Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Εθνικό Μετσόβιο Πολυτεχνείο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Hamiltonian MCMC και Εφαρμογές σε STAN

Ιωάννης Ρωτούς

Επιβλέπων Καθηγητής: Φουσκάκης Δημήτριος

Τριμελής Επιτροπή
Φουσκάκης Δημήτριος
Λουλάκης Μιχαήλ
Ντζούφρας Ιωάννης

Αθήνα, Οκτώβριος 2019

Περιεχόμενα

Περίληψη

Τα συνηθέστερα ερωτήματα που πρέπει να απαντήσει κάποιος ερευνητής τις περισσότερες φορές μεταφράζονται σε υπολογισμούς ολοκληρωμάτων, για παράδειγμα, ο υπολογισμός της αναμενόμενης τιμής ή της συνδιακύμανσης τυχαίων μεταβλητών.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x dF_X(x).$$

$$cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Συνεπώς η ικανότητα να παίρνουμε δείγμα από κατανομές πιθανότητας με αποτελεσματικό τρόπο είναι αναγκαία. Ας δούμε όμως μερικές από τις μεθόδους οι οποίες χρησιμοποιούνται συνηθέστερα. Αρχικά στον $\mathbb{R}^n$ για $n = 1$ έχουμε για παράδειγμα τις μεθόδους της Αντιστροφής, Importance Sampling , Rejection Sampling και άλλες, που όμως όταν οι διαστάσεις του προβλήματος αυξάνονται, $n > 1$, οι μέθοδοι αυτοί δεν γενικεύονται ή αρχίζουν να γίνονται προβληματικές. Γι΄ αυτό τον λόγο χρησιμοποιήθηκαν οι Markov Chain Monte Carlo (MCMC) μέθοδοι οι οποίοι καταφέρνουν να παράγουν δείγμα από περίπλοκες κατανομές και να αντιμετωπίσουν με αποτελεσματικότητα το πρόβλημα των μεγάλων διαστάσεων.
Παρόλα αυτά και οι MCMC αλγόριθμοι όπως ο Metropolis-Hastings και ειδικότερα ο Random Walk Metropolis για μεγάλες διαστάσεις εμφανίζουν κάποιες φορές παθολογικά προβλήματα.
Αρχικά για να κατανοήσουμε καλύτερα μια πτυχή αυτών των προβλημάτων θα πρέπει να καταλάβουμε πως συμπεριφέρεται η μάζα των κατανομών σε μεγάλες διαστάσεις. Για παράδειγμα γνωρίζουμε ότι η κανονική κατανομή $N(0,1)$ στον $\mathbb{R}^n$ για $n = 1$ έχει το μεγαλύτερο ποσοστό της μάζας της στο κέντρο. Αντίθετα όταν $n >> 1$ η κανονική κατανομή συγκεντρώνει το μεγαλύτερο ποσοστό της μάζας της στις ουρές και αυτό θα το διαπιστώσουμε άμεσα.

•Έστω χωρίς βλάβη της γενικότητας μια πολυμεταβλητή κανονική κατανομή $\mathcal{N}(0, I_p)$ στον $\mathbb{R}^p$ με συνάρτηση πυκνότητας πιθανότητας $g_p(\boldsymbol{x}) = (2\pi)^{-p/2} \frac{\|\boldsymbol{x}\|^2}{2}$.
•Επίσης η μέγιστη τιμή που μπορεί να πάρει η $g_p(x)$ είναι για $x = 0$ που όμως με την αύξηση των διαστάσεων γίνεται αντιληπτό ότι η μάζα της κατανομής τείνει στο 0, $g_p(0) = (2\pi)^{-\pi/2} \overset{p\to\infty}{\to} 0$.
•Στην συνέχεια υπολογίζουμε την μάζα στην ΄καμπάνα΄ της κατανομής όπου η πιθανότητα είναι μεγαλύτερη για αυτές τις παρατηρήσεις. Έστω $\delta > 0$ με $\delta \in \mathbb{R}$ και κλειστή μπάλα

$$B_{p,\delta} = \{\boldsymbol{x} \in \mathbb{R}^p : g_p(\boldsymbol{x}) \geq \delta g_p(0)\} = \left\{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|^2 \leq 2log(\delta^{-1})\right\}$$

δηλαδή στην $B_{p,\delta}$ θα ανήκουν εκείνα τα σημεία για τα οποία η πιθανότητα τους ανήκει στο διάστημα $[\delta g_p(0), g_p(0)]$.

$$\mathbb{P}(X \in B_{p,\delta}) = \mathbb{P}(e^{\frac{-\|\boldsymbol{x}\|^2}{2}} \geq \delta) \leq \frac{1}{\delta}\mathbb{E}[e^{\frac{-\|\boldsymbol{x}\|^2}{2}}] \quad \text{(Markov Inequality)}$$

$$= \frac{1}{\delta}\int_{\boldsymbol{x}\in\mathbb{R}^p} e^{\|\boldsymbol{x}\|^2} \frac{dx}{(2\pi)^{p/2}} = \frac{1}{\delta 2^{\pi/2}} \overset{p\to\infty}{\to} 0.$$

Συνεπώς το μεγαλύτερο ποσοστό της μάζας της πολυμεταβλητής κανονικής κατανομής είναι συγκεντρωμένο στις ουρές. Γενικότερα είναι συχνό φαινόμενο η μάζα των γεωμετρικών αντικειμένων με την αύξηση των διαστάσεων να συγκεντρώνεται στα εξωτερικά σημεία, γι΄ αυτό τον λόγο θα χρειαστεί μια μέθοδος που να αντιλαμβάνεται πως να κινηθεί στην κατανομή που μας ενδιαφέρει αξιοποιώντας

την γεωμετρία της.
Στους MCMC αλγορίθμους αυτό το φαινόμενο θα έχει ως αποτέλεσμα αν για παράδειγμα κάποιος χρησιμοποιήσει Random Walk Metropolis

$$a(q^{'}|q) = min(1, \frac{\pi(q^{'})}{\pi(q)}),$$

σχεδόν ποτέ να μην εξερευνήσει τις ουρές, μιας και θα έχουν πολύ μικρή πιθανότητα αποδοχής, αν και οι ουρές είναι οι περιοχές με την μεγαλύτερη μάζα πιθανότητας. Επιπλέον να τονίσουμε ότι λόγω της random-walk συμπεριφοράς αυτών των αλγορίθμων εξαιτίας της έλλειψης συστηματικού τρόπου εξερεύνησης της κατανομής, υπάρχει μεγάλη πιθανότητα περιοχές της κατανομής να αφεθούν εκτός δείγματος ή να έχουν ελλειπή αντιπροσώπευση από το δείγμα.
Επίσης ένα ακόμα πρόβλημα που αντιμετωπίζουν οι αλγόριθμοι MCMC είναι ότι έχουν μεγάλη δυσκολία να παράγουν αντικειμενικά δείγματα από περιοχές με μεγάλη καμπυλότητα. Συνηθέστερα τέτοιες περιοχές συναντιούνται στις περιπτώσεις που χρησιμοποιούμε ιεραρχικά μοντέλα. Ως αποτέλεσμα στην προσπάθεια τους να εξερευνήσουν αυτές τις περιοχές μεγάλης καμπυλότητας, παίρνουμε πληθώρα προτεινόμενων παρατηρήσεων οι οποίες βρίσκονται πάνω στο σύνορο αυτών των περιοχών. Κατά συνέπεια το δείγμα που θα πάρουμε δεν θα είναι αντιπροσωπευτικό της κατανομής, μιας και θα περιέχει ένα σημαντικό ποσοστό παρατηρήσεων από την περιοχή μεγάλης καμπυλότητας και θα εισάγει μεροληψία στην συμπερασματολογία μας.
Γίνεται κατανοητό ότι ακόμα και οι αλγόριθμοι MCMC υστερούν και δεν είναι αποτελεσματικοί σε απαιτητικά προβλήματα. Γι΄ αυτό τον λόγο ερευνητές όπως ο Radford M. Neal, Michael Betancourt, Andrew Gelman, Mathew Hoffman και άλλοι, εισήγαγαν και συνέβαλαν ο καθένας από την δική του οπτική γωνία στην δημιουργία και αποτελεσματική υλοποίηση του αλγορίθμου Hamiltonian Monte Carlo και εξήγησαν τα πλεονεκτήματα του. Μέσω του οποίου όχι μόνο ξεπέρασαν τις παθολογίες που αναφέραμε αλλά για παράδειγμα έκαναν πιο αποτελεσματική την εξερεύνηση κατανομών, με μεγαλύτερα και όχι τυχαία βήματα όπως του Random Walk Metropolis.Κάνοντας εύκολη και ουσιαστικά αυτόματη την λήψη δειγμάτων από περίπλοκες κατανομές με την χρήση της πλατφόρμας Stan και του σύγχρονου αλγορίθμου NUTS .

Κεφάλαιο 1

Εισαγωγή στην Μπεϋζιανή στατιστική

1.1 Εισαγωγή

Το 1701 στο Λονδίνο της Αγγλίας γεννιέται ο Tomas Bayes ο οποίος πρόκειται να παίξει σημαντικό ρόλο στην εξέλιξη της στατιστικής συμπερασματολογίας. Σπούδασε θεολογία και μαθηματικά στο Πανεπιστήμιο του Εδιμβούργου και ήταν ο πρώτος ο οποίος έκανε χρήση πιθανοτήτων επαγωγικά και αυτό τεκμηριώνεται από την διατύπωση του Θεωρήματος Bayes γύρω στο 1740. Επίσης το 1763 εξέδωσε τα ευρήματα του στο Philosophical Transactions του Royal Society με τίτλο "Essay Towards Solving a Problem in the Doctrine of Chances". Από εκείνη την στιγμή και έπειτα η συγκεκριμένη εργασία του Thomas Bayes έγινε το έναυσμα για την ανάπτυξη της Μπευζιανής συμπερασματολογίας. Με το πέρασμα των αιώνων αρχίζει να παίρνει μορφή και υπόσταση, η στατιστική προσέγγιση που εισήχθει μέσω της υποκειμενικότητας και της φυσικής ερμηνείας των παραμέτρων ως τυχαίων μεταβλητών. Παρόλα αυτά η υπολογιστική ικανότητα της εποχής ήταν πολύ μικρή και για πολλά χρόνια οι ιδέες αυτές θα παρέμεναν στο παρασκήνιο. Τελικά, από την δεκαετία του 1950 και έπειτα που οι ηλεκτρονικοί υπολογιστές άρχισαν να αναπτύσσονται ραγδαία, μπόρεσαν οι επιστήμονες να αναπτύξουν αυτές τις πρώιμες ιδέες σε αυτό που ονομάζουμε στις μέρες μας Μπεϋζιανή Στατιστική.

1.1.1 Κλασική ή Μπεϋζιανή Στατιστική .

Η κυρίαρχη διαφορά μεταξύ της Κλασικής και Μπεϋζιανής στατιστικής είναι ότι στην πρώτη οι παράμετροι θεωρούνται ώς σταθερές ενώ στην δεύτερη περίπτωση ώς άγνωστες τυχαίες μεταβλητές. Κατά συνέπεια στην Κλασική στατιστική θα έχουμε σημειακές εκτιμήσεις για τις παραμέτρους (εκτιμήτριες Μέγιστης Πιθανοφάνειας) ενώ στην Μπεϋζιανή θα κάνουμε χρήση πρότερων (prior)κατανομών. Επίσης, γίνεται αντιληπτό απο τον ορισμό των δύο προσεγγίσεων ότι η ερμηνεία στην πρώτη περίπτωση θα είναι αρκετά πιο δύσκολα κατανοήσιμη σε σύγκριση με την φυσική ερμηνεία που εμπεριέχεται στην Μπεϋζιανή προσέγγιση. Για παράδειγμα, έστω θ μια παράμετρος ενδιαφέροντος, στην Κλασική στατιστική ένα 95% διάστημα εμπιστοσύνης θα ερμηνεύεται ώς εξής:

Αν κατασκευάσουμε 100 διαστήματα εμπιστοσύνης τα 95 απο αυτά θα περιέχουν την παράμετρο θ,

αντίθετα στην Μπεϋζιανή στατιστική θα έχουμε ότι ένα διάστημα εμπιστοσύνης 95% θα ερμηνεύεται ώς εξής:

Η παράμετρος θ θα ανήκει στο διάστημα εμπιστοσύνης με πιθανότητα 0.95.

Συνεπώς γίνεται αντιληπτό ότι η χρήση πιθανοτήτων για αντικείμενα που περιέχουν αβεβαιότητα βοηθάει στην κατανόηση και επικοινωνία τους.

Μπεϋζιανή Συμπερασματολογία:

- Χρησιμοποιεί πιθανότητες για την μοντελοποίηση στατιστικών υποθέσεων,παραμέτρων,μοντέλων και δεδομένων.
- Η συμπερασματολογία εξαρτάται απο την επιλογή της πρότερης κατανομής και της συνάρτησης πιθανοφάνειας.
- Δεν υπάρχει μεθοδολογία για την επιλογή της πρότερης κατανομής.
- Μπορεί να χρειαστούν αρκετοί υπολογιστικοί πόροι για τον υπολογισμό πολλαπλών ολοκληρωμάτων.
- Αντιμετωπίζει δυσκολία στις περιπτώσεις που έχουμε λίγα δεδομένα.
- Η ερμηνεία των αποτελεσμάτων είναι απλή,για παράδειγμα μπορείς να απαντήσεις με χρήση πιθανοτήτων πόσο πιθανό είναι να έχει κάποιος μια ασθένεια.
- Μπορεί να υπάρχει συνεχής ανανέωση της πεποίθησης μας όσο έρχονται καινούργια δεδομένα.

Κλασική Συμπερασματολογία:

- Δύσκολη ερμηνεία αποτελεσμάτων.
- Εξαρτάται απο την συνάρτηση πιθανοφάνειας.
- Υπολογιστικά είναι πιο εύκολη.
- Το στατιστικό πείραμα περιγράφετε εξολοκλήρου εξαρχής.

1.1.2 Θεώρημα Bayes

Στην συνέχεια θα αναφερθούμε στο θεώρημα Bayes και πως αυτό συμβάλει στην Μπεϋζιανή στατιστική.

Θεώρημα 1.1.1 *Έστω A, B ενδεχόμενα σε έναν δειγματικό χώρο Ω με πιθανότητα $\mathbb{P}(B) \neq 0$. Τότε θα ισχύει:*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Μπορεί να γίνει εύκολα κατανοητό ότι το θεώρημα του Bayes μπορεί να χρησιμοποιηθεί για να ανανεώνουμε επαγωγικά την υπάρχουσα πληροφορία που έχουμε για το εκάστοτε πρόβλημα.
Για παράδειγμα, διεξάγουμε έρευνα για τον υπολογισμό της πιθανότητας κάποιος χρήστη να λάβει spam e-mail.Έστω A το ενδεχόμενο ένα e-mail να είναι spam και B το ενδεχόμενο το e-mail να περιέχει συγκεκριμένες λέξεις που το καθορίζουν ώς spam.Επίσης ας υποθέσουμε ότι αρχικά πιστεύουμε ότι η πιθανότητα ένα e-mail να είναι spam είναι 20% (πρότερη κατανομή) και εξετάζουμε ένα καινούργιο e-mail μέσω του οποίου βρίσκουμε ότι το 45% των λέξεων ενδείκνυνται για spam e-mail,με δεσμευμένη πιθανότητα $\mathbb{P}(B|A) = 0.7$ (συνάρτηση πιθανοφάνειας).Τότε εφαρμόζοντας το θεώρημα του Bayes θα έχουμε:

$$\mathbb{P}(A|B) = \frac{0.7 * 0.2}{0.45} = 0.31$$

Εξετάζοντας ένα καινούργιο e-mail , δηλαδή εισάγοντας καινούργια πληροφορία, η πιθανότητα να λαμβάνει κάποιος spam e-mail απο 0.2 έγινε 0.3.Σε αυτό το πλαίσιο λογικής βασίζεται και η Μπεϋζιανή συμπερασματολογία .

Τέλος, να σημειώσουμε ότι όλη η πληροφορία για το ενδεχόμενο A βρίσκεται μόνο στον αριθμητή, για τον λόγο αυτό θα μπορούμε να αναφερόμαστε στην πιθανότητα $\mathbb{P}(A|B)$ (ύστερη κατανομή) ως:

$$\mathbb{P}(A|B) \propto \mathbb{P}(B|A)\mathbb{P}(A).$$

Αυτή η αναλογία αναφέρεται ώς κανόνας του Bayes και προαναφέρθηκε απο τον Piere Laplace (1814).

1.1.3 Συμπερασματολογία

Έχουμε ήδη εισάγει την Μπεϋζιανή προσέγγιση και το Θεώρημα του Bayes,εν συνεχεία θα δούμε πώς υλοποιούνται τα προαναφερθέντα στην πράξη. Έστω θ η παράμετρος ενδιαφέροντος και $\boldsymbol{x} = (x_1, ..., x_n)$ με $f(x_i|\theta)$ τα δεδομένα που έχουν συλλεχθεί. Θα μπορούσαμε να χωρίσουμε την όλη διαδικασία σε τέσσερα βήματα:

1. διαλέγουμε πρότερη κατανομή, $\pi(\theta)$, για την παράμετρο θ,
2. με συνάρτηση συνάρτηση πιθανοφάνειας $L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$,
3. υπολογίζουμε τον κανόνα του Bayes $\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta) \cdot \pi(\theta)$ και παίρνουμε την ύστερη κατανομή του θ,
4. για καινούργιο $\boldsymbol{x}' = (x_1', ..., x_n')$ ξανά υπολογίζουμε τον κανόνα του Bayes με πρότερη κατανομή αυτή την φορά την $\pi(\theta|\boldsymbol{x})$, δηλαδή $\pi(\theta|\boldsymbol{x}') \propto f(\boldsymbol{x}'|\theta) \cdot \pi(\theta|\boldsymbol{x})$.

Παρατηρούμε δηλαδή ότι η συμπερασματολογία για το θ βασίζετε εξολοκλήρου στην πρότερη κατανομή και στην συνάρτηση πιθανοφάνειας, με απλά λόγια έχουμε ότι

$$\text{Ύστερη} = \text{Πεποίθηση} + \text{Παρατηρηθέντα Δεδομένα} .$$

Τέλος, όπως και η πεποίθηση μας μπορεί να κυμαίνεται μεταξύ πλήρης αβεβαιότητας και πλήρης σιγουριάς έτσι θα συμπεριφέρεται και η πρότερη κατανομή που θα επιλέγουμε κάθε φορά. Δηλαδή αν θα παρέχει κάποια πληροφορία ή αν θα είναι μή πληροφοριακή λόγω της αβεβαιότητας μας.

1.2 Πρότερες Κατανομές

Όπως έχει γίνει αντιληπτό μέχρι στιγμής η ποσοτικοποίηση της αβεβαιότητας για την παράμετρο θ προέρχεται μέσω της πρότερης κατανοής,μιας και η παράμετρος είναι τυχαία μεταβλητή. Ώς εκ τούτου η Μπεϋζιανή συμπερασματολογία έχει δεχτεί μεγάλη κριτική διότι για την επιλογή της πρότερης κατανομής δεν ακολουθείται κάποιο κριτήριο και υπάρχει πληθώρα πρότερων κατανομών που μπορούν να επιλεχθούν. Συνεπώς κατά την διαδικασία επιλογής πρότερης κατανομής θα πρέπει να είμαστε πολύ προσεκτικοί. Θα πρέπει το πεδίο ορισμού να συνάδει με τις τιμές που μπορεί να πάρει η παράμετρος θ. Ακόμα θα πρέπει κάποιος να λαμβάνει υπόψιν της φύση της παραμέτρου, για παράδειγμα αν μας απασχολεί η μοντελοποίηση ενός ποσοστού, θα πρέπει να χρησιμοποιήσουμε μια Βήτα κατανομή και όχι μια Κανονική κατανομή. Στην συνέχεια θα αναφερθούμε σε δύο μεγάλες κατηγορίες πρότερων κατανομών τις πληροφοριακές και τις μη πληροφοριακές και τι αποτελέσματα έχουν στην συμπερασματολογία της ύστερης κατανομής.

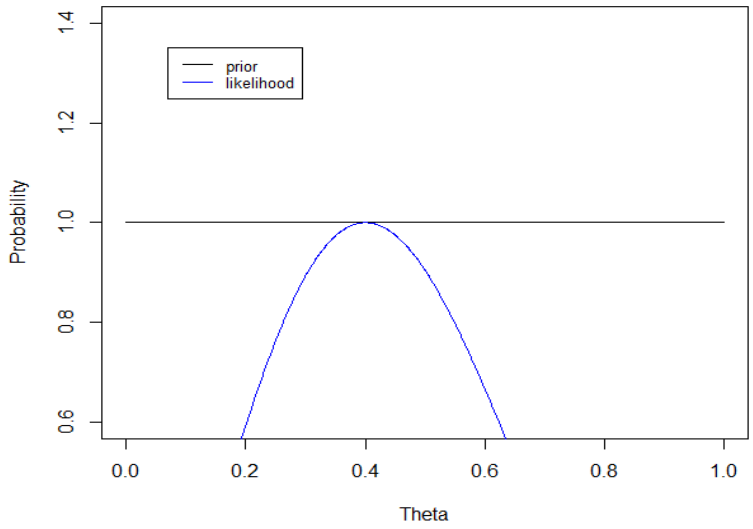
1.2.1 Μη Πληροφοριακές Πρότερες Κατανομές

Πολλές φορές στα προβλήματα που μελετάμε είναι πιθανό να μην έχουμε κάποια εκ των προτέρων πληροφορία. Αυτή η έλλειψη πληροφορίας για την παράμετρο θ θα πρέπει να εκφραστεί μέσω μιας μή πληροφοριακής πρότερης κατανομής. Ουσιαστικά θα είναι μια κατανομή η οποία θα δίνει σχεδόν το ίδιο βάρος σε όλες τις πιθανές τιμές της παραμέτρου θ με απώτερο σκοπό η συμπερασματολογία να επηρεαστεί κατα κύριο λόγο απο τα δεδομένα. Παρόλα αυτά ας έχουμε υπόψιν μας την ακόλουθη σημείωση για τις μή πληροφοριακές πρότερες κατανομές οι οποίες τείνουν να είναι επίπεδες. Δηλαδή αυτές οι οποίες δίνουν το ίδιο βάρος σε όλες τις πιθανές τιμές της παραμέτρου στο πεδίο ορισμού τους και άπειρο βάρος σε αυτές εκτός του πεδίου ορισμού τους. Για παράδειγμα, μια ομοιόμορφη στο $(0, 1)$ για την παράμετρο θ θα έχει πολύ μεγαλύτερη μάζα στα σημεία που βρίσκονται εκτός του $0 \leq \theta \leq 1$ με αποτέλεσμα η συμπερασματολογία μας να ευνοεί ακραίες τιμές του θ ,για αυτό τον λόγο θα πρέπει να είμαστε πολύ προσεκτικοί στην επιλογή των πρότερων κατανομών.

Επιπλέον, τέτοιου είδους κατανομές μπορούν να είναι μια Ομοιόμορφη κατανομή, μια Κανονική κατανομή με μεγάλη διακύμανση, μια Βήτα κατανομή με ίδιες παραμέτρους κτλ. Ας δούμε το ακόλουθο απλό παράδειγμα:

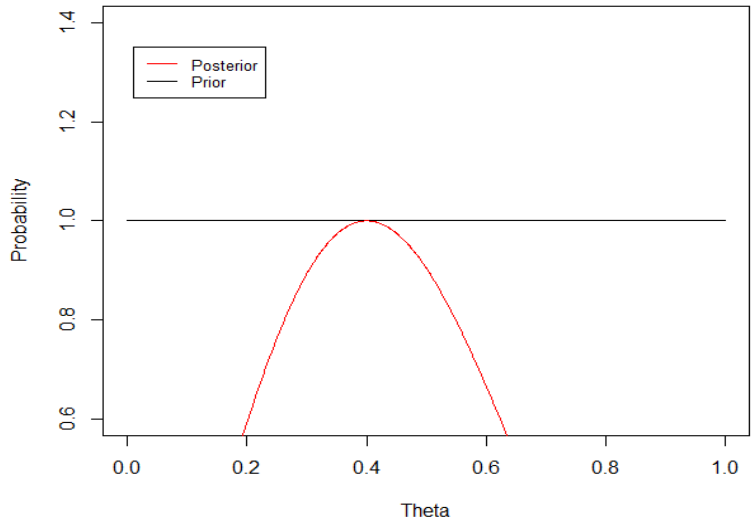
$$\theta \sim Beta(a = 1, b = 1)$$

$$\boldsymbol{x}|\theta \sim Binomial(N = 5, \theta)$$
$$\theta|\boldsymbol{x} \sim Beta(a + \sum x_i, b + \sum N - \sum x_i)$$



Διάγραμμα 1.1: Απεικόνιση της μη πληροφοριακής πρότερης κατανομής και της συνάρτησης πιθανοφάνειας των δεδομένων.

Παρατηρούμε ότι στην περιοχή που η συνάρτηση πιθανοφάνειας των δεδομένων δίνει βάρος, η πρότερη κατανομή είναι επίπεδη με αποτέλεσμα να αφήνει τα δεδομένα να επηρεάσουν εξολοκλήρου την ύστερη κατανομή.



Διάγραμμα 1.2: Απεικόνιση της μη πληροφοριακής πρότερης κατανομής και της ύστερης κατανομής.

Επίσης, παρατηρούμε ότι η ύστερη κατανομή ταυτίζεται με την συνάρτηση πιθανοφάνειας για τον λόγο τον οποίο αναφέραμε προηγουμένως.
Αξίζει να σημειωθεί ότι η πρότερη και η ύστερη κατανομή ανήκουν στην οικογένεια των Βήτα κατανομών. Αυτό ισχύει διότι η Βήτα καντανομή είναι συζυγής πρότερη κατανομή της Διωνυμικής κατανομής. Η χρήση συζυγής πρότερης κατανομής βοηθάει στον να αποφύγουμε τον υπολογισμό του παρανομαστή στο Θεώρημα Bayes , που μπορεί ο υπολογισμός του να γίνει πολύ απαιτητικός.

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta}.$$

Συνάρτηση Πιθανοφάνειας	Συζυγής Πρότερη Κατανομή	Παράμετροι Ύστερης Κατανομής
Binomial(p)	Beta(a,b)	$a+\sum_{i=1}^{n} x_i, b+n-\sum_{i=1}^{n} x_i$
Negative Binomial(r,p) με γνωστό r	Beta(a,b)	$a+\sum_{i=1}^{n} x_i, b+rn$
Poisson(λ)	Gamma(a,b)	$a+\sum_{i=1}^{n} x_i, b+n$
Exponential(λ)	Gamma(a,b)	$a+n, b+\sum_{i=1}^{n} x_i$
Normal(μ,σ) με σ γνωστό	Normal(μ_0,σ_0)	$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$

Τέλος, να σημειώσουμε ότι οι περισσότερες μή πληροφοριακές πρότερες κατανομές δεν είναι αναλλοίωτες σε μετασχηματισμούς. Αυτό σημαίνει ότι για μια μή πληροφοριακή $\pi(\theta)$ αν χρησιμοποιήσουμε μια παραμετροποίηση $k=g(\theta)$ τότε η $\pi(k)$ είναι πολύ πιθανό να είναι πληροφοριακή. Στην περίπτωση που θέλουμε να αποφύγουμε αυτό το ιδιάζον φαινόμενο μπορούμε να χρησιμοποιήσουμε την Jeffreys prior η οποία ορίζεται μέσω της πληροφορίας του Fisher

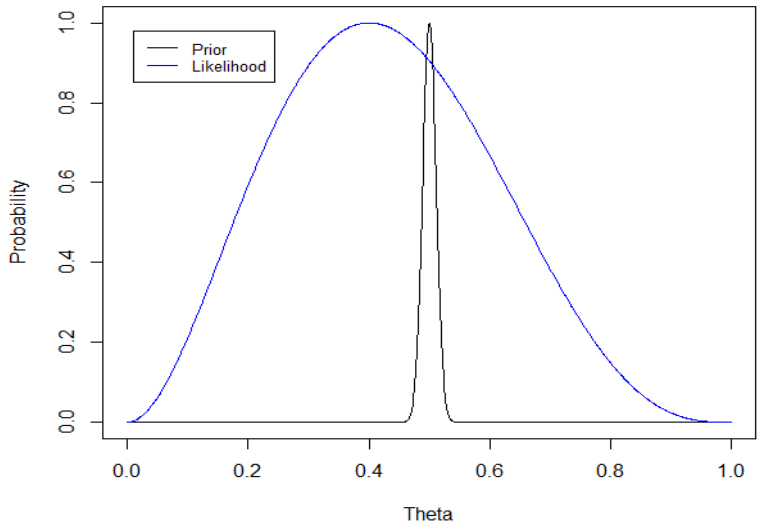
$$\pi_J(\theta) \propto I(\theta)^{\frac{1}{2}},$$

με $I(\theta) = -\mathbb{E}_\theta[\frac{d^2 log f(x|\theta)}{d\theta^2}]$. Θα πρέπει όμως να προσέξουμε η πρότερη κατανομή που θα παράξουμε όντως να ολοκληρώνει στην μονάδα.

1.2.2 Πληροφοριακές Πρότερες Κατανομές

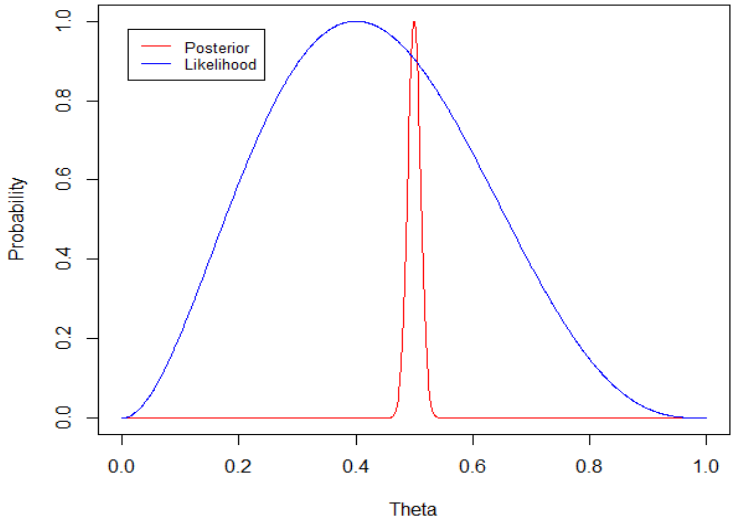
Αντίθετα στην περίπτωση που έχουμε πληροφορία για την τυχαία παράμετρο θ, τότε μπορούμε να διαλέξουμε μια πρότερη κατανομή η οποία να είναι σε αρμονία με τους περιορισμούς της παραμέτρου και δίνει βάρος στην περιοχή για την οποία έχουμε πληροφορία. Θα πρέπει όμως να είμαστε προσεκτικοί και να μην χρησιμοποιήσουμε πρότερες κατανομές οι οποίες τείνουν να είναι σημειακές διότι σε αυτή την περίπτωση τα δεδομένα που έχουν συλλεχθεί δεν θα ληφθούν υπόψιν στην συμπερασματολογία. Θα χρησιμοποιήσουμε το ίδιο παράδειγμα με προηγουμένως απλά αυτή την φορά για πρότερη κατανομή θα έχουμε:

$$\theta \sim Beta(a=1000, b=1000).$$



Διάγραμμα 1.3: Απεικόνιση της πληροφοριακής πρότερης κατανομής και της συνάρτησης πιθανοφάνειας των δεδομένων.

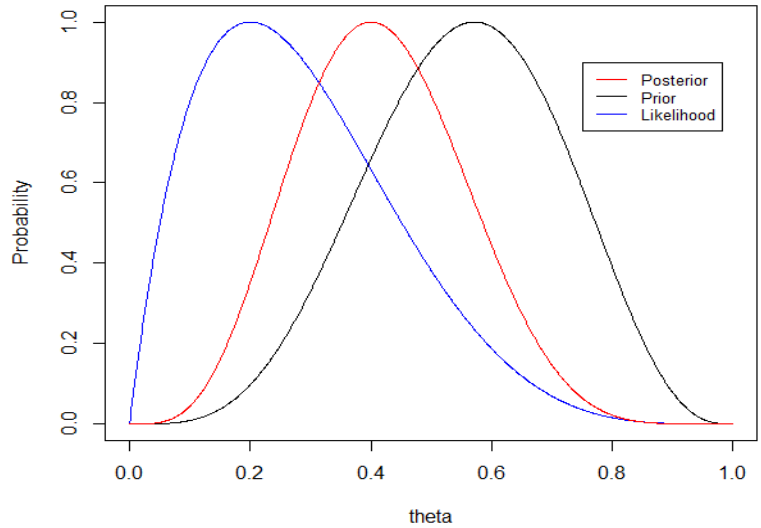
Παρατηρούμε ότι η πρότερη κατανομή είναι υπερπληροφοριακή παρόλα αυτά δίνει βάρος εκεί που δίνει και η συνάρτηση πιθανοφάνειας αυτό όμως θα επηρεάσει την μορφή της ύστερης κατανομής.



Διάγραμμα 1.4: Απεικόνιση της ύστερης κατανομής κατανομής και της συνάρτησης πιθανοφάνειας των δεδομένων.

Η ύστερη κατανομή δεν πήρε καθόλου πληροφορία απο τα δεδομένα για αυτό είναι ίδια με την πρότερη κατανομή.
Συνεπώς γίνεται αντιληπτό ότι η καλύτερη επιλογή για πρότερη κατανομή βρίσκεται στο ενδιάμεσο των δύο προηγούμενων παραδειγμάτων, ώστε να μπορέσει η ύστερη κατανομή να σταθμίσει αποτελεσματικά την πληροφορία απο τα δεδομένα και απο την πρότερη κατανομή. Για παράδειγμα θα μπορούσαμε να πάρουμε για πρότερη κατανομή μια

$$\theta \sim Beta(5, 4).$$



Διάγραμμα 1.5: Απεικόνιση και σύγκριση των πρότερων, ύστερων κατανομών και της συνάρτησης πιθανοφάνειας.

1.2.3 Μπεϋζιανή Συμπερασματολογία

Μέσω του Θεωρήματος Bayes μπορέσαμε να εξασφαλίσουμε την ύστερη κατανομή για την παράμετρο θ. Αυτό μας κάνει κατανοητό ότι για να απαντήσουμε οποιοδήποτε στατιστικό ερώτημα θα πρέπει να κάνουμε χρήση αναγκαστικά της ύστερης κατανομής.

Ας ξεκινήσουμε παραθέτοντας απλούς υπολογισμούς της διαμέσου, του μέσου και της διακύμανσης.

$$\int_{\infty}^{m(\theta)} \pi(\theta^{'}|\boldsymbol{x})d\theta^{'},$$

$$\mathbb{E}(\theta|\boldsymbol{x}) = \int \theta\pi(\theta|\boldsymbol{x})d\theta,$$

$$Var(\theta|\boldsymbol{x}) = \int [\theta - \mathbb{E}(\theta|\boldsymbol{x})]^2\pi(\theta|\boldsymbol{x})d\theta.$$

Επίσης μπορούμε να κατασκευάσουμε και αντίστοιχα διαστήματα εμπιστοσύνης τα ονομαζόμενα σύνολα αξιοπιστίας (credible sets).Ένα $100(1-a)\%$ σύνολο αξιοπιστίας είναι ένα υποσύνολο $\mathcal{T}$ του πεδίου ορισμού Θ, με σύνολο Θ να περιέχει όλες της πιθανές τιμές της παραμέτρου θ, τέτοιο ώστε να ισχύει ότι

$$\int_T \pi(\theta|\boldsymbol{x})d\theta = 1 - a,$$

το οποίο ερμηνεύεται ώς η πιθανότητα το $\theta \in \mathcal{T}$ να είναι ίση με $1 - a$. Συνηθέστερα αντί για το σύνολο αξιοπιστίας χρησιμοποιείται το Highest Posterior Density Interval το οποίο περιέχει τις πιο πιθανές τιμές της παραμέτρου θ. Ένα $100(1 - a)\%$ HPD Interval θα είναι ένα υποσύνολο $\mathcal{T}$ του Θ ορισμένο ώς

$$\mathcal{T} = \{\theta : \pi(\theta|\boldsymbol{x}) \geq \gamma\},$$

όπου γ είναι ο μεγαλύτερος αριθμός τέτοιος ώστε

$$\int_{\theta:\pi(\theta|\boldsymbol{x})\geq\gamma} \pi(\theta|\boldsymbol{x})d\theta = 1 - a.$$

Επιπλέον, μπορούμε να υλοποιήσουμε στατιστικούς ελέγχους υποθέσεων. Μόνο που αυτή την φορά μπορούμε να αναθέσουμε πιθανότητες στην κάθε υπόθεση που θέλουμε να ελέγξουμε.
Έστω H_0 και H_1 οι υποθέσεις που θέλουμε να ελέγξουμε. Όπως έχουμε αναφέρει πολλές φορές, στην Μπεϋζιανή προσέγγιση πάντα μοντελοποιούμε τις άγνωστες παραμέτρους μέσω πρότερων κατανομών. Έτσι και εδώ θα έχουμε $\mathbb{P}(H_0)$ και $\mathbb{P}(H_1)$ με $\mathbb{P}(H_0) + \mathbb{P}(H_1) = 1$. Τώρα το επόμενο βήμα είναι ο υπολογισμός των συναρτήσεων πιθανοφάνειας $f(x|H_0)$ και $f(x|H_1)$, δηλαδή πόσο πιθανό είναι τα δεδομένα να συνάδουν με τον ισχυρισμό της κάθε υπόθεσης. Τέλος, αφού κάνουμε χρήση του Θεωρήματος Bayes και υπολογίσουμε τις ύστερες κατανομές, θα δούμε πια υπόθεση είναι πιο πιθανή θα ελέγχοντας την ακόλουθη ανισότητα,

$$\mathbb{P}(H_0|x) > \mathbb{P}(H_1|x),$$

και θα επιλέξουμε την υπόθεση με την μεγαλύτερη ύστερη πιθανότητα.

Τέλος, μπορεί να θέλουμε να κάνουμε πρόβλεψη για μελλοντικές παρατηρήσεις αξιοποιώντας την πληροφορίας απο τις υπάρχουσες παρατηρήσεις. Για να το επιτύχουμε κάνουμε χρήση της προβλεπτικής κατανομής

$$f(y|\boldsymbol{x}) = \int f(y|\theta)\pi(\theta|\boldsymbol{x})d\theta.$$

Να σημειώσουμε ότι επειδή για την παράμετρο θ χρησιμοποιούμε την ύστερη κατανομή που εισάγει περισσότερη αβεβαιότητα θα έχουμε πιο παχιά ύστερη προβλεπτική κατανομή.

1.3 Υπολογιστικη Στατιστική

Στις προηγούμενες ενότητες παρατηρήσαμε ότι όλα τα στατιστικά ερωτήματα όπως διαστήματα εμπιστοσύνης, μέσοι κ.τ.λ. εμπεριέχουν τον υπολογισμό κάποιους ολοκληρώματος. Τα ολοκληρώματα αυτά βασίζονται σε παρατηρήσεις οι οποίες προέρχονται απο την ύστερη κατανομή της παραμέτρου ενδιαφέροντος. Το δύσκολο κομμάτι της Μπεϋζιανής συμπερασματολογίας είναι ο υπολογισμός αυτών

των ολοκληρωμάτων που μπορεί να γίνει πολύ απαιτητικός είτε γιατί μπορεί η παράμετρος ενδιαφέροντος να έχει πολλές διαστάσεις είτε γιατί το ολοκλήρωμα μπορεί να μην λύνεται με τις κλασσικές γνωστές προσεγγίσεις. Συνεπώς γίνεται αντιληπτό ότι για τον υπολογισμό τέτοιων ολοκληρωμάτων θα χρειαστούμε πιο αποτελεσματικές μεθόδους οι οποίες μπορεί να είναι είτε ντετερμινιστικές είτε στοχαστικές (Monte Carlo). Στις στοχαστικές μεθόδους με τις οποίες και θα ασχοληθούμε, ανήκουν διαδικασίες όπως Important Sampling και Markov Chain Monte Carlo .

1.3.1 Monte Carlo

Η ιδέα στις Monte Carlo μεθόδους είναι να χειριστούμε τον υπολογισμό του ολοκληρώματος, σαν τον υπολογισμό μιας αναμενόμενης τιμής κάποιας κατανομής, που στην δικιά μας περίπτωση θα είναι η ύστερη κατανομή της παραμέτρου ενδιαφέροντος. Έστω ότι θέλουμε να υπολογίσουμε το ακόλουθο ολοκλήρωμα

$$I = \int f(\theta)\pi(\theta)d\theta,$$

το οποίο μπορούμε να το δούμε ώς την αναμενόμενη τιμή $\mathbb{E}[f]_\pi$. Τότε αρκεί να πάρουμε σε μέγεθος N ανεξάρτητα και ισόνομα τυχαία δείγματα $\{\theta_n\}_{n=1}^N$ απο την ύστερη κατανομή $\pi(\theta)$ και να υπολογίσουμε τον αμερόληπτο εκτιμητή

$$\hat{I}_N = \frac{1}{N}\sum_{n=1}^{N} f(\theta_n).$$

Απο τον ισχυρό νόμο των μεγάλων αριθμών, ο εκτιμητής $\hat{I}_N$ θα συγκλίνει σχεδόν βέβαια στο I όσο μεγαλώνουμε το μέγεθος του δείγματος N. Επίσης, αν έχουμε ότι η διακύμανση της $f(\theta)$ είναι σ_f^2 τότε απο το κεντρικό οριακό θεώρημα θα έχουμε ότι το σφάλμα της εκτίμησης $\sqrt{N}(\hat{I}_N - I)$ θα συγκλίνει σε μια τυχαία μεταβλητή με κατανομή $\mathcal{N}(0, \sigma_f^2)$ και η διακύμανση του εκτιμητή μας θα είναι $Var(\hat{I}_N) = \frac{\sigma_f^2}{N}$.

1.3.2 Importance Sampling

Πολλές φορές όμως η δειγματοληψία απο την ύστερη κατανομή είναι αρκετά δύσκολη, με αποτέλεσμα για να εξασφαλίσουμε υπολογιστικούς πόρους να παίρνουμε δείγμα απο κάποια άλλη πιο εύκολη για δειγματοληψία κατανομή g. Επίσης στην περίπτωση που γνωρίζουμε τις σταθερές κανονικοποίησης των κατανομών θα έχουμε ότι

$$\pi = \frac{p}{Z} \quad g = \frac{q}{Z_g}.$$

Με την σειρά του το ολοκλήρωμα θα γίνεται

$$\mathbb{E}[f]_\pi = I = \int f(\theta)\pi(\theta)d\theta = \int f(\theta)\frac{p(\theta)}{Z}d\theta$$

$$= \frac{\int f(\theta)\frac{p(\theta)}{Z}}{\int \frac{p(\theta)}{Z}} = \frac{\frac{Z}{Z_g}\int f(\theta)\frac{p(\theta)}{Z}d\theta}{\frac{Z}{Z_g}\int \frac{p(\theta)}{Z}d\theta}$$

$$= \frac{\int f(\theta)\frac{p(\theta)}{q(\theta)}\frac{q(\theta)}{Z_g}d\theta}{\int \frac{p(\theta)}{q(\theta)}\frac{q(\theta)}{Z_g}d\theta} = \frac{\mathbb{E}[wf]}{\mathbb{E}[w]},$$

όπου $w(\theta) = \frac{p(\theta)}{q(\theta)}$ θα ονομάζεται important weight . Ο Monte Carlo εκτιμητής θα είναι

$$\hat{I}_N = \frac{\sum_{n=1}^N w(\theta_n)f(\theta_n)}{\sum_{n=1}^N w(\theta_n)}, \ \theta_n \sim g.$$

Θα έχουμε πάλι ότι η εκτιμήτρια $\hat{I}_N$ θα συγκλίνει σχεδόν βέβαιο στο ολοκλήρωμα I όσο αυξάνεται το μέγεθος δείγματος N. Για να έχουμε όμως έναν αποτελεσματικό importance sampler θα πρέπει

η κατανομή απο την οποία θα πάρουμε δείγμα να είναι συμβατή με την ύστερη κατανομή που μας ενδιαφέρει, δηλαδή

$$\pi(\theta) > 0 \Rightarrow g(\theta) > 0$$

και η δειγματοληψία απο την κατανομή g να είναι πραγματικά πιο απλή απο το να παίρναμε δείγμα απο την ύστερη κατανομή. Παρόλη την ευκολία της υλοποίησης αυτής της μεθόδου υπάρχουν και μειονεκτήματα, όπως στην περίπτωση που πάρουμε κάποιες μη αντιπροσωπευτικές παρατηρήσεις απο την g οι οποίες έχουν μεγάλο importance weight τότε ο εκτιμητής $\hat{I}$ μπορεί να αποκλίνει δραματικά απο το ολοκλήρωμα I που θέλουμε να εκτιμήσουμε. Επίσης θα έχουμε πάλι πρόβλημα στην περίπτωση που έχουμε μεγάλη διακύμανση για τα importance weights τα οποία με την σειρά τους θα εισάγουν μεγάλη διακύμανση στην εκτίμηση $\hat{I}_N$. Τέλος, η μέθοδος importance sampling αντιμετωπίζει μεγάλη δυσκολία όταν η παράμετρος θ είναι πολυδιάστατη, για τον λόγο αυτό χρησιμοποιούμε άλλες μεθόδους οι οποίες εισάγουν όμως συσχέτιση μεταξύ των διαδοχικών παρατηρήσεων όπως είναι οι μέθοδοι Markov Chain Monte Carlo (MCMC).

