθε φορά στις πιθανότητες ένταξης με αποτέλεσμα, όπως φαίνεται και στον Πίνακα 6.4, να επιλέγει κάθε φορά το λάθος μοντέλο.

Από τον Πίνακα 6.5 μπορούμε να δούμε τα τρία μοντέλα που τελικά επιλέχθηκαν από τους αλγορίθμους. Παρατηρούμε πλέον πως σε κάθε περίπτωση όλοι οι αλγόριθμοι εκτός από τον SSVS στις δύο τελευταίες περιπτώσεις, επιλέγουν το πραγματικό μοντέλο και μάλιστα με σημαντικά υψηλότερη εκ των υστέρων πιθανότητα. Ο SSVS αλγόριθμος στην απαιτητική περίπτωση 1 κατατάσσει το πραγματικό μοντέλο ως το δεύτερο πιο πιθανό ενώ στην απαιτητική περίπτωση 2 το κατατάσσει ως το πέμπτο. Για τον Πίνακα 6.6, υπάρχει μια γενική παρατήρηση για τις μέσες πιθανότητες ένταξης. Η μονή περίπτωση όπου οι μέσες πιθανότητες ένταξης φαίνονται να είναι πολύ κοντά στην πραγματικότητα $n = 5,000$ είναι όταν είμαστε σε περιπτώσεις όπως στον Πίνακα 6.6 (α'), δηλαδή παίρνουμε κάθε φορά την σωστή ψήφο. Ωστόσο, στις υπόλοιπες περιπτώσεις οι πιθανότητες ένταξης φαίνεται απλώς να αντανακλούν το σύστημα ψηφοφορίας που έχουμε κατασκευάσει για τις μεταβλητές και όχι τις πραγματικές πιθανότητες ένταξης για $n = 5,000$. Παρ' όλο αυτά, θα μπορούσαν να είναι χρήσιμες για να πάρουμε αποφάσεις σχετικά με το τελικό μοντέλο που θα σχολιάσουμε στη συνέχεια.

Από τον Πίνακα 6.6 (α'), παρατηρούμε την ιδανική συμπεριφορά από όλες τις μεθόδους. Συγκεκριμένα, μπορούμε να δούμε πως και στα δέκα τυχαία υποσύνολα δεδομένων, οι πραγματικές μεταβλητές ψηφίστηκαν πάντα ενώ οι ασήμαντες καμία φορά. Αυτό φαίνεται και από τις μέσες τιμές των πιθανοτήτων ένταξης $\bar{\gamma_j}$ καθώς είναι πολύ κοντά με τις "πραγματικές" του $n = 5,000$.

Στον Πίνακα 6.6 (β') έχουμε επίσης αρκετά ικανοποιητικά αποτελέσματα όπως και στο (α'). Η μονή περίπτωση που ξεχωρίζει είναι του SSVS αλγορίθμου καθώς η μεταβλητή X_4 εμφανίστηκε στα μισά υποσύνολα δεδομένων. Το πιο δημοφιλές κριτήριο σε αυτή την περίπτωση είναι να δεχτούμε εκείνες τις μεταβλητές που εμφανίζονται πάνω από πέντε φορές. Σε αυτή την περίπτωση δεν θα είχαμε λάβει το πραγματικό μοντέλο. Ωστόσο, αν υιοθετήσουμε τις μέσες τιμές των πιθανοτήτων ένταξης $\bar{\gamma_j}$ θα δούμε πως η $\bar{\gamma_4} > 0.5$ με αποτέλεσμα να επιλέξαμε το πραγματικό μοντέλο.

Πίνακας 6.6: Αποτελέσματα μεθόδων για $n_k = 500$ σε κάθε περίπτωση.

