ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

Μπεϋζιανή επιλογή μεταβλητών με χρήση προσεγγιστικών Μπεϋζιανών μεθόδων

Ελπίδα Σ. Πάττα

Τριμελής επιτροπή

Επιβλέπων: Παπασταμούλης Παναγιώτης, ΟΠΑ
Ντζούφρας Ιωάννης, ΟΠΑ
Φουσκάκης Δημήτριος, ΕΜΠ

Αθήνα, Δεκέμβριος 2020

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΚΑΤΕΥΘΥΝΣΗ
ΣΤΑΤΙΣΤΙΚΗ & ΠΙΘΑΝΟΤΗΤΕΣ

Μπεϋζιανή επιλογή μεταβλητών με χρήση προσεγγιστικών Μπεϋζιανών μεθόδων

Ελπίδα Σ. Πάττα
Επιβλέπων : Παναγιώτης Παπασταμούλης

Διπλωματική εργασία που υποβλήθηκε στη σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου,
ως μέρος των απαιτήσεων για την απόκτηση του μεταπτυχιακού διπλώματος στις
Εφαρμοσμένες Μαθηματικές Επιστήμες.

Αθήνα, Δεκέμβριος 2020

Ευχαριστίες

Στα πλαίσια απόκτησης του μεταπτυχιακού διπλώματος στις Εφαρμοσμένες Μαθηματικές Επιστήμες που απονέμει η σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου περιλαμβάνεται η κατάθεση της παρούσας διπλωματικής εργασίας, η ολοκλήρωση της οποίας θα ήταν αδύνατη χωρίς τη στήριξη των καθηγητών μου και των ατόμων του κοντινού μου περιβάλλοντος.

Καθόλη τη διάρκεια της ακαδημαϊκής μου πορείας, καθηγητές και καθηγήτριες συνέβαλαν με τις γνώσεις τους και την ανιδιοτελή προσφορά τους στη διαμόρφωση της σκέψης και της αντίληψής μου. Η όρεξη που τους διέκρινε για μετάδοση γνώσεων με ώθησε σε μία ολοκληρώμενη προσέγγιση κάθε αντικειμένου και σε ουσιαστική εμβάθυνση στον κλάδο. Εκτός όμως από την επιστομηνική και πνευματική υποστήριξη των καθηγητών και καθηγητριών μου, η ψυχολογική υποστήριξη που μου παρείχαν, συνειδητά ή μη, τα άτομα από το κοντινό μου περιβάλλον αποτέλεσε την κινητήρια δύναμή μου.

Περίληψη

Οι μέθοδοι προσέγγισης Μπεϋζιανών υπολογισμών, ευρέως γνωστές και ως approximate Bayesian computational methods ή αλλιώς likelihood-free methods, αποτελούν τα τελευταία χρόνια ένα ιδιαίτερα χρήσιμο εργαλείο για την ανάλυση σύνθετων στοχαστικών μοντέλων. Θεωρούνται ως η πιο ικανοποιητική προσέγγιση σε προβλήματα όπου ο υπλογισμός της πιθανοφάνειας καθίσταται ανέφικτος κάνοντας αισθητή την τεράστια συνεισφορά τους για πρώτη φορά στον κλάδο της γενετικής.

Υπό την Μπεϋζιανή σκοπιά, οι κυριότερες προκλήσεις είναι ο καθορισμός της εκ των προτέρων κατανομής των παραμέτρων ενός μοντέλου και ο υπολογισμός της εκ των υστέρων κατανομής του μοντέλου. Η εκ των υστέρων κατανομή του μοντέλου είναι ανάλογη με την κατανομή των διαθέσιμων δεδομένων, η λεγόμενη πιθανοφάνεια, πολλαπλασιασμένη με την εκ των προτέρων κατανομή των παραμέτρων. Συνεπώς, η Μπεϋζιανή συμπερασματολογία είναι άμεσα εξαρτημένη από την συνάρτηση της πιθανοφάνειας.

Γι΄ αυτό, παρουσιάστηκε η ανάγκη ανάπτυξης μεθόδων αποφυγής του υπολογισμού της πιθανοφάνειας σε σύνθετα υπολογιστικά μοντέλα. Ο υπολογισμός της πιθανοφάνειας αντικαθίσταται από συνεχή προσομοίωση των δεδομένων από το μοντέλο και η τεχνική αυτή ενσωματώθηκε σε πλήθος διαδεδομένων υπολογιστικών αλγορίθμων. Η επιτυχία των αλγορίθμων ABC στηρίζεται στην εύρεση τιμών για τις παραμέτρους του μοντέλου οι οποίες παράγουν προσομοιωμένα δεδομένα που βρίσκονται κοντά στα διαθέσιμα δεδομένα.