1.3.3 Markov Chain Monte Carlo

Η μέθοδος η οποία χρησιμοποιείται κατά κόρον για δειγματοληψία στην Μπεϋζιανη στατιστική κάνει χρήση Μαρκοβιανών αλυσίδων. Η χρήση MCMC αλγορίθμων καταφέρνει να είναι αποτελεσματική ακόμα και όταν έχουμε πολυδιάστατες παραμέτρους και επίσης για να υλοποιηθεί δεν χρειάζεται η γνώση κάποιας σταθεράς κανονικοποίησης. Η ιδέα είναι να κατασκευαστεί μια Μαρκοβιανή αλυσίδα η οποία θα έχει ώς αναλλοίωτη κατανομή την ύστερη κατανομή, έτσι ώστε σε κάθε βήμα να μετακινούμαστε σε παρατηρήσεις που έχουν κατανομή την ύστερη κατανομή. Επίσης, γνωρίζουμε ότι οι Μαρκοβιανές αλυσίδες βασίζονται σε γνώση του ακριβώς προηγούμενου χρόνου έτσι και τώρα θα έχουμε ότι

$$\theta_{n+1} \sim K(\theta_{n+1}|\theta_n),$$

όπου $K(\cdot|\cdot)$ είναι ένας πυρήνας μετάβασης. Όμως για να έχουμε την ιδιότητα του αναλλοίωτου της ύστερης κατανομής θέλουμε να ισχύει για κάθε σημείο $\theta^{'}$ ότι

$$\pi(\theta^{'}) = \int K(\theta^{'}|\theta)\pi(\theta)d\theta.$$

Δηλαδή, όποια κίνηση και να υλοποιήσουμε μέσο του $K(\cdot|\cdot)$ να διατηρούμε την ύστερη κατανομή.
Ακόμα την περίπτωση που η Μαρκοβιανή αλυσίδα είναι μη υποβιβάσιμη, δηλαδή ότι μπορούμε να μεταβούμε σε όλες τις πιθανές καταστάσεις σε πεπερασμένο χρονικό διάστημα και είναι και απεριοδική δηλαδή ότι μπορούμε να επισκεπτόμαστε καταστάσεις σε ακανόνιστο χρόνο τότε η αλυσίδα έχει μοναδική αναλλοίωτη κατανομή η οποία θα συγκλίνει στην ύστερη κατανομή για $n \to \infty$. Επίσης ως συνέπεια του Εργοδικού Θεωρήματος επιτυγχάνεται η εργοδικότητα της αλυσίδας και η σύγκλιση στο ολοκλήρωμα I για $n \to \infty$.
Τέλος μια πολύ σημαντική ιδιότητα για τις Μαρκοβιανές αλυσίδες είναι αυτή της ακριβής ισορροπίας δηλαδή θα πρέπει να ισχύει ότι

$$\pi(\theta)K(\theta|\theta^{'}) = \pi(\theta^{'})K(\theta^{'}|\theta),$$

και στην περίπτωση που ισχύει έχουμε ότι η αλυσίδα είναι αντιστρέψιμη. Η σημαντικότητα αυτής της ιδιότητας στέκεται στο ότι στην περίπτωση που θέλουμε να κατασκευάσουμε έναν MCMC αλγόριθμο που να έχει αναλλοίωτη κατανομή την ύστερη κατανομή αρκεί να ελέγξουμε ότι ισχύει η ιδιότητα της ακριβής ισορροπίας.

Ο πρώτος MCMC αλγόριθμος ο οποίος χρησιμοποιήθηκε είναι ό Metropolis-Hastings . Σε κάθε επανάληψη γεννάμε μια υποψήφια παράμετρο απο την προτεινόμενη κατανομή $q(\cdot|\cdot)$, η οποία εξαρτάται απο την παράμετρο στον προηγούμενο χρόνο, και είτε την αποδεχόμαστε είτε όχι ανάλογα με την την τιμή που παίρνει η πιθανότητα αποδοχής. Η επιλογή της προτεινόμενης κατανομής χωρίζεται γενικότερα σε συμμετρικές ή ασύμμετρες κατανομές. Στην περίπτωση που χρησιμοποιήσουμε συμμετρική κατανομή π.χ Κανονική,Ομοιόμορφη κ.τ.λ. τότε κάνουμε χρήση του αλγορίθμου Metropolis ή αλλιώς Τυχαίος Περίπατος (Random Walk). Επίσης πρέπει να σημειωθεί ότι η επιλογή της προτεινόμενης κατανομής παίζει σημαντικό ρόλο στην αποτελεσματικότητα του αλγορίθμου. Θα πρέπει οι προτεινόμενες κατανομές οι οποίες θα χρησιμοποιούμε να είναι σε πλήρη συμφωνία με το πεδίο ορισμού της

ύστερης κατανομής. Ακόμα στην περίπτωση που χρησιμοποιούμε προτεινόμενη κατανομή με παχιές ουρές μας δίνεται η δυνατότητα να εξερευνήσουμε την ύστερη κατανομή πιο αποτελεσματικά μιας και θα έχουμε την δυνατότητα να εξερευνήσουμε μεγαλύτερο εύρος τιμών. Το αρνητικό είναι ότι για πολύ απομακρυσμένες τιμές απο τον μέσο της προτεινόμενης κατανομής θα έχουμε μεγάλη πιθανότητα απόρριψης. Αντίθετα γίνεται κατανοητό ότι στην περίπτωση που χρησιμοποιήσουμε προτεινόμενη κατανομή με πολύ μικρή διακύμανση τότε θα αποδεχόμαστε με μεγάλη πιθανότητα τις παρατηρήσεις και το αρνητικό σε αυτή την περίπτωση είναι ότι θα έχουμε αργή εξερεύνηση του χώρου. Στον Αλγόριθμο 1, περιγράφουμε τα βήματα τα οποία ακολουθούνται για την υλοποίηση του αλγορίθμου Metropolis-Hastings ώστε να πάρουμε δείγμα από την ύστερη κατανομή.

Αλγόριθμος 1 Metropolis-Hastings Algorithm

$\theta^{(0)} \sim q(\theta)$
for *i=1,2,...* **do**
  $\theta^{cand} \sim q(\theta^{(i)}|\theta^{(i-1)})$
  $\alpha(\theta^{cand}|\theta^{(i-1)}) = min\left\{1, \frac{q(\theta^{(i-1)}|\theta^{cand})\pi(\theta^{cand})}{q(\theta^{cand}|\theta^{(i-1)})\pi(\theta^{(i-1)})}\right\}$
  $u \sim Uniform(u; 0, 1)$
  if $u < \alpha$ **then**
    $\theta^{(i)} \leftarrow \theta^{cand}$
  else
    $\theta^{(i)} \leftarrow \theta^{(i-1)}$
  end
end

Τέλος, λόγω της συμπεριφοράς του τυχαίου περιπάτου του αλγορίθμου έχουμε ώς αποτέλεσμα την αργή εξερεύνηση της ύστερης κατανομής, για παράδειγμα η αλυσίδα μπορεί να κολλήσει σε κάποιο τοπικό ακρότατο. Επίσης είναι συνηθισμένο να χρειάζονται αρκετές επαναλήψεις έτσι ώστε να πάρουμε ένα αντιπροσωπευτικό δείγμα, σε αυτή την περίπτωση παρατηρήσεις θα έχουν μεγάλη συσχέτιση μεταξύ τους που μας δημιουργεί πρόβλημα. Για την αντιμετώπιση αυτών των προβλημάτων έχουν αναπτυχθεί αλγόριθμοι με σημαντικότερο τον Hamiltonian Monte Carlo (HMC) ο οποίος παίρνει δείγμα απο την κατανομή $\pi(\theta, p)$ όπου p είναι μια βοηθητική μεταβλητή. Στις επόμενες ενότητες θα ακολουθήσει περιγραφή του συγκεκριμένου αλγορίθμου και πώς ξεπερνάει τα προαναφερθέντα προβλήματα.

Κεφάλαιο 2

Hamiltonian Monte Carlo

Στην συνέχεια θα αναφερθούμε στον αλγόριθμο Hamiltonian Monte Carlo ο οποίος μας δίνει την ικανότητα να εξερευνήσουμε την ύστερη κατανομή με συστηματικό τρόπο και όχι με συμπεριφορές τυχαίου περίπατου. Ο αλγόριθμος HMC για την υλοποίηση του βασίζεται στις εξισώσεις Hamilton οι οποίες μπορούν να χρησιμοποιηθούν στα περισσότερα προβλήματα τα οποία αφορούν συνεχείς κατανομές. Η ιδιαιτερότητα του αλγορίθμου βρίσκεται στο ότι εισάγεται στο πρόβλημα μια βοηθητική μεταβλητή p που θα ερμηνεύεται ως η ορμή του συστήματος στο πρόβλημα που θα περιγράψουμε στην συνέχεια.

2.1 Hamiltonian Monte Carlo

2.1.1 Τυπικό Σύνολο

Αρχικά θα δώσουμε τον ορισμό του Τυπικού Συνόλου μέσω του οποίου θα διαπιστώσουμε ότι για μια καλή προσέγγιση ενός ολοκληρώματος αρκεί να χρησιμοποιήσουμε ένα αντιπροσωπευτικό υποσύνολο της ύστερης κατανομής.

Ορισμός 2.1.1 *Έστω $x_1, x_2, ..., x_n$ μια ακολουθία από ανεξάρτητες και ισόνομες τιμές της τυχαίας μεταβλητής $P(X)$. Η ακολουθία $\boldsymbol{x} = (x_1, ..., x_n)$ θα ανήκει στο τυπικό σύνολο A_ϵ^n για την κατανομή $P(X)$ αν:*

$$\left| \frac{1}{n} \sum_{i=1}^{n} log(x_i) - H[X] \right| \leq \epsilon.$$

όπου $H[X]$ είναι η διαφορική εντροπία της τυχαίας μεταβλητής $\boldsymbol{x}$.

Ορισμός 2.1.2 *Έστω $\boldsymbol{x}$ μια τυχαία μεταβλητή με συνάρτηση κατανομής πιθανότητας f όπου το πεδίο ορισμού της είναι το σύνολο $\mathcal{X}$. Τότε ώς διαφορική εντροπία $H[X]$ θα ορίζουμε την*

$$H[X] = -\int_{\mathcal{X}} f(x) log f(x) dx.$$

Συνεπώς αφού στην Μπεϋζιανή στατιστική τα περισσότερα ερωτήματα αφορούν υπολογισμούς ολοκληρωμάτων της ύστερης κατανομής, κάνοντας χρήση του Τυπικού Συνόλου, έστω $T \subset \boldsymbol{\Theta}$ με $\boldsymbol{\Theta}$ το πολυδιάστατο πεδίο τιμών της παραμέτρου θ, θα έχουμε προσεγγίσεις της μορφής:

$$\int_{\boldsymbol{\Theta}} f(\theta) p(\theta|data) d\theta \approx \int_{T} f(\theta) p(\theta|data) d\theta.$$

Δηλαδή για τον αλγόριθμο MCMC αρκεί να εξερευνήσουμε ένα ΄καλό΄ υποσύνολο της ύστερης κατανομής.

Για να καταλάβουμε όμως πραγματικά τι είναι ένα τυπικό σύνολο ας δούμε το ακόλουθο παράδειγμα:

Έστω $y_n \sim Bernoulli(0.80)$ με $n = 1, ..., 100$ για τις οποίες μας δίνεται το ερώτημα ΄Τι είναι πιο

πιθανό να έχουμε μια ακολουθία από 100 συνεχόμενες επιτυχίες ή μια οποιαδήποτε ακολουθία που περιέχει 80 επιτυχίες∴.
Αφού γνωρίζουμε ότι η πιθανότητα επιτυχίας είναι 0.8 θα έχουμε ότι μια ακολουθία με 100 συνεχόμενες επιτυχίες, έχει μάζα πιθανότητας να εμφανιστεί 0.8^{100}. Αντίθετα για μια οποιαδήποτε ακολουθία με 80 επιτυχίες θα έχουμε ότι η πιθανότητα εμφάνισης της είναι $0.8^{80} \cdot 0.2^{20}$. Δηλαδή, παρατηρούμε ότι είναι πιο πιθανό να έχουμε 100 συνεχόμενες επιτυχίες αντί οποιουδήποτε συνδυασμού 80 επιτυχιών. Παρόλα αυτά το πλήθος των ακολουθιών με 80 επιτυχίες και 20 αποτυχίες είναι $\binom{80}{20}$ συγκριτικά πολύ μεγαλύτερο από αυτό των 100 επιτυχιών. Με αποτέλεσμα η μάζα πιθανότητας να είναι πολύ μεγαλύτερη για τις 80 επιτυχίες.

$$\binom{80}{20} 0.8^{80} \cdot 0.2^{20} >> \binom{100}{100} 0.8^{100}.$$

Για να βεβαιωθούμε για τους προηγούμενους ισχυρισμούς ας προσομοιώσουμε ένα εκατομμύριο παρατηρήσεις από την $Binomial(N = 100, p = 0.8)$ κατανομή.

Min.	1st Quarter	Median	Mean	3rd Quarter	Max.
61	77	80	80	83	97

Παρατηρούμε ότι ο τυπικός αριθμός επιτυχιών αντιστοιχεί στις 80. Ενώ αν πάρουμε ένα 99% διάστημα εμπιστοσύνης

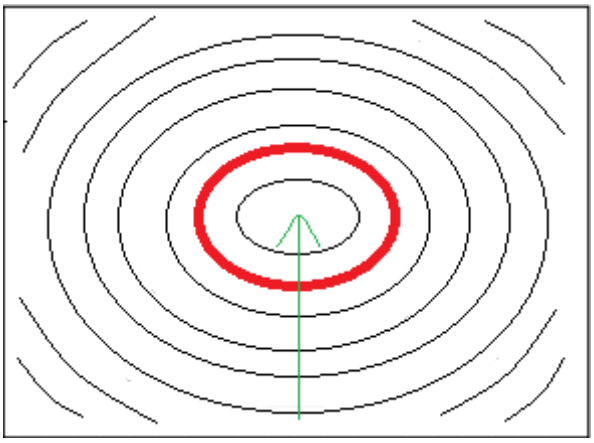
0.5%	99.5%
69	90

θα κατανοήσουμε ότι ακόμα και η πιο πιθανή ακολουθία επιτυχιών δεν ανήκει στο τυπικό σύνολο, δηλαδή η συνεισφορά της σε υπολογισμούς είναι πολύ μικρή. Γι΄ αυτό τον λόγο το Τυπικό Σύνολο πρέπει να υπολογίζεται με προσοχή έτσι ώστε να γίνονται οι υπολογισμοί μας πιο γρήγοροι και αποτελεσματικοί.

2.1.2 Πεδίο Παραγώγων

Σκοπός μας τώρα είναι να βρούμε έναν τρόπο έτσι ώστε να μπορέσουμε να κινηθούμε στο Τυπικό Σύνολο χωρίς να επιστρατεύσουμε καμία συμπεριφορά τυχαίου περίπατου. Στην συνεχή περίπτωση ένας τρόπος κωδικοποίησης τέτοιων συνόλων είναι μέσω διανυσματικών πεδίων τα οποία μας δίνουν κατευθύνσεις που μπορούμε να κινηθούμε στον χώρο. Για να κατασκευάσουμε όμως ένα τέτοιο διανυσματικό πεδίο θα πρέπει να εισάγουμε πληροφορία που αφορά την κατανομή που μας ενδιαφέρει και ο μόνος τρόπος να το καταστήσουμε εφικτό είναι μέσω της παραγώγου της κατανομής. Όμως όπως γνωρίζουμε η παράγωγος, κατευθύνει τις συναρτήσεις προς τις μεγιστικές τιμές τους. Στην δική μας περίπτωση θα κατευθύνει την κατανομή προς το mode της και όχι προς το Τυπικό Σύνολο που μας ενδιαφέρει.

Για να καταφέρουμε να ευθυγραμμίσουμε το Τυπικό Σύνολο με το πεδίο παραγώγων θα πρέπει να εισάγουμε μια καινούργια μεταβλητή $\boldsymbol{p}$ στο πρόβλημα. Παρόλα αυτά για να καταλάβουμε λίγο καλύτερα τι προσπαθούμε να καταφέρουμε, ας σκεφτούμε το εξής πρόβλημα: Πώς θα μπορούσαμε να βάλουμε σε τροχιά έναν δορυφόρο γύρω από την γη· Λόγω της βαρυτικής ενέργειας θα έπρεπε να εισάγουμε ορμή στον δορυφόρο ώστε να μπορέσει να αντισταθμίσει την βαρυτική έλξη και να μπορέσει να μπει σε τροχιά. Θα πρέπει όμως να αποφύγουμε να δώσουμε μεγάλη ή μικρή ορμή διότι και στις δύο περιπτώσεις δεν θα καταφέρουμε να βάλουμε τον δορυφόρο σε τροχιά γύρω από την γη. Στην περίπτωση όμως που δώσουμε ακριβώς την κατάλληλη ορμή τότε θα υπάρξει ισορροπία μεταξύ της βαρυτικής και της κινητικής ενέργειας (την οποία εισάγει η ορμή) και ο δορυφόρος θα αρχίσει να κινείται γύρω από την γη δημιουργώντας ένα συντηρητικό σύστημα στο οποίο θα έχουμε συμπληρωματικές αυξομειώσεις της κινητικής και βαρυτικής ενέργειας ανάλογα με την θέση του δορυφόρου. Ακόμα, καταλαβαίνουμε ότι υπάρχουν πολλές πιθανές τροχιές γύρω από την γη που σε κάθε μια μπορούμε να μεταβούμε χρησιμοποιώντας διαφορετική ορμή. Σε κάθε περίπτωση όμως το επίπεδο ενέργειας της



Διάγραμμα 2.1: Με κόκκινο συμβολίζουμε το εκάστοτε Τυπικό Σύνολο ενώ με πράσινο την κατεύθυνση του πεδίου παραγώγων.

κάθε τροχιάς είναι διαφορετικό αλλά σταθερό αφού το σύστημα που περιγράψαμε είναι συντηρητικό. Ακριβώς ένα τέτοιο σύστημα μπορεί να περιγραφεί μέσω των εξισώσεων Hamilton και του αλγορίθμου που θα περιγράψουμε στην συνέχεια HMC.

2.1.3 Εξισώσεις Hamilton

Έστω $\boldsymbol{\theta} \in \mathbb{R}^d$ η μεταβλητή ενδιαφέροντος, τότε εισάγοντας την ορμή θα πάρουμε ένα σύστημα από τον $\mathbb{R}^d$ στον $\mathbb{R}^d \times \mathbb{R}^d$:

$$\boldsymbol{\theta} \to (\boldsymbol{\theta}, \boldsymbol{p}),$$

όπου η $\boldsymbol{\theta}$ θα ερμηνεύεται ως η θέση στο σύστημα και ως ύστερη κατανομή θα έχουμε την από κοινού

$$\pi(\boldsymbol{\theta}, \boldsymbol{p}) = \pi(\boldsymbol{p}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Παρατηρούμε ότι στην περίπτωση που θέλουμε να πάρουμε ένα δείγμα από την κατανομή ενδιαφέροντος το μόνο που χρειάζεται να κάνουμε είναι να ολοκληρώσουμε ως προς την ορμή $\boldsymbol{p}$.
Παρόλα αυτά όπως έχουμε ήδη αναφέρει η κάθε τροχιά που παράγεται έχει και μια συγκεκριμένη σταθερή ενέργεια ανάλογα με την επιλογή της $\boldsymbol{p}$ (αφού μέσω της $\boldsymbol{p}$ μεταπηδάμε σε διαφορετικές τροχιές). Η ενέργεια αυτή χαρακτηρίζεται από την συνάρτηση ενέργειας Hamilton, $H(\boldsymbol{\theta}, \boldsymbol{p})$ η οποία είναι σταθερή για όλα τα σημεία της τροχιάς. Επίσης μπορούμε να συσχετίσουμε την κατανομή ενδιαφέροντος με την συνάρτηση ενέργειας $H(\boldsymbol{\theta}, \boldsymbol{p})$ κάνοντας χρήση ενός canonical ensemble

$$\pi(\boldsymbol{\theta}, \boldsymbol{p}) = e^{-H(\boldsymbol{\theta}, \boldsymbol{p})}.$$

Επιπλέον η συνάρτηση $H(\boldsymbol{\theta}, \boldsymbol{p})$ μπορεί να διασπαστεί σε δύο μέρη:

$$H(\boldsymbol{\theta}, \boldsymbol{p}) = -log\ \pi(\boldsymbol{\theta}, \boldsymbol{p}) = -log\ \pi(\boldsymbol{p}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

$$= -log\ \pi(\boldsymbol{p}|\boldsymbol{\theta}) - log\ \pi(\boldsymbol{\theta}) = K(\boldsymbol{p}, \boldsymbol{\theta}) + V(\boldsymbol{\theta}).$$

Με $K(\boldsymbol{p}, \boldsymbol{\theta})$ θα συμβολίζουμε την κινητική ενέργεια του συστήματος ενώ με $V(\boldsymbol{\theta})$ την βαρυτική. Τώρα βασιζόμενοι στις εξισώσεις Hamilton

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial H}{\partial \boldsymbol{p}} = \frac{\partial K}{\partial \boldsymbol{p}}, \quad (1)$$
$$\frac{d\boldsymbol{p}}{dt} = -\frac{\partial H}{\partial \boldsymbol{\theta}} = -\frac{\partial K}{\partial \boldsymbol{\theta}} - \frac{\partial V}{\partial \boldsymbol{\theta}}, \quad (2)$$

θα έχουμε την δυνατότητα να παράγουμε τις επιθυμητές τροχιές για το κάθε επίπεδο ενέργειας και να εξερευνούμε αποτελεσματικά την από κοινού κατανομή αφού πλέον θα έχουμε έναν ντετερμινιστικό τρόπο μετακινήσεων από το ένα σημείο στο επόμενο, χωρίς την χρήση συμπεριφορών τυχαίου περιπάτου. Να σημειώσουμε επίσης ότι οι τροχιές που παράγονται από τις εξισώσεις Hamilton έχουν

την ιδιότητα της διατήρησης του όγκου και για να επιτευχθεί κάτι τέτοιο η ορμή και η θέση συμπεριφέρονται αντιστρόφως ανάλογα. Ας δούμε όμως το θεώρημα Liouville το οποίο κάνοντας χρήση των εξισώσεων Hamilton αποδεικνύει ότι σημεία πάνω στην παραγόμενη τροχιά έχουν σταθερό όγκο. Αρχικά θα εισάγουμε τους ακόλουθους συμβολισμούς οι οποίοι θα μας βοηθήσουν στην απόδειξη του θεωρήματος Liouville

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} \frac{d\dot{\theta}}{d\theta} & \frac{d\dot{\theta}}{dp} \\ \frac{d\dot{p}}{d\theta} & \frac{\dot{p}}{dp} \end{bmatrix},$$

όπου $\dot{\theta} = \frac{dH}{d\boldsymbol{p}}$ και $\dot{p} = -\frac{H}{d\boldsymbol{\theta}}$.

Θεώρημα 2.1.1 *Θεώρημα Liouville : Έστω ότι παίρνουμε ένα κομμάτι από τον χώρο $(\boldsymbol{\theta}, \boldsymbol{p})$ όγκου υ. Επίσης έστω t_0 μια τυχαία χρονική στιγμή μαζί με μια τυχαία περιοχή D_0 του $(\boldsymbol{\theta}, \boldsymbol{p})$. Χωρίς βλάβη της γενικότητας θα υποθέσουμε ότι $t_0 = 0$ και θα ορίζουμε $D_{t_0} = D_0$. Ο όγκος που καταλαμβάνει η περιοχή D_t θα είναι*

$$\upsilon(t) = \int_{D_t} d\boldsymbol{\theta}^{'} d\boldsymbol{p}^{'} = \int_{D_0} det(I + tM) d\boldsymbol{\theta} d\boldsymbol{p} + O(t),$$

όπου χρησιμοποιήσαμε τον ακόλουθο μετασχηματισμό:

$$\begin{aligned} \boldsymbol{\theta}^{'} &= \boldsymbol{\theta} + t\frac{d\boldsymbol{\theta}}{dt} + o(t), \\ \boldsymbol{p}^{'} &= \boldsymbol{p} + t\frac{d\boldsymbol{p}}{dt} + o(t). \end{aligned}$$

Παρατηρούμε ότι λόγω του μετασχηματισμού που πραγματοποιήσαμε στο δεύτερο ολοκλήρωμα θα πρέπει να υπολογίσουμε και την αντίστοιχη Jacobian.Δηλαδή θα έχουμε

$$J = \begin{vmatrix} \frac{d\theta^{'}}{d\theta} & \frac{d\theta^{'}}{dp} \\ \frac{dp^{'}}{d\theta} & \frac{dp^{'}}{dp} \end{vmatrix} = \begin{vmatrix} 1 + t\frac{d\dot{\theta}}{d\theta} & t\frac{\dot{\theta}}{dp} \\ t\frac{d\dot{p}}{d\theta} & 1 + t\frac{d\dot{\theta}}{dp} \end{vmatrix} = \left| \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + t \begin{bmatrix} \frac{d\dot{\theta}}{d\boldsymbol{\theta}} & \frac{d\dot{\theta}}{d\boldsymbol{p}} \\ \frac{d\dot{p}}{d\boldsymbol{\theta}} & \frac{\dot{p}}{d\boldsymbol{p}} \end{bmatrix} \right| = det(I_2 + tM_2).$$

Από ιδιότητες όμως της γραμμικής άλγεβρας γνωρίζουμε ότι ισχύει

$$det(I_2 + tM_2) = 1 + t \cdot trace(M_2) + o(t),$$

και υπολογίζοντας το ίχνος (trace) του πίνακα M_2 θα έχουμε

$$trace(M_2) = \frac{d\dot{\theta}}{d\boldsymbol{\theta}} + \frac{d\dot{p}}{d\boldsymbol{p}} = \frac{d}{d\boldsymbol{\theta}}\frac{dH}{d\boldsymbol{p}} + \frac{d}{d\boldsymbol{p}}\frac{-dH}{d\boldsymbol{\theta}} = 0.$$

Άρα θα έχουμε

$$\upsilon(t) = \int_{D_0} (1 + 0) d\boldsymbol{\theta} d\boldsymbol{p} + o(t) = \upsilon(0) + o(t),$$

με

$$\frac{d}{dt}\upsilon(t)|_0 = 0$$

και αφού η επιλογή του $t = 0$ είναι τυχαία θα έχουμε ότι ο όγκος θα είναι σταθερός για οποιονδήποτε χρόνο, αρκεί να κάνουμε χρήση των Hamiltonian εξισώσεων. □

Αυτή η ιδιότητα έχει ως αποτέλεσμα ότι για την προσομοίωση των τροχιών θα πρέπει να χρησιμοποιήσουμε αριθμητικές μεθόδους που να έχουν με την σειρά τους και αυτές την ιδιότητα της διατήρησης του όγκου. Να σημειώσουμε επίσης, ότι η σταθερή ενέργεια κάθε τροχιάς είναι ισοδύναμη με την ιδιότητα της διατήρησης του όγκου. Συνεπώς, προσομοιωμένες τιμές από αριθμητικές μεθόδους που συμβαδίζουν με τα προαναφερθέντα θα τείνουν να παράγουν τροχιές οι οποίες θα έχουν μικρή απόκλιση από τις πραγματικές, διότι το επίπεδο ενέργειας τους θα είναι πολύ κοντά στο πραγματικό. Ακόμα στην δεύτερη εξίσωση, (2), παρατηρούμε ότι εμπλέκεται το διαφορικό της βαρυτικής ενέργειας. Δηλαδή γίνεται αντιληπτό ότι η ορμή εξαρτάται από την θέση που βρισκόμαστε και ως αποτέλεσμα η ορμή θα πρέπει να παίρνει κατάλληλες τιμές ανάλογα με την θέση, $\boldsymbol{\theta}$, έτσι ώστε να παραμένουμε στην

ίδια τροχιά από όπου ξεκινήσαμε.

Τώρα, όπως είδαμε και προηγουμένως η συνάρτηση ενέργειας $H(\boldsymbol{\theta}, \boldsymbol{p})$ μπορεί να περιγράψει εξολοκλήρου την από κοινού κατανομή των $(\boldsymbol{\theta}, \boldsymbol{p})$. Συνεπώς, θα ήταν εύλογο να εκφράσουμε τα σημεία της κατανομής εκμεταλλευόμενοι την ενέργεια της κάθε τροχιάς.
Για αυτό τον λόγο ορίζουμε ως

$$H^{-1}(E) = \{\boldsymbol{\theta}, \boldsymbol{p} : H(\boldsymbol{\theta}, \boldsymbol{p}) = E\},$$

το σύνολο που περιέχει όλα εκείνα τα σημεία που ανήκουν σε μια συγκεκριμένη τροχιά, δηλαδή ένα συγκεκριμένο επίπεδο ενέργειας. Τότε μέσω αυτής της προσέγγισης μπορούμε να εκφράσουμε την από κοινού κατανομή ως

$$\pi(\boldsymbol{\theta}, \boldsymbol{p}) = \pi(\omega_E|E)\pi(E),$$

όπου ω_E είναι η θέση στην τροχιά με ενέργεια E. Διαισθητικά η κατανομή ενός σημείου είναι ίση με το γινόμενο της επιλογής ενός επιπέδου ενέργειας και της επιλογής ενός σημείου από αυτό το επιλεχθέν επίπεδο ενέργειας. Επίσης γίνεται αντιληπτό ότι η όλη διαδικασία που έχουμε περιγράψει μέχρι στιγμής χωρίζεται σε ένα στοχαστικό και ντετερμινιστικό κομμάτι. Το στοχαστικό κομμάτι αποτελείται από την τυχαία επιλογή επιπέδου ενέργειας, δηλαδή από την λήψη $\boldsymbol{p} \sim \pi(p|\boldsymbol{\theta})$, ενώ το ντετερμινιστικό αποτελείται από τον υπολογισμό των εξισώσεων Hamilton για να διασχίσουμε την εκάστοτε τροχιά ώστε να φτάσουμε στο επόμενο σημείο. Τέλος, αφού η επιλογή επιπέδου ενέργειας εξαρτάται από την επιλογή του $\boldsymbol{p}$ θα θέλουμε η κατανομή της ενέργειας $\pi(E|\boldsymbol{p})$ που εισάγεται μέσω της ορμής p και η κατανομή των ενεργειών $\pi(E)$ να είναι όσο το δυνατόν όμοιες έτσι ώστε ο αλγόριθμος μας να είναι αποτελεσματικός. Στην περίπτωση που είναι αρκετά κοντά θα μεταπηδούμε σε καινούργια επίπεδα ενέργειας πολύ γρήγορα. Αντίθετα στην περίπτωση που αποκλίνουν αρκετά και ακόμα χειρότερα στην περίπτωση που η $\pi(E)$ είναι πολύ πιο πλατιά από την $\pi(E|\boldsymbol{p})$ τότε θα έχουμε αργή εξερεύνηση μεταξύ των τροχιών.

2.2 Επιλογές Παραμέτρων Αλγορίθμου

Αφού έχουμε περιγράψει τον τρόπο λειτουργίας του αλγόριθμου HMC είναι εύλογο εν συνεχεία να αναφερθούμε στις επιλογές των παραμέτρων που επηρεάζουν την αποτελεσματικότητα του, όπως είναι το μέγεθος του βήματος ϵ, το μήκος της τροχιάς L και η επιλογή κατανομής για την παράμετρο της ορμής p.

2.2.1 Επιλογή βήματος ϵ και μήκος τροχιάς L

Αρχικά ένα πολύ μεγάλο βήμα ϵ θα οδηγήσει σε μικρή πιθανότητα αποδοχής για τις προτεινόμενες παρατηρήσεις της παραγόμενης τροχιάς μιας και αυτές θα βρίσκονται πολύ μακριά από τις αρχικές μας παρατηρήσεις. Αντίθετα ένα αρκετά μικρό βήμα ϵ, θα οδηγήσει σε σπατάλη υπολογιστικών πόρων μιας και θα εξερευνούμε τον χώρο πολύ αργά διότι οι παρατηρήσεις που θα παίρνουμε σε κάθε επανάληψη θα είναι πολύ κοντά. Αρκεί να δούμε πως συμπεριφέρεται το $\epsilon \cdot L$, για αρκετά μικρό ϵ αν δεν έχουμε κατάλληλα μεγάλο L τότε το μήκος της τροχιάς, $\epsilon \cdot L$ θα είναι επικίνδυνα μικρό. Επιπλέον στην περίπτωση που από την προσομοίωση της τροχιάς έχουμε μεγάλο σφάλμα (π.χ. μπορεί να αποκλίνουμε από το προκαθορισμένο επίπεδο ενέργειας) τότε η επιλογή μικρής τιμής για το βήμα ϵ μπορεί να οδηγήσει σε καλύτερα αποτελέσματα από ότι ένα μεγάλο βήμα ϵ που θα μπορούσε να εντείνει το σφάλμα και να απομακρύνει την τροχιά από το αρχικό επίπεδο ενέργειας της.

Από την άλλη, η επιλογή του μεγέθους της τροχιάς L εξαρτάται κατά κύριο λόγο από την μορφή της κατανομής που θέλουμε να εξερευνήσουμε. Για παράδειγμα αν μια κατανομή είναι συγκεντρωμένη σε σημεία τα οποία είναι μακριά το ένα από το άλλο τότε καλή επιλογή θα ήταν να πάρουμε μεγάλο L έτσι ώστε να μπορούμε να μεταβούμε από την μια περιοχή στην άλλη. Παρόλα αυτά αν το L είναι πολύ μεγάλο τότε διακινδυνεύουμε να γυρίσουμε στο σημείο από όπου ξεκινήσαμε και να πάρουμε μηδενική καινούργια πληροφορία.

Δηλαδή κατανοητό ότι η επιλογή του ϵ και του L θέλουν μεγάλη προσοχή για την αποτελεσματικότητα του αλγορίθμου. Παρόλα αυτά οι τροχιές οι οποίες θα εξερευνήσουμε είναι πολύ πιθανό να

χρειάζονται διαφορετικό χρόνο εξερεύνησης, δηλαδή κάποια τροχιά μπορεί να εξερευνηθεί πλήρως με ένα μικρό L ενώ μια άλλη μπορεί να χρειάζεται μεγαλύτερο χρόνο εξερεύνησης. Γι΄ αυτό τον λόγο οι Matthew D.Hoffman και Andrew Gelman προτείνουν τον αλγόριθμο No-U-Turn-Sample (NUTS) που ξεπερνάει αυτά τα προβλήματα κάνοντας χρήση δυναμικών μεθόδων για την εξερεύνηση των τροχιών, χωρίς να χρειάζεται να επιλέξουμε ένα σταθερό L για τις τροχιές αλλά προσομοιώνοντας παρατηρήσεις μέχρι να επιστρέψουμε σε περιοχές της τροχιάς που έχουμε ήδη εξερευνήσει.

2.2.2 Επιλογή Κατανομής p

Επιπλέον στην προηγούμενη ενότητα παρατηρήσαμε ότι η λήψη της ορμής $\boldsymbol{p}$ παίζει πολύ σημαντικό ρόλο στην αποτελεσματικότητα του αλγορίθμου. Μια καλή επιλογή της κατανομής της $\boldsymbol{p}$ μπορεί να κάνει τα επίπεδα ενέργειας E όσο το δυνατόν πιο ομοιόμορφα γίνεται μιας και $H^{-1}(E) = \{\boldsymbol{\theta}, \boldsymbol{p} : H(\boldsymbol{\theta}, \boldsymbol{p}) = E\}$, ενώ παράλληλα η μετάβαση μεταξύ των επιπέδων ενέργειας μπορεί να γίνει πληροφοριακή επιτυγχάνοντας συνάφεια μεταξύ των κατανομών $\pi(E|\boldsymbol{p})$ και $\pi(E)$. Η συνηθέστερη επιλογή είναι η Ευκλείδεια Κανονική κατανομή.
Έστω g η Ευκλείδεια μετρική μέσω της οποίας μπορούμε να μετρήσουμε την απόσταση ανάμεσα σε οποιαδήποτε δύο σημεία $\boldsymbol{\theta}, \boldsymbol{\theta}^{'}$. Τότε οι αποστάσεις μπορούν να αποθηκευτούν σε έναν πίνακα διαστάσεων $D \times D$.

$$\Delta(\boldsymbol{\theta}, \boldsymbol{\theta}^{'}) = (\boldsymbol{\theta} - \boldsymbol{\theta}^{'})^T \cdot g \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^{'}).$$

Επίσης μπορούμε να γενικεύσουμε την Ευκλείδεια μετρική g δημιουργώντας μια οικογένεια μετρικών:

$$M = R \cdot S \cdot g \cdot S^T \cdot R^T,$$

όπου S και R θα είναι διαγώνιοι και ορθογώνιοι πίνακες αντίστοιχα. Όμως όπως έχουμε αναφέρει το $\boldsymbol{\theta}$ είναι αντιστρόφως ανάλογο του $\boldsymbol{p}$ δηλαδή όποια παραμετροποίηση εφαρμόζεται σε μια από τις δύο μεταβλητές τότε στην άλλη εφαρμόζεται η ακριβώς αντίθετη (λόγω της διατήρησης του όγκου). Άρα μπορούμε να μετρήσουμε αποστάσεις και για την $\boldsymbol{p}$:

$$\Delta(\boldsymbol{p}, \boldsymbol{p}^{'}) = (\boldsymbol{p} - \boldsymbol{p}^{'})^T M^{-1} (\boldsymbol{p} - \boldsymbol{p}^{'}).$$

Συνεπώς, κάνοντας χρήση της μετρικής M μπορούμε να κατασκευάσουμε για την $\boldsymbol{p}$ την κατανομή:

$$\pi(\boldsymbol{p}, |\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{p}|0, M),$$

που παρατηρούμε ότι δεν εξαρτάται από την παράμετρο θ. Ακόλουθα η κινητική ενέργεια θα είναι της μορφής

$$K(\boldsymbol{\theta}, \boldsymbol{p}) = \frac{1}{2} \boldsymbol{p}^T M^{-1} \boldsymbol{p} + log\,|M| + \sigma\tau\alpha\theta.$$

Παρόλα αυτά για να έχουμε ομοιόμορφα επίπεδα ενέργειας, E, θα πρέπει ο πίνακας M^{-1} να είναι όσο πιο κοντά γίνεται στον πίνακα συνδιακύμανσης της κατανομής ενδιαφέροντος. Ένας τρόπος να το πετύχουμε είναι να θέσουμε την μετρική ως

$$M^{-1} = \mathbb{E}_\pi[(\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^T].$$

Τέλος, επειδή η επιλογή της κατανομής της ορμής $\boldsymbol{p}$ παίζει πολύ σημαντικό ρόλο στην μορφή του χώρου $(\boldsymbol{\theta}, \boldsymbol{p})$ λόγω του πίνακα M^{-1} θα πρέπει να είμαστε προσεκτικοί έτσι ώστε να μην αλλάζουμε την γεωμετρία της κατανομής $\boldsymbol{\theta}$ σε ακραίο βαθμό.

2.3 Δημιουργία της τροχιάς στην πράξη

Όπως προαναφέραμε για να δημιουργήσουμε τις τροχιές αρκεί να χρησιμοποιήσουμε και να λύσουμε τις εξισώσεις Hamilton.Κάτι τέτοιο όμως είναι πολύ δύσκολο να γίνει αναλυτικά, για αυτό χρησιμοποιούμε προσεγγιστικές αριθμητικές μεθόδους. Λαμβάνοντας όμως υπόψιν την ¨φύση' των τροχιών που παράγουν οι εξισώσεις Hamilton θα πρέπει και η αριθμητική μέθοδος που θα προτείνουμε να βρίσκεται σε αρμονία με την ιδιότητα διατήρησης του όγκου. Αυτό θα έχει ως συνέπεια οι παραγόμενες τροχιές να μην αποκλίνουν ανησυχητικά από το επίπεδο ενέργειας από το οποίο ξεκίνησαν

μιας και θα μπορούσαμε να πούμε ότι η ιδιότητα της διατήρησης του όγκου είναι ισοδύναμη με την σταθερότητα της ενέργειας στη τροχιά.
Μια μεγάλη οικογένεια τέτοιων αριθμητικών μεθόδων ονομάζεται symplectic integrators (δηλαδή στους παραγόμενους μετασχηματισμούς των μεταβλητών για την κάθε επανάληψη η ορίζουσα του Jacobian πίνακα είναι 1). Συγκεκριμένα επειδή η κατανομή της ορμής είναι Ευκλείδεια Κανονική και κατά συνέπεια η ορμή είναι ανεξάρτητη της θέσης $\boldsymbol{\theta}$ θα χρησιμοποιήσουμε τον leapfrog integrator.Ο οποίος για συγκεκριμένο βήμα ϵ λειτουργεί ως εξής:

Αλγόριθμος 2 Leapfrog intergrator

$\theta_0 \leftarrow \theta, p_0 \leftarrow p$
for $0 \leq n \leq L$ **do**
  $p_{n+\frac{1}{2}} \leftarrow p_n - \frac{\epsilon}{2}\frac{dV}{d\theta}(\theta_n)$
  $\theta_{n+1} \leftarrow \theta_n + \epsilon p_{n+\frac{1}{2}}$
  $p_{n+1} \leftarrow p_{n+\frac{1}{2}} - \frac{\epsilon}{2}\frac{dV}{d\theta}(\theta_{n+1})$
end

Αρχικά ξεκινάμε κάνοντας μισό βήμα για την ορμή $\boldsymbol{p}$, στην συνέχεια κάνουμε ένα ολόκληρο βήμα για την θέση $\boldsymbol{\theta}$ και τέλος υλοποιούμε ακόμα ένα μισό βήμα για την ορμή. Μέσω αυτών των δύο μισών βημάτων για την ορμή καταφέρνουμε να εξασφαλίσουμε την ιδιότητα διατήρησης του όγκου διότι έχουμε έναν shear μετασχηματισμό.

Παρ' όλη την αποτελεσματικότητα της δημιουργίας και της εξερεύνησης της κατανομής υπάρχουν περιπτώσεις που δημιουργούν πρόβλημα στην αποτελεσματικότητα του αλγορίθμου HMC.Στην περίπτωση των ιεραρχικών μοντέλων, για παράδειγμα:

$$y_i \sim \mathcal{N}(\theta_i, \sigma^2),$$

$$\theta_i \sim \mathcal{N}(\mu, \tau^2).$$

παρατηρείται ότι μικρές αλλαγές στις υπερ-παραμέτρους δηλαδή μ και τ μπορούν να επιφέρουν σημαντικές αλλαγές στην θ_i, δημιουργώντας έτσι περιοχές υψηλής καμπυλότητας. Ένας τρόπος αντιμετώπισης τέτοιου είδους περιοχών είναι να μικρύνουμε το βήμα και να χρησιμοποιήσουμε μη κεντροποιημένη μοντελοποίηση. Θα δούμε όμως μετέπειτα εκτενέστερα πώς μπορούμε να διορθώσουμε αυτά τα παθολογικά προβλήματα.

Συνεπώς καταλαβαίνουμε ότι παρόλο που αυτού του είδους αριθμητικές προσεγγίσεις έχουν καλά αποτελέσματα μπορούν πολλές φορές να αποκλίνουν και να εισάγουν μεροληψίες στις μεταβάσεις μεταξύ των σημείων. Ένας φυσικός τρόπος για να λύσουμε αυτό το πρόβλημα είναι να αντιμετωπίσουμε τις μετακινήσεις που περιγράψαμε προηγουμένως ως προτεινόμενη κατανομή ενός Metropolis-Hastings σχηματισμού.
Δηλαδή, έστω προτεινόμενη κατανομή $\mathbb{Q}(\boldsymbol{\theta}', \boldsymbol{p}'|\boldsymbol{\theta}_0, \boldsymbol{p}_0) = \delta(\boldsymbol{\theta}' - \boldsymbol{\theta}_L)\delta(\boldsymbol{p}' - \boldsymbol{p}_L)$, με $\delta(\cdot)$ συναρτηση dirac ,

$$\delta(\boldsymbol{\theta}' - \boldsymbol{\theta}_L) = \begin{cases} 1, & \boldsymbol{\theta}' = \boldsymbol{\theta}_L, \\ 0, & \boldsymbol{\theta}' \neq \boldsymbol{\theta}_L, \end{cases}$$

όπου $\boldsymbol{\theta}_L$ το σημείο στο τέλος της τροχιάς με μήκος L. Για να είναι όμως μια προτεινόμενη κατανομή έγκυρη θα πρέπει να έχουμε την ιδιότητα της αντιστρεψιμότητας. Επειδή όμως προτείνουμε σημεία προσομοιώνοντας προς τα μπρος της τροχιάς θα έχουμε ότι ισχύει

$$\frac{\mathbb{Q}(\boldsymbol{\theta}_0, \boldsymbol{p}_0|\boldsymbol{\theta}_L, \boldsymbol{p}_L)}{\mathbb{Q}(\boldsymbol{\theta}_L, \boldsymbol{p}_L|\boldsymbol{\theta}_0, \boldsymbol{p}_0)} = \frac{0}{1}.$$

Συνεπώς για να πετύχουμε την απαιτούμενη ιδιότητα της αντιστρεψιμότητας αρκεί να αλλάξουμε το πρόσημο της ορμής που δώσαμε στο σύστημα αρχικά. Διαισθητικά αρκεί να σκεφτούμε τι θα συνέβαινε στην περίπτωση που σε ένα μπολ χωρίς τριβή θέλαμε να μετακινήσουμε μια μπάλα στην αρχική της

θέση. Απλά θα δίναμε την ίδια ποσότητα ορμής προς την αντίθετη κατεύθυνση.
Άρα η πιθανότητα αποδοχής για τον αντίστοιχο Metropolis-Hastings θα είναι

$$\alpha(\boldsymbol{\theta}_L, -\boldsymbol{p}_L|\boldsymbol{\theta}_0, \boldsymbol{p}_0) = min(1, \frac{\mathbb{Q}(\boldsymbol{\theta}_0, \boldsymbol{p}_0|\boldsymbol{\theta}_L, \boldsymbol{p}_L)\pi(\boldsymbol{\theta}_L, \boldsymbol{p}_L)}{\mathbb{Q}(\boldsymbol{\theta}_L, \boldsymbol{p}_L|\boldsymbol{\theta}_0, \boldsymbol{p}_0)\pi(\boldsymbol{\theta}_0, \boldsymbol{p}_0)})$$

$$= min(1, \frac{\delta(\boldsymbol{\theta}_L - \boldsymbol{\theta}_L)\delta(-\boldsymbol{\theta}_L + \boldsymbol{\theta}_L)\pi(\boldsymbol{\theta}_L, -\boldsymbol{p}_L)}{\delta(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0)\delta(-\boldsymbol{\theta}_0 + \boldsymbol{\theta}_0)\pi(\boldsymbol{\theta}_0, -\boldsymbol{p}_0)})$$

$$= min(1, \frac{\pi(\boldsymbol{\theta}_L, -\boldsymbol{p}_L)}{\pi(\boldsymbol{\theta}_0, \boldsymbol{p}_0)})$$

$$= min(1, \frac{exp(-H(\boldsymbol{\theta}_L, -\boldsymbol{p}_L))}{exp(-H(\boldsymbol{\theta}_0, \boldsymbol{p}_0))})$$

$$= min(1, exp(-H(\boldsymbol{\theta}_L, -\boldsymbol{p}_L) + H(\boldsymbol{\theta}_0, \boldsymbol{p}_0))).$$

Συνεπώς τα σημεία τα οποία θα είναι κοντά στην τροχιά θα έχουν μεγάλη πιθανότητα αποδοχής μιας και η συνάρτηση ενέργειας, $H(\cdot, \cdot)$, θα είναι σχεδόν ίδια με των αρχικών σημείων, αφού έχουμε την ιδιότητα διατήρησης του όγκου και κατά συνέπεια της ενέργειας. Αντίθετα αν έχουμε αποκλίνει πολύ η διαφορά μεταξύ των ενεργειών θα είναι μεγάλη συνεπώς θα έχουμε και μικρή πιθανότητα αποδοχής.