	CA	GVS	SSVS	BAS
X_0	10	10	10	10
X_1	10	10	10	10
X_2	10	10	10	10
X_3	0	0	0	0
X_4	10	10	10	10
X_5	0	0	0	0
$\bar{\gamma_0}$	1	1	1	1
$\bar{\gamma_1}$	1	1	1	1
$\bar{\gamma_2}$	1	1	1	1
$\bar{\gamma_3}$	0.077	0.079	0.1033	0.077
$\bar{\gamma_4}$	1	1	0.9986	1
$\bar{\gamma_5}$	0.059	0.060	0.102	0.059

(α') Απλή δυσκολία

	CA	GVS	SSVS	BAS
X_0	10	10	10	10
X_1	10	10	10	10
X_2	10	10	10	10
X_3	0	0	0	0
X_4	10	10	5	10
X_5	0	0	0	0
$\bar{\gamma_0}$	1	1	1	1
$\bar{\gamma_1}$	1	1	1	1
$\bar{\gamma_2}$	1	1	1	1
$\bar{\gamma_3}$	0.060	0.066	0.070	0.060
$\bar{\gamma_4}$	1	1	0.568	1
$\bar{\gamma_5}$	0.067	0.066	0.067	0.067

(β') Μέτρια δυσκολία

	CA	GVS	SSVS	BAS
X_0	10	10	10	10
X_1	2	1	0	2
X_2	7	5	0	7
X_3	1	1	0	1
X_4	0	0	0	0
X_5	1	1	0	1
$\bar{\gamma_0}$	1	1	0.999	1
$\bar{\gamma_1}$	0.229	0.221	0.121	0.229
$\bar{\gamma_2}$	0.573	0.545	0.254	0.573
$\bar{\gamma_3}$	0.199	0.225	0.200	0.199
$\bar{\gamma_4}$	0.080	0.080	0.141	0.080
$\bar{\gamma_5}$	0.178	0.186	0.224	0.178

(γ') Απαιτητική δυσκολία 1

	CA	GVS	SSVS	BAS
X_0	10	10	10	10
X_1	8	8	7	8
X_2	2	2	0	2
X_3	0	0	0	0
X_4	10	10	10	10
X_5	0	0	0	0
$\bar{\gamma_0}$	1	1	0.999	1
$\bar{\gamma_1}$	0.778	0.783	0.572	0.778
$\bar{\gamma_2}$	0.375	0.365	0.192	0.375
$\bar{\gamma_3}$	0.093	0.096	0.141	0.093
$\bar{\gamma_4}$	1	1	1	1
$\bar{\gamma_5}$	0.103	0.107	0.118	0.103

(δ') Απαιτητική δυσκολία 2

Παρομοίως, αν συγκρίνουμε με τον Πίνακα 6.5 (γ') θα δούμε πως υπάρχει συμφωνία με τον Πίνακα 6.6 (γ'). Τέλος, τα αποτελέσματα στον Πίνακα 6.6 (δ') πέφτουν εκτός σε κάθε αλγόριθμο. Σε αυτή την περίπτωση το πιο πιθανό μοντέλο θα ήταν το 110010 με το δεύτερο πιο πιθανό το 111010 το οποίο είναι το πραγματικό που μας γέννησε τα δεδομένα.

Τώρα συγκεντρωνόμαστε στο υπολογιστικό κόστος των αλγορίθμων σε κάθε περίπτωση προσομοιωμένων δεδομένων για $n = 5,000$ και τις συνολικές δέκα επαναλήψεις των $n_k = 500$. Όλοι οι υπολογισμοί έχουν γίνει σε έναν υπολογιστή με 2.3 GHz Intel 14-Core I7 επεξεργαστή και 16GB μνήμη χρησιμοποιώντας την γλώσσα προγραμματισμού R (για το CA και BAS) και WinBUGS (για το GVS και SSVS).

Πίνακας 6.7: Χρόνοι εκτέλεσης για $n = 5,000$ και συνολικά όλα τα $n_k = 500$ σε δευτερόλεπτα.