Επίσης, αντί να παίρνουμε ως σημείο μετάβασης το τελευταίο σημείο της τροχιάς θα ήταν πιο εύλογο να παράγουμε σημεία ομοιόμορφα από κάθε τροχιά που δημιουργούμε. Μέσω αυτού του μηχανισμού θα προσεγγίζουμε πιο αποτελεσματικά την κατανομή ενδιαφέροντος (πολύ απλά στην περίπτωση που προσομοιώσουμε για μεγάλο διάστημα υπάρχει πιθανότητα να γυρίσουμε στο σημείο από όπου ξεκινήσαμε, συνεπώς συμφέρει περισσότερο να παίρνουμε προτεινόμενα σημεία από όλο το φάσμα της τροχιάς). Για τον σκοπό αυτό θα έχουμε ως proposal την

$$\mathbb{Q}(\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}|\boldsymbol{\theta}_0, \boldsymbol{p}_0) = \frac{1}{L}\sum_{l=0}^{L}\delta(\boldsymbol{\theta}^{'} - \boldsymbol{\theta}_l)\delta(\boldsymbol{p}^{'} + \boldsymbol{p}_l),$$

που προτείνει για μετάβαση όλα τα σημεία της τροχιάς και ως γενικευμένη πιθανότητα αποδοχής

$$\alpha(\boldsymbol{\theta}_l, -\boldsymbol{p}_l|\boldsymbol{\theta}_0, \boldsymbol{p}_0) = min(1, exp(-H(\boldsymbol{\theta}_l, -\boldsymbol{p}_l) + H(\boldsymbol{\theta}_0, \boldsymbol{p}_0)),$$

για $0 \leq l \leq L$. Αυτή η προσέγγιση έχει και μειονεκτήματα διότι προτείνει μεταβάσεις ανεξάρτητες από την πιθανότητα αποδοχής τους. Δηλαδή μπορεί ομοιόμορφα να διαλέξουμε κάποιο σημείο της τροχιάς που να έχει μεγάλο σφάλμα και να αναγκαστούμε να το απορρίψουμε χωρίς να λαμβάνουμε υπόψιν άλλα σημεία της τροχιάς τα οποία έχουν μικρότερο σφάλμα και μεγαλύτερη πιθανότητα αποδοχής. Για την αποφυγή τέτοιων προβλημάτων έχουν κατασκευαστεί μέθοδοι όπως ο αλγόριθμος NUTS για τον οποίο θα μιλήσουμε στην συνέχεια.

2.4 Διαγνωστικοί Έλεγχοι

Αφού έχουμε επιλέξει το βήμα ϵ, το μήκος L και την κατανομή της ορμής $\boldsymbol{p}$ υλοποιούμε τον αλγόριθμο HMC. Πώς όμως θα μπορούμε να ελέγχουμε ότι έχουμε συγκλίνει και ότι γενικότερα δεν υπάρχει κάποιο πρόβλημα στην αποτελεσματικότητα του αλγορίθμου;

2.4.1 Γεωμετρική Εργοδικότητα

Αρχικά θα κάνουμε έλεγχο της γεωμετρικής εργοδικότητας των αλυσίδων. Επειδή όμως πολλές φορές κάποιος είναι πολύ δύσκολο έως αδύνατο να ελέγξει την γεωμετρική εργοδικότητα αναλυτικά, χρησιμοποιούμε εμπειρικά στατιστικά όπως το $\hat{R}$.

Ορισμός 2.4.1 *Ο βαθμός σύγκλισης μιας Μαρκοβιανής Αλυσίδας μπορεί να εκτιμηθεί χρησιμοποιώντας το στατιστικό του Gleman Rubin, $\hat{R}$. Βασίζεται στην ισορροπία μεταξύ της διακύμανσης εντός και μεταξύ των αλυσίδων (των αλυσίδων που χρησιμοποιούνται). Τιμές του στατιστικού $\hat{R}$ κοντά στην τιμή 1 θα υποδηλώνουν σύγκλιση στην ζητούμενη κατανομή.*

Οι επαναληπτικές μέθοδοι όπως ο αλγόριθμος HMC εισάγουν δύο κύρια προβλήματα: Πρώτον, αν δεν τρέξουμε για αρκετές επαναλήψεις τον αλγόριθμο θα πάρουμε μη αντιπροσωπευτικό δείγμα από την κατανομή ενδιαφέροντος. Επιπρόσθετα στην περίπτωση που τον τρέξουμε για αρκετές επαναλήψεις θα πρέπει να είμαστε προσεκτικοί στην επιλογή των παρατηρήσεων που θα χρησιμοποιήσουμε για την εκτίμηση των στατιστικών ερωτημάτων. Οι αρχικές προσομοιωμένες παρατηρήσεις θα εισάγουν μεροληψία στην συμπερασματολογία μας, μιας και αυτές δεν θα είναι αντιπροσωπευτικές της κατανομής ενδιαφέροντος αφού είναι οι παρατηρήσεις οι οποίες δεν έχουν συγκλίνει ακόμα στο Τυπικό Σύνολο. Δεύτερον, εισάγεται συσχέτιση μεταξύ των διαδοχικών παρατηρήσεων δηλαδή μπορεί να αποκτούμε αργά καινούργια πληροφορία και το δείγμα που παίρνουμε από τον αλγόριθμο συγκριτικά με το πλήθος των επαναλήψεων του αλγορίθμου να είναι μή πληροφοριακό. Συνεπώς τα προβλήματα συνοψίζονται στην διαπίστωση της σύγκλισης της αλυσίδας στην κατανομή ενδιαφέροντος και στον εντοπισμό του κατάλληλου δείγματος και στην αποτελεσματικότητα του αλγορίθμου να παράγει καινούργια πληροφορία με κάθε επανάληψη.
Για την αντιμετώπιση τους επιστρατεύουμε τρία βήματα: Πρώτον, τρέχουμε αρκετές ακολουθίες από διαφορετικές αρχικές τιμές, αυτό θα έχει ως αποτέλεσμα στην περίπτωση που μια από τις ακολουθίες δεν συγκλίνει να συγκλίνουν οι υπόλοιπες ή ακόμα πιο σημαντικά στην αναγνώριση πολλαπλών κορυφών της κατανομής ενδιαφέροντος. Δεύτερον, παρακολουθούμε την διακύμανση μεταξύ των ακολουθιών και εντός των ακολουθιών, ώσπου να έχουμε ότι η εντός και η μεταξύ τους διακύμανση είναι σχεδόν ίσες. Μέσω αυτού μπορούμε να διαπιστώσουμε ότι όλες οι ακολουθίες εξερευνούν την ίδια περιοχή και ότι το δείγμα που παράγουν είναι αντιπροσωπευτικό της κατανομής ενδιαφέροντος. Τρίτον, σε περίπτωση που η υλοποίηση των επαναλήψεων και γενικότερα του αλγορίθμου χρειάζεται υπερβολικά αρκετό χρόνο τότε θα πρέπει να διαφοροποιήσουμε τον αλγόριθμο ώστε να συγκλίνει σε πεπερασμένο χρονικό διάστημα.
Αρχικά, κρατάμε εκτός της συμπερασματολογίας τις πρώτες παρατηρήσεις που συνήθως το πλήθος τους ανέρχεται στο μισό των επαναλήψεων της κάθε αλυσίδας. Αυτές οι παρατηρήσεις θα ονομάζονται burn-in ή warm-up οι οποίες δεν είναι αντιπροσωπευτικές τιμές της κατανομής ενδιαφέροντος μιας και δεν έχουν συγκλίνει στο Τυπικό Σύνολο. Επιπλέον όταν έχουμε παράξει την ακολουθία που θέλουμε, μπορούμε να υπολογίσουμε την τάξη των αυτοσυσχετίσεων που έχουν μεταξύ τους οι παρατηρήσεις και αντίστοιχα να επιλέξουμε ένα κατάλληλο βήμα (ανάλογα με τον βαθμό της αυτοσυσχέτισης) και να πάρουμε παρατηρήσεις από την ακουλουθία που ήδη έχουμε παράξει. Αυτό όμως θα οδηγήσει σε μικρό πλήθος παρατηρήσεων, για αυτό τον λόγο εύλογο είναι να αυξήσουμε το συνολικό πλήθος των επαναλήψεων.
Τέλος, παράγουμε τουλάχιστον δυο αλυσίδες έτσι ώστε να μπορέσουμε να συγκρίνουμε τις εντός και μεταξύ διακυμάνσεις των αλυσίδων, μέσω των οποίων θα μπορέσουμε να δούμε αν έχει επιτευχθεί σύγκλιση και η εξερεύνηση της ίδιας περιοχής της κατανομής ενδιαφέροντος από όλες τις αλυσίδες.

Έστω ότι παράγουμε m ακολουθίες κάθε μία μήκους L. Επίσης, έστω ότι η μεταβλητή ενδιαφέροντος είναι η $\boldsymbol{\theta}$ με παραγόμενες παρατηρήσεις $\theta_{i,j}$ μέ $i = 1, ..., L$ και $j = 1, ..., m$. Τότε οι εντός και μεταξύ διακυμάνσεις θα υπολογίζονται ως εξής:

$$B = \frac{L}{m-1}\sum_{j=1}^{m}(\bar{\theta_{.j}} - \bar{\theta_{..}})^2, \quad \bar{\theta_{.j}} = \frac{1}{L}\sum_{i=1}^{L}\theta_{i,j}, \quad \bar{\theta_{..}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta_{.j}},$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad s_j^2 = \frac{1}{L}\sum_{i=1}^{L}(\theta_{i,j} - \bar{\theta_{.j}})^2.$$

Τέλος, μπορούμε να υπολογίσουμε την διακύμανση

$$\hat{Var}(\boldsymbol{\theta}) = \frac{n-1}{n}W + \frac{1}{n}B.$$

Παρατηρούμε ότι όσο το $n \to \infty$ η εντός διακύμανση όλο και πλησιάζει την εκτιμώμενη διακύμανση του $\boldsymbol{\theta}$ αυτό είναι κάτι το οποίο περιμέναμε. Τέλος, το στατιστικό που χρησιμοποιούμε για την γεωμετρική εργοδικότητα βασιζόμενο σε όλα τα προηγούμενα που είπαμε είναι

$$\hat{R} = \sqrt{\frac{\hat{Var}(\boldsymbol{\theta})}{W}}.$$

Συνεπώς, τιμές του $\hat{R}$ κοντά στο 1 υποδηλώνουν καλή σύγκλιση και μίξη.

2.4.2 Έλεγχος Καταλληλότητας Κατανομής p

Επίσης έλεγχος θα πρέπει να γίνει και στην επιλογή της κατανομής της ορμής $\boldsymbol{p}$. Για παράδειγμα, μια παθολογική συμπεριφορά που εμποδίζει την αποτελεσματικότητα του αλγορίθμου είναι οι περιπτώσεις που η κατανομή ενδιαφέροντος έχει παχιές ουρές. Τότε το τυπικό σύνολο θα παίρνει πληθώρα τιμών στις ουρές, με αποτέλεσμα στην προσπάθεια μας να μετακινηθούμε σε αυτές τις περιοχές να χρειαζόμαστε πολύ χρόνο. Συνεπώς μια καλή επιλογή κινητικής ενέργειας μπορεί να κάνει πολύ πιο γρήγορη και αποτελεσματική εξερεύνηση της κατανομής ενδιαφέροντος, πολύ απλά εξομαλύνοντας αυτές τις περιοχές (αν σκεφτούμε ότι ο πίνακας διακύμανσης της $\boldsymbol{p}$ επηρεάζει την γεωμετρία του χώρου των σημείων $(\boldsymbol{\theta}, \boldsymbol{p})$. Βασιζόμενοι πάντα στην απαίτηση που έχουμε ότι η κατανομή των ενεργειών, $\pi(E)$, που περιγράφει την κάθε τροχιά να είναι σε συμφωνία και να μην αποκλίνει από την κατανομή μεταβάσεων μεταξύ των ενεργειών, $\pi(E|\boldsymbol{p})$. Συνεπώς, αυτή την απόκλιση μπορούμε να την ποσοτικοποιήσουμε μέσω του ακόλουθου στατιστικού:

$$E - BFMI = \frac{\mathbb{E}_{\pi}\left[Var_{\pi_{E|\boldsymbol{p}}}\right]}{Var_{\pi_E}(E)} \approx \widehat{E - BFMI} = \frac{\sum_{n=1}^{N}(E_n - E_{n-1})}{\sum_{n=0}^{N}(E_n - \bar{E})^2}.$$

Εμπειρικά τιμές του $E - BFMI$ μικρότερες του 0.3 έχουν αποδειχθεί ότι είναι προβληματικές.

2.4.3 Έλεγχος Περιοχών Καμπυλότητας

Τέλος, θα πρέπει να γίνεται και έλεγχος για περιοχές που έχουμε μεγάλη καμπυλότητα. Τέτοιες περιοχές συνηθέστερα συναντώνται στις περιπτώσεις που χρησιμοποιούμε ιεραρχικά μοντέλα. Τις περισσότερες φορές στην προσπάθεια του ο αλγόριθμος να εξερευνήσει αυτές τις περιοχές προσκολλάται στο σύνορο τους και παίρνει πληθώρα παρατηρήσεων με αποτέλεσμα οι παρατηρήσεις που θα παίρνουμε να μην είναι αντιπροσωπευτικές. Επίσης, τέτοιες περιοχές υψηλής καμπυλότητας εξαναγκάζουν τον αλγόριθμο και ειδικότερα την αριθμητική μέθοδο leapfrog να αποκλίνει με το που περάσει το σύνορο τους. Δηλαδή το αμέσως επόμενο παραγόμενο σημείο αφότου περάσουμε το παθολογικό σύνορο αποκλίνει από το εκάστοτε επίπεδο ενέργειας, το οποίο διαπιστώνεται εύκολα. Δηλαδή οι παραγόμενες τροχιές από την αριθμητική μέθοδο leapfrog προσφέρουν έναν τρόπο διάγνωσης τέτοιων περιοχών, διότι αποκλίνουν πολύ γρήγορα από το επίπεδο ενέργειας τους. Ένας τρόπος διόρθωσης αυτού του προβλήματος είναι μέσω της μείωσης του μεγέθους του βήματος ϵ. Παρόλα αυτά σε επόμενη ενότητα θα αναλύσουμε εκτενέστερα την λύση αυτού του προβλήματος παρουσιάζοντας αριθμητικά αποτελέσματα για ιεραρχικά μοντέλα των Betancourt και Girolami.

Κεφάλαιο 3

Αλγόριθμος No-U-Turn Hamiltonian Monte Carlo

3.1 Δυναμική Επιλογή του Μήκους L

Έχοντας περιγράψει τον αλγόριθμο HMC γίνεται κατανοητό ότι η σωστή επιλογή του μήκους της τροχιάς L (ή αλλιώς του πλήθους των κόμβων) και του μεγέθους του βήματος ϵ παίζουν σημαντικό ρόλο στην αποτελεσματικότητα του αλγορίθμου. Κακή επιλογή αυτών των δύο παραμέτρων μπορεί να οδηγήσει σε σπατάλη υπολογιστικών πόρων και παραγωγή μη αντικειμενικών παρατηρήσεων για την κατανομή.
Αρχικά θα πρέπει να βρούμε έναν τρόπο έτσι ώστε να μπορέσουμε να διατηρήσουμε τα ποιοτικά χαρακτηριστικά που μας δίνει ο HMC αλγόριθμος, κάνοντας χρήση τροχιών και αποφεύγοντας random-walk συμπεριφορές, χωρίς να χρειάζεται να επιλέξουμε τον αριθμό των βημάτων L. Ουσιαστικά θέλουμε να κατασκευάσουμε μια δυναμική μέθοδο που θα σταματάει την δειγματοληψία αυτόματα με το που ξεκινάει η τροχιά να γυρίζει προς το σημείο που ξεκίνησε (σε περιοχές που ήδη έχουμε εξερευνήσει).

Ένα βολικό κριτήριο για τον σκοπό αυτό, είναι το γινόμενο της ορμής $\boldsymbol{p}$ με την διαφορά μεταξύ του αρχικού και τωρινού σημείου $(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, το οποίο περιγράφεται μέσω του διαφορικού της μέσης τετραγωνικής διαφοράς των $\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_0$:

$$\frac{d}{dt}\frac{(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\cdot(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}{2} = (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\frac{d}{dt}(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\cdot\boldsymbol{p},$$

όπου $\boldsymbol{\theta}_0$ το αρχικό μας σημείο. Ουσιαστικά μέσω αυτού του κριτηρίου θέλουμε να κατασκευάσουμε έναν αλγόριθμο που θα υλοποιεί leapfrog βήματα μέχρι να γίνει η ποσότητα $(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\cdot\boldsymbol{p}$ αρνητική, δηλαδή μέχρι η ποσότητα $\bar{\boldsymbol{\theta}}$ να αρχίζει να κινείται πίσω προς το $\boldsymbol{\theta}_0$. Παρόλα αυτά ένας τέτοιος αλγόριθμος δεν μας εγγυάται την ιδιότητα της αντιστρεψιμότητας και κατά συνέπεια μπορεί να μην συγκλίνουμε στην πραγματική κατανομή. Όμως αυτό το πρόβλημα μπορεί να αποφευχθεί ακολουθώντας μια δυναμική διαδικασία διπλασιασμού των σημείων σε κάθε επανάληψη προς τα εμπρός ή προς τα πίσω. Αυτή την ιδέα την εισήγαγε ο Randford Neal το 2003 για την υλοποίηση slice sampling.

Αρχικά θα ορίσουμε μια slice μεταβλητή u, με δεσμευμένη κατανομή

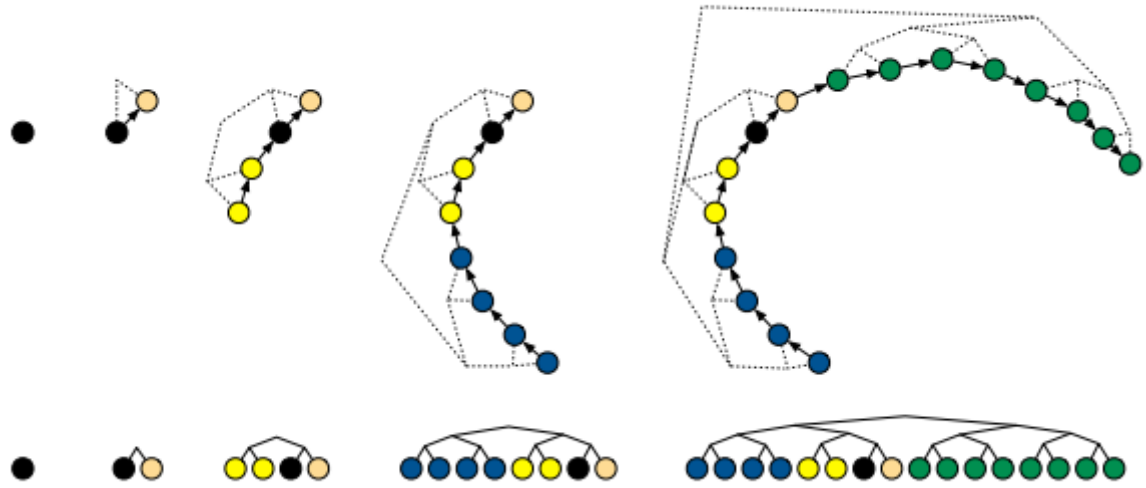
$$p(u|\boldsymbol{\theta}, \boldsymbol{p}) = Uniform(u; [0, exp\left\{V(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{p}\cdot\boldsymbol{p}\right\}]),$$

μέσω της οποίας οδηγούμαστε στην από κοινού κατανομή

$$p(\boldsymbol{\theta}, \boldsymbol{p}|u) = Uniform(\boldsymbol{\theta}, \boldsymbol{p}; \left\{\boldsymbol{\theta}', \boldsymbol{p}'|exp\left\{V(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{p}\cdot\boldsymbol{p}\right\} \geq u\right\}).$$

Γενικότερα χρησιμοποιούμε την slice μεταβλητή u για να διευκολυνθούμε στην υλοποίηση του αλγορίθμου NUTS.

Μετά την λήψη δείγματος $(u|\boldsymbol{\theta}, \boldsymbol{p})$, ο αλγόριθμος θα χρησιμοποιεί την αριθμητική μέθοδο leapfrog για να μετακινηθούμε είτε προς τα εμπρός είτε προς τα πίσω στην τροχιά. Η επιλογή της κατεύθυνσης συνηθέστερα γίνεται με χρήση της διακριτής ομοιόμορφης κατανομής $Unif\{-1, 1\}$. Σε κάθε επανάληψη θα διπλασιάζουμε το πλήθος των σημείων μας με κατεύθυνση προς τα εμπρος ή προς τα πίσω. Συνεπώς, θα υλοποιούμε δυο φορές περισσότερα leapfrog βήματα από το πλήθος των σημείων που είχαμε στην προηγούμενη επανάληψη. Μέσω αυτής της διαδικασίας δημιουργείται ένα ισορροπημένο δυαδικό δέντρο όπου τα φύλα του αντιστοιχούν στα σημεία $(\boldsymbol{\theta}, \boldsymbol{p})$.



Διάγραμμα 3.1: Διπλασιασμός και κατασκευή του ισορροπημένου δένδρου.

Γενικά η διαδικασία θα σταματάει στην περίπτωση που σε κάποιον διπλασιασμό ξεκινήσουμε να κατευθυνόμαστε προς το αρχικό μας σημείο, δηλαδή σε περιοχές που έχουμε ήδη εξερευνήσει (αυτή η κίνηση χαρακτηρίζεται και ως $U-turn$). Τέλος, όταν σταματήσει ο αλγόριθμος θα πάρουμε ομοιόμορφο δείγμα από τα σημεία που έχουμε προσομοιώσει.
Στην συνέχεια, ορίζουμε την από κοινού κατανομή των $\boldsymbol{\theta}, u, \boldsymbol{p}$

$$p(\boldsymbol{\theta}, \boldsymbol{p}, u) = \frac{p(u|\boldsymbol{\theta}, \boldsymbol{p})}{p(\boldsymbol{\theta}, \boldsymbol{p})} = \frac{1}{e^{V(\boldsymbol{\theta})-K(\boldsymbol{p})}} \mathbb{I}(u)_{\{0, e^{V(\boldsymbol{\theta})-K(\boldsymbol{p})}\}} \frac{e^{V(\boldsymbol{\theta})-K(\boldsymbol{p})}}{K} \propto \mathbb{I}(u)_{\{0, e^{V(\boldsymbol{\theta})-K(\boldsymbol{p})}\}},$$

όπου K είναι μια σταθερά κανονικοποίησης.
Ακόμα, θα ορίσουμε τα σύνολα $\mathcal{C}$ και $\mathcal{B}$ όπου $\mathcal{C} \subset \mathcal{B}$. Το σύνολο $\mathcal{B}$ θα περιέχει όλα εκείνα τα σημεία τα οποία προσομοιώθηκαν από την αριθμητική μέθοδο leapfrog από τους προς τα εμπρός και προς τα πίσω διπλασιασμούς μέ σκοπό την εξερεύνηση της τροχιάς. Το σύνολο $\mathcal{C}$ θα περιέχει εκείνα τα σημεία τα οποία επιλέχθηκαν ντετερμινιστικά από το σύνολο $\mathcal{B}$ και διατηρούν την ιδιότητα της αντιστρεψιμότητας. Μέσω αυτής της τυχαίας διαδικασίας δημιουργείται η δεσμευμένη από κοινού κατανομή, των $\mathcal{B}, \mathcal{C}$, $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon)$ πάνω στην οποία θέτουμε τις ακόλουθες συνθήκες:

$C.1$: Όλα τα σημεία τα οποία ανήκουν στο $\mathcal{C}$ θα πρέπει να έχουν την ιδιότητα διατήρησης του όγκου. Μέσω αυτού πετυχαίνουμε ότι για κάθε παρατήρηση που ανήκει στο $\mathcal{C}$, ισχύει ότι

$$p(\boldsymbol{\theta}, \boldsymbol{p}|(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}) \propto p(\boldsymbol{\theta}, \boldsymbol{p}),$$

δηλαδή κάθε σημείο που παράγουμε λόγω της διατήρησης του όγκου δεν θα μπορεί να αποκλίνει σημαντικά από την εκάστοτε τροχιά συνεπώς η κατανομή του θα είναι ανάλογη της κατανομής ενδιαφέροντος $p(\boldsymbol{\theta}, \boldsymbol{p})$.

$C.2$: $p((\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon) = 1$. Δηλαδή η παρούσα θέση που βρισκόμαστε θα πρέπει να ανήκει στο $\mathcal{C}$.

$C.3$: $p(u \leq exp\left\{V(\boldsymbol{\theta}^{'}) - \frac{1}{2}\boldsymbol{p}^{'} \cdot \boldsymbol{p}^{'}\right\} | (\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}) \in \mathcal{C}) = 1$. Δηλαδή, όλα τα σημεία για τα οποία $(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}$ θα πρέπει να ανήκουν στην περιοχή που ορίζει ο slice sampler και να έχουν ίδια δεσμευμένη πιθανότητα $p(\boldsymbol{\theta}, \boldsymbol{p}|u)$.

$C.4$: Αν $(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}$ και $(\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}) \in \mathcal{C}$ τότε για οποιοδήποτε σύνολο $\mathcal{B}$ να ισχύει ότι $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon) = p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}^{'}, \overset{'}{\surd}, u, \epsilon)$. Δηλαδή, από οποιοδήποτε σημείο του $\mathcal{C}$ και να ξεκινήσουμε θα πρέπει τα $\mathcal{B}, \mathcal{C}$ να έχουν ίδια πιθανότητα να παραχθούν.

Τέλος, θα δείξουμε ότι η διαδικασία που περιγράψαμε προηγουμένως αφήνει την κατανομή $p(\boldsymbol{\theta}, \boldsymbol{p}, u, \mathcal{B}, \mathcal{C}|\epsilon)$ αναλλοίωτη.

1. $\boldsymbol{p} \sim \mathcal{N}(0, I)$,
2. $u \sim Uniform([0, exp\left\{V(\boldsymbol{\theta}^t) - \frac{1}{2}\boldsymbol{p} \cdot \boldsymbol{p}\right\}])$,
3. $\mathcal{B}, \mathcal{C} \sim p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}^t, \boldsymbol{p}, u, \epsilon)$,
4. $\boldsymbol{\theta}^{t+1}, \boldsymbol{p} \sim T(\boldsymbol{\theta}^t, \boldsymbol{p}, \mathcal{C})$,

όπου $T(\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}|\boldsymbol{\theta}, \boldsymbol{p}, C)$ είναι ένας πυρήνας μετάβασης όπου αφήνει αναλλοίωτη την ομοιόμορφη κατανομή του C, δηλαδή ισχύει ότι

$$\frac{1}{|C|} \sum_{(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}} T(\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}|\boldsymbol{\theta}, \boldsymbol{p}, \mathcal{C}) = \frac{\mathbb{I}[(\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}) \in \mathcal{C}]}{|C|}.$$

Επίσης, το βήμα 4 είναι έγκυρο διότι η από κοινού δεσμευμένη κατανομή των $\boldsymbol{\theta}, \boldsymbol{p}$ είναι ομοιόμορφη στα στοιχεία της $\mathcal{C}$. Δηλαδή οι καταστάσεις στις οποίες μεταβαίνουμε αφήνουν αναλλοίωτη την κατανομή της $\mathcal{C}$.

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{p}|u, \mathcal{B}, \mathcal{C}, \epsilon) &\propto p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon) p(\boldsymbol{\theta}, \boldsymbol{p}|u), \\ &\propto p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon) \mathbb{I}[u \leq exp\left\{V(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{p} \cdot \boldsymbol{p}\right\}], \\ &\propto \mathbb{I}[(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}], \end{aligned}$$

άρα είμαστε ελεύθεροι να επιλέξουμε οποιοδήποτε $(\boldsymbol{\theta}^{t+1}, \boldsymbol{p}^{t+1})$ αρκεί να διαλέξουμε ένα πυρήνα μετάβασης ο οποίος αφήνει αναλλοίωτη την κατανομή του $\mathcal{C}$.
Μέσω της ιδιότητας $C.1$ και του τρόπου που έχει οριστεί η κατανομή $p(\theta, p|u)$ μπορούμε να γράψουμε ότι :

$$p(\boldsymbol{\theta}, \boldsymbol{p}|u) \propto \mathbb{I}[u \leq exp\left\{V(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{p} \cdot \boldsymbol{p}\right\}].$$

Επίσης, μέσω των ιδιοτήτων $C.2$ και $C.4$ έχουμε ότι:

$$p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon) \propto \mathbb{I}[(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}].$$

Διότι, από $C.2$ έχουμε $(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}$ και από $C.4$ γνωρίζουμε ότι η πιθανότητα $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon)$ είναι σταθερή ως προς $\boldsymbol{\theta}$ και $\boldsymbol{p}$ όσο ισχύει $(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}$. Τέλος, λόγω της ιδιότητας $C.3$ παίρνουμε ότι

$$\mathbb{I}[u \leq exp\left\{V(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{p} \cdot \boldsymbol{p}\right\}] = 1,$$

και άρα το ζητούμενο.
Συνεπώς, η από κοινού κατανομή των $(\boldsymbol{\theta}, \boldsymbol{p})$ δεδομένου των u και $\mathcal{C}$ θα είναι ομοιόμορφη στα στοιχεία του $\mathcal{C}$ και θα επιλέγουμε οποιοδήποτε $(\boldsymbol{\theta}^{t+1}, \boldsymbol{p}^{t+1})$ όσο διαλέγουμε πυρήνα μετάβασης ο οποίος να αφήνει αναλλοίωτη την ομοιόμορφη κατανομή του $\mathcal{C}$.

Στην συνέχεια θα ασχοληθούμε με την κατανομή $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon)$. Όπως έχουμε προαναφέρει το

σύνολο $\mathcal{B}$ δημιουργείται διπλασιάζοντας κάθε φορά προς τα εμπρός ή προς τα πίσω το σύνολο των προηγούμενων κόμβων δημιουργώντας ένα ισορροπημένο δέντρο με φύλλα που αντιστοιχούν στα σημεία $(\boldsymbol{\theta}, \boldsymbol{p})$. Αρχικά έχουμε ένα δέντρο με ένα κόμβο δηλαδή το αρχικό μας σημείο. Στην συνέχεια επιλέγουμε τυχαία μια κατεύθυνση $v_j \sim Unif\{-1, 1\}$ και υλοποιούμε 2^j leapfrog βήματα μεγέθους $v_j\epsilon$ με j να είναι το ύψος του τωρινού δέντρου. Επίσης ξεκινώντας από ένα σημείο $(\boldsymbol{\theta}, \boldsymbol{p})$ μπορούμε να κατασκευάσουμε 2^j διαφορετικά δέντρα ύψους j. Δηλαδή η πιθανότητα να κατασκευάσεις ένα συγκεκριμένο δέντρο ξεκινώντας από ένα ¨φύλο' του δέντρου είναι ίσο με 2^{-j}.

Τώρα όσο αναφορά τις συνθήκες που πρέπει να τηρούνται από το δέντρο ώστε να τερματίσει ο αλγόριθμος υπάρχουν δύο κριτήρια τα οποία μας υποδηλώνουν πότε πρέπει να σταματήσουμε τον διπλασιασμό του δέντρου. Το πρώτο κριτήριο τερματισμού αφορά την αύξηση του σφάλματος του αλγορίθμου λόγω της προσομοίωσης ακατάλληλων παρατηρήσεων (η αριθμητική προσομοίωση της τροχιάς αποκλίνει από την ζητούμενη τροχιά και όσο περισσότερο αποκλίνουμε τόσο περισσότερο μεγαλώνει το σφάλμα, αυτό για παράδειγμα μπορεί να συμβεί κάνοντας χρήση μεγάλου βήματος ϵ) οι οποίες θα έχουν υπερβολικά πολύ μικρή πιθανότητα αποδοχής (μιας και η πιθανότητα αποδοχής εξαρτάται από το πόσο κοντά είναι τα επίπεδα ενέργειας). Το δεύτερο κριτήριο αφορά όλα τα πιθανά υποδέντρα του συνόλου $\mathcal{B}$ και ελέγχει αν για τα ακραία αριστερά και δεξιά σημεία τους έχουμε ένα $U-turn$ δηλαδή αν ξεκινάμε να προσομοιώνουμε σημεία σε είδη εξερευνημένες περιοχές.

Όπως αναφέραμε το πρώτο κριτήριο δηλώνει ότι σταματάμε τον διπλασιασμό στην περίπτωση που σε κάποια επανάληψη πάμε να συμπεριλάβουμε έναν κόμβο ο οποίος αυξάνει το σφάλμα του αλγορίθμου με αποτέλεσμα να έχουμε αύξηση του επιπέδου ενέργειας και συνεπώς να μην έχουμε σε ισχύει την επιθυμητή ανισότητα του slice sampling $u \leq e^{-H(\boldsymbol{\theta},\boldsymbol{p})}$. Σε αυτή την περίπτωση σταματάμε τον αλγόριθμο NUTS όταν επιτευχθεί ότι

$$V(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{p}\cdot\boldsymbol{p} - logu < -\Delta_{max}, \quad (1)$$

για Δ_{max} μια μεγάλη μη αρνητική τιμή. Το δεύτερο κριτήριο αφορά υποδέντρα. Έστω $2^j - 1$ ισορροπημένα δέντρα ύψους j με $j > 0$. Ο αλγόριθμος NUTS θα σταματάει τον διπλασιασμό όταν για κάποιο από αυτά τα υποδέντρα ισχύει ότι

$$(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-)\cdot\boldsymbol{p}^- \quad \text{ή} \quad (\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-)\cdot\boldsymbol{p} < 0. \quad (2)$$

Τα $(\boldsymbol{\theta}^+, \boldsymbol{p}^+)$ και $(\boldsymbol{\theta}^-, \boldsymbol{p}^-)$ είναι τα ακραία σημεία του εκάστοτε υποδέντρου. Δηλαδή, στον τελευταίο διπλασιασμό ελέγχουμε όλα τα υποδέντρα τα οποία προστέθηκαν και στην περίπτωση που κάποιο υποδέντρο παραβιάζει το κριτήριο τότε σταματάμε τον αλγόριθμο NUTS.

Όλη η διαδικασία διπλασιασμού που έχουμε περιγράψει παράγει μια κατανομή $p(\mathcal{B}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon)$. Τώρα όμως θα πρέπει να ορίσουμε έναν ντετερμινιστικό τρόπο έτσι ώστε να μπορούμε να κατασκευάζουμε το σύνολο $\mathcal{C}$ προσέχοντας πάντα να τηρούνται οι συνθήκες $C1 - C4$ για την $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon)$.
Η συνθήκη $C1$ θα ικανοποιείται αυτόματα αφού η αριθμητική μέθοδος leapfrog έχει την ιδιότητα της διατήρησης του όγκου. Επίσης η $C2$ ικανοποιείται αρκεί να εμπεριέχεται στο σύνολο $\mathcal{C}$ η αρχική κατάσταση της αλυσίδας. Η ιδιότητα $C3$ ικανοποιείται αν δεν συμπεριλάβουμε τα σημεία για τα οποία ισχύει ότι $exp\left\{V(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{p}\cdot\boldsymbol{p}\right\} < u$. Για την ιδιότητα $C4$ γνωρίζουμε ότι πρέπει να ισχύει $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon) = p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}, u, \epsilon)$. Επίσης γνωρίζουμε ότι η πιθανότητα να παράγουμε οποιοδήποτε δέντρο από ένα σημείο $(\boldsymbol{\theta}, \boldsymbol{p})$ είναι 2^{-j} άρα θα έχουμε ότι $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}, \boldsymbol{p}, u, \epsilon) = 2^{-j} = p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}, u, \epsilon)$ ή $p(\mathcal{B}, \mathcal{C}|\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}, u, \epsilon) = 0$ (δηλαδή να μην μπορεί να παραχθεί το δέντρο από το σημείο $(\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'})$ διότι μπορεί να οδηγεί σε πρόωρο τερματισμό της διαδικασίας διπλασιασμού). Άρα η ιδιότητα $C4$ θα ικανοποιείται όσο δεν εμπεριέχουμε σημεία στο $\mathcal{C}$ τα οποία δεν θα μπορούσαν να παράγουν το $\mathcal{B}$. Ένα τέτοιο σημείο μπορεί να παραχθεί αν ξεκινώντας από το $\boldsymbol{\theta}^{'}, \boldsymbol{p}^{'}$ επιτευχθεί κάποιο κριτήριο τερματισμού πριν ολοκληρωθεί το ολικό δέντρο. Μπορούμε να σκεφτούμε δύο πιθανά σενάρια:

1. Η διαδικασία διπλασιασμού σταματάει επειδή επιτεύχθηκε μια εξίσωση από τις (1) και (2) για κάποιο κόμβο ή υποδέντρο. Σε αυτή την περίπτωση θα πρέπει να αφαιρέσουμε από το σύνολο $\mathcal{C}$ όποιο σημείο εντάχθηκε στον τελευταίο διπλασιασμό. Διότι ξεκινώντας τον διπλασιασμό από ένα από αυτά τα σημεία θα οδηγούσε σε τερματικό κριτήριο πριν παραχθεί ολόκληρο το δέντρο που αντιστοιχεί στο

σύνολο $\mathcal{B}$.

2. Η διαδικασία σταματάει διότι για το πλήρες δέντρο που περιέχει όλα τα σημεία του $\mathcal{B}$ επιτεύχθηκε η εξίσωση (2) για τα δύο ακραία σημεία του δέντρου. Σε αυτή την περίπτωση δεν είχαμε κάποιο κριτήριο τερματισμού για κανέναν κόμβο ή υποδέντρο συνεπώς η ιδιότητα $C.4$ επιτυγχάνεται αυτόματα.
Στον Αλγόριθμο 3 θα παρουσιάσουμε τον τρόπο και τα βήματα τα οποία ακολουθούνται έτσι ώστε να κατασκευαστεί το σύνολο $\mathcal{C}$.

Αλγόριθμος 3 Naive No-U-Turn Sampler

Δοθέντος θ^0, ϵ, V, M:
for $m = 1 : M$ **do**
  Πάρε $p^0 \sim \mathcal{N}(0, I)$
  Πάρε $u \sim Uniform([0, exp\{V(\theta^{m-1} - \frac{1}{2}p^0 \cdot p^0)\}])$
  Με $\theta^- = \theta^{m-1}, \theta^+ = \theta^{m-1}, p^- = p^0, p^+ = p^0, j = 0, \mathcal{C} = \{(\theta^{m-1}, p^0)\}, s = 1.$
  while *s=1* **do**
    Πάρε κατεύθυνση $u_j \sim Uniform(\{-1, 1\})$.
    if $v_j = -1$ **then**
      $\theta^-, p^-, -, -, \mathcal{C}', s' \leftarrow$ BuildTree$(\theta^-, p^-, u, v_j, j, \epsilon)$
    else
      $-, -, \theta^+, p^+, \mathcal{C}', s' \leftarrow$ BuildTree$(\theta^+, p^+, u, v_j, j, \epsilon)$
    end
    if $s' = 1$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}'$
    end
    $s \leftarrow s' \mathbb{I}[(\theta^+ - \theta^-) \cdot p^- \geq 0] \mathbb{I}[(\theta^+ - \theta^-) \cdot p^+ \geq 0].$
    $j \leftarrow j + 1$
  end
  Πάρε δείγμα (θ^m, p) ομοιόμορφα από το σύνολο $\mathcal{C}$.
end
function Buildtree$(\theta, p, u, v, j, \epsilon)$ **if** $j = 0$ **then**
  Κάνε ένα leapfrog βήμα με κατεύθυνση v
  $\theta', p' \leftarrow$ Leapfrog$(\theta, p, v\epsilon)$
  $\mathcal{C}' \leftarrow \begin{cases} \left\{(\theta', p')\right\} & if \quad u < exp\left\{V(\theta') - \frac{1}{2}p' \cdot p'\right\} \\ \emptyset & else \end{cases}$
  $s' \leftarrow \mathbb{I}[u < exp\left\{\Delta_{max} + V(\theta') - \frac{1}{2}p' \cdot p'\right\}]$
  return $\theta', p', \mathcal{C}', s'$
else
  $\theta^-, p^-, \theta^+, p^+, \mathcal{C}', s' \leftarrow$ BuildTree$(\theta, p, u, v, j - 1, \epsilon)$
  if $v = -1$ **then**
    $\theta^-, p^-, -, -, \mathcal{C}'', s'' \leftarrow$ BuildTree$(\theta^-, p^-, u, v, j - 1, \epsilon)$
  else
    $-, -, \theta^+, p^+, \mathcal{C}'', s'' \leftarrow$ BuildTree$(\theta^+, p^+, u, v, j - 1, \epsilon)$
  end
  $s' \leftarrow s' s'' \mathbb{I}[(\theta^+ - \theta^-) \cdot p^- \geq 0] \mathbb{I}[(\theta^+ - \theta^-) \cdot p^+ \geq 0].$
  $\mathcal{C}' \leftarrow \mathcal{C}' \cup \mathcal{C}''$
  return$(\theta^-, p^-, \theta^+, p^+, \mathcal{C}', s')$
end

Όταν $s \neq 0$ όταν δεν έχει επιτευχθεί κάποιο κριτήριο τερματισμού. Αρχικά να επισημάνουμε ότι το αρχικό σημείο $\boldsymbol{\theta}^{m-1}, \boldsymbol{p}^0$ περιλαμβάνεται στο $\mathcal{C}$ για την ικανοποίηση της ιδιότητας $C.2$. Επίσης διαλέγουμε σημεία από το σύνολο $\mathcal{C}$ ομοιόμορφα, συνεπώς όπως έχουμε δει και προηγουμένως αυτά τα σημεία αφήνουν την κατανομή της $\mathcal{C}$ αναλλοίωτη. Τέλος, μέσω του αλγορίθμου NUTS μας δίνεται η δυνατότητα να ελέγξουμε το ύψος j του δέντρου. Στην περίπτωση που ο αλγόριθμος σταματήσει

πρόωρα δηλαδή πριν ξεκινήσει να κάνει ένα U-turn θα έχουμε ως αποτέλεσμα ότι δεν έχουμε αξιοποιήσει πλήρως την πληροφορία της τροχιάς. Συνεπώς, ο αλγόριθμος NUTS (μέσω της πλατφόρμας STAN) δίνει την δυνατότητα να ελέγξουμε το ύψος του δέντρου και στην προκειμένη περίπτωση να το αυξήσουμε. Παρόλα αυτά πρέπει να είμαστε προσεκτικοί διότι με την αύξηση του j, αυξάνουμε και το πλήθος των $\frac{dV}{dt}$ που θα πρέπει να υπολογίσουμε, κάνοντας έτσι τον αλγόριθμο σημαντικά πιο αργό.