	CA	GVS	SSVS	BAS
$n = 5,000$	8956.54	0.27	0.23	0.00
$n_k = 500$	28.90	1.78	0.64	0.06

(α') Απλή δυσκολία

	CA	GVS	SSVS	BAS
$n = 5,000$	8935.06	0.28	0.24	0.02
$n_k = 500$	28.64	1.75	0.47	0.08

(β') Μέτρια δυσκολία

	CA	GVS	SSVS	BAS
$n = 5,000$	8372.47	0.23	0.16	0.00
$n_k = 500$	27.11	1.63	0.63	0.05

(γ') Απαιτητική δυσκολία 1

	CA	GVS	SSVS	BAS
$n = 5,000$	4392.61	0.32	0.1	0.00
$n_k = 500$	28.32	1.86	0.55	0.00

(δ') Απαιτητική δυσκολία 2

Ο Πίνακας 6.7 μας δείχνει τους χρόνους εκτέλεση (σε δευτερόλεπτα) που αφιέρωσε ο επεξεργαστής για να εκτελεστούν. Ο πιο χρονοβόρος αλγόριθμος ήταν αυτό του CA για τα συνολικά δεδομένα $n = 5,000$. Από όσο φαίνεται σε αυτή την περίπτωση η απλοποίηση του προβλήματος σε μικρότερα υποσύνολα βοήθησε σημαντικά στον χρόνο εκτέλεσης του αλγορίθμου. Ωστόσο, το ίδιο πράγμα δεν ισχύει για τους υπόλοιπους αλγορίθμους. Βλέπουμε πως ο χρόνος εκτέλεσης στα υποσύνολα δεδομένα ήταν μεγαλύτερος από των συνολικών δεδομένων.

6.3 Συμπεράσματα Κεφαλαίου

Σε αυτό το Κεφάλαιο εφαρμόσαμε τους αλγορίθμους επιλογής μοντέλου κάτω από διαφορετικές συνθήκες προσομοιωμένων δεδομένων. Εξετάστηκε αρχικά η συμπεριφορά των αλγορίθμων κάτω από ένα σχετικά μικρό αριθμό παρατηρήσεων και στην συνέχεια κάτω από ένα "μεγάλο" αριθμό παρατηρήσεων. Για τον μικρότερο αριθμό παρατηρήσεων, είδαμε πως σε γενικές γραμμές οι αλγόριθμοι είχαν καλή συμπεριφορά, ειδικά στις δύο πιο απλές περιπτώσεις. Από την άλλη στις υπόλοιπες περιπτώσεις αντιμετώπιζαν ιδιαίτερες δυσκολίες. Ο αλγόριθμος ο οποίος δυσκολευόταν περισσότερο σε σύγκριση με τις υπόλοιπες ήταν ο SSVS αλγόριθμος ο οποίος είχε πρόβλημα στην πολυσυγγραμμικότητα. Τέλος, στην περίπτωση των μεγάλων δεδομένων, προσπαθήσαμε να σπάσουμε τα δεδομένα σε υποσύνολα δεδομένων. Σε κάθε υποσύνολο δεδομένων εφαρμόσαμε τους αλγορίθμους, προσπαθήσαμε να συνθέσουμε τα αποτελέσματα τους και μετρήσαμε συνολικά τον χρόνο εκτέλεσης τους. Τα αποτελέσματα ήταν αρκετά ικανοποιητικά, ωστόσο οι χρόνοι εκτέλεσης τους ήταν σχετικά χειρότεροι.

Κεφάλαιο 7

Συζήτηση και μελλοντική έρευνα

Αυτή η διατριβή χρησιμεύει ως μια γενική επισκόπηση στο πρόβλημα Μπεϋζιανής επιλογή μοντέλου. Μετά από μια ανάλυση των θεμελιωδών εννοιών της Μπεϋζιανής θεωρίας, μας οδήγησε στην καλύτερη κατανόηση των αλγορίθμων επιλογής μοντέλου και εφαρμογή τους μέσω προσομοιώσεων. Η ανάγκη για ισχυρές μεθόδους επιλογής μοντέλου ωθεί στην δημιουργία νέων προσεγγίσεων όπου αντιμετωπίζουν το πρόβλημα από διαφορετικές οπτικές. Η πολύτιμη πληροφορία της Μπεϋζιανής ανάλυσης να ποσοτικοποιεί την αβεβαιότητα των μοντέλων, αποτελεί χρήσιμο εργαλείο στην επιλογή ενός μοντέλου ή ακόμα και σύνθεση των μοντέλων σε ένα, μέσω του BMA.

Βάσει των διαφορετικών περιπτώσεων προσομοίωσης, είχαμε την δυνατότητα να δούμε την συμπεριφορά κάθε αλγορίθμου. Τα αποτέλεσμα ήταν στην πλειοψηφία ικανοποιητικά με την έννοια ότι εντόπιζαν σε γενικές γραμμές βάσει του MAP το πραγματικό μοντέλο με αρκετά υψηλή πιθανότητα. Ακόμα και στις δυσκολότερες περιπτώσεις, παρ' όλο που δεν ήταν τόσο επιτυχής, η αβεβαιότητα των αποτελεσμάτων ήταν πολύ υψηλή ακόμα και για το MAP μοντέλο. Το κύριο πρόβλημα το εμφάνιζε ο αλγόριθμος SSVS ο οποίος είχε ιδιαίτερη δυσκολία στην πολυσυγγραμμικότητα, θόρυβο και ασθενής συντελεστές. Το σημαντικότερο μειονέκτημα των παραπάνω μεθόδων είναι η εξάρτηση του από την εκ των προτέρων κατανομή και τις αντίστοιχες υπερ-παραμέτρους της, όπου τα τελικά αποτελέσματα, θα αλλάζουν για διαφορετικές επιλογές τους. Έτσι, για τα συγκεκριμένα προβλήματα προσομοίωσης θα ήταν ενδιαφέρον να εξεταστεί η επιλογή διαφορετικών εκ των προτέρων κατανομών και την συμπεριφορά τους με διαφορετικές υπερ-παραμέτρους. Μέσα από αυτά, θα μπορούσαμε να βρούμε τρόπους για τον βέλτιστο ορισμό μιας εκ των προτέρων κατανομής και των αντίστοιχων υπερ-παραμέτρων ανάλογα το πρόβλημα το οποίο αντιμετωπίζουμε. Προφανώς, επειδή ο κάθε αλγόριθμος λειτουργεί διαφορετικά, είναι επίσης πολύ πιθανό διαφορετικές εκ των προτέρων κατανομές και υπερ-παράμετροι να χρειάζονται για να ληφθεί το επιθυμητό αποτέλεσμα για το ίδιο πρόβλημα.

Επιπροσθέτως, στο πρόβλημα του μεγάλου όγκου παρατηρήσεων, είδαμε την συμπεριφορά των παραπάνω αλγορίθμων όταν θεωρήσουμε πως δεν έχουμε την δυνατότητα να εργαστούμε με αυτά. Για αυτόν τον λόγο, σπάσαμε το πρόβλημα σε υποσύνολα δεδομένων με σκοπό την εκτέλεση των αλγορίθμων και έπειτα έγινε η προσπάθεια σύνθεσης των αποτελεσμάτων. Παρατηρήθηκε, πως ο κάθε αλγόριθμος στις πρώτες τρεις περιπτώσεις προσομοιωμένων δεδομένων, συμφωνούν με την επιλογή του MAP μοντέλου των συνολικών δεδομένων. Η μονή εξαίρεση ήταν η τελευταία περίπτωση η οποία κατατάσσει βάσει πλειοψηφίας το MAP μοντέλο των συνολικών δεδομένων ως το δεύτερο πιο πιθανό. Σε αυτό το σημείο θα ήταν ενδιαφέρουσα η εξερεύνηση διαφορετικών τρόπων χειρισμού μεγάλων δεδομένων. Ένας τρόπος είναι η επιλογή ενός υποσυνόλου των δεδομένων που παρέχουν την βέλτιστη πληροφορία από τα συνολικά δεδομένα. Στην βιβλιογραφία υπάρχουν πολλοί τρόποι αντιμετώπισης του παραπάνω προβλήματος όπως Wang, H et al (2019) και Wang, L et al (2021). Ένα υποψήφιο κριτήριο για αυτή την επιλογή του βέλτιστου υποσυνόλου είναι να χρησιμοποιήσουμε τον πλήρες πίνακα σχεδιασμού με αποτέλεσμα κάθε φωλιασμένο μοντέλο να μην χάνει πληροφορία. Τέλος, αξίζει να σημειωθεί πως το πρόβλημα των μεγάλων δεδομένων γίνετε όλο και πιο σύνθετο όταν λαμβάνουμε υπόψη έναν μεγάλο αριθμό αριθμό μεταβλητών. Παρομοίως, υπάρχουν και εδώ πολλοί τρόποι αντιμετώπισης που θα ήταν επίσης ένα αντικείμενο ενασχόλησης για να εξεταστούν σε συνδυασμό με το μεγάλο όγκο παρατηρήσεων.