3.1.1 Αποτελεσματικός Αλγόριθμος NUTS

Παρατηρούμε ότι παρόλο που ο αλγόριθμος μας έχει έναν δυναμικό τρόπο για να καταλαβαίνει πότε πρέπει να τερματίσει υπάρχουν ακόμα μειονεκτήματα. Αρχικά για την υλοποίηση του χρειάζεται να υπολογίσει $2^j - 1$ φορές την συνάρτηση $V(\boldsymbol{\theta})$ και το διαφορικό της που ακόμα και στα πιο τετριμμένα προβλήματα είναι ανεπίτρεπτα μεγάλο το υπολογιστικό κόστος. Επίσης ο αλγόριθμος NUTS θα πρέπει να κρατήσει στην μνήμη του 2^j θέσεις $(\boldsymbol{\theta}, \boldsymbol{p})$ από τις οποίες κάθε φορά θα παίρνει σημεία ομοιόμορφα και για τα οποία σημεία δεν είναι εξασφαλισμένο ότι θα βρίσκονται σε πληροφοριακές περιοχές δηλαδή μακριά από το αρχικό σημείο. Ακόμα αυτές οι 2^j παρατηρήσεις μπορεί να είναι ένας αποτρεπτικά μεγάλος όγκος δεδομένων για την μνήμη που χρησιμοποιούμε. Τέλος, στην περίπτωση που επιτευχθεί ένα κριτήριο τερματισμού στην τελευταία επανάληψη δεν υπάρχει λόγος για τον υπολογισμό του τελευταίου υποσυνόλου του $\mathcal{C}$.
Αρχικά για την αντιμετώπιση του τελευταίου προβλήματος αρκεί να σταματήσουμε τον αλγόριθμο την στιγμή που έχουμε μηδενική τιμή για τον δείκτη s, σε αυτή την περίπτωση δεν θα συνεχιστούν οι υπολογισμοί και θα έχουν σωθεί οι επαναλήψεις πριν το τερματισμό του αλγορίθμου. Επίσης, για το πρόβλημα της μνήμης αντί να αποθηκεύουμε όλα τα σημεία σε κάθε διπλασιασμό και στο τέλος από 2^j σημεία (όπου στην συγκεκριμένη περίπτωση j είναι οι φορές που καλούμε την BuildTree συνάρτηση) να παίρνουμε δείγμα ομοιόμορφα θα έχουμε ότι σε κάθε διπλασιασμό που κάνουμε θα μετακινούμαστε σε κάποιο σημείο του καινούργιου υποδέντρου, έτσι θα αποθηκεύουμε μόνο τα σημεία στα οποία μετακινούμαστε συνεπώς το μέγεθος των παρατηρήσεων που θα έχουμε θα είναι της τάξεως $O(j)$. Τέλος, μέσω αυτής της διαδικασίας είναι πιο πιθανό το σημείο που θα πάρουμε στο τέλος να είναι αρκετά μακριά από το αρχικό μας σημείο, μιας και σε κάθε επανάληψη θα περνάμε σε καινούργιο υποδέντρο. Για την υλοποίηση όμως αυτής της σκέψης θα χρειαστούμε τον ακόλουθο πυρήνα μετάβασης

$$T(w^{'}|w, C) = \begin{cases} \frac{\mathbb{I}[w^{'} \in C^{new}]}{|C^{new}|} & \text{if } |\mathcal{C}^{new}| \geq \left|\mathcal{C}^{old}\right| \\ \frac{|C^{new}|}{|C^{old}|}\frac{\mathbb{I}[w^{'} \in \mathcal{C}^{new}]}{|\mathcal{C}^{new}|} + (1 - \frac{|\mathcal{C}^{new}|}{|\mathcal{C}^{old}|})\mathbb{I}[w^{'} = w] & \text{if } |\mathcal{C}^{new}| \leq \left|\mathcal{C}^{old}\right| \end{cases}$$

όπου $w \in \mathcal{C}^{old}$ και $w^{'} \in \mathcal{C}^{new}$ με $\mathcal{C}^{old} \cup \mathcal{C}^{new} = \mathcal{C}$ και $\mathcal{C}^{old} \cap \mathcal{C}^{new} = \mathcal{C}$. Ουσιαστικά τον πυρήνα μετάβασης T προτείνει μια μετακίνηση από το σύνολο $\mathcal{C}^{old}$ (το δέντρο που είχαμε στην τελευταία επανάληψη j) στο $\mathcal{C}^{new}$ (το δέντρο το οποίο πρότεινε η χρήση της συνάρτησης BuildTree) και γενικά θα αποδεχτούμε μια τέτοια κίνηση με πιθανότητα $\frac{|\mathcal{C}^{new}|}{|\mathcal{C}^{old}|}$ (αυτή η πιθανότητα δεν ταυτίζεται με την πιθανότητα να μετακινηθούμε από ένα σημείο του $\mathcal{C}^{old}$ σε ένα του $\mathcal{C}^{new}$). Δηλαδή αντί στο τέλος του αλγορίθμου να παίρνουμε ένα σημείο ομοιόμορφα από το $\mathcal{C}$, τώρα μέσω του πυρήνα μετάβασης T σε κάθε διπλασιασμό θα κάνουμε μια μετακίνηση στο καινούργιο μισό δέντρου που προτείνεται. Τέλος, αυτός ο πυρήνας μετάβασης επιτυγχάνει την ακριβή ισορροπία για την ομοιόμορφη κατανομή του συνόλου $\mathcal{C}$,

$$p(w|\mathcal{C})T(w^{'}|w, \mathcal{C}) = p(w^{'}|\mathcal{C})T(w|w^{'}, \mathcal{C}),$$

και συνεπώς αφήνει την κατανομή του $\mathcal{C}$ αναλλοίωτη.

Ακόμα με τον πυρήνα μετάβασης που ορίσαμε προηγουμένως θα πρέπει να είμαστε σε θέση να παράγουμε δείγμα ομοιόμορφο από το $\mathcal{C}^{'}$ που μπορεί να περιέχει μέχρι και 2^{j-1} (σε κάθε διπλασιασμό που υλοποιούμε το καινούργιο μισό δέντρο μπορεί να έχει το πολύ 2^{j-1} σημεία ανάλογα με το πλήθος που ικανοποιεί τις συνθήκες $C.1 - C.4$). Υπάρχει η δυνατότητα να παράγουμε δείγματα από το $\mathcal{C}^{'}$ χωρίς να χρειάζεται να αποθηκεύσουμε όλο το σύνολο κάνοντας χρήση της δυαδικής μορφής που έχει το δέντρο που παράγουμε. Έστω το υποδέντρο $\mathcal{C}_{subtree}$ του συνόλου $\mathcal{C}$, για το οποίο μπορούμε να έχουμε ότι ένα σημείο $(\boldsymbol{\theta}, \boldsymbol{p}) \in \mathcal{C}_{subtree}$ θα επιλεχθεί ομοιόμορφα από το $\mathcal{C}^{'}$ με πιθανότητα:

$$p(\theta, p|\mathcal{C}^{'}) = \frac{1}{|\mathcal{C}^{'}|} == \frac{|\mathcal{C}_{subtree}|}{|\mathcal{C}^{'}|}\frac{1}{|\mathcal{C}_{subtree}|}$$

$$p((\boldsymbol{\theta},\boldsymbol{p}) \in \mathcal{C}_{subtree}|\mathcal{C}^{'})p(\boldsymbol{\theta},\boldsymbol{p}|(\boldsymbol{\theta},\boldsymbol{p}) \in \mathcal{C}_{subtree},\mathcal{C}^{'}).$$

Δηλαδή, η πιθανότητα $p(\boldsymbol{\theta},\boldsymbol{p}|\mathcal{C}^{'})$ είναι το γινόμενο να επιλέξουμε ένα σημείο από το υποδέντρο $\mathcal{C}_{subtree}$ και να επιλεχθούν τα θ, p ομοιόμορφα από το $\mathcal{C}_{subtree}$. Κάθε υποδέντρο θα χαρακτηρίζεται από τα σημεία του, δηλαδή από την πιθανότητα $p(\boldsymbol{\theta},\boldsymbol{p}|(\boldsymbol{\theta},\boldsymbol{p}) \in \mathcal{C}_{subtree})$. Επίσης κάθε διπλασιασμό θα παράγει ένα καινούργιο υποδέντρο $\mathcal{C}^{'}$ το οποίο όμως θα αποτελείται κάθε φορά από δύο μικρότερα υποδέντρα $\mathcal{C}_{subtree}$. Για κάθε ένα από αυτά τα δύο υποδέντρα παίρνουμε σημείο $(\boldsymbol{\theta},\boldsymbol{p})$ από το $p(\boldsymbol{\theta},\boldsymbol{p}|(\boldsymbol{\theta},\boldsymbol{p}) \in \mathcal{C}_{subtree})$ το οποίο σημείο θα χαρακτηρίζει και το υποδέντρο. Στην συνέχεια θα διαλέγουμε ένα από τα δύο σημεία ανάλογα με το βάρος που δίνει το κάθε υποδέντρο στο σημείο του, το οποίο είναι ο λόγος του πλήθους των σημείων του υποδέντρου $\mathcal{C}_{subtree}$ προς το $\mathcal{C}^{'}$. Μέσω αυτής της διαδικασίας θα μπορούμε να παίρνουμε ένα δείγμα $\boldsymbol{\theta}^{'}$ από το σύνολο $\mathcal{C}^{'}$ και έναν ακέραιο $n^{'}$ το οποίο κωδικοποιεί το μέγεθος του $\mathcal{C}^{'}$, μέσω του οποίου θα μπορούμε να υλοποιούμε και να υπολογίζουμε τον πυρήνα μετάβασης T. Συνεπώς γίνεται κατανοητό ότι αρχικά παράγουμε ομοιόμορφα ένα δείγμα από το καινούργιο μισό δέντρο $\mathcal{C}^{'}$ και στην συνέχεια μέσω του πυρήνα μετάβασης T μπορούμε και μετακινούμαστε σε ένα καινούργιο σημείο και αυτή την διαδικασία την υλοποιούμε για κάθε διπλασιασμό που κάνουμε. Μέσω αυτής της διαδικασίας χρειάζεται να καταχωρίσουμε μόνο $O(j)$ σημεία στην μνήμη αντί για $O(2^j)$ και επίσης επειδή μετακινούμαστε σε κάθε διπλασιασμό έχουμε μεγαλύτερη πιθανότητα τελικά να βρεθούμε σε ένα σημείο μακριά από την αρχική μας τιμή. Στον Αλγόριθμο 4 περιγράφουμε την διαδικασία που ακολουθείτε για την κατασκευή του αποτελεσματικού αλγορίθμου μέσω της μεταπήδησης απο μισό δέντρο σε μισό δέντρο.

Αλγόριθμος 4 Efficient No-U-Turn Sampler

Δοθέντος θ^0, ϵ, V, M:
for $m = 1 : M$ **do**
  $p^0 \sim \mathcal{N}(0, I)$
  $u \sim Uniform([0, exp\{V(\theta^{m-1}) - \frac{1}{2}p^0 \cdot p^0\}])$ Με $\theta^- = \theta^{m-1}, \theta^+ = \theta^{m-1}, p^- = p^0, p^+ = p^0, j = 0, \theta^m = \theta^{m-1}, n = 1, s = 1.$
  while $s = 1$ **do**
    $v_j \sim Uniform(\{-1, 1\})$
    if $v_j = -1$ **then**
      $\theta^-, p^-, -, -, \theta^{'}, n^{'}, s^{'} \leftarrow BuildTree(\theta^-, p^-, u, v_j, j, \epsilon).$
    else
      $-, -, \theta^+, p^+, \theta^{'}, n^{'}, s^{'} \leftarrow BuilTree(\theta^+, p^+, u, v_j, \epsilon).$
    end
    if $s^{'} = 1$ **then**
      Με πιθανότητα $min\left\{1, \frac{n^{'}}{n}\right\}$, $\theta^m \leftarrow \theta^{'}$.
    end
    $n \leftarrow n + n^{'}$
    $s \leftarrow s^{'}\mathbb{I}[(\theta^+ - \theta^-) \cdot p^- \geq 0]\mathbb{I}[(\theta^+ - \theta^-) \cdot p^+ \geq 0]$
    $j \leftarrow j + 1$
  end
end
function $BuildTree(\theta, p, u, v, j, \epsilon)$
if $j = 0$ **then**
  $\theta^{'}, p^{'} \leftarrow Leapfrog(\theta, p, v\epsilon)$
  $n^{'} \leftarrow \mathbb{I}[u \leq exp\left\{V(\theta) - \frac{1}{2}p^{'} \cdot p^{'}\right\}]$
  $s^{'} \leftarrow \mathbb{I}[u < exp\left\{\Delta_{max} + V(\theta^{'}) - \frac{1}{2}p^{'} \cdot p^{'}\right\}]$
  return $\theta^{'}, p^{'}, n^{'}, s^{'}$
else
  $\theta^-, p^-, \theta^+, p^+, \theta^{'}, n^{'}, s^{'} \leftarrow BuildTree(\theta, p, u, v, j-1, \epsilon)$ **if** $s^{'} = 1$ **then**
    if $v = -1$ **then**
      $\theta^-, p^-, -, -, \theta^{''}, n^{''}, s^{''} \leftarrow BuildTree(\theta^-, p^-, u, v, j-1, \epsilon)$
    else
      $-, -, \theta^+, p^+, \theta^{''}, n^{''}, s^{''} \leftarrow BuildTree(\theta^+, p^+, u, v, j-1, \epsilon)$
    end
    Με πιθανότητα $\frac{n^{''}}{n^{'}+n^{''}}$, $\theta^{'} \leftarrow \theta^{''}$
    $s^{'} \leftarrow s^{''}\mathbb{I}[(\theta^+ - \theta^-) \cdot p^- \geq 0]\mathbb{I}[(\theta^+ - \theta^-) \cdot p^+ \geq 0]$
    $n^{'} \leftarrow n^{'} + n^{''}$
  end
  return $\theta^-, p^-, \theta^+, p^+, \theta^{'}, n^{'}, s^{'}$
end

3.2 Επιλογή μεγέθους βήματος ϵ

Έχοντας συζητήσει τον τρόπο με τον οποίο κάνουμε αυτόματη την επιλογή του μήκους L, τώρα θα συζητήσουμε και τον τρόπο με τον οποίο επιλέγεται το βήμα ϵ από τον αλγόριθμο NUTS . Θα βασιστούμε στην ιδέα ότι διαλέγοντας το επίπεδο πιθανότητας αποδοχής θα καθορίσουμε και το αντίστοιχο βήμα. Εμπειρικά έχει αποδειχθεί ότι με πιθανότητα αποδοχής 0.65 έχουμε έναν αποτελεσματικό αλγόριθμο. Για να καταφέρουμε να βρούμε το βήμα ϵ θα χρησιμοποιήσουμε στοχαστικές μεγιστοποιήσεις και θα διαλέξουμε εκείνο το βήμα ϵ το οποίο μας δίνει πιθανότητα αποδοχής 0.65.

Έστω ότι έχουμε μια ποσότητα H_t η οποία μεταβάλλεται με κάθε επανάληψη $t \geq 1$ του MCMC

αλγορίθμου. Μέσω της οποίας μπορούμε να ορίσουμε την αναμενόμενη τιμή

$$h(x) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[H_t|x]$$

όπου $x \in \mathbb{R}$ είναι μια μεταβλητή του MCMC αλγορίθμου.

Για παράδειγμα, έστω ότι a_t είναι η Metropolis πιθανότητα αποδοχής και ορίζουμε $H_t := \delta - a_t$ όπου δ είναι η επιθυμητή πιθανότητα αποδοχής (στην περίπτωση μας 0.65). Τότε για την συγκεκριμένη H_t μπορούμε να ορίσουμε την αντίστοιχη $h(x)$. Στην περίπτωση που συνθήκες όπως τα x_t να είναι φραγμένα, η h είναι μή φθίνουσα και άλλες ιδιότητες, που αναφέρονται στο paper Andrieu and Thoms (2008), τηρούνται τότε μέσω επαναληπτικών διαδικασιών θα έχουμε ότι:

$$x_{t+1} \leftarrow x_t - \eta_t H_t,$$

και αναγκαστικά η $h(x_t)$ θα συγκλίνει στο 0 αρκεί για το η_t να ισχύουν τα ακόλουθα:

$$\sum_t \eta_t = \infty \qquad \sum_t \eta_t^2 < \infty.$$

Τέτοιες συνθήκες ικανοποιούνται από η_t της μορφής $\eta_t = t^{-k}$ με $k \in (0.5, 1]$.

Παρόλα αυτά πρέπει να λάβουμε υπόψιν ότι οι βέλτιστες τιμές για τον αλγόριθμο MCMC διαφέρουν σημαντικά μεταξύ των φάσεων του burn-in και της στασιμότητας. Συνεπώς, θα χρησιμοποιήσουμε το dual averaging για να συμπεριλάβουμε αυτή την ιδιαιτερότητα για την εύρεση του βήματος ϵ.
Η μέθοδος dual averaging για το συγκεκριμένο πρόβλημα εφαρμόζεται ως εξής:

$$x_{t+1} \leftarrow \mu - \frac{\sqrt{t}}{\gamma} \frac{1}{t+t_0} \sum_{i=1}^{t} H_i, \qquad (*)$$

$$\bar{x}_{t+1} \leftarrow \eta_t x_{t+1} + (1-\eta_t)\bar{x}_t.$$

Όπου μ είναι σημείο προς το οποίο συγκλίνει η x_t, ενώ το $\gamma > 0$ κοντρολάρει την σύγκλιση προς το μ. Επίσης, το $t_0 \geq 0$ σταθεροποιεί τις αρχικές παρατηρήσεις με $\eta_t = t^{-k}$ να είναι το βήμα που είχαμε ορίσει αρχικά για την επαναληπτική διαδικασία και ισχύει ότι $\bar{x}_1 = x_1$.

Στον αλγόριθμο HMC θέλουμε βήμα ϵ το οποίο να μην είναι μικρό (διότι θα σπαταλήσουμε υπολογιστικούς πόρους) ούτε μεγάλο (διότι θα απορρίπτουμε πολλές προτεινόμενα σημεία). Όπως αναφέραμε και προηγουμένως επιλέγουμε ϵ τέτοιο ώστε η Metropolis πιθανότητα αποδοχής να είναι ίση με δ δηλαδή 0.65. Για τον υπολογισμό του ϵ μέσω του αλγορίθμου HMC ορίζουμε

$$H_t^{HMC} = min\left\{1, \frac{\pi(\boldsymbol{\theta}^t, \boldsymbol{p}^t)}{\pi(\boldsymbol{\theta}^{t-1}, \boldsymbol{p}^{t-1})}\right\} \quad h^{HMC}(\epsilon) = \mathbb{E}_t[H_t^{HMC}|\epsilon]$$

όπου H_t^{HMC} είναι η πιθανότητα αποδοχής για την t επανάληψη και h^{HMC} είναι η αναμενόμενη πιθανότητα αποδοχής εξαρτώμενη από το βήμα ϵ. Τώρα υποθέτοντας ότι η h^{HMC} είναι μή φθίνουσα ως προς ϵ και ότι τηρούνται όλες οι απαραίτητες προυποθέσεις μπορούμε υλοποιήσουμε την εξίσωση (*) για $H_t = \delta - H_t^{HMC}$ που θα μας οδηγήσει στο $h^{HMC} = \delta$ με $\delta \in (0, 1)$.

Επίσης για τον αλγόριθμο NUTS γνωρίζουμε ότι επιλέγουμε στοιχεία κάνοντας δειγματοληψία από το σύνολο $\mathcal{C}$. Για τον λόγο αυτό θα ορίσουμε μια εναλλακτική Metropolis πιθανότητα αποδοχής. Ορίζουμε ως H_t^{NUTS} και h^{NUTS}

$$H_t^{NUTS} = \frac{1}{\left|\mathcal{B}_t^{final}\right|} \sum_{(\boldsymbol{\theta},\boldsymbol{p})\in\mathcal{B}_t^{final}} min\left\{1, \frac{\pi(\boldsymbol{\theta}, \boldsymbol{p})}{\pi(\boldsymbol{\theta}^{t-1}, \boldsymbol{p}^{t-1})}\right\} \quad h^{NUTS} = \mathbb{E}_t[H_t^{NUTS}]$$

όπου $\mathcal{B}_t^{final}$ είναι το σύνολο των σημείων που παράχθηκαν στον τελευταίο διπλασιασμό της επανάληψης t. Πάλι θα πρέπει να τηρούνται όλες οι απαραίτητες συνθήκες για την h^{NUTS} με $H_t = \delta - H^{NUTS}$ τότε $h^{NUTS} = \delta$ και $\delta \in (0, 1)$.

Κεφάλαιο 4

Εισαγωγή στο STAN με Χρήση R

Το STAN είναι μια σύγχρονη πλατφόρμα για την υλοποίηση απαιτητικών Μπευζιανών στατιστικών υπολογισμών. Δημιουργήθηκε από τους Andrew Gelman και Bob Carpenter και πήρε την ονομασία του από τον Stanislaw Ulam συνεφευρέτη της μεθόδου Markov Chain Monte Carlo .
Μοιάζει αρκετά με το προγραμματιστικό πακέτο BUGS αλλά περιέχει περισσότερες διαδικασίες που πρέπει να υλοποιηθούν για την αποφυγή λαθών στην μοντελοποίηση για παράδειγμα απαιτεί εισαγωγή πεδίων ορισμού, ιδίως μεταβλητών κ.τ.λ. Το κύριο πλεονέκτημα του STAN είναι ότι υλοποιεί MCMC δειγματοληψία βασιζόμενος στα χαρακτηριστικά του HMC αλγορίθμου αξιοποιώντας όμως την αποτελεσματικότητα του αλγορίθμου NUTS για την δημιουργία τροχιών.
Για να χρησιμοποιήσει όμως κάποιος την πλατφόρμα STAN στην R θα πρέπει να εγκαταστήσει το πακέτο rstan κάνοντας χρήση της εντολής

```
install.packages("rstan"),
```

και στην συνέχεια να τρέξει την βιβλιοθήκη

```
library(rstan).
```

4.1 Σύνταξη Μοντέλου στο STAN

Αρχικά για να συντάξουμε ένα μοντέλο θα πρέπει να δημιουργήσουμε ένα αρχείο STAN , αλλάζοντας απλά την συντομογραφία του script που χρησιμοποιούμε στην R σε stan και εν συνεχεία αποθηκεύοντας το.
Ένα μοντέλο σε STAN μπορεί να συνταχθεί μέσω της χρήσης έξη προγραμματιστικών block.

- Data(required)
- Transformed Data
- Parameters(required)
- Transformes Parameters
- Model(required)
- Generated quantiles

Για να καταλάβουμε καλύτερα πως χρησιμοποιούνται τα προαναφερθέντα block ας δούμε πώς μοντελοποιούμε την ακόλουθη γραμμική παλινδρόμηση.

$$y_n = a + bx_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma),$$

η οποία είναι ισοδύναμη με

$$y_n \sim \mathcal{N}(a + bx_n, \sigma).$$

Στο Data block εισάγουμε τα δεδομένα του προβλήματος, ορίζοντας το είδος τους, το μέγεθος τους και το πεδίο ορισμού τους.

```
data{
   int<lower=0> N;
   vector[N] x;
   vector[N] y;
}
```

Παρατηρούμε ότι για να εισάγουμε το μέγεθος του δείγματος δηλώνουμε την φύση της μεταβλητής ότι είναι ακέραιος (integer)και το πεδίο ορισμού της, δηλαδή θα παίρνει τιμές μεγαλύτερες του μηδενός. Επίσης για τις μεταβλητές x και y δηλώνουμε ότι είναι διανύσματα (vector) μεγέθους N όσο και το μέγεθος του δείγματός μας. Επειδή όμως οι δύο ποσότητες ανήκουν στους πραγματικούς αριθμούς, μπορούμε να παραλείψουμε την εισαγωγή του αντίστοιχου περιορισμού (real). Να τονίσουμε ότι μετά από κάθε αυτόνομη γραμμή κώδικα που εισάγουμε είναι απαραίτητη και η χρήση του ελληνικού ερωτηματικού ";" .

Ακόμα θα μπορούσαμε να μετασχηματίσουμε τα δεδομένα μας δηλαδή τα N, x, y στην περίπτωση που θέλαμε να δηλώσουμε κάτι τέτοιο θα έπρεπε να το εισάγουμε στο ακόλουθο block. Για παράδειγμα μπορεί να θέλαμε να ορίσουμε ότι η επεξηγηματική μεταβλητή y είναι διπλάσια της ανεξάρτητης x.

```
transformed data{
   vector y[N];
   for(i in 1:N){
          y[i] = 2*x[i];
   }
  }
```

Στην συνέχεια θα εισάγουμε τις παραμέτρους του μοντέλου δηλαδή όλες εκείνες τις μεταβλητές για τις οποίες θέλουμε να εξάγουμε συμπερασματολογία

```
parameters{
   real alpha;
   real beta;
   real <lower=0> sigma;
 }
```

Παρατηρούμε ότι για να δηλώσουμε ότι το $a \in \mathbb{R}$ αρκεί να γράψουμε ότι είναι πραγματικός αριθμός(real),όμοια και για την παράμετρο b. Μέχρι στιγμής θα πρέπει να έχει γίνει κατανοητό ότι για την μοντελοποίηση της διακύμανσης που είναι μια θετική πραγματική παράμετρος αρκεί να συνδυάσουμε τα προαναφερθέντα, δηλαδή ότι ανήκει στους πραγματικούς αριθμούς και παίρνει μόνο θετικές τιμές. Να σημειωθεί ότι στην περίπτωση που θέλουμε να εισάγουμε ένα άνω φράγμα έστω το 1 αρκεί πολύ απλά να γράψουμε $< lower = 0, upper = 1 >$.

Επιπλέον στην περίπτωση που μας ενδιαφέρει να μετασχηματίσουμε κάποια παράμετρο αρκεί να την εισάγουμε στο transformed parameters block . Για παράδειγμα θα μπορούσαμε για την μοντελοποίηση της διακύμανσης να χρησιμοποιούσαμε μια precision μεταβλητή .

```
transformed parameters{
   real<lower=0> sig;
   sig=1/tau ;
   }
```

όπου το *tau* θα ορίζεται στο parameters block . Επίσης, πρέπει να συγκεκριμενοποιήσουμε και τις κατανομές που υπεισέρχονται στο μοντέλο (συνάρτηση πιθανοφάνεια, πρότερη κατανομή)

```
model{
y ~ normal(alpha+beta*x,sigma);
}
```

Παρατηρούμε ότι για τις παραμέτρους a, b, σ δεν έχουμε ορίσει κάποιες πρότερες κατανομές, σε αυτή την περίπτωση αυτόματα από την πλατφόρμα του STAN επιλέγει ομοιόμορφες μη πληροφοριακές πρότερες κατανομές. Αν θέλαμε να εισάγουμε κάποια συγκεκριμένη πρότερη κατανομή αρκεί να γράψουμε

```
variable ~ F(v,u);
```

όπου F μια οποιαδήποτε κατανομή με παραμέτρους v, u οι οποίες με την σειρά τους θα πρέπει να οριστούν στο block των παραμέτρων στην περίπτωση που θέλουμε να τις θεωρήσουμε και αυτές άγνωστες (ιεραρχικό μοντέλο).

Τέλος, αν θέλουμε μπορούμε να χρησιμοποιήσουμε το block generated quantiles μέσω του οποίου ζητάμε από την ύστερη κατανομή να υπολογίσει κάποια ποσότητα ενδιαφέροντος (postprocessing).Για παράδειγμα θα μπορούσαμε να ζητήσουμε να υπολογίσει κάποια πρόβλεψη κ.τ.λ.

```
generated quantities{
  vector[N] y_pred;
  y_pred = alpha+x*beta;
}
```

Αφού έχουμε ορίσει τα απαραίτητα για το πρόβλημα blocks τα αποθηκεύουμε σε ένα αρχείο STAN ή ακόμα μπορούμε να αποθηκεύσουμε τα blocks,κάνοντας χρήση του κλασσικού script της R, σε μια μεταβλητή της αρεσκείας μας ως εξής:

$$code < -'\Sigma\acute{\upsilon}\nu\text{ολο των blocks }'$$

Σε κάθε περίπτωση για να τρέξουμε τον αλγόριθμο NUTS θα διαβάζουμε αρχικά την βιβλιοθήκη *rstan* και στην συνέχεια θα χρησιμοποιήσουμε μια από τις ακόλουθες εντολές ανάλογα με τον τρόπο που έχουμε αποθηκεύσει τα blocks

```
stan(model_code = 'code' , data=c("y","N","x"), chains = 4 , warmup = 1000 , iter = 2000),

 stan(file = 'file.stan', data=c("y","N","x"), chains = 4 , warmup = 1000 , iter = 2000).
```

Στην πρώτη περίπτωση χρησιμοποιούμε την μεταβλητή που αποθηκεύσαμε τα blocks,ενώ στην δεύτερη χρησιμοποιούμε το αρχείο *.stan* στο οποίο αποθηκεύσαμε τα blocks.Επίσης, όπως παρατηρούμε εισάγαμε όλα τα απαραίτητα δεδομένα (y, x, N) και τρέξαμε τον αλγόριθμο για 4 αλυσίδες (γενικά είναι καλό να τρέχουμε περισσότερες από μια αλυσίδες) με burn-in ή αλλιώς warm-up 1000 παρατηρήσεις από τις συνολικές 2000 που θα τρέξει η κάθε αλυσίδα ξεχωριστά. Ακόμα έστω ότι έχουμε αποθηκεύσει και τρέξει τον αλγόριθμο NUTS σε μια μεταβλητή fit

```
fit <- stan(model_code = 'code' , data=c("y","N","x"), chains = 4 ,

 warmup = 1000 , iter = 2000).
```

Μέσω της εντολής print(fit) μπορούμε να πάρουμε αποτελέσματα που αφορούν μέσες τιμές, τυπικά σφάλματα κ.τ.λ. των παραμέτρων ενδιαφέροντος.

```
Inference for Stan model: 8de990b2608ae646acb6a6dd2c903f05.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

        mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
a       0.32    0.01 0.28  -0.24   0.13   0.31   0.51   0.88  1837    1
b       0.27    0.00 0.07   0.15   0.23   0.27   0.31   0.40  1897    1
sigma   1.04    0.00 0.08   0.90   0.99   1.04   1.09   1.20  2260    1
lp__  -53.32    0.03 1.22 -56.42 -53.89 -52.98 -52.41 -51.94  1382    1

Samples were drawn using NUTS(diag_e) at Fri Jul 12 19:39:30 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Επίσης, τα αποτελέσματα μπορούμε να τα εκφράσουμε και μέσω Shiny.Αρκεί να τρέξουμε τις εντολές

```
install.packages("shinystan")

library("shinystan")
```

μέσω της οποίας θα εγκαταστήσουμε και θα εισάγουμε την βιβλιοθήκη της Shiny στα πλαίσια όμως της πλατφόρμας STAN.Στην συνέχεια μπορούμε να τρέξουμε την εντολή

```
launch_shinystan(fit)
```

μέσω της οποίας έχουμε πρόσβαση σε διαγνωστικούς ελέγχους, αποτελέσματα και γενικότερα σε διαγράμματα που αφορούν την καλή προσαρμογή του αλγορίθμου.

4.2 Είδη Μεταβλητών και Αριθμητικών Δεδομένων

Όπως ήδη έχουμε καταλάβει υπάρχουν πληθώρα αριθμητικών δεδομένων τα οποία μπορούν να χρησιμοποιηθούν στην πλατφόρμα STAN.

• Scalar

$int\ N;$, $int < lower = 0, upper = 1 >\ cond;$

• Vector

Παίρνουν μόνο πραγματικές τιμές.

$vector < lower = 0 > [3]\ u;$

$simplex[5]\ theta;$ Πέντε συνιστώσες οι οποίες είναι μη αρνητικές και αθροίζουν στην μονάδα.

$unit\_vector[5]\ theta;$ Διάνυσμα με νόρμα 1 .

$ordered[5]\ c;$ Οι συντεταγμένες του διανύσματος σε αύξουσα σειρά.

$positive\_ordered[5]\ d;$

$row_vector < lower = -1, upper = 1 > [10]u;$ Διάνυσμα γραμμή.

• Matrix

$matrix < upper = 0 > [3, 4]\ B$;

$corr\_matrix[3]\ Sigma$; Πίνακας με τιμές από -1 έως 1.

$cov\_matrix[K]\ Omega$; Συμμετρικός και θετικά ορισμένος πίνακας.

Κεφάλαιο 5

Αριθμητικά Πειράματα

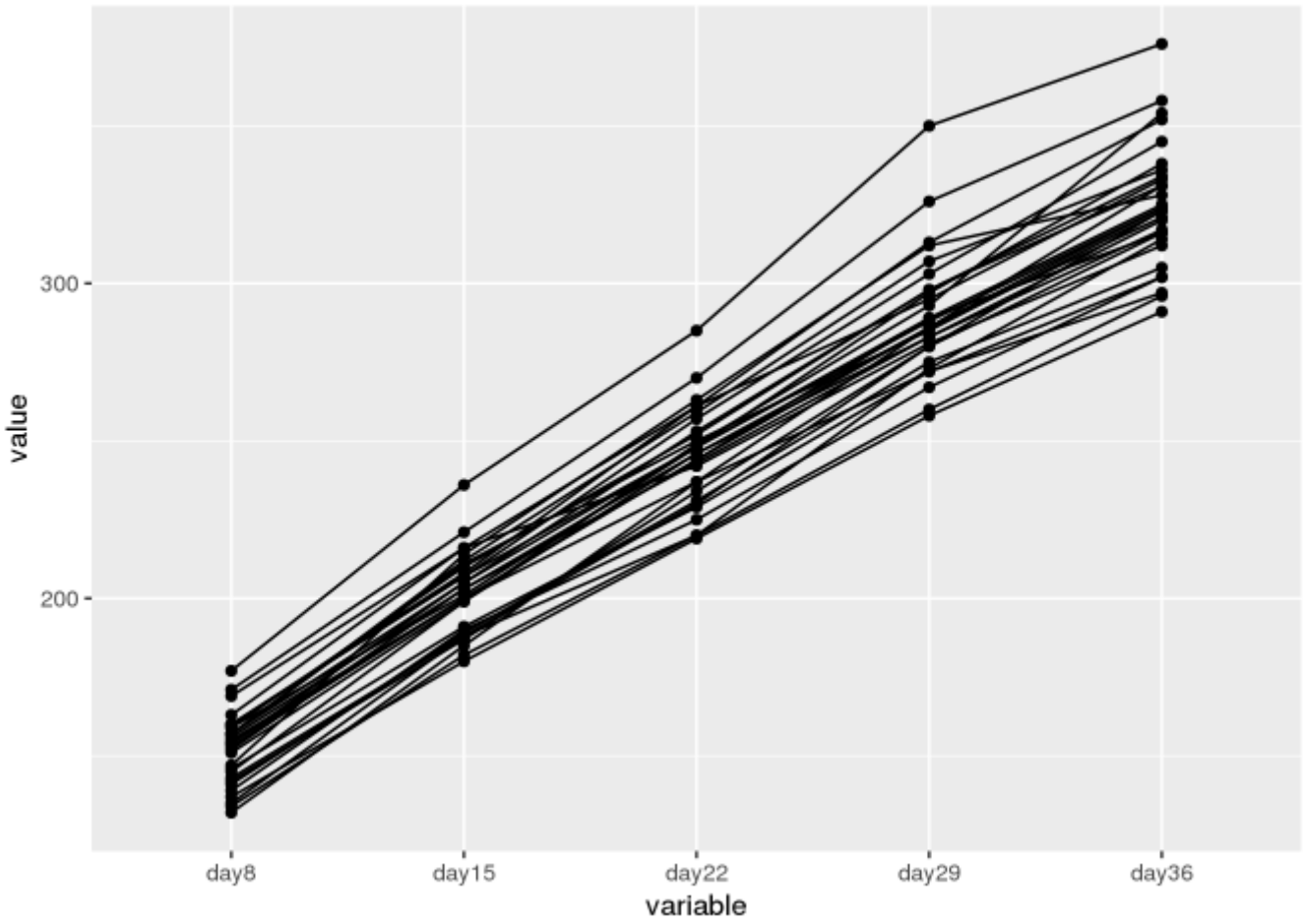
5.1 Πρόβλεψη Βάρους Ποντικιών

Μετρήθηκε το βάρος 30 ποντικιών σε διάστημα 5 εβδομάδων. Στον ακόλουθο Πίνακα 5.1 παρουσιάζουμε τα δεδομένα τα οποία συλλέχθηκαν και στο τέλος κάθε βδομάδας παρουσιάζουμε το βάρος κάθε ποντικιού.

8η μέρα	15η μέρα	22η μέρα	29η μέρα	36η μέρα
151	199	246	283	320
145	199	249	293	354
147	214	263	312	328
155	200	237	272	297
135	188	230	280	323
159	210	252	298	331
141	189	231	275	305
159	201	248	297	338
177	236	285	350	376
134	182	220	260	296
160	208	261	313	352
143	188	220	273	314
154	200	244	289	325
171	221	270	326	358
163	216	242	281	312
160	207	248	288	324
142	187	234	280	316
156	203	243	283	317
157	212	259	307	336
152	203	246	286	321
154	205	253	298	334
139	190	225	267	302
146	191	229	272	302
157	211	250	285	323
132	185	237	286	331
160	207	257	303	345
169	216	261	295	333
157	205	248	289	316
137	180	219	258	291
153	200	244	286	324

Πίνακας 5.1: Πίνακας εβδομαδιαίου βάρους ποντικιών

Βάση αυτών των δεδομένων στο Διάγραμμα 5.1 παραθέτουμε και παρατηρούμε την εξέλιξη του βάρους των ποντικιών στο πέρας των εβδομάδων.



Διάγραμμα 5.1: Εξέλιξη του βάρους στο πέρας των βδομάδων.

Γίνεται αντιληπτό ότι υπάρχει μια γραμμική εξάρτηση μεταξύ βάρους και χρόνου. Επίσης, παρατηρούμε ότι για κάθε ποντίκι έχουμε διαφορετικό αρχικό βάρος όπως και διαφορετική ανάπτυξη βάρους με αποτέλεσμα να χρησιμοποιήσουμε ένα ιεραρχικό μοντέλο έτσι ώστε να δώσουμε την δύνατοτητα στο μοντέλο που θα χρησιμοποιήσουμε να λάβει υπόψιν αυτή την πληροφορία. Αυτό το μοντέλο το οποίο θα εκφράσει την γραμμική εξάρτηση που προαναφέραμε στα πλαίσια ενός ιεραρχικού μοντέλου είναι το ακόλουθο.

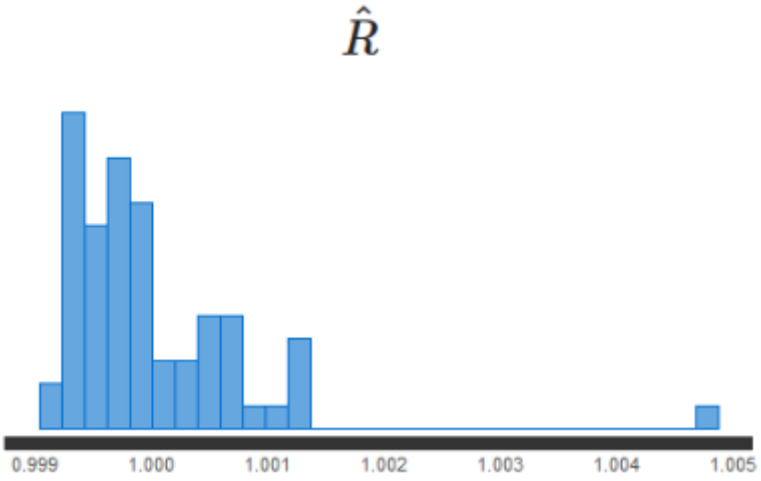
$$Y_{ij} \sim Normal(a_i + b_i(x_j - \bar{x}), \sigma_Y),$$
$$a_i \sim Normal(\mu_a, \sigma_a),$$
$$b_i \sim Normal(\mu_b, \sigma_b).$$

Με a_i να συμβολίζει την σταθερά του μοντέλου ενώ με b_i την κλίση του μοντέλου. Ακόμα, παρατηρούμε ότι κεντροποιούμε την μοντελοποίηση γύρο απο το $\bar{x} = 22$ που είναι η διάμεσος της μεταβλητής x που αντιστοιχεί στις μέρες. Για τις υπερ-παραμέτρους $\mu_a, \mu_b, \sigma_a, \sigma_b$ και την παράμετρο σ_Y θα χρησιμοποιήσουμε τις ακόλουθες πρότερες κατανομές.

$$\mu_a \sim Normal(0, 100),$$
$$\mu_b \sim Normal(0, 100),$$
$$\sigma_a \sim Inv - Gamma(0.001, 0.001),$$
$$\sigma_b \sim Inv - Gamma(0.001, 0.001),$$

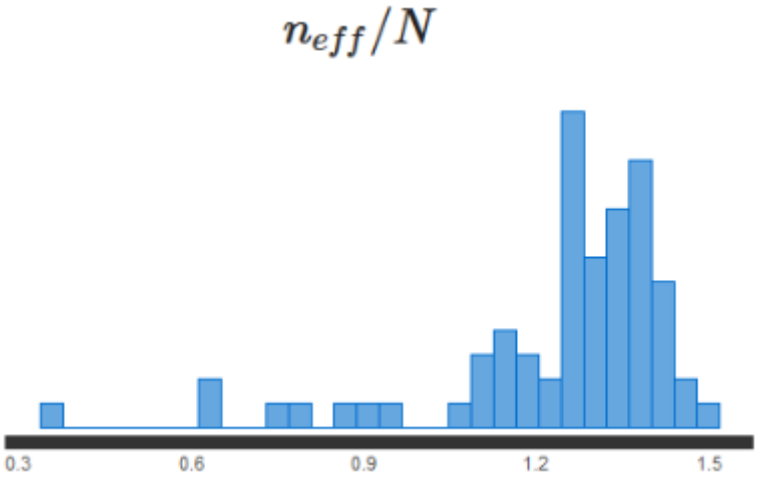
$$\sigma_Y \sim Inv - Gamma(0.001, 0.001).$$

Όπου η πρότερη κατανομή Inv-Gamma(Αντίστροφη Γάμμα) είναι συζυγής κατανομή της Κανονικής κατανομής. Στην συνέχεια θα τρέξουμε τον αλγόριθμο NUTS για 8000 επαναλήψεις όπου οι 4000 πρώτες θα χαρακτηριστούν ως burn-inμε χρήση 4 αλυσίδων. Αρχικά θα ελέγξουμε το $\hat{R}$ στατιστικό ώστε να διαπιστώσουμε ότι όλες οι αλυσίδες για κάθε παράμετρο εξερευνούν την ίδια περιοχή. Στο Διάγραμμα 5.2 παρουσιάζουμε την τιμή που παίρνει το στατιστικό $\hat{R}$ για κάθε παράμετρο. Όσο πιο κοντά στην τιμή 1 παίρνει τιμές το στατιστικό τόσο πιο συμπαγής και καλή ήταν η εξερεύνηση του χώρου. Επιπλέον, έλεγχος θα πρέπει να γίνει και για το αποτελεσματικό δείγμα που πήραμε απο τον



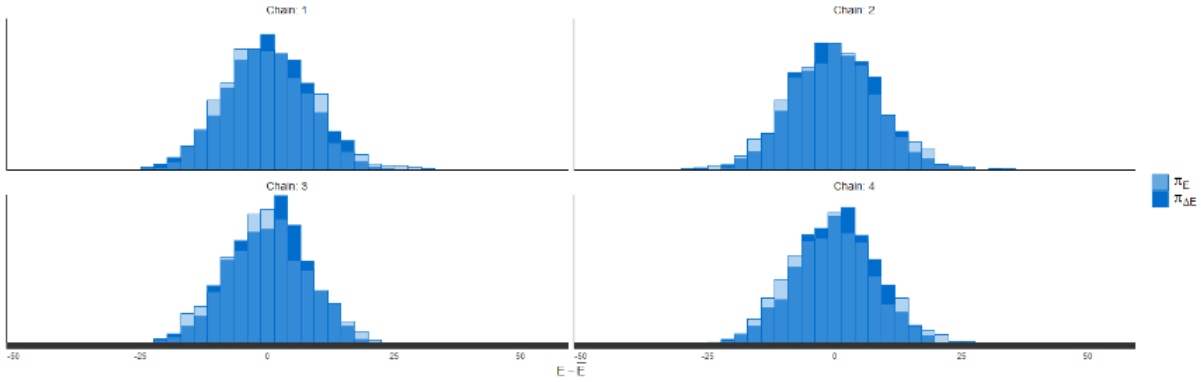
Διάγραμμα 5.2: Η κατανομή του $\hat{R}$ για κάθε παράμετρο.

αλγόριθμο NUTS.Στο Διάγραμμα 5.3 θα παρουσιάζουμε τον λόγο μεταξύ αποτελεσματικού δείγματος και του συνόλου των επαναλήψεων (εκτός των burn-in).Όσο παίρνουμε τιμές μεγαλύτερες του 1 τόσο περισσότερη πληροφορία πήραμε απο την κάθε επανάληψη για κάθε παράμετρο. Όπως έχουμε αναφέρει



Διάγραμμα 5.3: Η κατανομή του λόγου αποτελεσματικού δείγματος προς το σύνολο των επαναλήψεων (εκτός των burn-in).

στον αλγόριθμο HMC και αναπόφευκτα στον αλγόριθμο NUTS μεγάλης σημασίας είναι η επιλογή της κατανομής της ορμής p η οποία προσδιορίζει στην αποτελεσματικότητα εξερεύνησης των τροχιών. Στο Διάγραμμα 5.4 παραθέτουμε την σύγκριση των κατανομών των παραγόμενων επιπέδων ενέργειας $\pi_{\Delta E}$ με τα πραγματικά επίπεδα ενέργειας π_E. Καταλαβαίνουμε ότι όσο πιο κοντά βρίσκονται αυτές οι κατανομές τόσο καλύτερη και γρηγορότερη εξερεύνηση του χώρου της παραμέτρου ενδιαφέροντος θα

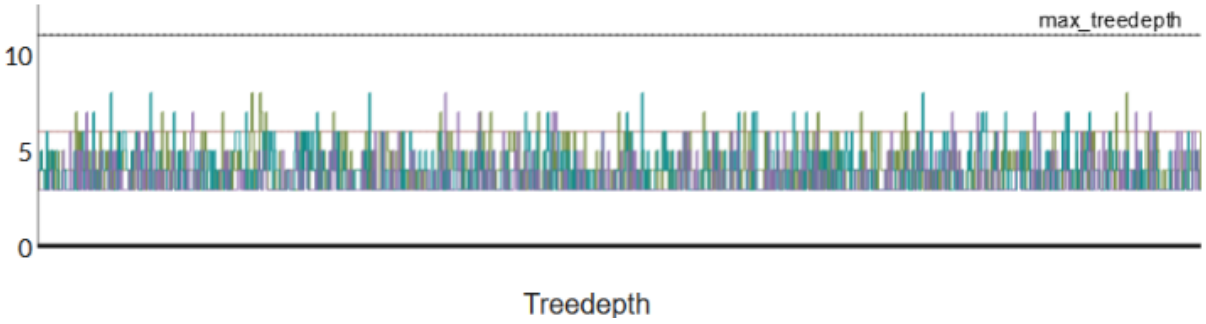


Διάγραμμα 5.4: Σύγκριση κατανομών επιπέδων ενέργειας.

έχουμε. Ακόμα, σε συνδυασμό με το προηγούμενο διάγραμμα παραθέτουμε τις τιμές του στατιστικού E-FBMI για κάθε αλυσίδα.

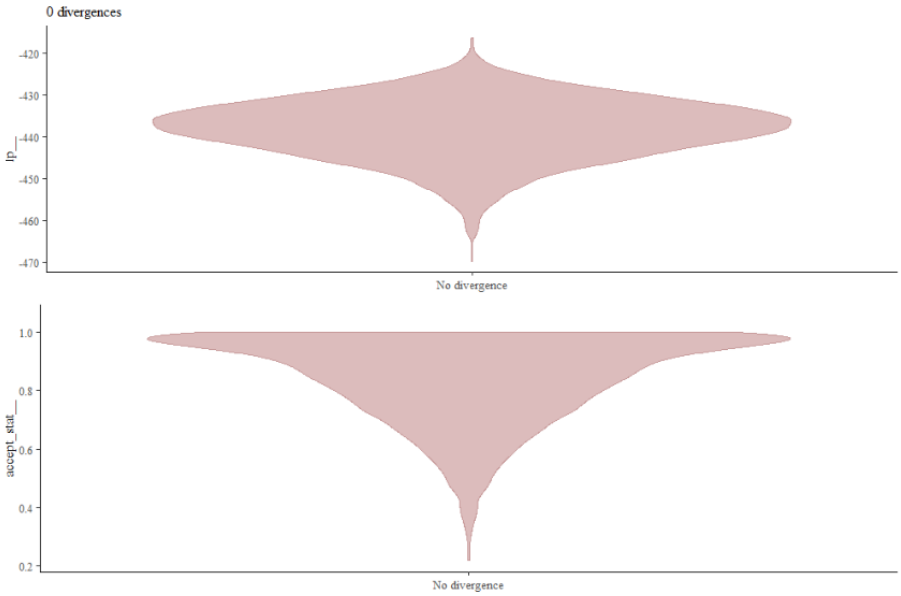
$E-FBMI_1$	$E-FBMI_2$	$E-FBMI_3$	$E-FBMI_4$
0.7769750	0.7623440	0.8748752	0.8338816

Παρατηρούμε ότι όλες οι τιμές του στατιστικού E-BFMI είναι μεγαλύτερες απο την εμπειρική τιμή 0.3 συνεπώς η κατανομή που χρησιμοποιήθηκε για την ορμή $\boldsymbol{p}$ είναι αποτελεσματική. Θα ελέγξουμε ακόμα και το βάθος του δέντρου ώστε να διαπιστώσουμε αν κάποια επανάληψη του αλγορίθμου τερμάτισε πρόωρα. Στο Διάγραμμα 5.5 παρουσιάζουμε για τις 4 αλυσίδες το βάθος του δέντρου το οποίο χρησιμοποιήθηκε για την εξερεύνηση κάθε τροχιάς. Απο προεπιλογή το φράγμα για το βάθος δέντρου που έχουμε προσδιορίσει είναι το 10, στην περίπτωση που το ξεπερνούσαμε θα έπρεπε να το αυξήσουμε ώστε να δώσουμε τον χρόνο στον αλγόριθμο να εξερευνήσει πλήρως την εκάστοτε τροχιά, θυσιάζοντας πάντα υπολογιστικούς πόρους.



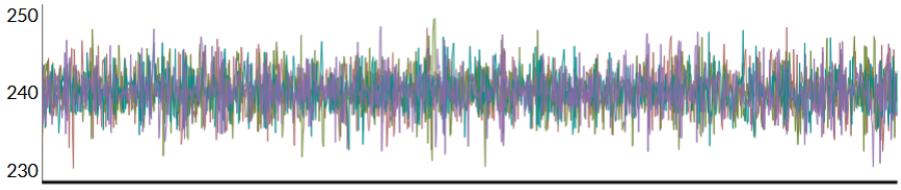
Διάγραμμα 5.5: Το βάθος δέντρου για κάθε επανάληψη για κάθε μια απο τις 4 αλυσίδες.

Στην συνέχεια κάνουμε έλεγχο για παρατηρήσεις οι οποίες μπορεί να απέκλιναν απο την προκαθορισμένη τροχιά τους. Στο Διάγραμμα 5.6 παρουσιάζουμε την κατανομή του λογαρίθμου της ύστερης κατανομής και την κατανομή της πιθανότητας αποδοχής κάθε παρατήρησης. Αρχικά γίνεται κατανοη-

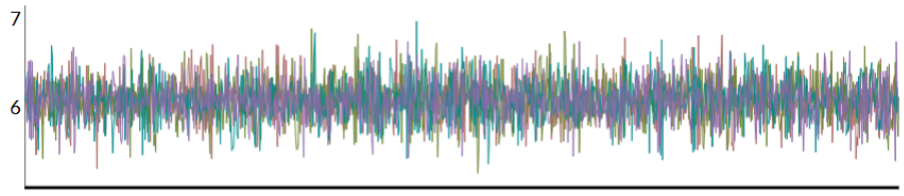


Διάγραμμα 5.6: Κατανομές του λογαρίθμου της ύστερης κατανομής και της πιθανότητας αποδοχής των παρατηρήσεων.

τό ότι η κατανομή του λογαρίθμου της ύστερης κατανομής δεν συγκεντρώνεται σε ένα συγκεκριμένο σημείο και απλώνεται σε όλο το φάσμα των τιμών της. Συνεπώς ο αλγόριθμος δεν προσκολλάται σε συγκεκριμένες παρατηρήσεις στην προσπάθεια του να τις εξερευνήσει διότι σε εκείνη την περίπτωση θα βλέπαμε ένα μεγάλο ποσοστό των τιμών του λογαρίθμου της ύστερης να είναι μαζεμένες σε συγκεκριμένα σημεία. Δεύτερον, παρατηρώντας την κατανομή της πιθανότητας αποδοχής αντιλαμβανόμαστε ότι δεν παίρνουμε μηδενικές τιμές δηλαδή δεν απορρίπτουμε κάποια υποψήφια παρατήρηση. Στην περίπτωση που παίρναμε μεγάλο πλήθος τιμών της πιθανότητας αποδοχής κοντά στο μηδέν ή και στο μηδέν θα αντιλαμβανόμασταν ότι ο αλγόριθμος στην προσπαθεί να εξερευνήσει κάποια περιοχή αλλά αποτυγχάνει με αποτέλεσμα να έχουμε αποκλίνουσες παρατηρήσεις. Τέλος, παραθέτουμε τα traceplot των παραμέτρων a_1 και b_1 για εξοικονόμηση χώρου ώστε να διαπιστώσουμε ότι σε συνδυασμό με όλα τα προαναφερθέντα έχει επιτευχθεί η σύγκλιση. Στα Διάγραμματα 5.7 και 5.8 παρουσιάζουμε τις τιμές των παραμέτρων a_1 και b_1 αντίστοιχα για κάθε επανάληψη.

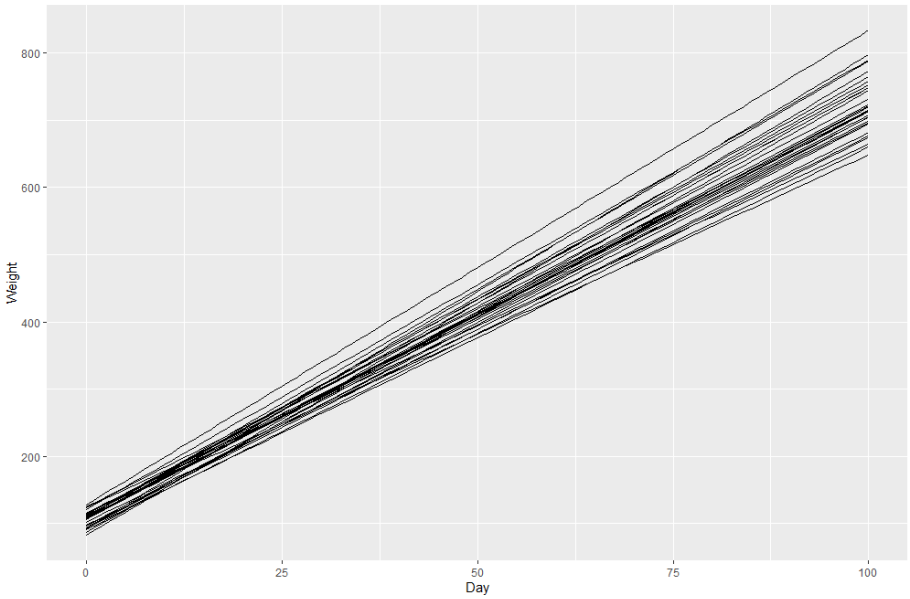


Διάγραμμα 5.7: Τιμή της παραμέτρου a_1 σε κάθε επανάληψη για κάθε αλυσίδα.



Διάγραμμα 5.8: Τιμή της παραμέτρου b_1 σε κάθε επανάληψη για κάθε αλυσίδα.

Συνεπώς, βασιζόμενοι σε όλους τους προηγούμενους διαγνωστικούς ελέγχους μπορούμε να υποθέσουμε ότι ο αλγόριθμος HMC και κατά συνέπεια ο αλγόριθμος NUTS έχουν συγκλίνει χωρίς να υπάρξει πρόβλημα. Τέλος, δίνουμε προβλέψεις για τα βάρη των ποντικιών βασιζόμενοι στις εκτιμήσεις των παραμέτρων που πήραμε απο τον αλγόριθμο NUTS.Στο Διάγραμμα 5.9 παραθέτουμε τις προβλέψεις των βαρών των ποντικιών.



Διάγραμμα 5.9: Προβλέψεις του βάρους των ποντικιών κάνοντας χρήση των εκτιμήσεων των παραμέτρων.

5.2 Ανάλυση Εισαγωγικών Τέστ για Οχτώ Πανεπιστήμια

Για οχτώ πανεπιστήμια μετρήσαμε την απόδοση των μαθητών σε ένα εισαγωγικό τέστ. Στην συνέχεια υπολογίσαμε την μέση απόδοση και την διακύμανση του κάθε πανεπιστημίου και την συμβολίσαμε ώς y_j και σ_j^2 αντίστοιχα, για $j = 1, ..., 8$. Στον ακόλουθο Πίνακα 5.2 παρουσιάζουμε τα δεδομένα που αφορούν την απόδοση των φοιτητών στο εισαγωγικό τέστ.

Σχολεία	Εκτίμηση Μέσης Απόδοσης y_j	Τυπικό Σφάλμα της Μέσης Απόδοσης σ_j
A	28	15
B	8	10
Γ	-3	16
Δ	7	11
E	-1	9
Z	1	11
H	18	10
Θ	12	18

Πίνακας 5.2: Πίνακας δεδομένων της απόδοσης των φοιτητών για τα 8 πανεπιστήμια.

Για αρχή παρατηρούμε ότι βασιζόμενοι μόνο στα y_j μπορούμε να διαφοροποιήσουμε την απόδοση των φοιτητών για το κάθε πανεπιστήμιο. Παρόλα αυτά συμπεριλαμβάνοντας και τις τυπικές αποκλίσεις σ_j γίνεται αντιληπτό ότι αυτές οι εκτιμήσεις y_j έχουν μεγάλη μεταβλητότητα. Επίσης, γνωρίζουμε ότι τα διαστήματα εμπιστοσύνης είναι ισοδύναμα με ελέγχους υποθέσεων συνεπώς λόγω της μεγάλης διακύμανσης τα διαστήματα εμπιστοσύνης θα αλληλοκαλύπτονται και επακόλουθα οι έλεγχοι υποθέσεων θα βγάζουν παραπλανητικά αποτελέσματα. Για αυτό τον λόγο θα κάνουμε χρήση ιεραρχικών μοντέλων, έτσι ώστε να μπορέσουμε να ελέγξουμε αυτή την αβεβαιότητα που προκαλείται λόγω της μεγάλης διακύμανσης.

$$\mu \sim \mathcal{N}(0, 5),$$
$$\tau \sim Half - Cauchy(0, 5),$$
$$\theta_n \sim \mathcal{N}(\mu, \tau),$$
$$y_n \sim \mathcal{N}(\theta_n, \sigma_n).$$

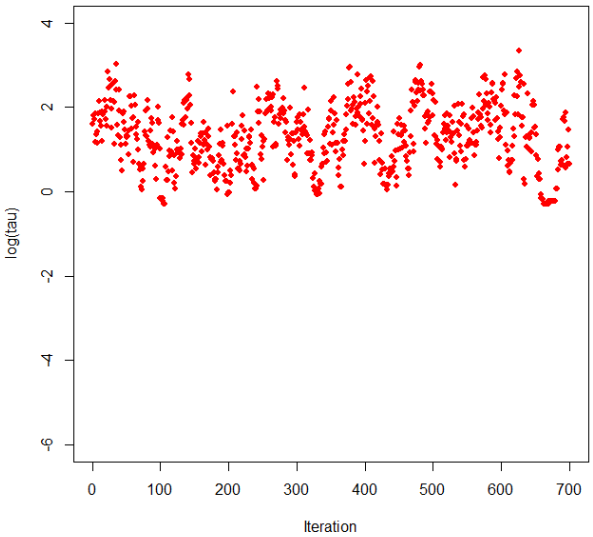
Ουσιαστικά τα y_j προέρχονται από την κανονική κατανομή με διακύμανση αυτή που έχει υπολογιστεί, σ_j^2. Παρόλα αυτά πιστεύουμε ότι οι μέσοι των y_j διαφέρουν (μιας και βλέπουμε μεγάλες αποκλίσεις μεταξύ των y_i που όμως εξαφανίζονται στην περίπτωση που χρησιμοποιήσουμε για παράδειγμα διαστήματα εμπιστοσύνης) και για να ποσοτικοποιήσουμε αυτή την μεταβλητότητα θα πρέπει να ορίσουμε μια επιπλέον κανονική κατανομή για το θ_j. Τέλος, για τις υπερ-παραμέτρους χρησιμοποιούμε μια ομοιόμορφη μη πληροφοριακή κατανομή. Παρατηρούμε ότι για την παράμετρο ενδιαφέροντος θ_j χρησιμοποιούμε έναν μη κεντροποιημένο μετασχηματισμό. Ένας πολύ γενικός κανόνας είναι ότι στην περίπτωση που κάνουμε χρήση ιεραρχικών μοντέλων με πρότερες κατανομές μη πληροφοριακές χρησιμοποιούμε μη κεντροποιημένους μετασχηματισμούς ενώ σε κάθε άλλη περίπτωση κεντροποιημένους (Michael Betancourt 2013).Για αρχή θα εφαρμόσουμε την ακόλουθη μοντελοποίηση,