Μετά από όλα αυτά, θα πρέπει να χρησιμοποιηθούν πραγματικά δεδομένα για την επικύρωση των αποτελεσμάτων και τη διασφάλιση της εγκυρότητας των ευρημάτων. Μπορεί να προκύψουν νέες παρατηρήσεις και ζητήματα, αλλά το ευρύ φάσμα έρευνας πάνω σε προσομοιωμένα δεδομένα μπορεί να οδηγήσει στην ανάπτυξη νέων μεθόδων ή την βελτίωση παλαιών, με έναν πιο αποτελεσματικό τρόπο.

Βιβλιογραφία

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). *Bayesian Data Analysis (3rd ed.)*. Chapman and Hall/CRC.

Art B. Owen and Peter W. Glynn. (2016). *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC* .Publisher: Springer International Publishing AG.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Studies in managerial economics. Publisher: Division of Research, Graduate School of Business Administration, Harvard University.

Daniel Fink. (1997). *A compendium of conjugate priors*. volume 46. Publisher: Citeseer.

Anthony O'Hagan. (2004). *Bayesian statistics: principles and benefits*.

Wagenmakers, EJ. (2007). *A practical solution to the pervasive problems of p values*. Psychonomic Bulletin & Review 14, 779–804.

Harold Sir Jeffreys. (1935). *Some Tests of Significance, Treated by the Theory of Probability*. Mathematical Proceedings of the Cambridge Philosophical Society. volume 31, 203–222.

Harold Sir Jeffreys. (1961). *Theory of Probability.* 3rd Edition, Clarendon Press, Oxford.

Kass, Robert E and Raftery, Adrian E. (1995). *Bayes factors*. Journal of the american statistical association. volume 90. Publisher: Taylor & Francis, 773-795.

Richard D. Morey and Jan-Willem Romeijn and Jeffrey N. Rouder. (2016). *The philosophy of Bayes factors and the quantification of statistical evidence*. Journal of Mathematical Psychology. volume 72, 6-18.

Dellaportas, Petros and Roberts, Gareth O. (2003). *An Introduction to MCMC*. Publisher: Springer New York, 1–41.

Eric C. Anderson. (1999). *Monte Carlo Methods and Importance Sampling*.

Mackay, D. J. C. (1998). *Introduction to Monte Carlo Methods*. Publisher:Springer Netherlands, pp 175–204.

Geman, Stuart and Geman, Donald. (1984). *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 721-741.

Tanner, M. A., & Wong, W. H. (1987). *The Calculation of Posterior Distributions by Data Augmentation.* Journal of the American Statistical Association, 528–40.

Gelfand, A. E., & Adrian F. M. Smith. (1990). *Sampling-Based Approaches to Calculating Marginal Densities.* Journal of the American Statistical Association, pp. 398–409.

Metropolis, N., et al. (1953). *Equation of State Calculations by Fast Computing Machines.* The Journal of Chemical Physics, 1263 - 1269.

Hastings, W. K. (1970). *Monte Carlo Sampling Methods Using Markov Chains and Their Applications.* Biometrika, 97–109.

Peter J. Green. (1995). *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination* Biometrika, Volume 82, 711-732.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., & Stine, R. A. (2001). *The Practical Implementation of Bayesian Model Selection.* Lecture Notes-Monograph Series, 38, 65–134

Carlin, Bradley P., and Siddhartha Chib. (1995). *Bayesian Model Choice via Markov Chain Monte Carlo Methods.* Journal of the Royal Statistical Society. Series B (Methodological) 57, pp 473–84.

Hoijtink, H., Klugkist, I. (2007). *Comparison of Hypothesis Testing and Bayesian Model Selection.* Qual Quant 41, 73–91

Inseok Park and Hemanth K. Amarchinta and Ramana V. Grandhi. (2010). *A Bayesian approach for quantification of model uncertainty.* Journal of Reliability Engineering & System Safety, volume 95, 777-785.

Hinne M, Gronau QF, van den Bergh D, Wagenmakers E-J. (2020). *A Conceptual Introduction to Bayesian Model Averaging.* Advances in Methods and Practices in Psychological Science, 200–215.

Ntzoufras I. (1999). *Aspects of Bayesian model and variable selection using MCMC.*

Green Peter J. (1995). *Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination.* Biometrika 82, no. 4, 711–32.

Patrick Rebeschini et al. (2018). *Advanced Simulation Methods Chapter 7 - Reversible Jump MCMC.*

Patrick Rebeschini et al. (2012). *Model choice using reversible jump Markov chain Monte Carlo.* Statistica Neerlandica 66.3: 309-338

George, Edward I., and Robert E. McCulloch. (1993). *Variable Selection Via Gibbs Sampling.* Journal of the American Statistical Association, vol. 88, no. 423, pp. 881–89

George, Edward I., and Robert E. McCulloch. (1997). *APPROACHES FOR BAYESIAN VARIABLE SELECTION.* Statistica Sinica, vol. 7, no. 2, pp. 339–73. JSTOR

George, Edward I., and Robert E. McCulloch. (1996). *Stochastic search variable selection.* Journal of Markov Chain Monte Carlo in Practice

Perrakis, Konstantinos and Ntzoufras, Ioannis. (2015). *Stochastic Search Variable Selection (SSVS).* Wiley StatsRef: Statistics Reference Online, 1-6.

Dellaportas, Petros and Forster, Jonathan and Ntzoufras, Ioannis. (1998). *Bayesian Variable Selection Using the Gibbs Sampler.* Journal of Generalized Linear Models: A Bayesian Perspective

Dellaportas, Petros, Jonathan J. Forster, and Ioannis Ntzoufras. (2002). *On Bayesian model and variable selection using MCMC.* Statistics and computing 12.1: 27-36.

Ntzoufras, Ioannis. (2002). *Gibbs variable selection using BUGS.* Journal of Statistical Software, 7, 1-19.

Ntzoufras, I. (2011). *Bayesian Modeling Using WinBUGS.* Wiley Series in Computational Statistics

Kuo, Lynn, and Bani Mallick. (1998). *Variable Selection for Regression Models.* The Indian Journal of Statistics, Series B (1960-2002), vol. 60, pp. 65–81.

Merlise A. Clyde, Joyee Ghosh and Michael L. Littman. (2011). *Bayesian Adaptive Sampling for Variable Selection and Model Averaging.* Journal of Computational and Graphical Statistics, volume 20, 80-101.

Selke, T., Bayarri, M., and Berger, J. (2001). *Calibration of P-values for testing precise null hypotheses.* The American Statistician, 55, 62–71

Gideon Schwarz. (1978). *Estimating the Dimension of a Model.* Ann. Statist, vol. 6, no. 2, pp. 461–64.

Wang, H., Yang, M., and Stufken, J. (2019). *Information-based optimal subdata selection for big data linear regression.* Journal of the American Statistical Association, 114(525):393–405.

Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021). *Orthogonal subsampling for big data linear regression.* Annals of Applied Statistics, 15(3):1273–1290.

Berger, James O., and Robert L. Wolpert. (1988). *The Likelihood Principle and Generalizations.* The likelihood principle. Vol. 6. Institute of Mathematical Statistics, 19-65.

Pedersen, J. G. (1978). *Fiducial Inference.* International Statistical Review, vol. 46, no. 2, 1978, pp. 147–70. JSTOR.

Δελλαπόρτας, Π και Τσιαμυρτζής, Π. (2004). *Στατιστική κατά Bayes*.

Παναγιώτης Παπασταμούλης. (2019). *Μέθοδοι Μπεϋζιανής Συμπερασματολογίας*.