```
data{
   int<lower=0> J;
   real y[J];
   real<lower=0> sigma[J];
}
parameters{
   real mu;
   real<lower=0> tau;
   real theta[J];
}
model{
   mu~ norma(0,5);
   tau~ cauchy(0,5);
   theta~ norma(mu,tau);
   y~ norma(theta,sigma);
}
```

η οποία ονομάζεται κεντροποιημένη διότι η παράμετρος θ είναι κεντραρισμένη γύρω από την πρότερη κατανομή με μέσο μ. Στην συνέχεια τρέχοντας μια αλυσίδα με 1200 επαναλήψεις και 500 επαναλήψεις και burn-in παίρνουμε τα ακόλουθα αποτελέσματα.

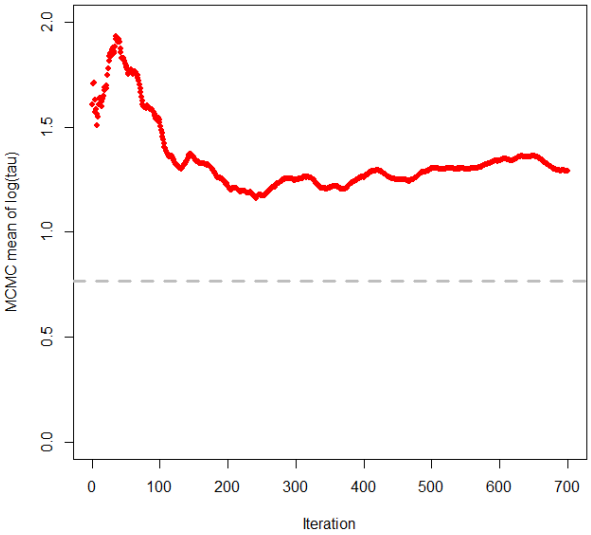
```
Inference for Stan model: 4e59963cd0b0547acb5a69ba59ccde54.
1 chains, each with iter=1200; warmup=500; thin=1;
post-warmup draws per chain=700, total post-warmup draws=700.

          mean se_mean   sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
mu        4.50    0.47 3.59  -1.54   1.62   4.32   7.14  11.59    58 1.05
tau       4.79    0.51 3.72   0.80   2.11   3.74   6.32  14.36    54 1.02
theta[1]  7.41    0.65 7.10  -2.86   2.36   6.44  10.31  25.97   118 1.00
theta[2]  5.39    0.36 5.77  -5.53   1.50   5.11   8.96  17.70   255 1.00
theta[3]  3.86    0.47 5.92  -8.28   0.42   4.12   7.70  14.31   159 1.02
theta[4]  5.22    0.45 5.20  -4.36   1.82   4.92   8.34  15.97   135 1.02
theta[5]  3.36    0.48 5.22  -7.93   0.33   3.45   6.82  13.26   118 1.03
theta[6]  3.92    0.45 5.68  -8.40   0.53   3.91   7.78  14.49   163 1.02
theta[7]  6.98    0.47 5.68  -2.63   2.75   6.78  10.50  20.26   148 1.01
theta[8]  5.23    0.53 6.19  -7.15   1.58   4.58   8.94  17.91   136 1.02
lp__    -16.92    0.94 5.82 -28.31 -20.86 -16.95 -12.68  -5.89    39 1.01

Samples were drawn using NUTS(diag_e) at Wed Apr 24 19:15:36 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Τα $\hat{R}$ φαίνονται αρκετά ικανοποιητικά αφού είναι κοντά στο 1 (γνωρίζουμε όμως από τη κατασκευή του $\hat{R}$ ότι για να είναι συνεπές ως προς τη συμπερασματολογία θα πρέπει να τρέξουμε περισσότερες από μια αλυσίδες), ενώ το αποτελεσματικό μέγεθος δείγματος είναι μικρό. Αυτό μας υποδηλώνει ότι σε σχέση με τις επαναλήψεις που έχουμε πάρει έχουμε πολύ λίγες ανεξάρτητες προσομοιωμένες τιμές. Τώρα έστω ότι θέλουμε να μελετήσουμε την παράμετρο τ. Γνωρίζουμε ότι παίρνει θετικές τιμές, συνεπώς θα ήταν εύλογο να χειριστούμε τον λογάριθμο της, για να μπορέσουμε να διακρίνουμε καλύτερα μικρές τιμές της τ. Στο Διάγραμμα 5.10 παρουσιάζουμε την τιμή του λογαρίθμου της παραμέτρου τ για κάθε επανάληψη του αλγορίθμου.



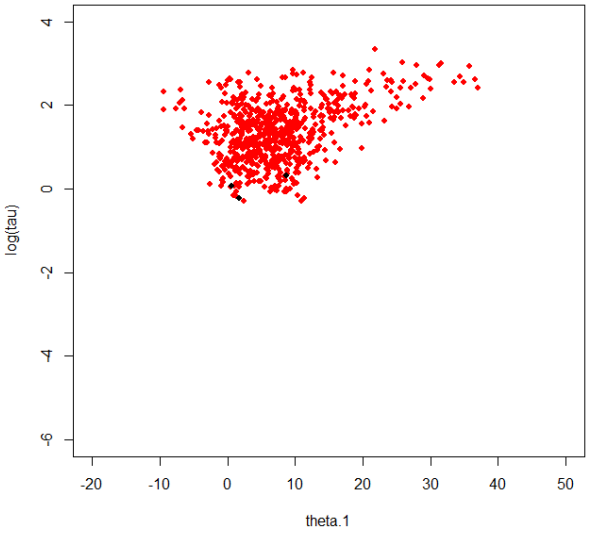
Διάγραμμα 5.10: Γράφημα των προσομοιωμένων τιμών για την μετασχηματισμένη μεταβλητή $log(\tau)$.

Το traceplot της $log(\tau)$ φαίνεται ικανοποιητικό ακόμα και για αυτές τις λίγες προσομοιωμένες τιμές μιας και δεν έχουμε προβληματικές περιοχές στις οποίες ο αλγόριθμος προσκολλάται. Αντί αυτού παρατηρούμε ότι ο ο αλγόριθμος εξερευνά και μικρές και μεγάλες τιμές. Ας δούμε τώρα πως συμπεριφέρεται ο εργοδικός της μέσος. Στο Διάγραμμα 5.11 παρουσιάζουμε την εξέλιξη του εργοδικού μέσου σε όλο το εύρος των επαναλήψεων.



Διάγραμμα 5.11: Γράφημα το οποίο μας παρουσιάζει την διαμόρφωση του εργοδικού μέσου σε κάθε επανάληψη.

Παρατηρούμε πρώτον ότι ο εργοδικός μέσος είναι πολύ μεγαλύτερος της πραγματικής τιμής της $log(\tau)$ και δεύτερον ότι δεν έχει καταφέρει να συγκλίνει σε οποιαδήποτε τιμή συνεπώς είναι ένδειξη ότι η δειγματοληψία μας έχει πρόβλημα. Επίσης, στις περιπτώσεις που προκύπτουν τέτοιου είδους προβλήματα συνιστάται να ελέγχουμε το πλήθος των παρατηρήσεων οι οποίες χαρακτηρίστηκαν ως αποκλίνουσες (δηλαδή που απέκλιναν πολύ από την πραγματική τροχιά). Θα αντιληφθούμε ότι μόνο το 0.5% από το σύνολο των παρατηρήσεων είναι αποκλίνουσες δηλαδή μόνο 4 στις 700. Παρόλα αυτά τέτοιες παρατηρήσεις μπορούν να εισάγουν μεροληψία στην εκτίμηση των παραμέτρων ενδιαφέροντος ειδικότερα στην περίπτωση μας που έχουμε λίγες προσομοιωμένες τιμές με χρήση μόνο μιας αλυσίδας. Ας δούμε τώρα σε ποια σημεία είναι μαζεμένες αυτές οι αποκλίνουσες παρατηρήσεις στον χώρο της παραμέτρου θ_n. Στο Διάγραμμα 5.12 παρουσιάζουμε την σχέση μεταξύ των παραμέτρων $log(\tau)$ και θ_1.



Διάγραμμα 5.12: Γράφημα μεταξύ των τυχαίων μεταβλητών $log(\tau)$ και θ_1 για την εξερεύνηση divergent παρατηρήσεων.

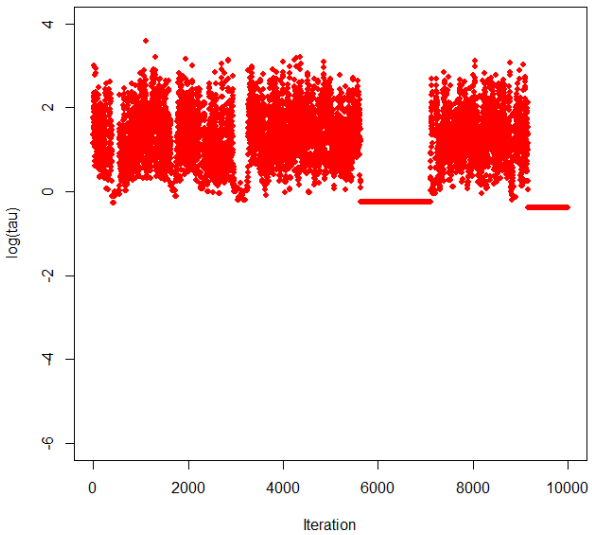
Παρατηρούμε ότι οι αποκλίνουσες παρατηρήσεις τείνουν να συγκεντρώνονται σε μικρές της τ, που κατά συνέπεια φέρνει κοντά όλες τις εκτιμήσεις των θ_n. Όμως η αλυσίδα αποκλίνει πριν ακόμα εξερευνήσει την περιοχή που περιέχει μικρές τιμές της τ. Αυτό το φαινόμενο είναι αποτέλεσμα της αδυναμίας

της αλυσίδας να εξερευνήσει περιοχές μεγάλης καμπυλότητας. Συνηθέστερα αυτό το πρόβλημα παρατηρείται σε ιεραρχικά μοντέλα.

Αρχικά, καλό θα ήταν να αυξήσουμε τις επαναλήψεις ώστε να μπορέσουμε να δούμε πώς αυτό το πρόβλημα χειροτερεύει και γίνεται πιο κατανοητό με την αύξηση του πλήθους των παρατηρήσεων. Έστω ότι τρέχουμε 11000 επαναλήψεις με burn-in 1000 επαναλήψεις, τότε θα πάρουμε τα ακόλουθα αποτελέσματα:

```
Inference for Stan model: 4e59963cd0b0547acb5a69ba59ccde54.
1 chains, each with iter=11000; warmup=1000; thin=1;
post-warmup draws per chain=10000, total post-warmup draws=10000.

          mean se_mean   sd   2.5%    25%    50%    75% 97.5% n_eff Rhat
mu        4.21    0.29 3.05  -1.70   2.25   4.04   6.04 10.45   114 1.01
tau       3.36    0.45 3.14   0.68   0.88   2.33   4.68 11.58    48 1.07
theta[1]  5.86    0.52 5.28  -2.98   2.00   5.46   7.99 19.31   105 1.02
theta[2]  4.85    0.11 4.38  -4.23   2.90   4.37   6.92 14.34  1482 1.00
theta[3]  3.84    0.32 4.99  -7.31   1.79   3.70   7.00 13.52   236 1.00
theta[4]  4.40    0.43 4.59  -4.70   1.26   4.25   6.91 14.20   114 1.02
theta[5]  3.67    0.16 4.38  -6.85   1.74   3.76   5.75 11.90   737 1.00
theta[6]  4.14    0.24 4.52  -6.36   2.18   3.71   7.39 12.69   348 1.00
theta[7]  5.85    0.53 4.84  -2.03   2.19   5.61   8.30 17.84    84 1.04
theta[8]  4.93    0.13 4.91  -5.55   2.82   4.38   7.03 15.77  1449 1.00
lp__    -14.15    1.07 6.10 -26.17 -18.74 -13.43 -10.84 -4.32    32 1.07

Samples were drawn using NUTS(diag_e) at Wed Apr 24 20:22:29 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

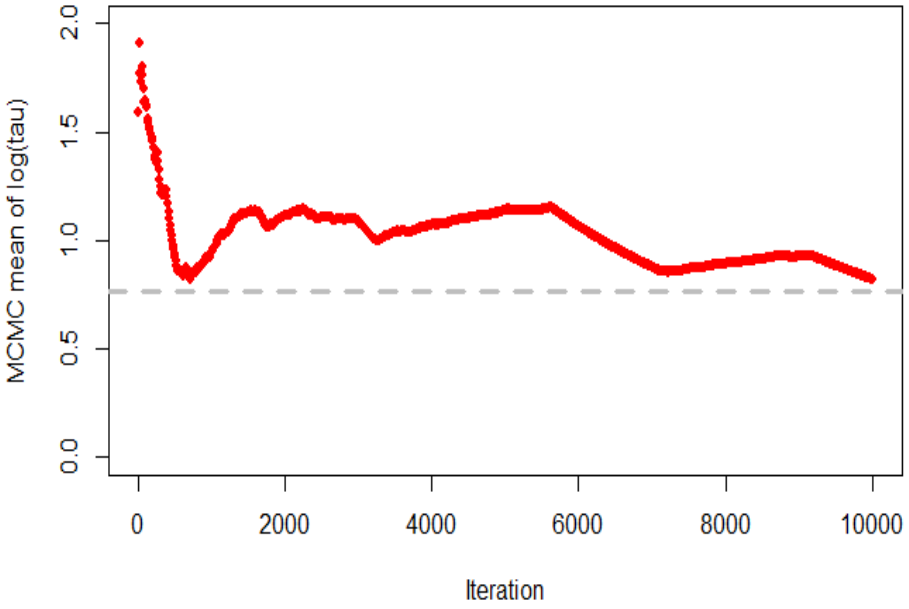
Παρατηρούμε ότι η δειγματοληψία έχει χειροτερέψει δραματικά διότι το αποτελεσματικό μέγεθος δείγματος είναι κατά πολύ μικρότερο του πλήθους των επαναλήψεων, δηλαδή η πληροφορία που παίρνουμε είναι πολύ αργή. Επίσης οι αποκλίνουσες παρατηρήσεις από 4 έχουν γίνει 1157. Αρκεί να ελέγξουμε το traceplot της $log(\tau)$ και θα διαπιστώσουμε ότι κάτι δεν πάει σωστά. Στο Διάγραμμα 5.13 παρουσιάζουμε τις τιμές της παραμέτρου $log(\tau)$ για όλο το εύρος των επαναλήψεων.



Διάγραμμα 5.13: Γράφημα των προσομοιωμένων τιμών για την μετασχηματισμένη μεταβλητή $log(\tau)$ στην περίπτωση που έχουμε προσομοιώσει 10000 επαναλήψεις.

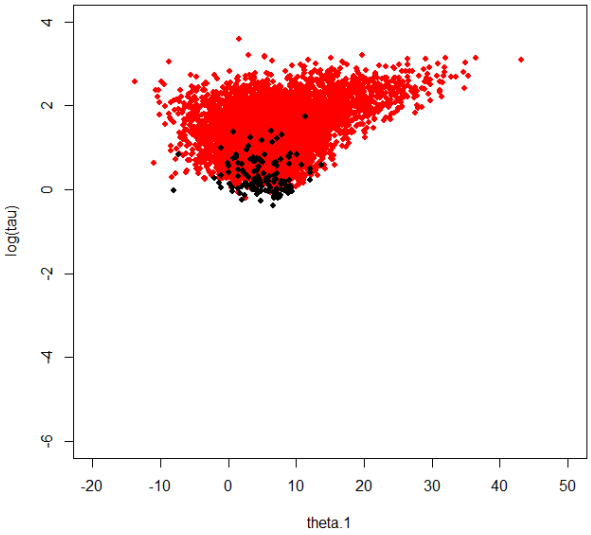
Παρατηρούμε ότι στην προσπάθεια του ο αλγόριθμος να εξερευνήσει τις περιοχές μεγάλης καμπυλότητας, προσκολλάται στο σύνορο τους με αποτέλεσμα να έχουμε για μεγάλα διαστήματα πα-

ρατηρήσεις από την ίδια περιοχή δημιουργώντας μεροληψία στους εκτιμητές. Στο Διάγραμμα 5.14 παραθέτουμε τον εργοδικό μέσο για όλο το εύρος των παρατηρήσεων που όμως αυτή την φορά θα αφορά τον αλγόριθμο με τις 10000 επαναλήψεις.



Διάγραμμα 5.14: Γράφημα το οποίο μας παρουσιάζει την διαμόρφωση του εργοδικού μέσου σε κάθε επανάληψη.

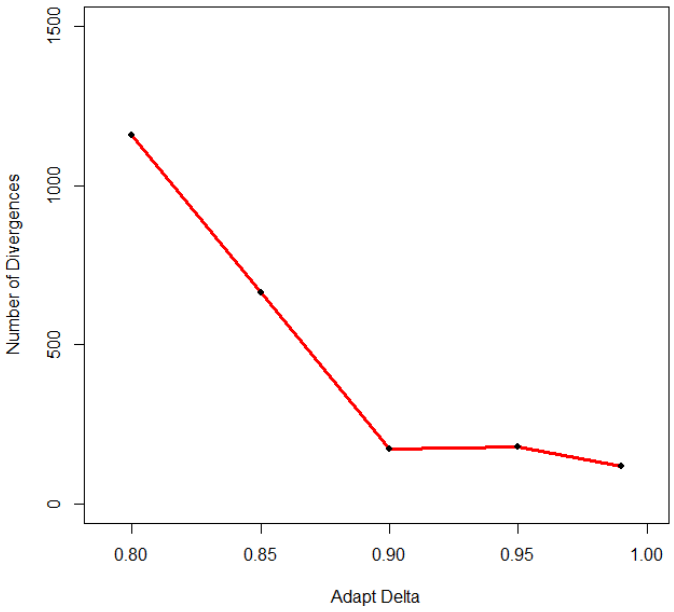
Ακόμα, επειδή αυτή την φορά υλοποιήσαμε τον αλγόριθμο για πολύ μεγαλύτερο πλήθος επαναλήψεων, βλέπουμε ότι ο εργοδικός μέσος προσπαθεί να πλησιάσει ανεπιτυχώς την πραγματική τιμή. Αυτή είναι η προσπάθεια σε κάθε επανάληψη να διορθωθεί η μεροληψία με το να δημιουργηθούν παρατηρήσεις κοντά στην περιοχή μεγάλης καμπυλότητας. Τέλος, οι αποκλίνουσες παρατηρήσεις είναι μαζεμένες στην περιοχή μεγάλης καμπυλότητας. Στο Διάγραμμα 5.15 παραθέτουμε την σχέση μεταξύ των παραμέτρων $log(\tau)$ και θ_1 μαζί και τις παρατηρήσεις οι οποίες χαρακτηρίστηκαν ώς αποκλίνουσες.



Διάγραμμα 5.15: Γράφημα μεταξύ των τυχαίων μεταβλητών $log(\tau)$ και θ_1 για την εξερεύνηση αποκλίνουσων παρατηρήσεων.

Ένας τρόπος για να μετριάσουμε αυτό το φαινόμενο είναι μέσω της μείωσης του βήματος ϵ έτσι ώστε να κάνουμε πιο ακριβείς μετακινήσεις, χωρίς να μας ενδιαφέρει ο παραπάνω χρόνος που θα χρεια-

στούμε για να υλοποιήσουμε όλες τις επαναλήψεις. Στο επόμενο διάγραμμα εύκολα αντιλαμβανόμαστε ότι όσο μικραίνουμε το ϵ τόσο λιγότερες αποκλίνουσες παρατηρήσεις έχουμε. Παρόλα αυτά από ένα σημείο και μετά όσο και να μικρύνουμε το βήμα δεν θα επιφέρει κάποια βελτίωση. Να σημειώσουμε ότι ο δείκτης delta αντιστοιχεί στο ποσοστό αποδοχής το οποίο είναι αντιστρόφως ανάλογο του βήματος, ϵ, δηλαδή όσο μεγαλύτερο ποσοστό αποδοχής θέλουμε να έχουμε τόσο πιο μικρό βήμα χρειαζόμαστε. Στο Διάγραμμα 5.16 παραθέτουμε την σχέση μεταξύ μεγέθους βήματος και αποκλίνουσων παρατηρήσεων.



Διάγραμμα 5.16: Γράφημα το οποίο παρουσιάζει το πλήθος των αποκλίνουσων παρατηρήσεων σε σχέση με το μέγεθος του βήματος.

Παρατηρούμε ότι ακόμα και με την μείωση του βήματος ϵ υπάρχουν ακόμα αποκλίνουσες παρατηρήσεις και ο χώρος δεν εξερευνάται πλήρως με αποτέλεσμα να έχουμε μεροληψία στην συμπερασματολογία μας. Επίσης όσον αφορά το ενδιαφέρον μας για την μεταβλητή $log(\tau)$, μπορούμε να διαπιστώσουμε ότι με μικρότερο βήμα, καταφέρνει να εξερευνήσει καλύτερα την περιοχή υψηλής καμπυλότητας και να προσομοιώσουμε αρκετό πλήθος παρατηρήσεων στον αρνητικό ημιάξονα. Στο Διάγραμμα 5.17 παρουσιάζουμε την σύγκριση μεταξύ των μεθόδων που χρησιμοποιήσαμε κάνοντας χρήση ιστογραμμάτων

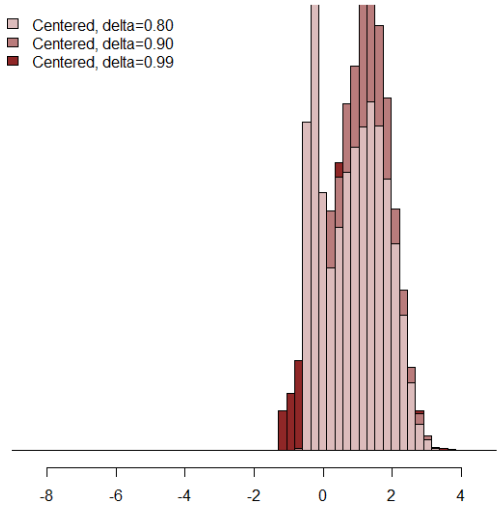
Ακόμα, όσο πιο βαθιά εξερευνούμε τις περιοχές υψηλής καμπυλότητας δηλαδή μικραίνοντας το βήμα παρατηρούμε ότι ακόμα και τότε έχουμε αποκλίνουσες παρατηρήσεις διότι η καμπυλότητα της περιοχής γίνεται όλο και πιο ακραία. Στο Διάγραμμα 5.18 παρουσιάζουμε την σχέση μεταξύ των παραμέτρων $log(\tau)$ και θ_1 μαζί με το πλήθος των αποκλίνουσων παρατηρήσεων.

Τέλος, μπορούμε να δούμε εύκολα αυτό που προαναφέραμε ότι με μικρότερο βήμα μπορούμε να εξερευνήσουμε καλύτερα τις περιοχές μεγάλης καμπυλότητας. Στο Διάγραμμα 5.19 παραθέτουμε την διαφορά των δύο μεθόδων όσο αναφορά τον χώρο που εξερευνούν.

Παρόλα αυτά ακόμα και με την μείωση του βήματος η μεροληψία υπάρχει ακόμα. Όπως προαναφέραμε ακόμα και μετά από την αλλαγή του ϵ εξακολουθούν να υπάρχουν αποκλίνουσες παρατηρήσεις που εισάγουν μεροληψία στις εκτιμήσεις και δημιουργούν πρόβλημα στην γεωμετρική εργοδικότητα. Για τον λόγο αυτό θα χρησιμοποιήσουμε την ακόλουθη μη κεντροποιημένη παραμετροποίηση για την θ_n :

$$\bar{\theta}_n \sim \mathcal{N}(0,1),$$
$$\theta_n = \mu + \tau \cdot \bar{\theta}_n.$$

Από τον τρόπο που ορίσαμε τις παραμέτρους θα πρέπει να περιμένουμε διαφορετική posterior κατανομή μιας και τώρα θα παίρνουμε δείγμα από την κατανομή της $\bar{\theta}_n$ για να υπολογίσουμε την θ_n. Θα υλοποιήσουμε μια αλυσίδα με 11000 επαναλήψεις (όπως και στην τελευταία προσπάθεια προηγουμένως) και 1000 επαναλήψεις warmup .



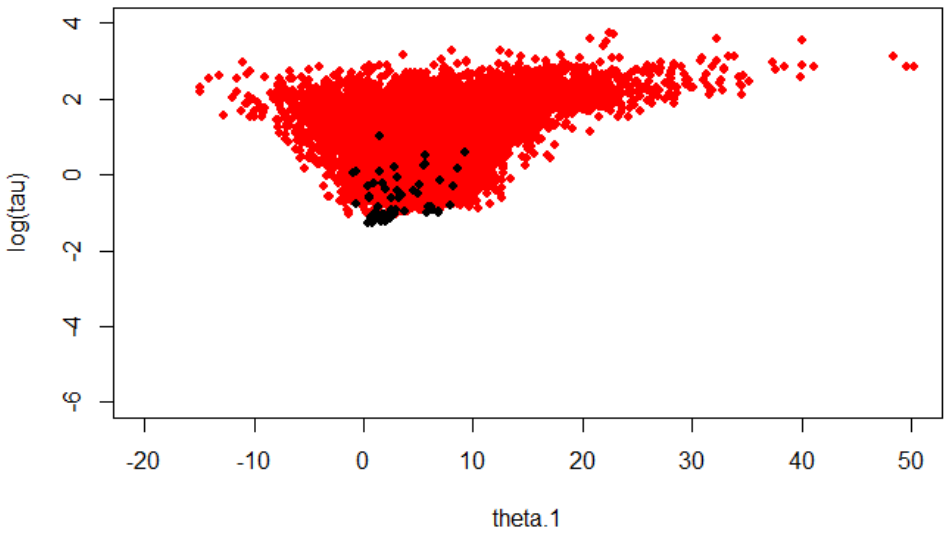
Διάγραμμα 5.17: Ιστόγραμμα της τυχαίας μεταβλητής $log(\tau)$ για διάφορες τιμές μεγέθους βήματος.

Παρατηρούμε ότι έχουμε δραματική βελτιστοποίηση του αποτελεσματικού μεγέθους δείγματος το οποίο μας πληροφορεί ότι εξερευνούμε τον χώρο πιο αποτελεσματικά. Επίσης το traceplot της $log(\tau)$ δεν έχει πλέον τα παθολογικά σημεία που εντοπίσαμε προηγουμένως. Στο Διάγραμμα 5.20 παρουσιάζουμε τις τιμές της παραμέτρου $log(\tau)$ σε κάθε επανάληψη και την βελτίωση του με την χρήση της μή κεντροποιημένης παραμετροποίησης.

Ακόμα να σημειώσουμε ότι σε αυτή την περίπτωση δεν υπάρχουν αποκλίνουσες παρατηρήσεις οι οποίες επηρεάζουν την εργοδικότητα του αλγορίθμου και πλέον ο αλγόριθμος κατάφερε να εξερευνήσει την περιοχή μεγάλης καμπυλότητας. Στο Διάγραμμα 5.21 παραθέτουμε τον χώρο που εξερευνεί ο αλγόριθμος και την αποτελεσματικότητα στο να το πετυχαίνει.

Ας δούμε επίσης πόσο βαθιά κατάφερε να εξερευνήσει η κάθε προσπάθεια την περιοχή μεγάλης καμπυλότητας. Στο Διάγραμμα 5.22 παραθέτουμε την σύγκριση μεταξύ των δύο μεθόδων που χρησιμοποιήσαμε και παρατηρούμε την διαφορά τους στην ικανότητα να εξερευνούν περιοχές μεγάλης καμπυλότητας.

Ακόμα, παρατηρούμε ότι μέσω της μη κεντροποιημένης παραμετροποίησης καταφέρνουμε να συγκλίνουμε προς την πραγματική τιμή της $log(\tau)$ και βλέπουμε ότι ο εργοδικός μέσος ασυμπτωτικά συγκλίνει στην πραγματική τιμή.



Διάγραμμα 5.18: Γράφημα μεταξύ των τυχαίων μεταβλητών $log(\tau)$ και θ_1 για την εξερεύνηση αποκλινουσών παρατηρήσεων.

Τέλος, καταλαβαίνουμε ότι παρόλο που μπορεί απο τα traceplot να μην διακρίνουμε μεγάλα προβλήματα στην σύγκλιση των αλυσίδων, ας σκεφτούμε ότι το $\hat{R}$ ήταν 1, πάντα θα πρέπει να ελέγχουμε τις περιοχές μεγάλης καμπυλότητας και την ύπαρξη αποκλινουσών παρατηρήσεων διότι εισάγουν μεροληψία στην μετέπειτα συμπερασματολογία μας.

5.3 Μεικτά Μοντέλα

Γνωρίζουμε ότι τα mixture models είναι πολύ χρήσιμα εργαλεία στις περιπτώσεις που κάποια παρατήρηση y μπορεί να παραχθεί από μια διαδικασία γέννησης δεδομένων ενός συνόλου K διαδικασιών. Για την μοντελοποίηση ενός τέτοιου μοντέλου χρησιμοποιούμε αρχικά μια μεταβλητή z η οποία μας υποδηλώνει από ποια παραγωγική διαδικασία επιλέχτηκε η παρατήρηση y,

$$z \in \{1, 2, ..., K\}.$$

Επίσης, η συνάρτηση πιθανοφάνειας του μοντέλου θα εκφράζεται ώς

$$\pi(y|\boldsymbol{\alpha}, z) = \pi_z(y|\boldsymbol{\alpha}),$$

με $\boldsymbol{\alpha} = (a_1, ..., a_K)$ οι παράμετροι της κάθε διαδικασίας γέννησης δεδομένων. Κάθε διαδικασία έχει μια πιθανότητα να προκύψει θ_k για την οποία ισχύει

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_K),\ 0 \leq \theta_k \leq 1,\ \sum_{i=1}^{K} \theta_k = 1,$$

έχοντας λοιπόν κατανομή για την μεταβλητή z

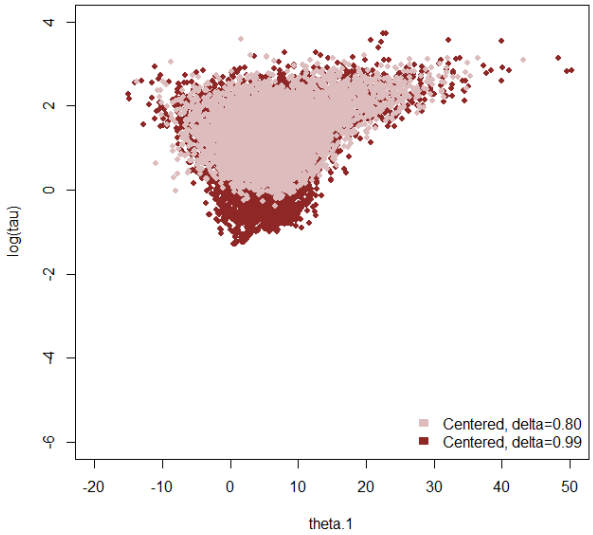
$$\pi(z|\boldsymbol{\theta}) = \theta_k.$$

Ακόμα η από κοινού συνάρτηση πιθανοφάνειας για τα y και z θα είναι

$$\pi(y, z|\boldsymbol{a}, \boldsymbol{\theta}) = \pi(y|\boldsymbol{a}, z)\pi(z|\boldsymbol{\theta}) = \pi_z(y|a_z)\theta_z.$$

Αθροίζοντας λοιπόν ως προς z βλέπουμε ότι η συνάρτηση πιθανοφάνειας του μοντέλου είναι ο γραμμικός συνδυασμός των K διαδικασιών γέννησης των δεδομένων.

$$\pi(y|\boldsymbol{a}, \boldsymbol{\theta}) = \sum_{z} \pi(y, z|\boldsymbol{a}, \boldsymbol{\theta})$$



Διάγραμμα 5.19: Σύγκριση μεταξύ του χώρου που εξερευνούν οι δύο μέθοδοι.

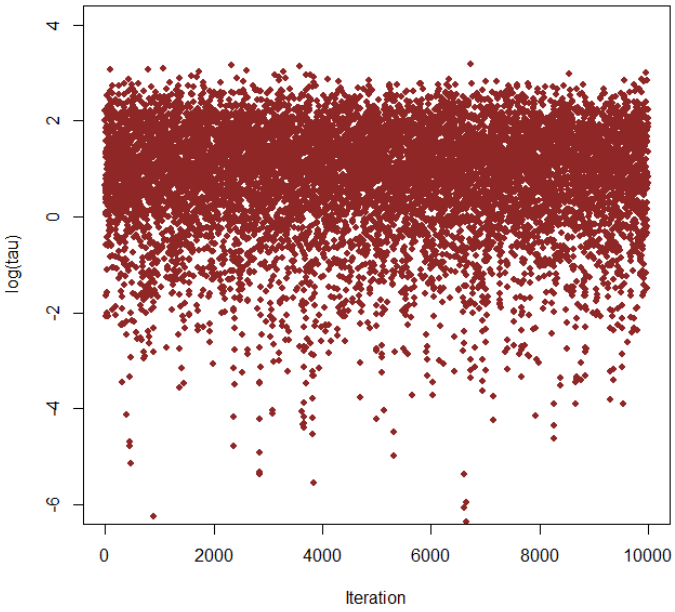
```
Inference for Stan model: 184c5e48db9c711e78f8d084fa104491.
1 chains, each with iter=11000; warmup=1000; thin=1;
post-warmup draws per chain=10000, total post-warmup draws=10000.

                mean se_mean   sd   2.5%   25%   50%   75% 97.5% n_eff Rhat
mu              4.41    0.03 3.34  -2.11  2.14  4.41  6.67 10.96 10165    1
tau             3.55    0.04 3.10   0.13  1.27  2.74  4.93 11.77  7410    1
theta_tilde[1]  0.31    0.01 0.98  -1.61 -0.34  0.31  0.98  2.21 12266    1
theta_tilde[2]  0.11    0.01 0.93  -1.70 -0.50  0.11  0.74  1.95 11129    1
theta_tilde[3] -0.09    0.01 0.96  -1.97 -0.74 -0.09  0.56  1.80 14975    1
theta_tilde[4]  0.07    0.01 0.93  -1.80 -0.54  0.06  0.69  1.90 11053    1
theta_tilde[5] -0.18    0.01 0.93  -2.01 -0.81 -0.19  0.44  1.72 11131    1
theta_tilde[6] -0.08    0.01 0.95  -1.95 -0.70 -0.08  0.55  1.82 11198    1
theta_tilde[7]  0.36    0.01 0.96  -1.53 -0.29  0.36  1.02  2.21  9049    1
theta_tilde[8]  0.07    0.01 0.98  -1.86 -0.60  0.08  0.74  1.99  9613    1
theta[1]        6.08    0.06 5.53  -3.25  2.60  5.57  8.76 18.92  9268    1
theta[2]        4.95    0.04 4.72  -4.13  1.98  4.83  7.76 14.76 12097    1
theta[3]        3.92    0.05 5.22  -7.82  1.08  4.12  7.14 13.54 11021    1
theta[4]        4.74    0.05 4.69  -4.43  1.87  4.74  7.48 14.44 10626    1
theta[5]        3.60    0.04 4.65  -6.52  0.91  3.85  6.58 12.25 11610    1
theta[6]        4.03    0.05 4.80  -6.06  1.17  4.22  7.11 13.00 11192    1
theta[7]        6.28    0.05 5.05  -2.34  3.03  5.87  8.91 17.96 10318    1
theta[8]        4.81    0.05 5.30  -5.80  1.77  4.76  7.81 15.80  9492    1
lp__           -6.94    0.04 2.29 -12.18 -8.28 -6.66 -5.27 -3.34  3984    1

Samples were drawn using NUTS(diag_e) at Fri Apr 26 13:10:52 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

$$= \sum_z \pi_z(y|a_z)\theta_z$$

$$= \sum_{k=1}^{K} \theta_k \pi_k(y|a_k).$$

Στην συμπερασματολογία των μικτών μοντέλων θα πρέπει να εκτιμήσουμε το βάρος θ_k και τις παραμέτρους a_k για την κάθε διαδικασία. Αυτό εισάγει την δυσκολία ότι αν οι παρατηρήσεις δεν μπορούν να διαχωριστούν μεταξύ των διαδικασιών τότε δεν μπορούν να διαχωριστούν και οι παράμετροι. Αν για παράδειγμα έχουμε δύο διαδικασίες μια κανονική και μια εκθετική κατανομή τότε λόγω των ξεχωριστών χαρακτηριστικών τους είναι εύκολο να διαχωριστούν και οι παράμετροι. Αντίθετα στην περίπτωση που έχουμε δύο πανομοιότυπες κανονικές κατανομές τότε ο διαχωρισμός των παραμέτρων γίνεται πολύ δύσκολος. Για να κατανοήσουμε καλύτερα το πρόβλημα ας ορίσουμε ως σ μια μετάθεση των δεικτών του μεικτού μοντέλου,

$$\sigma(1, ..., K) \mapsto (\sigma(1), ..., \sigma(K)),$$



Διάγραμμα 5.20: Γράφημα των προσομοιωμένων τιμών για την μετασχηματισμένη μεταβλητή $log(\tau)$ στην περίπτωση που έχουμε προσομοιώσει 10000 επαναλήψεις.

δηλαδή μπορούμε να έχουμε μεταθέσεις των παραμέτρων της μορφής

$$\sigma(\boldsymbol{a}) = \sigma(a_1, ..., a_K) \mapsto (a_{\sigma(1)}, ..., a_{\sigma(K)}).$$

Στην συνέχεια θα υπολογίσουμε την συνάρτηση πιθανοφάνειας του μοντέλου για πανομοιότυπες διαδικασίες και θα παρατηρήσουμε ότι μένει αναλλοίωτη σε οποιαδήποτε μετάθεση των δεικτών.

$$\pi(y|\sigma(\boldsymbol{a}), \sigma(\boldsymbol{\theta})) = \sum_{k=1}^{K} \theta_{\sigma(k)} \pi_{\sigma(k)}(y|a_{\sigma(k)})$$

$$\sum_{k'=1}^{K} \theta_{k'} \pi_{k'}(y|a_{k'})$$

$$= \pi(y|\boldsymbol{a}, \boldsymbol{\theta}).$$

Επιπλέον αν και οι πρότερες κατανομές για τα $\boldsymbol{a}$ και $\boldsymbol{\theta}$ είναι αναλλοίωτες (exchangable)στις μεταθέσεις τότε η ύστερη κατανομή με την σειρά της θα είναι αναλλοίωτη στις μεταθέσεις.

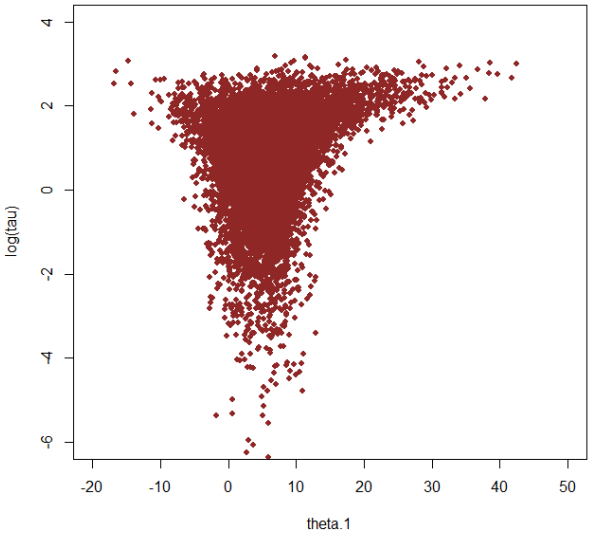
$$\pi(\sigma(\boldsymbol{a}), \sigma(\boldsymbol{\theta})) \propto \pi(\sigma(\boldsymbol{a}))\pi(\sigma(\boldsymbol{\theta})) \sum_{k=1}^{K} \theta_{\sigma(k)} \pi_{\sigma(k)}(y|a_{\sigma(k)})$$

$$\propto \pi(\sigma(\boldsymbol{a}))\pi(\sigma(\boldsymbol{\theta})) \sum_{k'}^{K} \theta_{k'} \pi_{k'}(y|a_{k'})$$

$$\propto \pi(\boldsymbol{a})\pi(\boldsymbol{\theta}) \sum_{k'}^{K} \theta_{k'} \pi_{k'}(y|a_{k'})$$

$$= \pi(\boldsymbol{a}, \boldsymbol{\theta}|y).$$

Συνεπώς τα συμπεράσματα που θα εξάγουμε θα είναι όλα τα ίδια ανεξάρτητα του τρόπου με τον οποίο αναθέτουμε δείκτες στις διαδικασίες. Αυτό θα έχει ώς αποτέλεσμα πολλές επαναλήψεις του αλγορίθμου για κάποια παράμετρο να είναι δεικτοδοτημένες αρχικά σε μια συγκεκριμένη διαδικασία και στην συνέχεια να αλλάζουν διαδικασία (μιας και η συμπερασματολογία ουσιαστικά όπως είδαμε δεν αλλάζει



Διάγραμμα 5.21: Γράφημα μεταξύ των τυχαίων μεταβλητών $log(\tau)$ και θ_1 για την εξερεύνηση αποκλίνουσων παρατηρήσεων.

από τις μεταθέσεις των δεικτών). Αυτό θα έχει ως αποτέλεσμα ότι για κάθε πιθανή δεικτοδότηση θα παίρνουμε και από μια κορυφή για την παράμετρο. Δηλαδή, αν έχουμε K παραμέτρους θα πρέπει να αντιμετωπίσουμε το πολύ $K!$ κορυφές, με την τρομακτική δυσκολία του να μπορέσουμε να μεταβούμε από την μία κορυφή στην άλλη.

Για να αντιμετωπίσουμε το πρόβλημα της πολυκόρυφης ύστερης κατανομής θα πρέπει να κατανοήσουμε από που κληρονομεί την ιδιότητα των αναλλοίωτων μεταθέσεων των δεικτών. Παρατηρήσαμε προηγουμένως ότι με την χρήση exchangable πρότερων κατανομών η ύστερη κατανομή αποκτάει αυτή την ιδιότητα, συνεπώς ένας εύκολος τρόπος ειδικά όταν οι παράμετροι των διαδικασιών διαφέρουν μεταξύ τους ερμηνευτικά (δηλαδή κάποιοι παράμετροι για παράδειγμα μπορεί να αφορούν κάποιο συγκεκριμένο κομμάτι κάποιου υποπληθισμού) είναι να χρησιμοποιήσουμε πρότερη κατανομή οι οποίες δεν μένουν αναλλοίωτες στις μεταθέσεις.

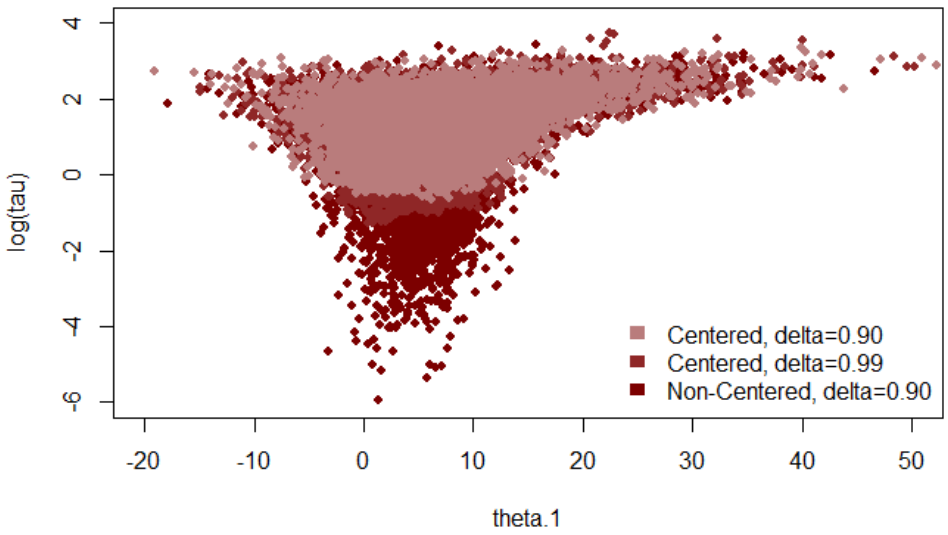
Στην συνέχεια θα αναλύσουμε την περίπτωση που έχουμε στην διάθεση μας μόνο exchangable πρότερη κατανομή και για να αντιμετωπίσουμε το πρόβλημα θα εκμεταλλευτούμε την γεωμετρία της ύστερης κατανομής. Αρχικά θα ορίσουμε μια δεικτοδότηση ως το σημείο αναφοράς όλων των μεταθέσεων των δεικτών και αυτή θα είναι

$$a_1 \leq ... \leq a_K$$

μέσω της οποίας με την χρήση του σ θα μπορούμε να υλοποιούμε οποιαδήποτε μετάθεση θέλουμε $a_{\sigma(1)} \leq ... \leq a_{\sigma(K)}$ και να ελέγχουμε πάντα αν συμπίπτει με την μετάθεση αναφοράς (μέσω των μεταθέσεων απλά αλλάζουν οι δείκτες όχι οι τιμές και η διάταξη που ορίσαμε προηγουμένως). Ακόμα αυτή η δεικτοδότηση συνοδεύεται και από μια ενδιαφέρουσα γεωμετρική ερμηνεία. Για κάθε μια μετάθεση ορίζεται και μια διαφορετική πυραμίδα με την κορυφή της στο 0 και παράλληλα αφού έχουμε $K!$ πυραμίδες μέσω κάθε μιας μετάθεσης (ή περιστροφής της πυραμίδας) θα έχουμε και την πολυκόρυφη κατανομή που μας ενδιαφέρει.

Ακόμα, όπως έχουμε αναφέρει όλα τα στατιστικά ερωτήματα περιστρέφονται γύρω από τον υπολογισμό ολοκληρωμάτων με συνέπεια της ύστερης κατανομής. Άρα θα περιοριστούμε σε υπολογισμούς συναρτήσεων οι οποίες είναι αναλλοίωτες στις μεταθέσεις $f(\sigma(\boldsymbol{a})) = f(\boldsymbol{a})$. Επίσης εφοδιάζοντας τον χώρο των παραμέτρων με την προαναφερθείσα διάταξη μπορεί να ερμηνευθεί και ως την προσπάθεια μας να κάνουμε μια exchangableπρότερη κατανομή να μην είναι exchangable .

$$\pi^{'}(\boldsymbol{a}) = \begin{cases} \pi(\boldsymbol{a}), & a_1 \leq ... \leq a_k \\ 0, & \text{διαφορετικά} \end{cases}$$



Διάγραμμα 5.22: Γράφημα μεταξύ των τυχαίων μεταβλητών $log(\tau)$ και θ_1 για την εξερεύνηση αποκλινουσών παρατηρήσεων συγκρίνοντας κεντροποιημένες και μη κεντροποιημένες μοντελοποιήσεις για διάφορες τιμές μεγέθους βήματος.

Βασιζόμενοι σε αυτά μπορούμε να αποδείξουμε ότι η συμπερασματολογία μας θα είναι συνεπής. Έστω ότι έχουμε δύο διαδικασίες με παραμέτρους a_1 και a_2 και βάρη $\theta_1, \theta_2 = 1 - \theta_1$. Τότε για τον υπολογισμό του ολοκληρώματος θα έχουμε

$$\mathbb{E}_\pi[f] = \int f(a_1, a_2)\pi(a_1, a_2, \theta_1, \theta_2)da_1 d_2 d\theta_1 d\theta_2$$

$$\propto \int f(a_1, a_2)\pi(a_1, a_2)\pi(\theta_1, \theta_2)(\theta_1\pi(y|a_1) + \theta_2\pi(y|a_2))da_1 d_2 d\theta_1 d\theta_2$$

$$\propto \int_{a_1<a_2} f(a_1, a_2)\pi(a_1, a_2)\pi(\theta_1, \theta_2)(\theta_1\pi(y|a_1) + \theta_2\pi(y|a_2))da_1 d_2 d\theta_1 d\theta_2$$

$$+ \int_{a_2<a_1} f(a_1, a_2)\pi(a_1, a_2)\pi(\theta_1, \theta_2)(\theta_1\pi(y|a_1) + \theta_2\pi(y|a_2))da_1 d_2 d\theta_1 d\theta_2.$$

Στην συνέχεια θα μεταχειριστούμε τον δεύτερο όρο έτσι ώστε να μετατρέψουμε τα όρια του ολοκληρώματος στην μορφή που είναι το πρώτο. Αυτό το πετυχαίνουμε υλοποιώντας μια μετάθεση $(a_1, a_2) \to (b_2, b_1)$ και $(\theta_1, \theta_2) \to (\lambda_1, \lambda_2)$ η οποία μας μεταφέρει από την πρώτη πυραμίδα στην δεύτερη.

$$\mathbb{E}_\pi[f] \propto \int_{a_1<a_2} f(a_1, a_2)\pi(a_1, a_2)\pi(\theta_1, \theta_2)(\theta_1\pi(y|a_1) + \theta_2\pi(y|a_2))da_1 d_2 d\theta_1 d\theta_2$$

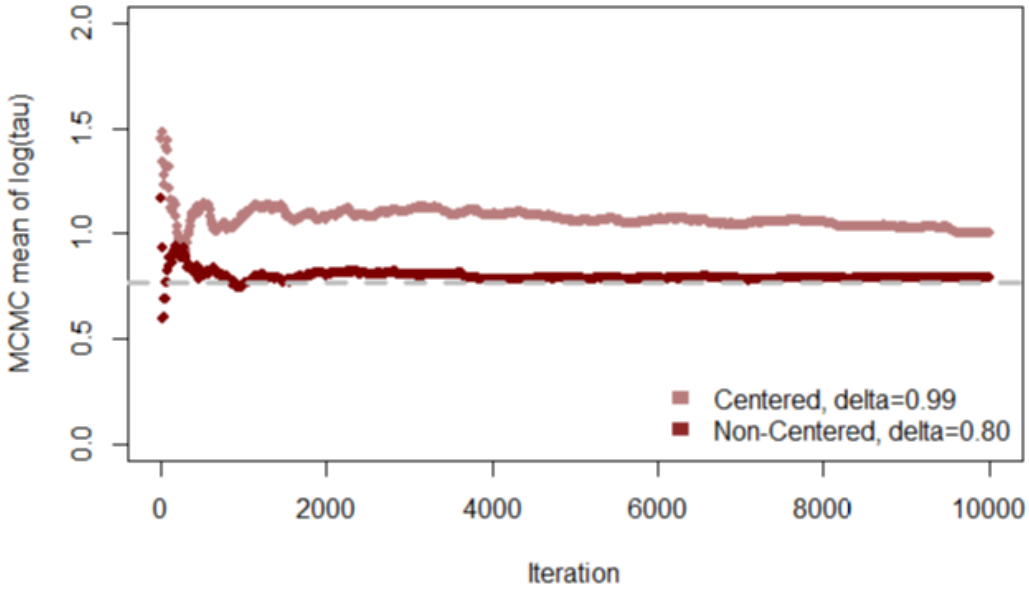
$$+ \int_{b_1<b_2} f(b_2, b_1)\pi(b_2, b_1)\pi(\lambda_2, \lambda_1)(\lambda_2\pi(y|b_2) + \lambda_1\pi(y|b_1))db_1 db_2 d\lambda_2 d\lambda_1$$

Τώρα κάνοντας χρήση της ιδιότητας ότι η $f(\cdot)$ είναι αναλλοίωτη στις μεταθέσεις θα έχουμε

$$\propto \int_{a_1<a_2} f(a_1, a_2)\pi(a_1, a_2)\pi(\theta_1, \theta_2)(\theta_1\pi(y|a_1) + \theta_2\pi(y|a_2))da_1 d_2 d\theta_1 d\theta_2$$

$$+ \int_{b_1<b_2} f(b_1, b_2)\pi(b_1, b_2)\pi(\lambda_1, \lambda_2)(\lambda_1)(\lambda_1\pi(y|b_1) + \lambda_2\pi(y|b_2))db_1 db_2 d\lambda_2 d\lambda_1$$

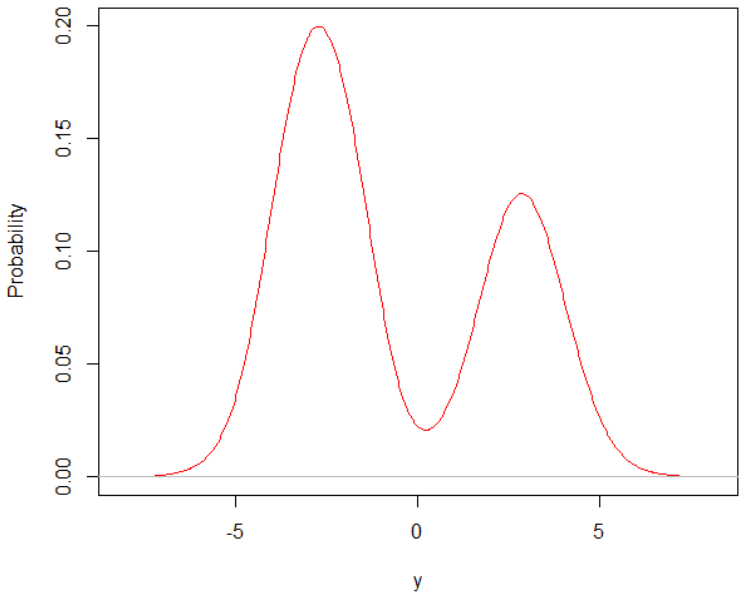
$$\propto 2\int_{a_1<a_2} f(a_1, a_2)\pi(a_1, a_2)\pi(\theta_1, \theta_2)(\theta_1\pi(y|a_1) + \theta_2\pi(y|a_2))da_1 d_2 d\theta_1 d\theta_2$$



$$\propto \int_{a_1<a_2} f(a_1,a_2)2\pi(a_1,a_2)\pi(\theta_1,\theta_2)2(\theta_1\pi(y|a_1)+\theta_2\pi(y|a_2))da_1d_2d\theta_1d\theta_2$$

$$= \int_{a_1<a_2} f(a_1,a_2)\pi^{'}(a_1,a_2,\theta_1,\theta_2|y),$$

όπου $\pi^{'}(a_1,a_2,\theta_1,\theta_2|y)$ είναι η ύστερη κατανομή περιορισμένη στην μετάθεση αναφοράς, συνεπώς έχουμε ότι $\mathbb{E}_{\pi}[f] = \mathbb{E}_{\pi'}[f]$. Για να κατανοήσουμε αυτά που αναλύσαμε προηγουμένως ας δούμε το ακόλουθο παράδειγμα. Έστω ότι έχουμε παράγει τα ακόλουθα 1000 δεδομένα y (για $\mu_1 = 2.78$ και $\mu_2 = -2.78$) για τα οποία παρατηρούμε το ακόλουθο δυκόρυφο διάγραμμα. Στο Διάγραμμα 5.21 παραθέτουμε την καμπύλη των δεδομένων.



Διάγραμμα 5.23: Δικόρυφη καμπύλη των δεδομένων.

Συνεπώς γίνεται αντιληπτό ότι για την μοντελοποίηση αυτών των δεδομένων θα μπορούσαμε να χρησιμοποιήσουμε έναν γραμμικό συνδυασμό κατανομών μιας και παρατηρούμε ότι έχουμε δύο

περιοχές που συγκεντρώνεται μεγάλη μάζα παρατηρήσεων.
Μια λογική αρχική επιλογή θα μπορούσε να είναι το ακόλουθο μοντέλο:

$$\pi(y_1, ..., y_n|\mu_1, \sigma_1, \mu_2, \sigma_2, \theta_1, \theta_2) = \sum_{n=1}^{N} \theta_1 \mathcal{N}_1(y_n|\mu_1, \sigma_1) + \theta_2 \mathcal{N}_2(y_n|\mu_2 \sigma_2),$$

επειδή παρατηρούμε να υπάρχει κάποια σχετική συμμετρία στις δύο ουρές και το σχήμα των κατανομών μοιάζει με καμπάνα κανονικής. Επίσης, ως θ_i για $i = 1, 2$ θα έχουμε την πιθανότητα επιλογής μιας από τις δύο κατανομές. Δηλαδή αφού η αριστερή καμπάνα έχει μεγαλύτερη μάζα παρατηρήσεων θα μπορούσαμε να υποθέσουμε ότι και η πιθανότητα θ_1 να επιλέγει μια παρατήρηση από την $\mathcal{N}(y|\mu_1, \sigma_1)$ είναι μεγαλύτερη από την θ_2.

Όπως έχουμε αναφέρει το πρόβλημα με την πολυκόρυφη ύστερη κατανομή μπορεί να προκληθεί με την χρήση exchangableπρότερης κατανομής. Τέτοιου είδους πρότερες κατανομές είναι οι ακόλουθες:

$$\mu_1, \mu_2 \sim \mathcal{N}(0, 2), \sigma_1, \sigma_2 \sim Half - \mathcal{N}(0, 2)$$

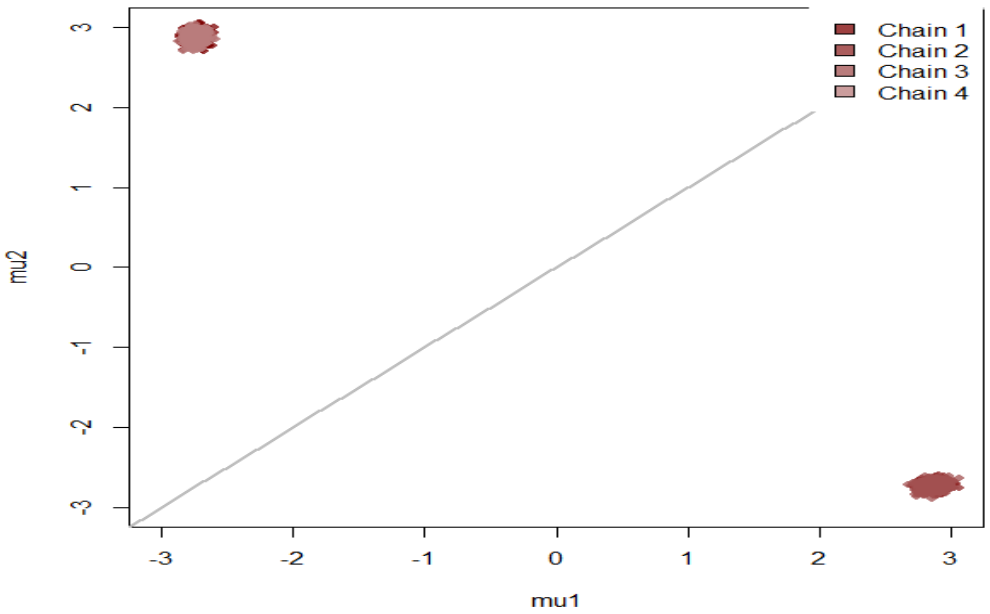
και συμμετρική Βήτα κατανομή για τα βάρη:

$$\theta_1 \sim Beta(5, 5).$$

Τώρα κάνοντας χρήση του αλγορίθμου NUTS για 4 αλυσίδες με 2000 επαναλήψεις η κάθε μια παίρνουμε τα ακόλουθα αποτελέσματα:

```
Inference for Stan model: 15e697c2b528b3cd91273c61d26be4a5.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

           mean se_mean   sd     2.5%      25%      50%      75%    97.5% n_eff  Rhat
mu[1]      0.07    1.98 2.80    -2.81    -2.73     0.04     2.87     2.95     2 63.16
mu[2]      0.07    1.98 2.80    -2.80    -2.73     0.05     2.87     2.96     2 63.04
sigma[1]   1.03    0.00 0.04     0.96     1.00     1.03     1.05     1.10  4296  1.00
sigma[2]   1.03    0.00 0.04     0.96     1.00     1.03     1.05     1.10  4982  1.00
theta      0.50    0.09 0.12     0.35     0.38     0.50     0.62     0.65     2  8.53
lp__    -2108.58    0.04 1.63 -2112.67 -2109.39 -2108.25 -2107.41 -2106.52  2112  1.00

Samples were drawn using NUTS(diag_e) at Tue May 21 18:58:10 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Παρατηρώντας το $\hat{R}$ το οποίο είναι υπερβολικά μεγαλύτερο της μονάδας, καταλαβαίνουμε ότι οι αλυσίδες εξερευνούν διαφορετικές περιοχές της ύστερης κατανομής. Στο Διάγραμμα 5.24 παραθέτουμε τις περιοχές για τις οποίες παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2.



Διάγραμμα 5.24: Γράφημα το οποίο παρουσιάζει την εξερεύνηση διαφορετικών περιοχών από τις αλυσίδες του αλγορίθμου.

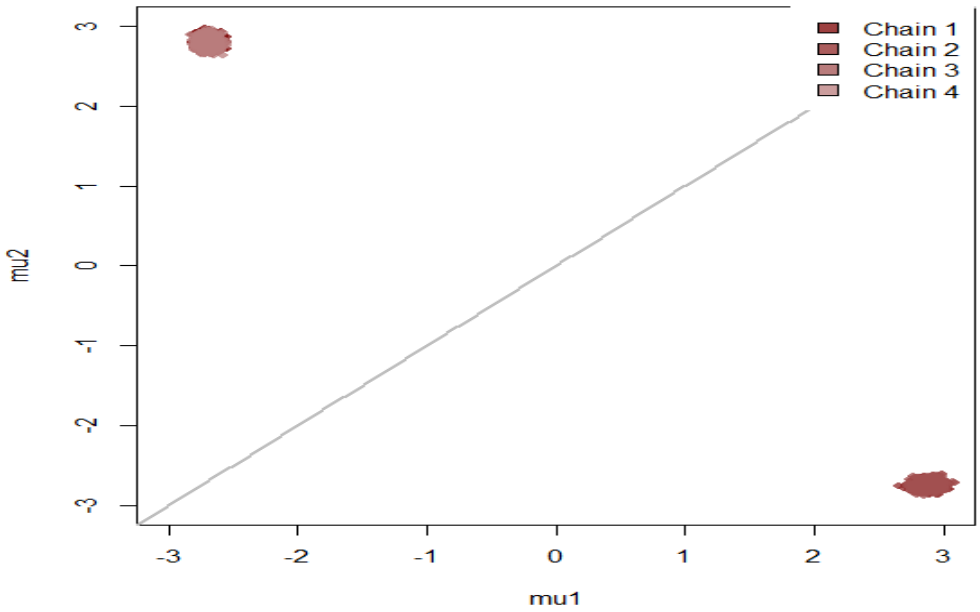
Αυτό το φαινόμενο είναι κάτι το οποίο περιμέναμε μιας και δεν έχουμε εξασφαλίσει μια μόνο κορυφή (αφού χρησιμοποιήσαμε exchangableπρότερη κατανομή), αλλά ξέρουμε ότι αφού έχουμε δύο διαδικασίες το πιθανό πλήθος κορυφών που θα εξερευνηθούν είναι $2! = 2$. Επίσης, καταλαβαίνουμε ότι γενικά είναι πολύ χρήσιμο να τρέχουμε πάνω από μια αλυσίδες, διότι στην περίπτωση που είχαμε κάνει χρήση μόνο μιας αλυσίδας δεν θα είχαμε ανακαλύψει τις δύο κορυφές. Αρχικά ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι μέσω μη exchangable πρότερων κατανομών γι΄ αυτό τον σκοπό θα ορίσουμε

$$\mu_1 \sim \mathcal{N}(4, 0.5),$$
$$\mu_2 \sim \mathcal{N}(-4, 0.5).$$

```
Inference for Stan model: d2a4a5a14fe8dda1d4ff82ae5f0f5f4a.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

          mean se_mean    sd     2.5%      25%      50%      75%    97.5% n_eff  Rhat
mu[1]     0.10    1.97  2.79    -2.76    -2.69     0.06     2.88     2.98     2 59.97
mu[2]     0.03    1.96  2.77    -2.81    -2.74     0.01     2.80     2.88     2 61.28
sigma[1]  1.03    0.00  0.04     0.96     1.00     1.03     1.05     1.10  4705  1.00
sigma[2]  1.03    0.00  0.04     0.96     1.00     1.03     1.05     1.11  4767  1.00
theta     0.50    0.09  0.12     0.35     0.38     0.50     0.62     0.65     2  8.45
lp__     -2201.27  62.85 88.92 -2293.20 -2289.81 -2204.87 -2112.00 -2110.41     2 57.79

Samples were drawn using NUTS(diag_e) at Tue May 21 19:57:49 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Ακόμα και με χρήση πληροφοριακών μη exchangable πρότερων κατανομών παρατηρούμε ότι το $\hat{R}$ είναι πολύ μακριά από την μονάδα. Στο Διάγραμμα 5.25 παραθέτουμε τις περιοχές για τις οποίες

παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2.

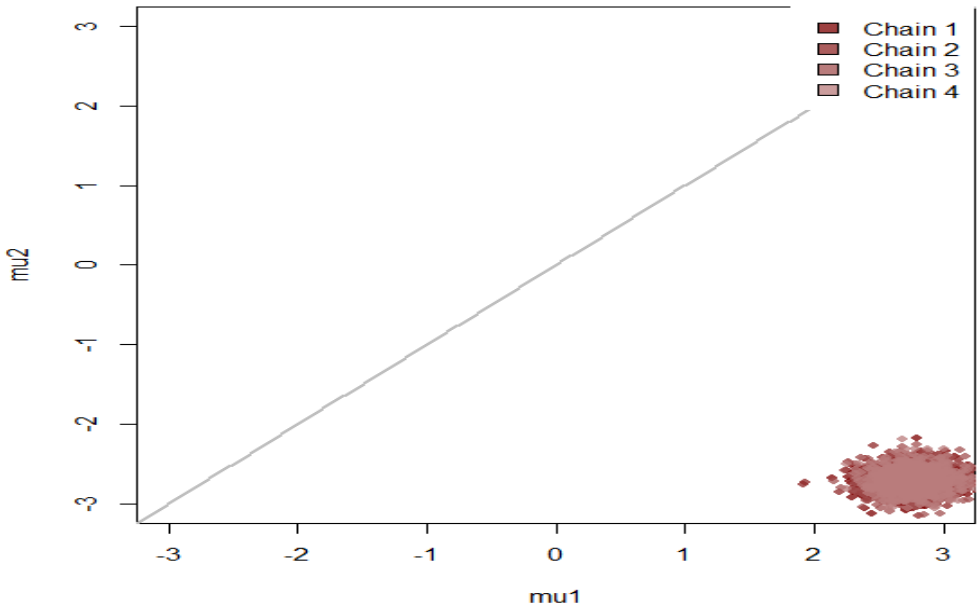


Διάγραμμα 5.25: Γράφημα το οποίο παρουσιάζει την εξερεύνηση διαφορετικών περιοχών από τις αλυσίδες του αλγορίθμου.

Αυτό συμβαίνει λόγω της πληθώρας των δεδομένων που έχουμε, με αποτέλεσμα η συνάρτηση πιθανοφάνειας να μην αφήνει την πρότερη κατανομή να δώσει πληροφορία στην ύστερη κατανομή. Ας δούμε τι αποτελέσματα θα παίρναμε στην περίπτωση που κάναμε χρήση λιγότερων δεδομένων έστω 100.

```
Inference for Stan model: d2a4a5a14fe8dda1d4ff82ae5f0f5f4a.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

          mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
mu[1]     2.77    0.00 0.18    2.41    2.65    2.76    2.88    3.13  4159    1
mu[2]    -2.71    0.00 0.13   -2.95   -2.79   -2.71   -2.62   -2.46  4529    1
sigma[1]  1.05    0.00 0.15    0.81    0.95    1.04    1.14    1.40  3307    1
sigma[2]  1.01    0.00 0.10    0.84    0.94    1.00    1.07    1.23  3827    1
theta     0.35    0.00 0.05    0.26    0.32    0.35    0.38    0.44  4635    1
lp__   -220.21    0.04 1.64 -224.22 -221.01 -219.90 -219.02 -218.05  1552    1

Samples were drawn using NUTS(diag_e) at Tue May 21 20:18:41 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Στο Διάγραμμα 5.26 παραθέτουμε τις περιοχές για τις οποίες παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2 όπου αυτή την φορά εξερευνήθηκαν με αποτελεσματικότητα.



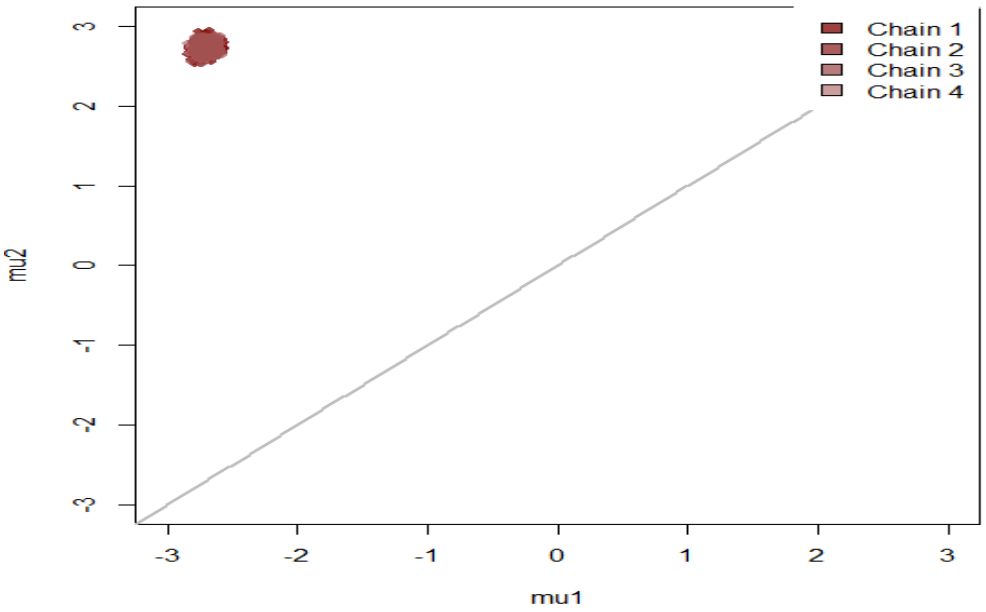
Διάγραμμα 5.26: Γράφημα το οποίο παρουσιάζει την εξερεύνηση της ίδιας περιοχής από τις αλυσίδες με χρήση μικρότερου δείγματος.

Μέσω της μείωσης των δεδομένων έχουμε και μείωση της επίδρασης της συνάρτησης πιθανοφάνειας με αποτέλεσμα η πρότερη κατανομή να μπορεί να επηρεάσει την ύστερη κατανομή και να συγκεντρώσει τις αλυσίδες στην περιοχή που θέλουμε.

Στην συνέχεια θα χρησιμοποιήσουμε την μέθοδο της διάταξης στον χώρο των παραμέτρων. Δηλαδή θα κρατήσουμε τις exchangeableπρότερες κατανομές και μέσω μιας διάταξης στα μ_i, $\mu_1 \leq \mu_2$, θα την μετατρέψουμε σε μη exchangable .

```
Inference for Stan model: 111b380ebe450c511197623bb2100600.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

          mean se_mean   sd     2.5%      25%      50%      75%    97.5% n_eff Rhat
mu[1]    -2.70    0.00 0.04    -2.77    -2.72    -2.70    -2.67    -2.62  4061    1
mu[2]     2.73    0.00 0.06     2.62     2.69     2.73     2.77     2.84  4876    1
sigma[1]  0.96    0.00 0.03     0.91     0.94     0.96     0.98     1.02  5027    1
sigma[2]  1.03    0.00 0.04     0.95     1.00     1.03     1.06     1.12  3994    1
theta     0.63    0.00 0.02     0.60     0.62     0.63     0.64     0.66  4223    1
lp__  -2063.35    0.04 1.64 -2067.35 -2064.16 -2063.02 -2062.17 -2061.23  1856    1

Samples were drawn using NUTS(diag_e) at Tue May 21 20:47:42 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

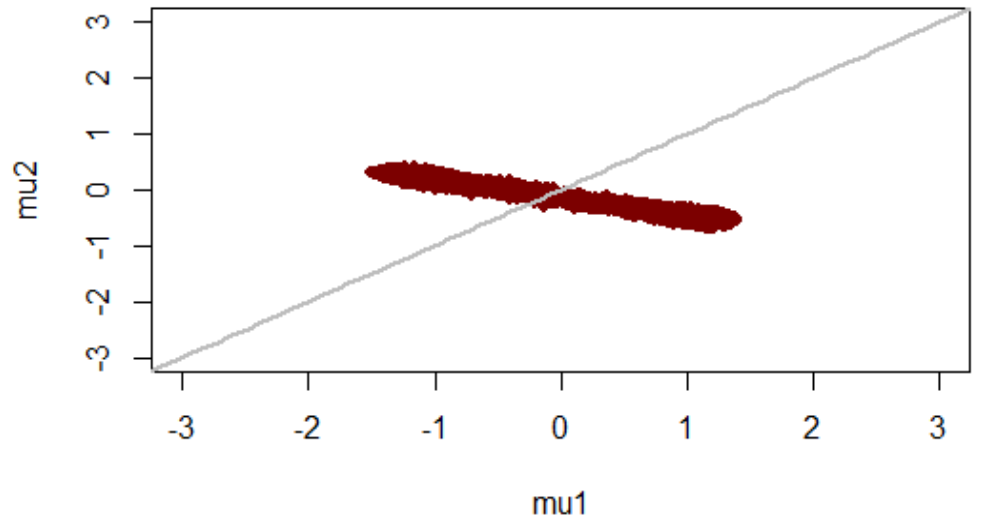
Στο Διάγραμμα 5.27 παραθέτουμε τις περιοχές για τις οποίες παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2 όπου αυτή την φορά εξερευνήθηκαν με αποτελεσματικότητα.



Διάγραμμα 5.27: Γράφημα το οποίο παρουσιάζει την εξερεύνηση της ίδιας περιοχής από τις αλυσίδες με χρήση διατεταγμένων παραμέτρων.

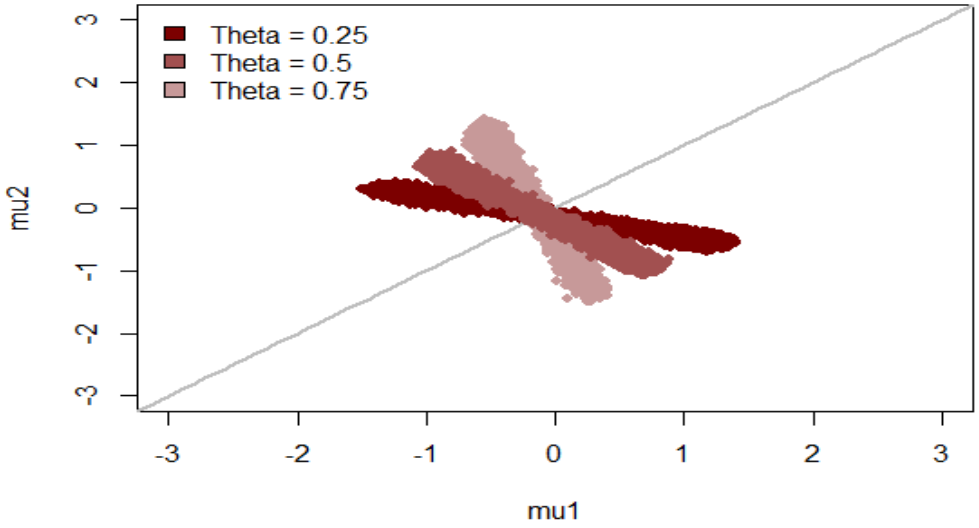
Παρατηρούμε ότι η ιδιότητα της διάταξης είναι πολύ ισχυρή για την αντιμετώπιση αυτού του φαινομένου σε σχέση με μη exchangableπρότερη κατανομή μέσω της οποίας μπορεί να μην καταφέρναμε να επηρεάσουμε την ύστερη κατανομή.

Τέλος, υπάρχει ένα ακόμα μεγάλο παθολογικό πρόβλημα στα μεικτά μοντέλα, στις περιπτώσεις που οι κατανομές των παραμέτρων αλληλοκαλύπτονται. Έστω για το ίδιο πρόβλημα με προηγουμένως, για συγκεκριμένο $\theta_1 = 0.25$ παίρνουμε το ακόλουθο διάγραμμα. Στο Διάγραμμα 5.28 παραθέτουμε τις περιοχές για τις οποίες παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2.



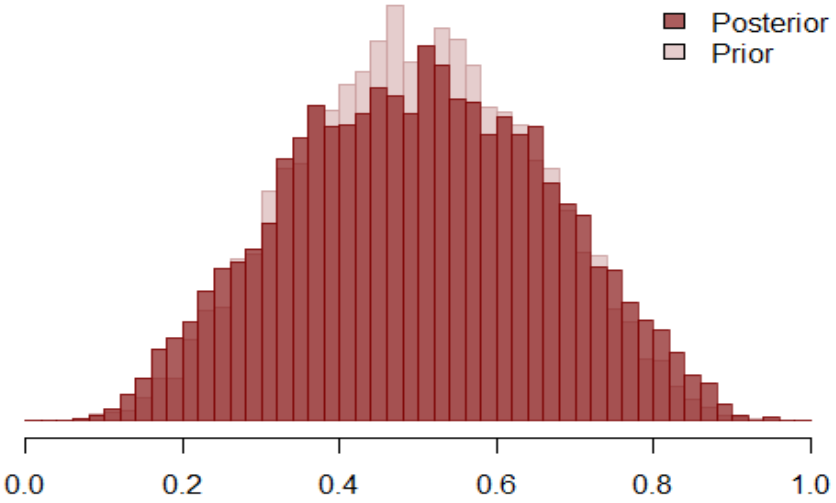
Διάγραμμα 5.28: Γράφημα μεταξύ των τυχαίων παραμέτρων μ_1 και μ_2 για βάρος $\theta_1 = 0.25$.

Επίσης έχει ενδιαφέρον ότι όσο αλλάζουμε το θ_1 παρατηρούμε μια περιστροφή. Στο Διάγραμμα 5.29 παραθέτουμε τις περιοχές για τις οποίες παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2 για τις οποίες χρησιμοποιήσαμε διαφορετικό βάρος για τις κατανομές.



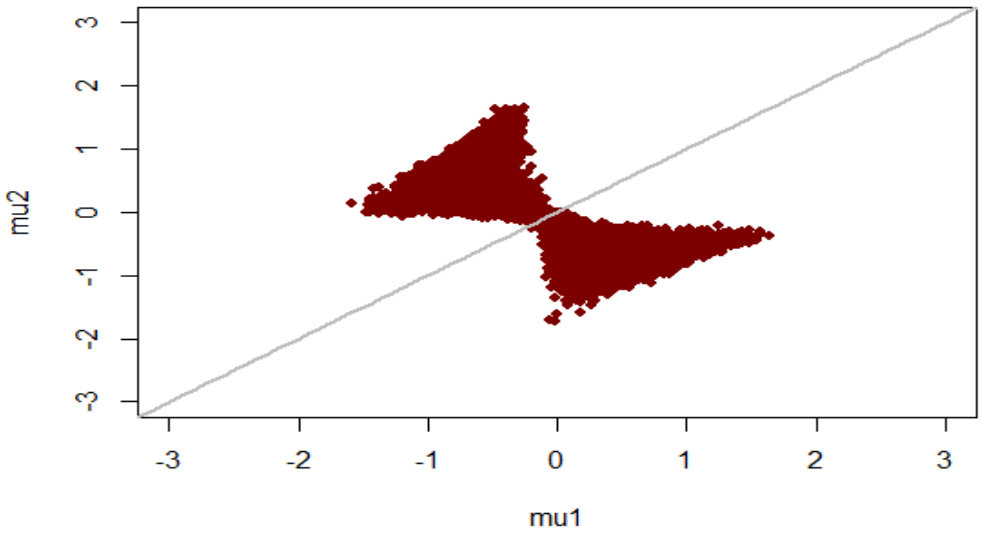
Διάγραμμα 5.29: Γράφημα μεταξύ των τυχαίων παραμέτρων μ_1 και μ_2 για διαφορετικές τιμές του βάρους θ_1.

Αυτό συμβαίνει διότι όσο αυξάνεται το θ_1 δίνουμε όλο και περισσότερο βάρος στις παρατηρήσεις που προέρχονται από την κατανομή της μ_1. Παρατηρώντας αυτή την ιδιαιτερότητα καταλαβαίνουμε ότι στην περίπτωση που εισάγουμε και την παράμετρο θ_1 ως άγνωστη στο μοντέλο ότι η συμπερασματολογία θα γίνει πολύ πιο προβληματική. Επίσης πρόβλημα είναι ότι επειδή οι παράμετροι μ_i αλληλοκαλύπτονται τα βάρη θ_i δεν παίρνουν αρκετή πληροφορία με αποτέλεσμα η ύστερη κατανομή να αλλάζει ελάχιστα από την πρότερη κατανομή. Στο Διάγραμμα 5.30 παραθέτουμε μέσο ιστογράμματος την διαφορά μεταξύ πρότερης και ύστερης κατανομής.



Διάγραμμα 5.30: Γράφημα της πρότερης κατανομής και της ύστερης κατανομής της τυχαίας μεταβλητής θ_1.

Συνεπώς βασιζόμενοι σε όλα τα προηγούμενα με $\theta \sim Beta(5,5)$ παίρνουμε το ακόλουθο διάγραμμα, για το οποίο παρατηρούμε ότι δεν εξερευνήθηκε σωστά η ύστερη κατανομή των μ_i. Στο Διάγραμμα 5.31 παραθέτουμε τις περιοχές για τις οποίες παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2 για τις οποίες όμως δεν παίρνουμε το ζητούμενο αποτέλεσμα.

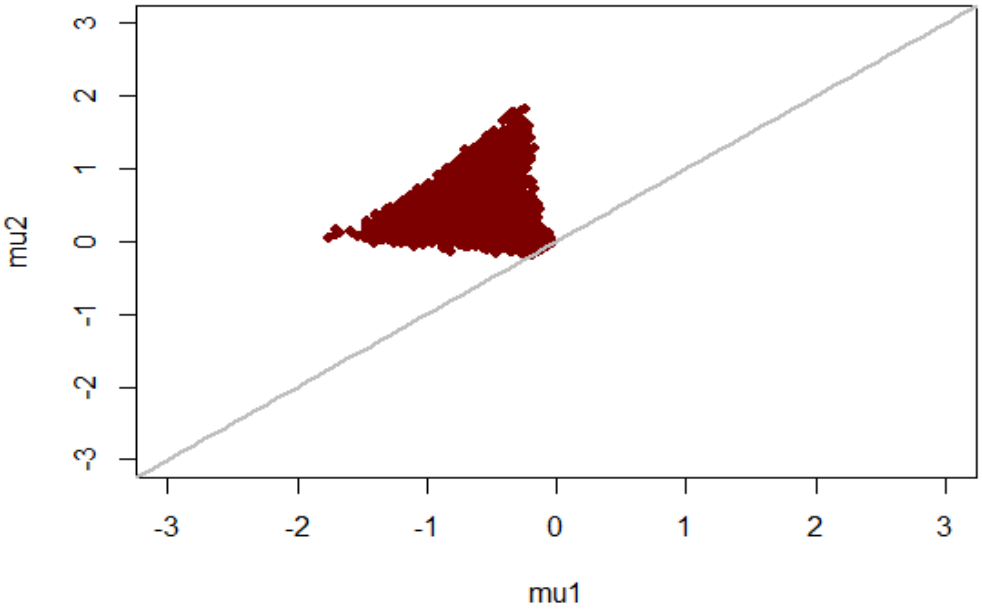


Διάγραμμα 5.31: Γράφημα μεταξύ των τυχαίων μεταβλητών μ_1 και μ_2 για τυχαίο βάρος θ_1.

Για το οποίο επίσης έχουμε αργή εξερεύνηση λόγω των γωνιών (περιοχές μεγάλης καμπυλότητας, παρόλα αυτά δεν έχουμε αποκλίνουσες παρατηρήσεις δηλαδή δεν έχουμε μεροληψία).

```
Inference for Stan model: 10f9e160127aaa61144c3abc66bb8bbe.
1 chains, each with iter=11000; warmup=1000; thin=1;
post-warmup draws per chain=10000, total post-warmup draws=10000.

          mean se_mean   sd     2.5%      25%      50%      75%    97.5% n_eff Rhat
mu[1]    -0.13    0.04 0.66    -1.10    -0.70    -0.24     0.45     1.12   294    1
mu[2]    -0.02    0.04 0.67    -1.08    -0.63     0.00     0.54     1.12   317    1
sigma[1]  1.06    0.00 0.13     0.80     0.97     1.05     1.15     1.30  1733    1
sigma[2]  1.06    0.00 0.13     0.81     0.97     1.05     1.15     1.30  1640    1
theta     0.50    0.00 0.17     0.19     0.38     0.50     0.63     0.82  1975    1
lp__  -1633.58    0.03 1.72 -1637.67 -1634.50 -1633.36 -1632.33 -1631.07  2492    1

Samples were drawn using NUTS(diag_e) at Thu Jun 20 19:11:31 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Παρατηρούμε ότι το πλήθος των αποτελεσματικών δειγμάτων είναι αρκετά μικρό. Συνεπώς για να το αντιμετωπίσουμε θα μπορούσαμε να τρέξουμε τον αλγόριθμο για περισσότερες επαναλήψεις ή ακόμα καλύτερα να τρέξουμε πολλαπλές αλυσίδες. Τέλος, όπως και προηγουμένως χρησιμοποιούμε μια διάταξη στις παραμέτρους μ_i. Στο Διάγραμμα 5.32 παραθέτουμε τις περιοχές για τις οποίες παράχθηκαν παρατηρήσεις για τις παραμέτρους μ_1 και μ_2 μέσο διάταξης και παρατηρούμε ότι τελικά μπορέσαμε να πιάσουμε τα ζητούμενα όρια για τις παραμέτρους.



Διάγραμμα 5.32: Γράφημα μεταξύ των τυχαίων μεταβλητών μ_1 και μ_2 για τυχαίο βάρος θ_1.

5.4 Συγκριτική Αξιολόγηση της Διδακτικής Ποιότητας

Κατά τα έτη $2009-2013$ διεξήχθη έρευνα σε πανεπιστήμιο για να υπολογιστεί η ποιότητα διδασκαλίας. Για τον σκοπό αυτό δόθηκαν ερωτηματολόγια στους φοιτητές για να βαθμολογήσουν διάφορες πτυχές της διδασκαλίας με μεταβλητές $X_1 - X_{20}$ και επίσης το μάθημα ως σύνολο με μεταβλητή Y. Όλες οι μεταβλητές οι οποίες αντιστοιχούν σε μια βαθμολογική ερώτηση ανήκουν στο διάστημα $[0, 1]$ ως

X_1	Σαφήνεια των στόχων του μαθήματος.(w_1)
X_2	Συνάφεια του περιεχομένου του μαθήματος με τους στόχους του μαθήματος.(w_2)
X_3	Συνάφεια του περιεχομένου του μαθήματος με τα καθορισμένα πρότυπα.(w_3)
X_4	Αξία υλικού μαθήματος.(w_4)
X_5	Το περιεχόμενο του μαθήματος ικανοποίησε τις ακαδημαϊκές διδακτικές απαιτήσεις.(w_5)
X_6	Σημασία της φοίτησης του μαθήματος στην εκμάθηση και την επιτυχή ολοκλήρωση του μαθήματος.(w_6)
X_7	Εξέταση του μαθήματος σε σχέση με τα υλικό του μαθήματος και τα διδαχθέντα χωρία.(w_7)
X_8	Η δυσκολία της εξέτασης και η συμβατότητα της με την δυσκολία του μαθήματος.(w_8)
X_9	Η εξέταση συνέβαλε στην ανάπτυξη των γνώσεων των φοιτητών.(w_9)
X_{10}	Ακρίβεια, αμεροληψία της εξέτασης.(w_{10})
X_{11}	Οργάνωση μαθημάτων διδάσκοντος.(w_{11})
X_{12}	Προετοιμασία καθηγητή.(w_{12})
X_{13}	Ελκυστικότητα της διδασκαλίας.(w_{13})
X_{14}	Δυνατότητα επικοινωνίας εκπαιδευτή.(w_{14})
X_{15}	Η γνώση του διδάσκοντος για το περιεχόμενο του μαθήματος.(w_{15})
X_{16}	Επικοινωνία με τους μαθητές, ενθάρρυνση της ομαδικής αλληλεπίδρασης.(w_{16})
X_{17}	Συνέπεια στο πρόγραμμα διδασκαλίας και διαθεσιμότητα σε ώρες γραφείου.(w_{17})
X_{18}	Δίκαιη μεταχείριση των φοιτητών.(w_{18})
X_{19}	Σεβασμός προς τους φοιτητές.(w_{19})
X_{20}	Νοητική δυσκολία και επέκταση γνώσεων στο τρέχω επιστημονικό πεδίο.(w_{20})
Y	Συνολική αξιολόγηση διδασκαλίας.(y)

Πίνακας 5.3: Το ερωτηματολόγιο το οποίο χρησιμοποιήθηκε στην έρευνα.

ποσοστά που εκφράζουν το βάρος που δίνει ο κάθε φοιτητής σε κάθε ερώτηση. Στον Πίνακα 5.3 παραθέτουμε τις ερωτήσεις οι οποίες χρησιμοποιήθηκαν στην έρευνα μαζί με τα βάρη που αντιστοιχούν στην κάθε μία.

Σκοπός μας είναι να δούμε ποια από τις μεταβλητές $X_1 - X_{20}$ έχει μεγαλύτερη επίδραση στο συνολικό σκόρ Y. Γι΄ αυτό τον λόγο θα υλοποιήσουμε μια Βήτα παλινδρόμηση με τους συντελεστές της να ακολουθούν μια Dirichlet κατανομή και να εκφράζουν διαφορετικό βάρος σε κάθε επεξηγηματική μεταβλητή. Ακόμα θα πρέπει να λάβουμε υπόψιν ότι μέσω των μεταβλητών που έχουμε συμπεριλάβει δεν έχουμε καταφέρει να εκφράσουμε όλη την αβεβαιότητα του προβλήματος. Συνεπώς είναι απαραίτητη η χρήση μιας μεταβλητής Z η οποία θα εκφράζει αυτή την αβεβαιότητα και η οποία αφού θέλουμε ποσοστά θα ακολουθεί μια Βήτα κατανομή. Την μεταβλητή Z θα την αποκαλούμε latent.

Για την μοντελοποίηση των δεδομένων έχουμε ότι για $i = 1, ..., n = 400$ και $j = 1, ..., 5$,

$$Y_{ij} \sim Beta(a_{ij}, b_{ij}),$$

$$a_{ij} = \frac{(1-\mu_{ij})\mu_{ij}^2 - \mu_{ij}\sigma_j^2}{\sigma_j^2},$$

$$b_{ij} = \frac{(1-\mu_{ij})(\mu_{ij} - \mu_{ij}^2 - \sigma_j^2)}{\sigma_j^2},$$

$$\mu_{ij} = w_{j,1}X_{ij,1} + ... + w_{j,k}X_{ij,k} + w_{j,k+1}Z_{ij},$$

όπου μ_{ij} και σ_j^2 είναι ο μέσος και η διακύμανση του Y_{ij}. Επίσης γνωρίζουμε ότι τα βάρη ακολουθούν μια Dirichlet κατανομή δηλαδή ισχύει ότι $0 \leq w_{j,l} \leq 1$ για $l = 1, ..., k+1$ και $\sum_{l=1}^{k+1} w_{j,l} = 1$ για $k = 20$ που αντιστοιχεί στο πλήθος των ερωτήσεων του ερωτηματολογίου. Τέλος, για την μοντελοποίηση θα ορίσουμε τις ακόλουθες πρότερες κατανομές:

$$\boldsymbol{w}_j = (w_{j,1}, ..., w_{j,k}, w_{j,k+1}) \sim Dirichlet(1, ..., 1),$$

δηλαδή μη πληροφοριακή κατανομή για το $\boldsymbol{w}_j$. Επίσης, για την τυχαία μεταβλητή σ_j θα χρησιμοποιήσουμε μια τυχαία μεταβλητή ακρίβειας τ μέσο της οποίας θα εκφράζουμε την διακύμανση ώς σ_j. Για

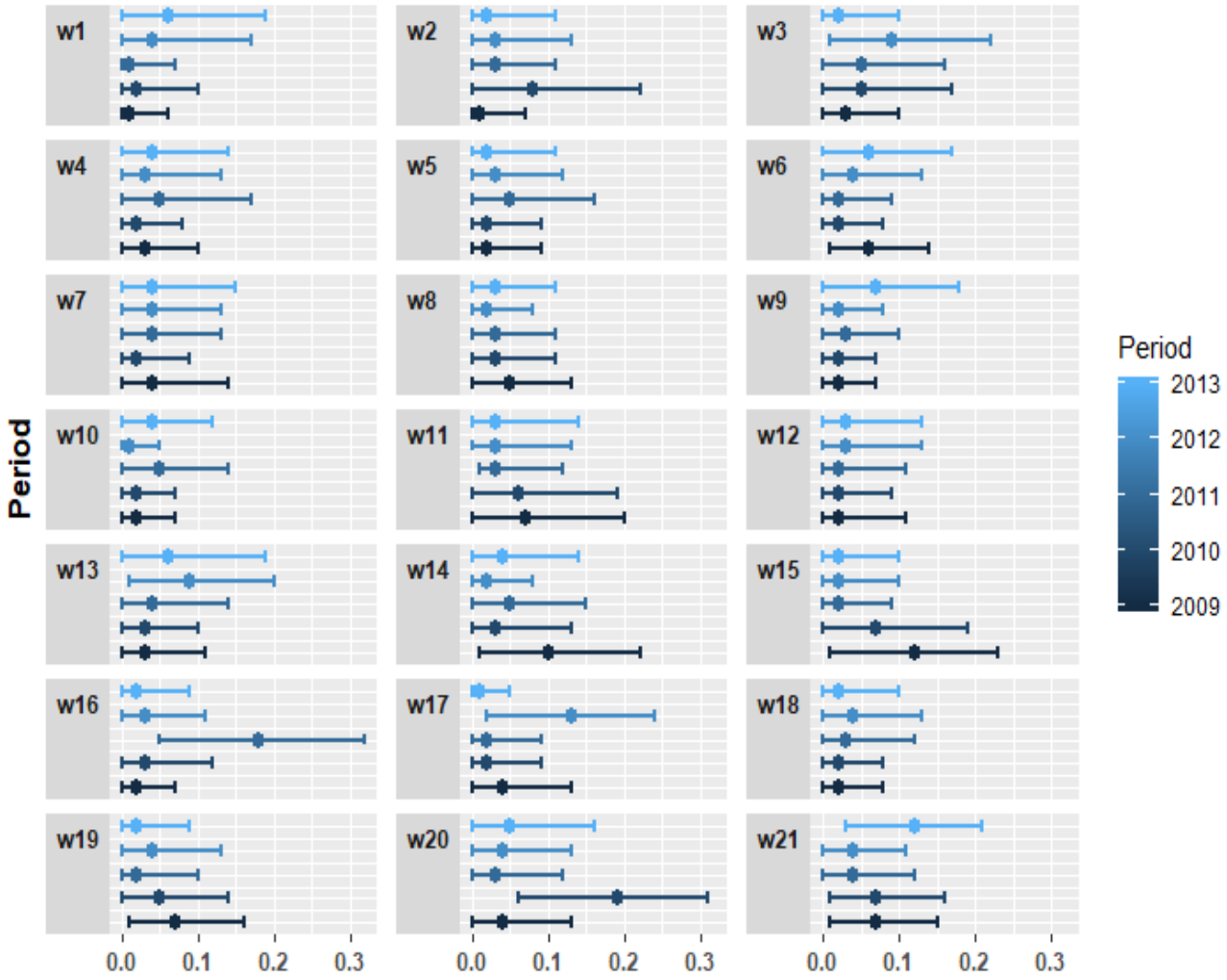
την παράμετρο $\tau = \tau_j = \sigma_j^{-1}$ θα ορίσουμε μια μη πληροφοριακή πρότερη κατανομή Γάμμα,

$$\tau \sim Gamma(0.5, 0.5).$$

Τέλος, για την latent μεταβλητή θα ορίσουμε μια μη πληροφοριακή μεταβλητή Βήτα

$$Z \sim Beta(0.5, 0.5).$$

Στην συνέχεια θα υλοποιήσουμε το μοντέλο το οποίο περιγράψαμε προηγουμένως κάνοντας χρήση του αλγορίθμου NUTS για 4 αλυσίδες με 4000 επαναλήψεις συνολικά όπου οι 2000 θα αντιστοιχούν στις burn-in επαναλήψεις. Επιπλέον, να αναφέρουμε ότι χρησιμοποιήθηκε πληθώρα διαγνωστικών ελέγχων για την διαπίστωση της σύγκλισης του αλγορίθμου. Οι σημαντικότεροι ήταν ο συνδιασμός των traceplot μαζί με το $\hat{R}$ στατιστικό, το E-BFMI στατιστικό και ο υπολογισμός των αποκλίνουσων επαναλήψεων οι οποίες μπορούν να παρεμποδίσουν σημαντικά την συμπεριφορά του αλγορίθμου. Στο Διάγραμμα 5.33 παραθέτουμε τα διαστήματα εμπιστοσύνης για τα βάρη για κάθε χρονιά. Συγκεκριμένα



Διάγραμμα 5.33: 95% διαστήματα εμπιστοσύνης με διαμέσους των posterior κατανομών για τα βάρη.

παρατηρούμε ότι τα βάρη τα οποία έχουν την μεγαλύτερη μεταβλητότητα μεταξύ των περιόδων $2009-2013$ είναι τα $2, 14, 16, 17$ και 20. Επιπλέον, μαζί με τα διαστήματα εμπιστοσύνης που παραθέσαμε είναι πολύ σημαντικό να υπολογίσουμε και τα βάρη για όλες τις περιόδους μαζί. Στην συνέχεια στον Πίνακα 5.4 παραθέτουμε τις εκτιμήσεις των βαρών για όλες τις χρονιές μαζί

Βάρη	Μέσοι	Τυπικό Σφάλμα	2.5%	Διάμεσος	97.5%
w_1	0.02	0.02	0.00	0.02	0.06
w_2	0.04	0.03	0.00	0.04	0.10
w_3	0.08	0.03	0.01	0.08	0.14
w_4	0.05	0.03	0.00	0.05	0.11
w_5	0.03	0.02	0.00	0.02	0.08
w_6	0.04	0.02	0.00	0.04	0.09
w_7	0.07	0.03	0.01	0.06	0.14
w_8	0.03	0.02	0.00	0.02	0.07
w_9	0.02	0.02	0.00	0.02	0.06
w_{10}	0.02	0.01	0.00	0.01	0.05
w_{11}	0.08	0.04	0.01	0.08	0.15
w_{12}	0.03	0.02	0.00	0.02	0.08
w_{13}	0.07	0.03	0.01	0.07	0.14
w_{14}	0.05	0.03	0.00	0.05	0.12
w_{15}	0.05	0.03	0.01	0.05	0.11
w_{16}	0.04	0.02	0.00	0.03	0.09
w_{17}	0.02	0.01	0.00	0.01	0.05
w_{18}	0.02	0.01	0.00	0.02	0.05
w_{19}	0.05	0.03	0.01	0.05	0.10
w_{20}	0.11	0.03	0.05	0.11	0.17
w_{21}	0.10	0.02	0.06	0.10	0.14

Πίνακας 5.4: Μέσοι και μέτρα θέσης για την ύστερη κατανομή για τις περιόδους $2009-2013$ από κοινού.

Για τα απο κοινού δεδομένα παρατηρούμε κάτι τελείως διαφορετικό απο την προηγούμενη ανάλυση μας. Οι ερωτήσεις με τα μεγαλύτερα βάρη ήταν οι κάτωθι $3, 11, 20$ και 21. Αυτό μπορεί να ερμηνευθεί ώς ότι αυτά τα βάρη έχουν σημαντική μακροχρόνια επίδραση για την βελτίωση της διεξαγωγής του μαθήματος. Επίσης βασιζόμενοι στα διαστήματα εμπιστοσύνης παρατηρούμε ότι τα βάρη 3 και 11 δεν είναι μεγάλης σημασίας, λανθασμένα, μιας και παίρνουν μικρές τιμές. Επιπλέον, επειδή τα ανα δύο συνεχόμενα χρόνια μπορεί να έχουν περισσότερα κοινά σε σχέση με άλλες ανα δύο ομάδες χρονιών , λογικό θα ήταν να παράξουμε συμπερασματολογία για τις ανα δύο χρονιές και να ψάξουμε για διαφορές. Για να υπολογίσουμε αυτές τις διαφορές θα χρησιμοποιήσουμε το π_0 στατιστικό, Ντζούφρας[25,σελ. 155], το οποίο έχει την ακόλουθη μορφή

$$\pi_0 = min\left\{f(w_{j-1,l} - w_{j,l} > 0|\boldsymbol{y}), f(w_{j-1,l} - w_{j,l} < 0|\boldsymbol{y})\right\}, \quad j = 1, ..., 5\ , l = 1, ..., 21.$$

Όταν οι διαφορές $w_{j-1,l} - w_{j,l}$ είναι κοντά στο μηδέν, τότε η τιμή του π_0 θα αναμένετε να βρίσκεται κοντά στο 0.5. Αντίθετα μικρές τιμές του π_0 (μικρότερες του 5%) θα υποδηλώνουν ότι υπάρχουν διαφορές μεταξύ των βαρών για τις διαδοχικές χρονιές. Στον Πίνακα 5.5 παραθέτουμε τις διαφορές των βαρών μεταξύ των δύο πρώτων χρονικών περιόδων.

Βάρη	Πρώτη Περίοδος	Δεύτερη Περίοδος	Διαφορές	π_0
w_1	0.02	0.03	-0.01	0.35125
w_2	0.02	0.09	-0.07	0.10550
w_3	0.03	0.06	-0.03	0.31925
w_4	0.03	0.03	0.00	0.41325
w_5	0.03	0.03	0.00	0.44075
w_6	0.06	0.02	0.04	0.17550
w_7	0.05	0.03	0.02	0.28200
w_8	0.05	0.03	0.02	0.32475
w_9	0.02	0.02	0.00	0.49025
w_{10}	0.02	0.02	0.00	0.47775
w_{11}	0.08	0.07	0.01	0.44350
w_{12}	0.03	0.03	0.00	0.45150
w_{13}	0.04	0.03	0.01	0.46775
w_{14}	0.11	0.04	0.07	0.15775
w_{15}	0.12	0.07	0.05	0.28375
w_{16}	0.02	0.04	-0.02	0.35725
w_{17}	0.05	0.03	0.02	0.31125
w_{18}	0.02	0.02	0.00	0.47025
w_{19}	0.07	0.06	0.01	0.40275
w_{20}	0.05	0.19	-0.14	0.03475
w_{21}	0.07	0.07	0.00	0.49425

Πίνακας 5.5: Μέσοι ύστερης κατανομής για τα βάρη των διαφορών για δύο συνεχόμενες χρονιές (πρώτη περίοδος - δεύτερη περίοδος).

Παρατηρούμε ότι για τα βάρη $2, 6, 14$ και 20 έχουμε π_0 μικρότερο 20% και ειδικότερα για το βάρος 20 έχουμε 3% . Συνεπώς, καταλαβαίνουμε ότι οι μόνες ερωτήσεις οι οποίες διαφέρουν δεσμευμένες στο χρόνο είναι $2, 6, 14$ και 20 ενώ οι υπόλοιπες παραμένουν σχετικά σταθερές. Ακολουθούμε την ίδια διαδικασία για τις υπόλοιπες ανά δύο συνεχόμενες χρονιές στον Πίνακα 5.6 παραθέτουμε τις διαφορές των βαρών μεταξύ των περιόδων 2010-2011.

Βάρη	Δεύτερη Περίοδος	Τρίτη Περίοδος	Διαφορές	π_0
w_1	0.03	0.02	0.01	0.36975
w_2	0.09	0.04	0.05	0.20275
w_3	0.06	0.06	0.00	0.48175
w_4	0.03	0.06	-0.03	0.24125
w_5	0.03	0.06	-0.03	0.26800
w_6	0.02	0.03	-0.01	0.42875
w_7	0.03	0.04	-0.01	0.34750
w_8	0.03	0.04	-0.01	0.46425
w_9	0.02	0.03	-0.01	0.39575
w_{10}	0.02	0.06	-0.04	0.16050
w_{11}	0.07	0.04	0.03	0.30325
w_{12}	0.03	0.03	0.00	0.45450
w_{13}	0.03	0.05	-0.02	0.41150
w_{14}	0.04	0.05	-0.01	0.40550
w_{15}	0.07	0.03	0.04	0.21550
w_{16}	0.04	0.18	-0.14	0.0300
w_{17}	0.03	0.03	0.00	0.46975
w_{18}	0.02	0.04	-0.02	0.30950
w_{19}	0.06	0.03	0.03	0.29000
w_{20}	0.19	0.04	0.15	0.02700
w_{21}	0.07	0.04	0.03	0.28150

Πίνακας 5.6: Μέσοι ύστερης κατανομής για τα βάρη των διαφορών για δύο συνεχόμενες χρονιές (δεύτερη περίοδος - τρίτη περίοδος).

Παρατηρούμε ότι τα βάρη $10, 16$ και 20 παίρνουμε π_0 μικρότερο 20% ειδικότερα για το βάρος 16 και 20 παίρνουμε τιμή μικρότερη του 3%. Συνεπώς καταλαβαίνουμε ότι οι μόνες ερωτήσεις οι οποίες δεσμευμένες στο χρόνο διαφέρουν είναι οι $10, 16$ και 20. Στον Πίνακα 5.7 παραθέτουμε τις διαφορές των βαρών μεταξύ των περιόδων 2011-2012.

Βάρη	Τρίτη Περίοδος	Τέταρτη Περίοδος	Διαφορές	π_0
w_1	0.02	0.05	-0.03	0.22000
w_2	0.04	0.04	0.00	0.44200
w_3	0.06	0.10	-0.04	0.30925
w_4	0.06	0.04	0.02	0.37625
w_5	0.06	0.04	0.02	0.38700
w_6	0.03	0.05	-0.02	0.32950
w_7	0.04	0.04	0.00	0.49250
w_8	0.04	0.03	0.01	0.36700
w_9	0.03	0.02	0.01	0.38800
w_{10}	0.06	0.02	0.04	0.13875
w_{11}	0.04	0.04	0.00	0.47775
w_{12}	0.03	0.04	-0.01	0.43025
w_{13}	0.05	0.09	-0.04	0.23950
w_{14}	0.05	0.02	0.03	0.25875
w_{15}	0.03	0.03	0.00	0.47100
w_{16}	0.18	0.04	0.14	0.03075
w_{17}	0.03	0.13	-0.10	0.04925
w_{18}	0.04	0.04	0.00	0.49550
w_{19}	0.03	0.05	-0.02	0.38125
w_{20}	0.04	0.04	0.00	0.47900
w_{21}	0.04	0.04	0.00	0.49475

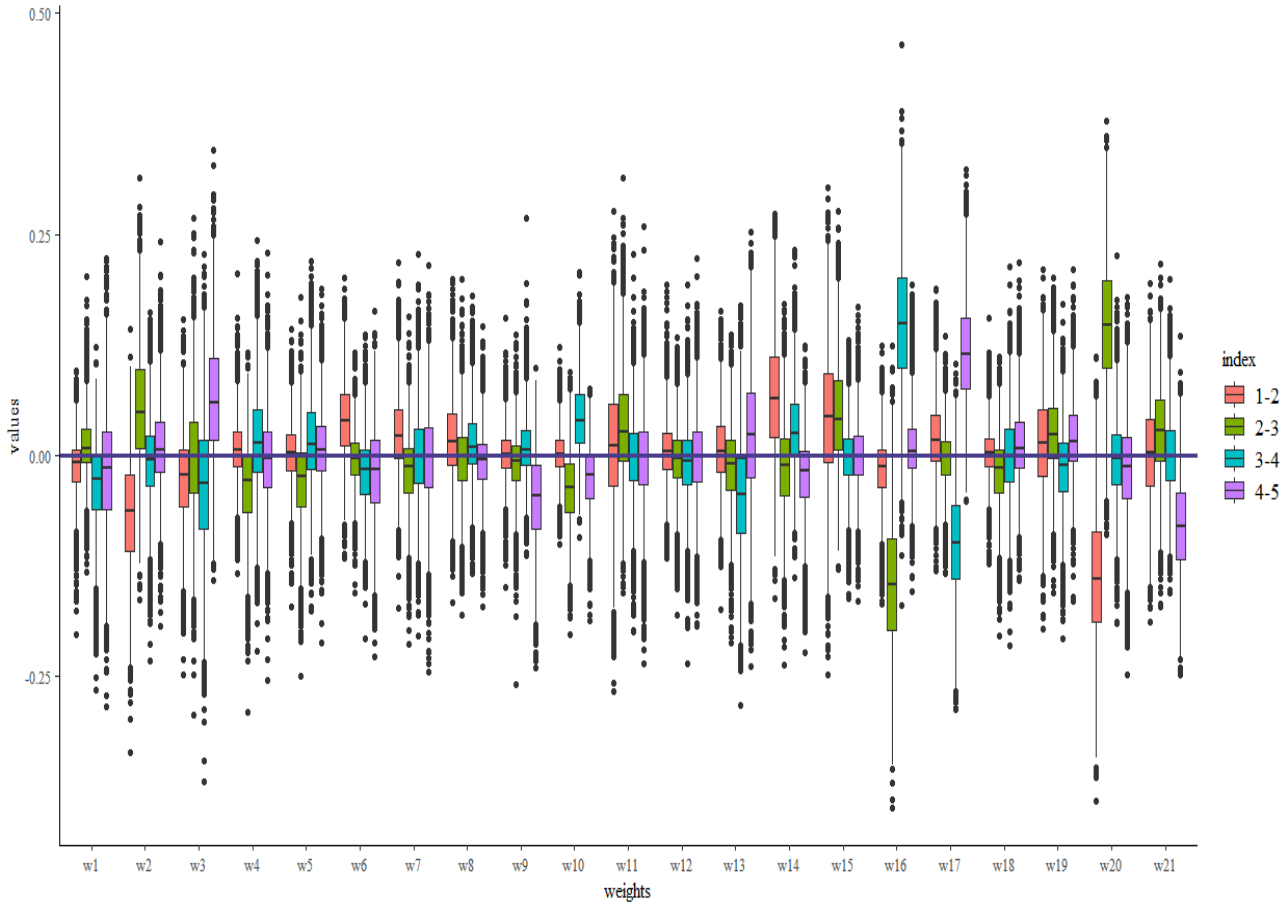
Πίνακας 5.7: Μέσοι ύστερης κατανομής για τα βάρη των διαφορών για δύο συνεχόμενες χρονιές (τρίτη περίοδος - τέταρτη περίοδος).

Παρατηρούμε ότι για τα βάρη $10, 16$ και 17 παίρνουμε π_0 μικρότερο του 20% και ειδικότερα για τα βάρη 16 και 17 παίρνει τιμή μικρότερη του 5%. Συνεπώς καταλαβαίνουμε ότι οι μόνες ερωτήσεις οι οποίες δεσμευμένες στο χρόνο διαφέρουν είναι οι $10, 16$ και 17. Στον Πίνακα 5.8 παραθέτουμε τις διαφορές των βαρών μεταξύ των περιόδων 2012-2013.

Βάρη	Τέταρτη Περίοδος	Πέμπτη Περίοδος	Διαφορές	π_0
w_1	0.05	0.07	-0.02	0.41400
w_2	0.04	0.03	0.01	0.41325
w_3	0.10	0.03	0.06	0.13825
w_4	0.04	0.05	-0.01	0.45750
w_5	0.04	0.03	0.01	0.44275
w_6	0.05	0.07	-0.02	0.37000
w_7	0.04	0.05	-0.01	0.47550
w_8	0.03	0.03	0.00	0.41150
w_9	0.02	0.07	-0.05	0.15050
w_{10}	0.02	0.04	-0.02	0.22025
w_{11}	0.04	0.04	0.00	0.47200
w_{12}	0.04	0.04	0.00	0.47500
w_{13}	0.09	0.07	0.02	0.37975
w_{14}	0.02	0.05	-0.03	0.30550
w_{15}	0.03	0.03	0.00	0.49800
w_{16}	0.04	0.03	0.01	0.43225
w_{17}	0.13	0.01	0.12	0.01400
w_{18}	0.04	0.03	0.01	0.38825
w_{19}	0.05	0.03	0.02	0.32125
w_{20}	0.04	0.06	-0.02	0.39175
w_{21}	0.04	0.12	-0.08	0.07450

Πίνακας 5.8: Μέσοι ύστερης κατανομής για τα βάρη των διαφορών για δύο συνεχόμενες χρονιές (τέταρτη περίοδος - περίοδος περίοδος).

Παρατηρούμε ότι για τα βάρη $3, 9, 17$ και 21 έχουμε π_0 μικρότερο του 20% και ειδικότερα για τα βάρη 17 και 21 παίρνουμε τιμή μικρότερη του 8%. Συνεπώς καταλαβαίνουμε ότι οι μόνες ερωτήσεις οι οποίες δεσμευμένες στο χρόνο διαφέρουν είναι οι $3, 9, 17$ και 21. Επιπλέον, σε συνδιασμό με τους προηγούμενους Πίνακες δίνουμε τα boxplot των διαφορών για όλες τις ανα δύο συνεχόμενες χρονιές. Στο Διάγραμμα 5.34 δίνουμε τα boxplot των διαφορών των βαρών για όλους τους διαδοχικούς συνδυασμούς περιόδων.



Διάγραμμα 5.34: Boxplots που περιγράφει όλες τις διαφορές για τα βάρη για όλες τις ανα δύο συνεχόμενες χρονιές(πρώτη περίοδος - δεύτερη περίοδος, δεύτερη περίοδος- τρίτη περίοδος, τρίτη περίοδος - τέταρτη περίοδος, τέταρτη περίοδος - πέμπτη περίοδος).

Παρατηρούμε ότι για τα boxplot τα οποία οι τιμές τους είναι μακριά απο το μηδέν (μηδενική γραμμή) μπορούμε να συμπεράνουμε ότι οι διαφορές των βαρών τους είναι στατιστικά σημαντικές. Για παράδειγμα , έχουμε τέσσερεις μεγάλες διαφορές , δύο για τα χαρακτηριστικά 16 και δύο για το βάρος 20 για τις χρονιές $2010-2011, 2011-2012$ και $2009-2010, 2010-2011$ αντίστοιχα. Τέλος ,δίνουμε τα αποτελέσματα για τα βάρη για τις χρονιές $2009-2013$. Στον Πίνακα 5.9 παραθέτουμε όλα τα εκτιμώμενα βάρη για όλες τις χρονιές.

Βάρη	Μέσος	Τυπικό Σφάλμα	Περίοδος	Βάρη	Μέσος	Τυπικό Σφάλμα	Περίοδος
	0.02	0.02	2009		0.03	0.03	2009
	0.03	0.03	2010		0.03	0.03	2010
w_1	0.02	0.02	2011	w_{12}	0.03	0.03	2011
	0.05	0.02	2012		0.04	0.03	2012
	0.07	0.05	2013		0.04	0.04	2013
	0.02	0.02	2009		0.04	0.03	2009
	0.09	0.06	2010		0.03	0.03	2010
w_2	0.04	0.03	2011	w_{13}	0.05	0.04	2011
	0.04	0.04	2012		0.09	0.05	2012
	0.03	0.03	2013		0.07	0.05	2013
	0.03	0.03	2009		0.11	0.06	2009
	0.06	0.04	2010		0.04	0.03	2010
w_3	0.06	0.04	2011	w_{14}	0.05	0.04	2011
	0.10	0.06	2012		0.02	0.02	2012
	0.03	0.03	2013		0.05	0.04	2013
	0.03	0.03	2009		0.12	0.05	2009
	0.03	0.02	2010		0.07	0.05	2010
w_4	0.06	0.04	2011	w_{15}	0.03	0.02	2011
	0.04	0.03	2012		0.03	0.03	2012
	0.05	0.04	2013		0.03	0.03	2013
	0.03	0.02	2009		0.02	0.02	2009
	0.03	0.02	2010		0.04	0.03	2010
w_5	0.06	0.04	2011	w_{16}	0.18	0.07	2011
	0.04	0.03	2012		0.04	0.03	2012
	0.03	0.03	2013		0.03	0.02	2013
	0.06	0.04	2009		0.05	0.03	2009
	0.02	0.02	2010		0.03	0.02	2010
w_6	0.03	0.02	2011	w_{17}	0.03	0.02	2011
	0.05	0.03	2012		0.13	0.06	2012
	0.07	0.04	2013		0.01	0.01	2013
	0.05	0.04	2009		0.02	0.02	2009
	0.03	0.02	2010		0.02	0.02	2010
w_7	0.04	0.03	2011	w_{18}	0.04	0.03	2011
	0.05	0.04	2012		0.04	0.03	2012
	0.05	0.04	2013		0.03	0.03	2013
	0.05	0.04	2009		0.07	0.04	2009
	0.03	0.03	2010		0.06	0.04	2010
w_8	0.04	0.03	2011	w_{19}	0.03	0.03	2011
	0.03	0.02	2012		0.05	0.04	2012
	0.03	0.03	2013		0.03	0.02	2013
	0.02	0.02	2009		0.05	0.04	2009
	0.02	0.02	2010		0.19	0.06	2010
w_9	0.03	0.03	2011	w_{20}	0.04	0.03	2011
	0.02	0.02	2012		0.04	0.03	2012
	0.08	0.05	2013		0.06	0.04	2013
	0.02	0.02	2009		0.07	0.04	2009
	0.02	0.02	2010		0.17	0.04	2010
w_{10}	0.06	0.04	2011	w_{21}	0.04	0.03	2011
	0.02	0.01	2012		0.04	0.03	2012
	0.05	0.03	2013		0.12	0.04	2013
	0.08	0.05	2009				
	0.07	0.05	2010				
w_{11}	0.04	0.03	2011				
	0.04	0.04	2012				
	0.04	0.04	2013				

Πίνακας 5.9: Μέσοι και τυπικά σφάλματα της ύστερης κατανομής για τα εκτιμημένα βάρη για τις περιόδους 2009-2013.

Κεφάλαιο 6

Παράρτημα

Τέλος θα παραθέσουμε τους κώδικες οι οποίοι χρησιμοποιήθηκαν στα αριθμητικά πειράματα συνοδευόμενοι με επεξηγήσεις.

• **Κώδικας** (5.1)

```
library(rstan)
setwd("D:/Desktop")
df<-read.table("data.txt",header=T,sep="")

prog<-'data {
int<lower=0> N; // the number of rats
int<lower=0> T; // the number of time points
real x[T]; // day at which measurement was taken
real y[N,T]; // matrix of weight times time
real xbar; // the median number of days in the time series
}
parameters {
real alpha[N]; // the intercepts of rat weights
real beta[N]; // the slopes of rat weights

real mu_alpha; // the mean intercept
real mu_beta; // the mean slope

real<lower=0> sigmasq_y;
real<lower=0> sigmasq_alpha;
real<lower=0> sigmasq_beta;
}
transformed parameters {
real<lower=0> sigma_y; // sd of rat weight
real<lower=0> sigma_alpha; // sd of intercept distribution
real<lower=0> sigma_beta; // sd of slope distribution

sigma_y <- sqrt(sigmasq_y);
sigma_alpha <- sqrt(sigmasq_alpha);
sigma_beta <- sqrt(sigmasq_beta);
}
model {
mu_alpha ~ normal(0, 100); // non-informative prior
mu_beta ~ normal(0, 100); // non-informative prior
sigmasq_y ~ inv_gamma(0.001, 0.001); // conjugate prior of normal
sigmasq_alpha ~ inv_gamma(0.001, 0.001); // conjugate prior of normal
```

```
sigmasq_beta ~ inv_gamma(0.001, 0.001); // conjugate prior of normal
alpha ~ normal(mu_alpha, sigma_alpha); // all intercepts are normal
beta ~ normal(mu_beta, sigma_beta);  // all slopes are normal
for (n in 1:N) // for each sample
for (t in 1:T)  // for each time point
y[n,t] ~ normal(alpha[n] + beta[n] * (x[t] - xbar), sigma_y);

}
generated quantities {
// determine the intercept at time 0 (birth weight)
real alpha0;
alpha0 <- mu_alpha - xbar * mu_beta;
}'


days <- as.numeric(regmatches(colnames(df), regexpr("[0-9]*$", colnames(df))))
rat.data <- list(N = nrow(df), T = ncol(df), x = days,
y = df, xbar = median(days))



rat.model = stan(
model_code  = prog,
data = rat.data)


library("bayesplot")
library("ggplot2")
library("rstan")

lp_ncp <- log_posterior(rat.model)
np_ncp <- nuts_params(rat.model)

color_scheme_set("red")
mcmc_nuts_divergence(np_ncp, lp_ncp)


library(shinystan)
launch_shinystan(rat.model)
install.packages("shinystan")




predict.rat.weight <- function(rat.model, newdays) {
# newdays: vector of time points to consider
rat.fit <- extract(rat.model)
alpha <- rat.fit$alpha
beta <- rat.fit$beta
xbar <- 22 # hardcoded since not stored in rat.model
y <- lapply(newdays, function(t) alpha + beta * (t - 22))
return(y)
}
newdays <- seq(0, 100)
pred.weights <- predict.rat.weight(rat.model, newdays)
# extract means and standard deviations from posterior samples
```

```
pred.means <- lapply(pred.weights, function(x) apply(x, 2, mean))
pred.sd <- lapply(pred.weights, function(x) apply(x, 2, sd))
# create plotting data frame with 95% CI interval from sd
pred.df <- data.frame(Weight = unlist(pred.means),
Upr_Weight = unlist(pred.means) + 1.96 * unlist(pred.sd),
Lwr_Weight = unlist(pred.means) - 1.96 * unlist(pred.sd),
Day = unlist(lapply(newdays, function(x) rep(x, 30))),
Rat = rep(seq(1,30), length(newdays)))
# predicted mean weight of all rats
ggplot(pred.df, aes(x = Day, y = Weight, group = Rat)) +
geom_line()
```

• **Κώδικας** (5.2)

```
library(rstan)

schools_data <- list(
J = 8,
y = c(28,  8, -3,  7, -1,  1, 18, 12),
sigma = c(15, 10, 16, 11,  9, 11, 10, 18)
)


prog<-'
data {
int<lower=0> J;
real y[J];
real<lower=0> sigma[J];
}

parameters {
real mu;
real<lower=0> tau;
real theta[J];
}

model {
mu ~ normal(0, 5);
tau ~ cauchy(0, 5);
theta ~ normal(mu, tau);
y ~ normal(theta, sigma);
}'



fit<-stan(model_code=prog,data=schools_data,iter=1200,warmup=500,chain=1,
seed=483892929, refresh=1200)

params_cp <- as.data.frame(extract(fit, permuted=FALSE))
names(params_cp) <- gsub("chain:1.", "", names(params_cp), fixed = TRUE)
names(params_cp) <- gsub("[", ".", names(params_cp), fixed = TRUE)
names(params_cp) <- gsub("]", "", names(params_cp), fixed = TRUE)
params_cp$iter <- 1:700

par(mar = c(4, 4, 0.5, 0.5))
plot(params_cp$iter, log(params_cp$tau), col=c_dark, pch=16, cex=0.8,
```

```
xlab="Iteration", ylab="log(tau)", ylim=c(-6, 4))



running_means <- sapply(params_cp$iter, function(n) mean(log(params_cp$tau)[1:n]))

par(mar = c(4, 4, 0.5, 0.5))
plot(params_cp$iter, running_means, col=c_dark, pch=16, cex=0.8, ylim=c(0, 2),
xlab="Iteration", ylab="MCMC mean of log(tau)")
abline(h=0.7657852, col="grey", lty="dashed", lwd=3)

divergent <- get_sampler_params(fit, inc_warmup=FALSE)[[1]][,'divergent__']
sum(divergent)

sum(divergent) / 700



params_cp$divergent <- divergent

div_params_cp <- params_cp[params_cp$divergent == 1,]
nondiv_params_cp <- params_cp[params_cp$divergent == 0,]

par(mar = c(4, 4, 0.5, 0.5))
plot(nondiv_params_cp$theta.1, log(nondiv_params_cp$tau),
col=c_dark, pch=16, cex=0.8, xlab="theta.1", ylab="log(tau)",
xlim=c(-20, 50), ylim=c(-6,4))
points(div_params_cp$theta.1, log(div_params_cp$tau),
col="green", pch=16, cex=0.8)

# Default step size delta = 0.80 , with many iterations

fit_cp80 <- stan(model_code=prog, data=schools_data,
iter=11000, warmup=1000, chains=1, seed=483892929,
refresh=11000)

params_cp80 <- as.data.frame(extract(fit_cp80, permuted=FALSE))
names(params_cp80) <- gsub("chain:1.", "", names(params_cp80), fixed = TRUE)
names(params_cp80) <- gsub("[", ".", names(params_cp80), fixed = TRUE)
names(params_cp80) <- gsub("]", "", names(params_cp80), fixed = TRUE)
params_cp80$iter <- 1:10000

par(mar = c(4, 4, 0.5, 0.5))
plot(params_cp80$iter, log(params_cp80$tau), col=c_dark, pch=16, cex=0.8,
xlab="Iteration", ylab="log(tau)", ylim=c(-6, 4))



running_means_cp80 <- sapply(1:1000, function(n) mean(log(params_cp80$tau)[1:(10*n)]))

par(mar = c(4, 4, 0.5, 0.5))
plot(10*(1:1000), running_means_cp80, col=c_dark, pch=16, cex=0.8, ylim=c(0, 2),
xlab="Iteration", ylab="MCMC mean of log(tau)")
abline(h=0.7657852, col="grey", lty="dashed", lwd=3)
```

```
divergent <- get_sampler_params(fit_cp80, inc_warmup=FALSE)[[1]][,'divergent__']
sum(divergent)

sum(divergent) / 10000


params_cp80$divergent <- divergent

div_params_cp <- params_cp80[params_cp80$divergent == 1,]
nondiv_params_cp <- params_cp80[params_cp80$divergent == 0,]

par(mar = c(4, 4, 0.5, 0.5))
plot(nondiv_params_cp$theta.1, log(nondiv_params_cp$tau),
col=c_dark, pch=16, cex=0.8, xlab="theta.1", ylab="log(tau)",
xlim=c(-20, 50), ylim=c(-6,4))
points(div_params_cp$theta.1, log(div_params_cp$tau),
col="green", pch=16, cex=0.8)


#Decrease step size  delta = 0.85 , with many iterations

fit_cp85 <- stan(model_code=prog, data=schools_data,
iter=11000, warmup=1000, chains=1, seed=483892929,
refresh=11000, control=list(adapt_delta=0.85))

#Decrease step size  delta = 0.90 , with many iterations

fit_cp90 <- stan(model_code=prog, data=schools_data,
iter=11000, warmup=1000, chains=1, seed=483892929,
refresh=11000, control=list(adapt_delta=0.90))

#Decrease step size  delta = 0.95 , with many iterations

fit_cp95 <- stan(model_code=prog, data=schools_data,
iter=11000, warmup=1000, chains=1, seed=483892929,
refresh=11000, control=list(adapt_delta=0.95))

#Decrease step size  delta = 0.99 , with many iterations

fit_cp99 <- stan(model_code=prog, data=schools_data,
iter=11000, warmup=1000, chains=1, seed=483892929,
refresh=11000, control=list(adapt_delta=0.99))

common_breaks=14 * (0:60) / 60 - 9

p_cp80 <- hist(log(extract(fit_cp80)$tau), breaks=common_breaks, plot=FALSE)
p_cp90 <- hist(log(extract(fit_cp90)$tau), breaks=common_breaks, plot=FALSE)
p_cp99 <- hist(log(extract(fit_cp99)$tau), breaks=common_breaks, plot=FALSE)

par(mar = c(4, 4, 0.5, 0.5))
plot(p_cp99, col=c_dark, main="", xlab="log(tau)", yaxt='n', ann=FALSE)
plot(p_cp90, col=c_mid, add=TRUE)
plot(p_cp80, col=c_light, add=TRUE)
legend("topleft",
c("Centered, delta=0.80", "Centered, delta=0.90", "Centered, delta=0.99"),
```

```
fill=c(c_light, c_mid, c_dark), bty="n")


params_cp99 <- as.data.frame(extract(fit_cp99, permuted=FALSE))
names(params_cp99) <- gsub("chain:1.", "", names(params_cp99), fixed = TRUE)
names(params_cp99) <- gsub("[", ".", names(params_cp99), fixed = TRUE)
names(params_cp99) <- gsub("]", "", names(params_cp99), fixed = TRUE)

divergent <- get_sampler_params(fit_cp99, inc_warmup=FALSE)[[1]][,'divergent__']
params_cp99$divergent <- divergent

div_params_cp99 <- params_cp99[params_cp99$divergent == 1,]
nondiv_params_cp99 <- params_cp99[params_cp99$divergent == 0,]

par(mar = c(4, 4, 0.5, 0.5))
plot(nondiv_params_cp99$theta.1, log(nondiv_params_cp99$tau),
xlab="theta.1", ylab="log(tau)", xlim=c(-20, 50), ylim=c(-6,4),
col=c_dark, pch=16, cex=0.8)
points(div_params_cp99$theta.1, log(div_params_cp99$tau),
col="green", pch=16, cex=0.8)


par(mar = c(4, 4, 0.5, 0.5))
plot(params_cp99$theta.1, log(params_cp99$tau),
xlab="theta.1", ylab="log(tau)", xlim=c(-20, 50), ylim=c(-6,4),
col=c_dark, pch=16, cex=0.8)
points(params_cp80$theta.1, log(params_cp80$tau), col=c_light, pch=16, cex=0.8)
legend("bottomright", c("Centered, delta=0.80", "Centered, delta=0.99"),
fill=c(c_light, c_dark), border="white", bty="n")



params_cp90 <- as.data.frame(extract(fit_cp90, permuted=FALSE))
names(params_cp90) <- gsub("chain:1.", "", names(params_cp90), fixed = TRUE)
names(params_cp90) <- gsub("[", ".", names(params_cp90), fixed = TRUE)
names(params_cp90) <- gsub("]", "", names(params_cp90), fixed = TRUE)

running_means_cp90 <- sapply(1:1000, function(n) mean(log(params_cp90$tau)[1:(10*n)]))
running_means_cp99 <- sapply(1:1000, function(n) mean(log(params_cp99$tau)[1:(10*n)]))

plot(10*(1:1000), running_means_cp80, col=c_light, pch=16, cex=0.8, ylim=c(0, 2),
xlab="Iteration", ylab="MCMC mean of log(tau)")
points(10*(1:1000), running_means_cp90, col=c_mid, pch=16, cex=0.8)
points(10*(1:1000), running_means_cp99, col=c_dark, pch=16, cex=0.8)
abline(h=0.7657852, col="grey", lty="dashed", lwd=3)
legend("bottomright",
c("Centered, delta=0.80", "Centered, delta=0.90", "Centered, delta=0.99"),
fill=c(c_light, c_mid, c_dark), border="white", bty="n")

# Non-centered parameterization

prod<-'data {
int<lower=0> J;
real y[J];
real<lower=0> sigma[J];
}
```

```
parameters {
real mu;
real<lower=0> tau;
real theta_tilde[J];
}

transformed parameters {
real theta[J];
for (j in 1:J)
theta[j] = mu + tau * theta_tilde[j];
}

model {
mu ~ normal(0, 5);
tau ~ cauchy(0, 5);
theta_tilde ~ normal(0, 1);
y ~ normal(theta, sigma);
}'


fit_ncp80 <- stan(model_code=prod, data=schools_data,
iter=11000, warmup=1000, chains=1, seed=483892929,
refresh=11000)

params_ncp80 <- as.data.frame(extract(fit_ncp80, permuted=FALSE))
names(params_ncp80) <- gsub("chain:1.", "", names(params_ncp80), fixed = TRUE)
names(params_ncp80) <- gsub("[", ".", names(params_ncp80), fixed = TRUE)
names(params_ncp80) <- gsub("]", "", names(params_ncp80), fixed = TRUE)
params_ncp80$iter <- 1:10000

par(mar = c(4, 4, 0.5, 0.5))
plot(params_ncp80$iter, log(params_ncp80$tau), col=c_dark, pch=16, cex=0.8,
xlab="Iteration", ylab="log(tau)", ylim=c(-6, 4))

divergent <- get_sampler_params(fit_ncp80, inc_warmup=FALSE)[[1]][,'divergent__']
sum(divergent)


sum(divergent) / 10000

divergent <- get_sampler_params(fit_ncp80, inc_warmup=FALSE)[[1]][,'divergent__']
params_ncp80$divergent <- divergent

div_params_ncp <- params_ncp80[params_ncp80$divergent == 1,]
nondiv_params_ncp <- params_ncp80[params_ncp80$divergent == 0,]

par(mar = c(4, 4, 0.5, 0.5))
plot(nondiv_params_ncp$theta.1, log(nondiv_params_ncp$tau),
xlab="theta.1", ylab="log(tau)", xlim=c(-20, 50), ylim=c(-6,4),
col=c_dark, pch=16, cex=0.8)
points(div_params_ncp$theta.1, log(div_params_ncp$tau),
col="green", pch=16, cex=0.8)
```

```
fit_ncp90 <- stan(model_code=prod, data=schools_data,
iter=11000, warmup=1000, chains=1, seed=483892929,
refresh=11000, control=list(adapt_delta=0.90))



params_ncp90 <- as.data.frame(extract(fit_ncp90, permuted=FALSE))
names(params_ncp90) <- gsub("chain:1.", "", names(params_ncp90), fixed = TRUE)
names(params_ncp90) <- gsub("[", ".", names(params_ncp90), fixed = TRUE)
names(params_ncp90) <- gsub("]", "", names(params_ncp90), fixed = TRUE)

par(mar = c(4, 4, 0.5, 0.5))
plot(params_ncp90$theta.1, log(params_ncp90$tau),
xlab="theta.1", ylab="log(tau)", xlim=c(-20, 50), ylim=c(-6,4),
col=c_dark_highlight, pch=16, cex=0.8)
points(params_cp99$theta.1, log(params_cp99$tau), col=c_dark, pch=16, cex=0.8)
points(params_cp90$theta.1, log(params_cp90$tau), col=c_mid, pch=16, cex=0.8)
legend("bottomright", c("Centered, delta=0.90", "Centered, delta=0.99",
"Non-Centered, delta=0.90"),
fill=c(c_mid, c_dark, c_dark_highlight), border="white", bty="n")



running_means_ncp <- sapply(1:1000, function(n) mean(log(params_ncp90$tau)[1:(10*n)]))

par(mar = c(4, 4, 0.5, 0.5))
plot(10*(1:1000), running_means_cp90, col=c_mid, pch=16, cex=0.8, ylim=c(0, 2),
xlab="Iteration", ylab="MCMC mean of log(tau)")
points(10*(1:1000), running_means_cp99, col=c_dark, pch=16, cex=0.8)
points(10*(1:1000), running_means_ncp, col=c_dark_highlight, pch=16, cex=0.8)
abline(h=0.7657852, col="grey", lty="dashed", lwd=3)
legend("bottomright", c("Centered, delta=0.90", "Centered, delta=0.99",
"Non-Centered, delta=0.90"),
fill=c(c_mid, c_dark, c_dark_highlight), border="white", bty="n")
```

• **Κώδικας** (5.3)

```
library(rstan)

mu <- c(-2.75, 2.75);
sigma <- c(1, 1);
lambda <- 0.4

set.seed(689934)

N <- 1000
z <- rbinom(N, 1, lambda) + 1;
y <- rnorm(N, mu[z], sigma[z]);


prog<-'data {
int<lower = 0> N;
vector[N] y;
}
```

```
parameters {
vector[2] mu;
real<lower=0> sigma[2];
real<lower=0, upper=1> theta;
}

model {
sigma ~ normal(0, 2);
mu ~ normal(0, 2);
theta ~ beta(5, 5);
for (n in 1:N)
target += log_mix(theta,
normal_lpdf(y[n] | mu[1], sigma[1]),
normal_lpdf(y[n] | mu[2], sigma[2]));
}
'


input_data <- c("N","y","z")

degenerate_fit <- stan(model_code=prog, data=input_data,
chains=4, seed=483892929, refresh=2000)

print(degenerate_fit)

params1 <- as.data.frame(extract(degenerate_fit, permuted=FALSE)[,1,])
params2 <- as.data.frame(extract(degenerate_fit, permuted=FALSE)[,2,])
params3 <- as.data.frame(extract(degenerate_fit, permuted=FALSE)[,3,])
params4 <- as.data.frame(extract(degenerate_fit, permuted=FALSE)[,4,])

par(mar = c(4, 4, 0.5, 0.5))
plot(params1$"mu[1]", params1$"mu[2]", col=c_dark_highlight_trans, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
points(params2$"mu[1]", params2$"mu[2]", col=c_dark_trans, pch=16, cex=0.8)
points(params3$"mu[1]", params3$"mu[2]", col=c_mid_highlight_trans, pch=16, cex=0.8)
points(params4$"mu[1]", params4$"mu[2]", col=c_mid_trans, pch=16, cex=0.8)
lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)
legend("topright", c("Chain 1", "Chain 2", "Chain 3", "Chain 4"),
fill=c(c_dark_highlight_trans, c_dark_trans,
c_mid_highlight_trans, c_mid_trans), box.lty=0, inset=0.0005)



prod<-'data {
int<lower = 0> N;
vector[N] y;
}

parameters {
vector[2] mu;
real<lower=0> sigma[2];
real<lower=0, upper=1> theta;
}
```

```
model {
sigma ~ normal(0, 2);
mu[1] ~ normal(4, 0.5);
mu[2] ~ normal(-4, 0.5);
theta ~ beta(5, 5);
for (n in 1:N)
target += log_mix(theta,
normal_lpdf(y[n] | mu[1], sigma[1]),
normal_lpdf(y[n] | mu[2], sigma[2]));
}'


asym_fit <- stan(model_code=prod, data=input_data,
chains=4, seed=483892929, refresh=2000)

print(asym_fit)


params1 <- as.data.frame(extract(asym_fit, permuted=FALSE)[,1,])
params2 <- as.data.frame(extract(asym_fit, permuted=FALSE)[,2,])
params3 <- as.data.frame(extract(asym_fit, permuted=FALSE)[,3,])
params4 <- as.data.frame(extract(asym_fit, permuted=FALSE)[,4,])

par(mar = c(4, 4, 0.5, 0.5))
plot(params1$"mu[1]", params1$"mu[2]", col=c_dark_highlight_trans, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
points(params2$"mu[1]", params2$"mu[2]", col=c_dark_trans, pch=16, cex=0.8)
points(params3$"mu[1]", params3$"mu[2]", col=c_mid_highlight_trans, pch=16, cex=0.8)
points(params4$"mu[1]", params4$"mu[2]", col=c_mid_trans, pch=16, cex=0.8)
lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)
legend("topright", c("Chain 1", "Chain 2", "Chain 3", "Chain 4"),
fill=c(c_dark_highlight_trans, c_dark_trans,
c_mid_highlight_trans, c_mid_trans), box.lty=0, inset=0.0005)



prof<-'data {
int<lower = 0> N;
vector[N] y;
}

parameters {
ordered[2] mu;
real<lower=0> sigma[2];
real<lower=0, upper=1> theta;
}

#Normal_lpdf uses the log-likelihood

model {
sigma ~ normal(0, 2);
mu ~ normal(0, 2);
theta ~ beta(5, 5);
for (n in 1:N)
target += log_mix(theta,
normal_lpdf(y[n] | mu[1], sigma[1]),
```

```
normal_lpdf(y[n] | mu[2], sigma[2]));
}'


ordered_fit <- stan(model_code=prof, data=input_data,
chains=4, seed=483892929, refresh=2000)


print(ordered_fit)


params1 <- as.data.frame(extract(ordered_fit, permuted=FALSE)[,1,])
params2 <- as.data.frame(extract(ordered_fit, permuted=FALSE)[,2,])
params3 <- as.data.frame(extract(ordered_fit, permuted=FALSE)[,3,])
params4 <- as.data.frame(extract(ordered_fit, permuted=FALSE)[,4,])

par(mar = c(4, 4, 0.5, 0.5))
plot(params1$"mu[1]", params1$"mu[2]", col=c_dark_highlight_trans, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
points(params2$"mu[1]", params2$"mu[2]", col=c_dark_trans, pch=16, cex=0.8)
points(params3$"mu[1]", params3$"mu[2]", col=c_mid_highlight_trans, pch=16, cex=0.8)
points(params4$"mu[1]", params4$"mu[2]", col=c_mid_trans, pch=16, cex=0.8)
lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)
legend("topright", c("Chain 1", "Chain 2", "Chain 3", "Chain 4"),
fill=c(c_dark_highlight_trans, c_dark_trans,
c_mid_highlight_trans, c_mid_trans), box.lty=0, inset=0.0005)




N <- 1000
mu <- c(-0.75, 0.75);
sigma <- c(1, 1);
lambda <- 0.4
z <- rbinom(N, 1, lambda) + 1;
y <- rnorm(N, mu[z], sigma[z]);
theta <- 0.25
stack<-'data {
int<lower = 0> N;
vector[N] y;
real<lower=0, upper=1> theta;
}

parameters {
vector[2] mu;
real<lower=0> sigma[2];
}

model {
sigma ~ normal(0, 2);
mu ~ normal(0, 2);
for (n in 1:N)
target += log_mix(theta,
normal_lpdf(y[n] | mu[1], sigma[1]),
normal_lpdf(y[n] | mu[2], sigma[2]));
```

```
}'


input_data=c("y","N","theta")


singular_fit <- stan(model_code=stack, data=input_data,
chains=1, iter=11000, warmup=1000, seed=483892929,
refresh=11000)

c_light <- c("#DCBCBC")
c_light_highlight <- c("#C79999")
c_mid <- c("#B97C7C")
c_mid_highlight <- c("#A25050")
c_dark <- c("#8F2727")
c_dark_highlight <- c("#7C0000")

params25 <- as.data.frame(extract(singular_fit, permuted=FALSE)[,1,])

par(mar = c(4, 4, 0.5, 0.5))
plot(params25$"mu[1]", params25$"mu[2]", col=c_dark_highlight, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)



theta <- 0.5
singular_fit <- stan(model_code=stack, data=input_data,
chains=1, iter=11000, warmup=1000, seed=483892929,
refresh=11000)


params25 <- as.data.frame(extract(singular_fit, permuted=FALSE)[,1,])

par(mar = c(4, 4, 0.5, 0.5))
plot(params25$"mu[1]", params25$"mu[2]", col=c_dark_highlight, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)


theta <- 0.5
input_data=c("y","N","theta")
singular_fit <- stan(model_code=stack, data=input_data,
chains=1, iter=11000, warmup=1000, seed=483892929,
refresh=11000)
params50 <- as.data.frame(extract(singular_fit, permuted=FALSE)[,1,])


theta <- 0.75
input_data=c("y","N","theta")
singular_fit <- stan(model_code=stack, data=input_data,
chains=1, iter=11000, warmup=1000, seed=483892929,
refresh=11000)
params75 <- as.data.frame(extract(singular_fit, permuted=FALSE)[,1,])

par(mar = c(4, 4, 0.5, 0.5))
```

```
plot(params25$"mu[1]", params25$"mu[2]", col=c_dark_highlight, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
points(params50$"mu[1]", params50$"mu[2]", col=c_mid_highlight, pch=16, cex=0.8)
points(params75$"mu[1]", params75$"mu[2]", col=c_light_highlight, pch=16, cex=0.8)

lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)

legend("topleft", c("Theta = 0.25", "Theta = 0.5", "Theta = 0.75"),
fill=c(c_dark_highlight, c_mid_highlight, c_light_highlight), bty="n")


input_data=c("y","N","z")
singular_fit <- stan(model_code=prog, data=input_data,
chains=1, iter=11000, warmup=1000, seed=483892929,
refresh=11000)
params1 <- as.data.frame(extract(singular_fit, permuted=FALSE)[,1,])

breaks=(0:50) / 50
prior_hist <- hist(rbeta(10000, 5, 5), breaks=breaks, plot=FALSE)
post_hist <- hist(params1$theta, breaks=breaks, plot=FALSE)

par(mar = c(4, 4, 0.5, 0.5))
plot(prior_hist, col=c_light_trans, border=c_light_highlight_trans,
main="", xlab="theta", yaxt='n', ann=FALSE)
plot(post_hist, col=c_dark_trans, border=c_dark_highlight_trans, add=TRUE)
legend("topright", c("Posterior", "Prior"),
fill=c(c_dark_trans, c_light_trans), bty="n")


par(mar = c(4, 4, 0.5, 0.5))
plot(params1$"mu[1]", params1$"mu[2]", col=c_dark_highlight, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)

print(singular_fit)

singular_fit <- stan(model_code = prof, data=input_data,
chains=1, iter=11000, warmup=1000, seed=483892929,
refresh=11000)

params1 <- as.data.frame(extract(singular_fit, permuted=FALSE)[,1,])

par(mar = c(4, 4, 0.5, 0.5))
plot(params1$"mu[1]", params1$"mu[2]", col=c_dark_highlight, pch=16, cex=0.8,
xlab="mu1", xlim=c(-3, 3), ylab="mu2", ylim=c(-3, 3))
lines(0.08*(1:100) - 4, 0.08*(1:100) - 4, col="grey", lw=2)
```

- **Κώδικας** (5.4)

```
library(MCMCpack)
library(rstan)
library("bayesplot")
library("rstanarm")
library("ggplot2")

my_data<-read.table("DATAFOUSK.txt", header = TRUE, sep = "")
```

```
Y<-my_data[,21]
X<-my_data[,1:20]

posterior <- as.matrix(fit2)
plot_title <- ggtitle("Posterior distributions",
"with medians and 80% intervals")
mcmc_areas(posterior,
pars = c("coef[1]", "coef[2]", "coef[3]", "coef[4]"),
prob = 0.8) + plot_title

fit2 <- stan_demo("eight_schools", warmup = 300, iter = 700)
posterior2 <- extract(fit2, inc_warmup = TRUE, permuted = FALSE)

color_scheme_set("mix-blue-pink")
p <- mcmc_trace(posterior,  pars = c("coef[1]", "coef[2]"), n_warmup = 1000,
facet_args = list(nrow = 2, labeller = label_parsed))
p + facet_text(size = 15)

color_scheme_set("red")
np <- nuts_params(fit2)
mcmc_nuts_energy(np) + ggtitle("NUTS Energy Diagnostic")
################################################
################################################
#######oles oi periodoi mazi############################
Y<-my_data[,21]
X<-my_data[,1:20]
N<-dim(X)[1]
K<-dim(X)[2]
C<-rep(1,length=(K+1))

stan_beta <- "
data {
int<lower=1> N;
int<lower=1> K;
vector<lower=1>[K+1] C;
vector<lower=0,upper=1>[N] Y;
matrix[N,K] X;
}

parameters {
simplex[K+1] beta;
real<lower=0> phi;
vector<lower=0,upper=1>[N] Z;

}

transformed parameters{
vector<lower=0,upper=1>[N] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N] A;             // parameter for beta distn
vector<lower=0>[N] B;             // parameter for beta distn

for (i in 1:N) {
mu[i]  = X[i,] * beta[1:K] + beta[K+1]*Z[i]  ;
}
sig = 1/phi  ;
```

```
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);
}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
beta ~ dirichlet(C);
// likelihood
Y ~ beta(A, B);
}
"


fit_all<-stan(model_code=stan_beta,data=c("N","Y","X","K","C"),iter=2000,warmup = 1000,chain=4)



##############################################
#beta<-c()
#beta<-rdirichlet(dim(X)[1],rep(1,length=21))
#z<-rbeta(dim(X)[1],0.5,0.5)

#for(i in 1:dim(X)[1]){
#mu[i]<- as.numeric(X[i,])%*% beta[1:20]+beta[21]*z[i]
#}
Y<-my_data[1:183,21]
X<-my_data[1:183,1:20]
N<-dim(X)[1]
K<-dim(X)[2]
C<-rep(1,length=(K+1))

stan_beta <- "
data {
int<lower=1> N;
int<lower=1> K;
vector<lower=1>[K+1] C;
vector<lower=0,upper=1>[N] Y;
matrix[N,K] X;
}

parameters {
simplex[K+1] beta;
real<lower=0> phi;
vector<lower=0,upper=1>[N] Z;

}

transformed parameters{
vector<lower=0,upper=1>[N] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N] A;             // parameter for beta distn
vector<lower=0>[N] B;             // parameter for beta distn

for (i in 1:N) {
```

```
mu[i]  = X[i,] * beta[1:K] + beta[K+1]*Z[i]  ;
}
sig = 1/phi  ;
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);
}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
beta ~ dirichlet(C);
// likelihood
Y ~ beta(A, B);
}
"


fit<-stan(model_code=stan_beta,data=c("N","Y","X","K","C"),iter=2000,warmup = 1000,chain=4)

launch_shinystan(fit)
np_ncp <- nuts_params(fit2)
color_scheme_set("red")
mcmc_nuts_energy(np_ncp)
rhats <- rhat(fit2)
color_scheme_set("brightblue") # see help("color_scheme_set")
mcmc_rhat(rhats[1:21])
color_scheme_set("red")
lp_ncp <- log_posterior(fit2)
mcmc_nuts_divergence(np_ncp, lp_ncp,chain=4)


####################MIA XRONIKH PERIODOS############
X_1<-my_data[which(my_data[,22]==0),1:20]
Y_1<-my_data[which(my_data[,22]==0),21]
N_1<-dim(X_1)[1]
K_1<-dim(X_1)[2]
C_1<-rep(1,length=(K_1+1))
which(X_1>1,arr.ind = TRUE)
X_1[4,18]<-mean(X[,18])
which(X_1<0,arr.ind = TRUE)

#precision variable,Z real

stan_beta_1 <- "
data {
int<lower=1> N_1;
int<lower=1> K_1;
vector<lower=1>[K_1+1] C_1;
vector<lower=0,upper=1>[N_1] Y_1;
matrix[N_1,K_1] X_1;
}

parameters {
simplex[K_1+1] beta;
real<lower=0> phi;
```

```
real<lower=0,upper=1> Z;

}

transformed parameters{
vector<lower=0,upper=1>[N_1] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N_1] A;             // parameter for beta distn
vector<lower=0>[N_1] B;             // parameter for beta distn

for (i in 1:N_1) {
mu[i]  = X_1[i,] * beta[1:K_1] + beta[K_1+1]*Z  ;
}
sig = 1/phi  ;
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);
}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
beta ~ dirichlet(C_1);
// likelihood
Y_1 ~ beta(A, B);
}
"

fit1<-stan(model_code=stan_beta_1,data=c("N_1","Y_1","X_1","K_1","C_1"),iter=3000,warmup = 1000,
control = list(adapt_delta = 0.85),chain=4)

#precesion variable ,vector Z
stan_beta_2 <- "
data {
int<lower=1> N_1;
int<lower=1> K_1;
vector<lower=1>[K_1+1] C_1;
vector<lower=0,upper=1>[N_1] Y_1;
matrix[N_1,K_1] X_1;
}

parameters {
simplex[K_1+1] coef;
real<lower=0> phi;
vector<lower=0,upper=1>[N_1] Z;

}

transformed parameters{
vector<lower=0,upper=1>[N_1] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N_1] A;             // parameter for beta distn
vector<lower=0>[N_1] B;             // parameter for beta distn


for (i in 1:N_1) {
```

```
mu[i]  = X_1[i,] * coef[1:K_1] + coef[K_1+1]*Z[i]  ;
}
sig = 1/phi  ;
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);

}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
coef ~ dirichlet(C_1);
// likelihood
Y_1 ~ beta(A, B);
}
"

fit2<-stan(model_code=stan_beta_2,data=c("N_1","Y_1","X_1","K_1","C_1"),iter=2000,chain=4)
extract(fit2, permuted = FALSE)
launch_shinystan(fit2)

np_ncp <- nuts_params(fit2)
color_scheme_set("red")
mcmc_nuts_energy(np_ncp)
rhats <- rhat(fit2)
color_scheme_set("brightblue") # see help("color_scheme_set")
mcmc_rhat(rhats[1:21])
color_scheme_set("red")
lp_ncp <- log_posterior(fit2)

mcmc_nuts_divergence(np_ncp, lp_ncp,chain=4)




parameters1<-extract(fit2, "coef[3]", permuted = TRUE, inc_warmup = FALSE,
include = TRUE)
plot(fit, plotfun = "hist", pars = "coef[1]", include = FALSE)


stan_hist(fit2, pars="coef[1]", include = TRUE, unconstrain = FALSE,
inc_warmup = FALSE) + geom_vline(xintercept = 0.06)


#################DEUTERH XRONIKH PERIODOS###########
X_2<-my_data[which(my_data[,22]==1),1:20]
Y_2<-my_data[which(my_data[,22]==1),21]
N_2<-dim(X_2)[1]
K_2<-dim(X_2)[2]
C_2<-rep(1,length=(K_2+1))
which(X_2>1,arr.ind = TRUE)
which(X_2<0,arr.ind = TRUE)
```

```
#precesion variable ,vector Z
stan_beta_2 <- "
data {
int<lower=1> N_2;
int<lower=1> K_2;
vector<lower=1>[K_2+1] C_2;
vector<lower=0,upper=1>[N_2] Y_2;
matrix[N_2,K_2] X_2;
}

parameters {
simplex[K_2+1] coef;
real<lower=0> phi;
vector<lower=0,upper=1>[N_2] Z;

}

transformed parameters{
vector<lower=0,upper=1>[N_2] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N_2] A;             // parameter for beta distn
vector<lower=0>[N_2] B;             // parameter for beta distn


for (i in 1:N_2) {
mu[i]  = X_2[i,] * coef[1:K_2] + coef[K_2+1]*Z[i]  ;
}
sig = 1/phi  ;
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);

}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
coef ~ dirichlet(C_2);
// likelihood
Y_2 ~ beta(A, B);
}
"

fit12<-stan(model_code=stan_beta_2,data=c("N_2","Y_2","X_2","K_2","C_2"),iter=2000,chain=4)
np_ncp <- nuts_params(fit2)

parameters2<-extract(fit12, "coef[1]", permuted = TRUE, inc_warmup = FALSE,
include = TRUE)
stan_hist(fit12, pars="coef[1]", include = TRUE, unconstrain = FALSE,
inc_warmup = FALSE) + geom_vline(xintercept = 0.06)

mean(which(parameters1-parameters2<0)

#######3trith periodo#####

X_3<-my_data[which(my_data[,22]==2),1:20]
```

```
Y_3<-my_data[which(my_data[,22]==2),21]
N_3<-dim(X_3)[1]
K_3<-dim(X_3)[2]
C_3<-rep(1,length=(K_3+1))
which(X_3>1,arr.ind = TRUE)
which(X_3<0,arr.ind = TRUE)



#precesion variable ,vector Z
stan_beta_2 <- "
data {
int<lower=1> N_3;
int<lower=1> K_3;
vector<lower=1>[K_3+1] C_3;
vector<lower=0,upper=1>[N_3] Y_3;
matrix[N_3,K_3] X_3;
}

parameters {
simplex[K_3+1] coef;
real<lower=0> phi;
vector<lower=0,upper=1>[N_3] Z;

}

transformed parameters{
vector<lower=0,upper=1>[N_3] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N_3] A;             // parameter for beta distn
vector<lower=0>[N_3] B;             // parameter for beta distn


for (i in 1:N_3) {
mu[i]  = X_3[i,] * coef[1:K_3] + coef[K_3+1]*Z[i]  ;
}
sig = 1/phi  ;
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);

}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
coef ~ dirichlet(C_3);
// likelihood
Y_3 ~ beta(A, B);
}
"
fit22<-stan(model_code=stan_beta_2,data=c("N_3","Y_3","X_3","K_3","C_3"),iter=2000,chain=4)



########Tetarth periodos#######
```

```
X_4<-my_data[which(my_data[,22]==3),1:20]
Y_4<-my_data[which(my_data[,22]==3),21]
N_4<-dim(X_4)[1]
K_4<-dim(X_4)[2]
C_4<-rep(1,length=(K_4+1))
which(X_4>1,arr.ind = TRUE)
which(X_4<0,arr.ind = TRUE)



#precesion variable ,vector Z
stan_beta_2 <- "
data {
int<lower=1> N_4;
int<lower=1> K_4;
vector<lower=1>[K_4+1] C_4;
vector<lower=0,upper=1>[N_4] Y_4;
matrix[N_4,K_4] X_4;
}

parameters {
simplex[K_4+1] coef;
real<lower=0> phi;
vector<lower=0,upper=1>[N_4] Z;

}

transformed parameters{
vector<lower=0,upper=1>[N_4] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N_4] A;             // parameter for beta distn
vector<lower=0>[N_4] B;             // parameter for beta distn


for (i in 1:N_4) {
mu[i]  = X_4[i,] * coef[1:K_4] + coef[K_4+1]*Z[i]  ;
}
sig = 1/phi  ;
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);

}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
coef ~ dirichlet(C_4);
// likelihood
Y_4 ~ beta(A, B);
}
"
fit32<-stan(model_code=stan_beta_2,data=c("N_4","Y_4","X_4","K_4","C_4"),iter=2000,chain=4)

library(msme)
dim(X_4)
```

```
Z=rbeta(87,0.5,0.5)
data_irls<-as.data.frame(cbind(cbind(X_4,Z),Y_4))
irls.poi<-irls(Y_4~.,family=c("gaussian"),link="identity", data=data_irls)


########Pempth periodos#######
X_5<-my_data[which(my_data[,22]==4),1:20]
Y_5<-my_data[which(my_data[,22]==4),21]
N_5<-dim(X_5)[1]
K_5<-dim(X_5)[2]
C_5<-rep(1,length=(K_5+1))
which(X_5>1,arr.ind = TRUE)
which(X_5<0,arr.ind = TRUE)



#precesion variable ,vector Z
stan_beta_2 <- "
data {
int<lower=1> N_5;
int<lower=1> K_5;
vector<lower=1>[K_5+1] C_5;
vector<lower=0,upper=1>[N_5] Y_5;
matrix[N_5,K_5] X_5;
}

parameters {
simplex[K_5+1] coef;
real<lower=0> phi;
vector<lower=0,upper=1>[N_5] Z;

}

transformed parameters{
vector<lower=0,upper=1>[N_5] mu;    // transformed linear predictor for mean of beta distribution
real<lower=0> sig;          // transformed linear predictor for precision of beta distribution
vector<lower=0>[N_5] A;             // parameter for beta distn
vector<lower=0>[N_5] B;             // parameter for beta distn


for (i in 1:N_5) {
mu[i]  = X_5[i,] * coef[1:K_5] + coef[K_5+1]*Z[i]  ;
}
sig = 1/phi  ;
A =((1-mu).* mu .* mu- mu*sig*sig)/(sig*sig);
B = (1-mu).*(mu-mu .* mu -sig*sig)/(sig*sig);

}

model {
// priors
phi ~ gamma(0.1,0.1);
Z ~ beta(0.5,0.5);
coef ~ dirichlet(C_5);
// likelihood
Y_5 ~ beta(A, B);
```

```
}
"
fit42<-stan(model_code=stan_beta_2,data=c("N_5","Y_5","X_5","K_5","C_5"),iter=2000,chain=4)


#####
pi<-c()
for(i in paste0(rep("coef[", 21),1:21,rep("]"))){
parameters1<-extract(fit2, i, permuted = TRUE, inc_warmup = FALSE,
include = TRUE)
parameters2<-extract(fit12, i, permuted = TRUE, inc_warmup = FALSE,
include = TRUE)
pi[i]<- min(mean((parameters1[[1]]-parameters2[[1]]>0)),mean((parameters1[[1]]-parameters2[[1]]<0)
}


#####################################3

w<-c("w1","w2","w3","w4","w5","w6","w7","w8","w9","w10","w11","w12",
"w13","w14","w15","w16","w17","w18","w19","w20","w21","w1","w2","w3","w4","w5","w6","w7","w8","w9"
"w13","w14","w15","w16","w17","w18","w19","w20","w21","w1","w2","w3","w4","w5","w6","w7","w8","w9"
"w13","w14","w15","w16","w17","w18","w19","w20","w21","w1","w2","w3","w4","w5","w6","w7","w8","w9"
"w13","w14","w15","w16","w17","w18","w19","w20","w21","w1","w2","w3","w4","w5","w6","w7","w8","w9"
"w13","w14","w15","w16","w17","w18","w19","w20","w21")
Period<-c(rep(2009,21),rep(2010,21),rep(2011,21),rep(2012,21),rep(2013,21))
medians<-c(0.01,0.01,0.03,0.03,0.02,0.06,0.04,0.05,0.02,0.02,0.07,
0.02,0.03,0.10,0.12,0.02,0.04,0.02,0.07,0.04,0.07,0.02,
0.08,0.05,0.02,0.02,0.02,0.02,0.03,0.02,0.02,0.06,0.02,0.03
,0.03,0.07,0.03,0.02,0.02,0.05,0.19,0.07,0.01,0.03,0.05,0.05
,0.05,0.02,0.04,0.03,0.03,0.05,0.03,0.02,0.04,0.05,0.02,0.18,
0.02,0.03,0.02,0.03,0.04,0.04,0.03,0.09,0.03,0.03,0.04,0.04
,0.02,0.02,0.01,0.03,0.03,0.09,0.02,0.02,0.03,0.13,0.04,
0.04,0.04,0.04,0.06,0.02,0.02,0.04,0.02,0.06,0.04,0.03,0.07,
0.04,0.03,0.03,0.06,0.04,0.02,0.02,0.01,0.02,0.02,0.05,0.12)

katw<-c(rep(0.00,5),0.01,rep(0.00,7),0.01,0.01,0.00,0.00,0.00,0.01
,0.00
,0.01,
rep(0.00,19),0.06,0.01
,rep(0.00,10),0.01,0.00,0.00,0.00,0.00,0.05,rep(0.00,5),
0.00,0.00,0.01,rep(0.00,9),0.01,rep(0.00,3),0.02,rep(0.00,4)
,rep(0.00,20),0.03)


anw<-c(0.06,0.07,0.10,0.10,0.09,0.14,0.14,0.13,0.07,0.07,0.20,0.11,0.11,
0.22,0.23,0.07,0.13,0.08,0.16,0.13,0.15,0.10,0.22,0.17,0.08,0.09,
0.08,0.09,0.11,0.07,0.07,0.19,0.09,0.10,0.13,0.19,0.12,0.09,
0.08,0.14,0.31,0.16,0.07,0.11,0.16,0.17,0.16,0.09,0.13,0.11,0.10,
0.14,0.12,0.11,0.14,0.15,0.09,0.32,0.09,0.12,0.10,0.12,0.12,
0.17,0.13,0.22,0.13,0.12,0.13,0.13,0.08,0.08,0.05,0.13,0.13,
0.20,0.08,0.10,0.11,0.24,0.13,0.13,0.13,0.11,0.19,0.11,0.10,
0.14,0.11,0.17,0.15,0.11,0.18,0.12,0.14,0.13,0.19,0.14,0.10,
0.09,0.05,0.10,0.09,0.16,0.21)


prits <- data.frame(w,medians,katw,anw,Period)
```

```
prits$w = factor(prits$w, levels = unique(prits$w))

p = ggplot(data=prits,
aes(x = Period,y = medians, ymin = katw, ymax = anw ))+
geom_pointrange(aes(col=Period))+

xlab('Period')+ ylab("95% Confidence Interval")+
geom_errorbar(aes(ymin=katw, ymax=anw,col=Period),width=0.5,cex=1)+
facet_wrap(~w,strip.position="left",nrow=9,scales = "free_y") +
theme(plot.title=element_text(size=16,face="bold"),
axis.text.y=element_blank(),
axis.ticks.y=element_blank(),
axis.text.x=element_text(face="bold"),
axis.title=element_text(size=12,face="bold"),
strip.text.y = element_text(hjust=0,vjust = 1,angle=180,face="bold"))+
coord_flip()


p

p+ scale_y_discrete(breaks=c(0,0.2))
```

```
#########################################################
#########################################################
stan_beta_3 <- "
data {
int<lower=1> N_1;
int<lower=1> K_1;
vector<lower=0>[K_1+1] C_1;
vector<lower=0,upper=1>[N_1] Y_1;
matrix[N_1,K_1] X_1;
}

parameters {
simplex[K_1+1] beta;
vector<lower=0,upper=1>[N_1] Z;
real <lower=0> sig;          // transformed linear predictor for precision of beta distribution


}

transformed parameters{
vector<lower=0,upper=1>[N_1] mu;    // transformed linear predictor for mean of beta distribution
vector<lower=0>[N_1] A;             // parameter for beta distn
vector<lower=0>[N_1] B;             // parameter for beta distn
real<lower=0> m;


for(i in 1:N_1){
```

```
mu[i]  = X_1[i,] * beta[1:K_1] + beta[K_1+1]*Z[i]  ;
}


m=min(mu-mu .*mu);




A =((1-mu).* mu .* mu- mu*sig)/(sig);
B = (1-mu).*(mu-mu .* mu -sig)/(sig);


}

model {
// priors
sig ~ uniform(0,1)T[,m];

Z ~ beta(0.5,0.5);
beta ~ dirichlet(C_1);
// likelihood
Y_1 ~ beta(A, B);
}
"
fit3<-stan(model_code=stan_beta_3,data=c("N_1","Y_1","X_1","K_1","C_1"),iter=2000,chain=4)
traceplot(fit, pars = "beta[15]", inc_warmup = FALSE)
check_n_eff(fit)
for(j in 1:K_1){
g[j]=X_1[i,j] * beta[j];
}
################################################
################################################
check_n_eff <- function(fit) {
fit_summary <- summary(fit, probs = c(0.5))$summary
N <- dim(fit_summary)[[1]]

iter <- dim(extract(fit)[[1]])[[1]]

no_warning <- TRUE
for (n in 1:N) {
ratio <- fit_summary[,5][n] / iter
if (ratio < 0.001) {
print(sprintf('n_eff / iter for parameter %s is %s!',
rownames(fit_summary)[n], ratio))
no_warning <- FALSE
}
}
if (no_warning)
print('n_eff / iter looks reasonable for all parameters')
else
print('  n_eff / iter below 0.001 indicates that the effective sample size has likely been overest
}
check_rhat <- function(fit) {
```

```
fit_summary <- summary(fit, probs = c(0.5))$summary
N <- dim(fit_summary)[[1]]

no_warning <- TRUE
for (n in 1:N) {
rhat <- fit_summary[,6][n]
if (rhat > 1.1 || is.infinite(rhat) || is.nan(rhat)) {
print(sprintf('Rhat for parameter %s is %s!',
rownames(fit_summary)[n], rhat))
no_warning <- FALSE
}
}
if (no_warning)
print('Rhat looks reasonable for all parameters')
else
print('  Rhat above 1.1 indicates that the chains very likely have not mixed')
}
check_div <- function(fit) {
sampler_params <- get_sampler_params(fit, inc_warmup=FALSE)
divergent <- do.call(rbind, sampler_params)[,'divergent__']
n = sum(divergent)
N = length(divergent)

print(sprintf('%s of %s iterations ended with a divergence (%s%%)',
n, N, 100 * n / N))
if (n > 0)
print('  Try running with larger adapt_delta to remove the divergences')
}
#####
```

Bibliography

[1] Albert, E. *Zur Elektrodynamik bewegter Körper.* (German) [*On the electrodynamics of moving bodies*]. Annalen der Physik, 322(10):891–921, 1905.

[2] Betancourt, M (2017), *A Conceptual Introduction to Hamiltonian Monte Carlo.* arXiv:1701.02434v2.

[3] Betancourt, M. (2013a). *A General Metric for Riemannian Hamiltonian Monte Carlo.* In First International Conference on the Geometric Science of Information (F. Nielsen and F. Barbaresco, eds.). Lecture Notes in Computer Science 8085. Springer. arXiv:1212.4693.

[4] Betancourt, M. (2013b). *Generalizing the No-U-Turn Sampler to Riemannian Manifolds.* arXiv:1304.1920 1304.1920

[5] Betancourt, M. (2016a). *Identifying the Optimal Integration Time in Hamiltonian Monte Carlo.* arXiv:1601.00225.

[6] Betancourt, M., Byrne, S. and Girolami, M. (2014). *Optimizing The Integrator Step Size for Hamiltonian Monte Carlo.* arXiv:1411.6669.

[7] Betancourt, M., Byrne, S., Livingstone, S. and Girolami, M. (2014). *The Geometric Foundations of Hamiltonian Monte Carlo.* arXiv:1410.5110 .

[8] Betancourt, M. and Girolami, M. (2015). *Hamiltonian Monte Carlo for Hierarchical Models.* arXiv:1312.0906

[9] Brooks, S., Gelman, A., Jones, G. L. and Meng, X.-L., eds. (2011). *Handbook of Markov Chain Monte Carlo.* CRC Press, New York.

[10] Christopher, N., Fredrik, L., Maurizio, F., James, H. (2019), *Pseudo-extended Markov Chain Monte Carlo.*

[11] Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987). *Hybrid Monte Carlo.* Physics Letters B 195 216 - 222.

[12] Hamiltonian Dynamics Sampling.
`http://probability.ca/jeff/ftpdir/ZhexinAofeiHMC.pdf`

[13] Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

[14] Holmes, S., Rubinstein-Salzedo, S., Seiler, C. (2014). *Curvature and Concentration of Hamiltonian Monte Carlo in High Dimensions.* arXiv:1407.1114.

[15] Knuth: Computers and Typesetting,
`http://www-cs-faculty.stanford.edu/~uno/abcde.html`

[16] Livingstone, S., Betancourt, M., Byrne, S. and Girolami, M. (2016). *On the Geometric Ergodicity of Hamiltonian Monte Carlo.* arXiv:1601.08057

[17] Matthew M. Graham,Amos J. Storkey (2017), *Continuously tempered hamiltonian monte carlo.* arXiv:1704.03338

[18] Michaelmas. (2006). *Advances MCMC Methods,University of Cambrige.* http://mlg.eng.cam.ac.uk/tutorials/06/im.pdf

[19] Brooks, S., Gelman, A., Jones, G. L., Meng X.-L. (2011). *In Handbook of Markov Chain Monte Carlo.* CRC Press, New York.

[20] Radford, M. N. (2012). *MCMC using Hamiltonian dynamics.* arXiv:1206.1901.

[21] Richard, M. (2015). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.*

[22] Stan Modelling Language https://mc-stan.org/

[23] UC Davis Mathematics. https://www.math.ucdavis.edu/ strohmer/courses/180BigData/180lecture1.pdf