Η αξιολόγηση της ομοιότητας των δύο συνόλων δεδομένων υλοποιείται μέσω σύγκρισης ενός κατάλληλου συνόλου στατιστικών περιγραφικών μέτρων. Ιδανικά, το σύνολο των στατιστικών περιγραφικών μέτρων είναι ελάχιστη επαρκής στατιστική συνάρτηση για την παράμετρο ενδιαφέροντος. Σε περίπτωση αδυναμίας εύρεσης επαρκούς στατιστικής συνάρτησης, γίνεται προσπάθεια εύρεσης στατιστικής συνάρτησης η οποία εξασφαλίζει μία ισορροπία ανάμεσα στην επάρκεια και στη μικρή διάσταση του συνόλου των περιγραφικών μέτρων.

Οι αλγόριθμοι ABC χαρακτηρίζονται κυρίως ως μία Μπεϋζιανή διαδικασία η οποία επιχειρεί να βρει τιμές των παραμέτρων από την εκ των προτέρων κατανομή οι οποίες παράγουν ένα σύνολο δεδομένων παρόμοιο με το πραγματικό. Υπό ορισμένες προϋποθέσεις, οι αλγόριθμοι χαρακτηρίζονται ικανοί και για την επιλογή του καταλληλότερου μοντέλου. Σε αυτή την περίπτωση αναδεικνύεται η σπουδαιότητα της επαρκούς στατιστικής συνάρτησης στον αλγόριθμο, όπως περιγράφεται αναλυτικά στο άρθρο των Robert et al., 2011.

Στην παρούσα διπλωματική εργασία, επιχειρείται να αξιολογηθεί η ικανότητα των αλγορίθμων ABC στην εκτίμηση των παραμέτρων ενός μοντέλου καθώς και στην επιλογή του καταλληλότερου μοντέλου. Στο 1ο Κεφάλαιο επικεντρωνόμαστε σε μοντέλα με συζυγείς εκ των προτέρων κατανομές, όπως είναι το Poisson και το Διωνυμικό μοντέλο, το Κανονικό μοντέλο με άγνωστη μέση τιμή και το Κανονικό μοντέλο με άγνωστες και τις δύο παραμέτρους. Το πλεονέκτημα σε αυτά τα μοντέλα είναι η δυνα-

τότητα υπολογισμού της εκ των υστέρων κατανομής σε αναλυτική μορφή. Έτσι, είμαστε σε θέση να κρίνουμε την ικανότητα του αλγορίθμου ABC στην εκτίμηση των άγνωστων παραμέτρων. Επίσης, ενσωματώνουμε likelihood-free τεχνικές στους αλγορίθμους MCMC για μείωση του υπολογιστικού κόστους βάσει του άρθρου των Marjoram et al, 2003.

Στο 2ο Κεφάλαιο μελετάται η ικανότητα των αλγορίθμων ABC στην εύρεση του καταλληλότερου μοντέλου. Τα μοντέλα προς μελέτη είναι το Διωνυμικό, το Κανονικό με άγνωστη μέση τιμή και το Κανονικό με άγνωστες και τις δύο παραμέτρους. Το πλεονέκτημα σε αυτά τα μοντέλα είναι ότι ανήκουν στην εκθετική οικογένεια κατανομών και έτσι είναι εφικτή η εύρεση επαρκών στατιστικών συναρτήσεων. Υλοποιώντας τον αλγόριθμο ABC-Model Choice παρατηρούμε ποιες είναι οι καταλληλότερες συνθήκες κάτω από τις οποίες ο αλγόριθμος έιναι πιο αποτελεσματικός. Παράλληλα, εντοπίζουμε και εξηγούμε τις αδυνάμιες του αλγορίθμου οι οποίες προκύπτουν εξαιτίας των μη επαρκών στατιστικών συναρτήσεων στηριζόμενοι στο άρθρο των Robert et al..

Κλείνοντας, στο 3ο Κεφάλαιο πραγματοποιείται εκτενής ανάλυση στα κανονικά γραμμικά μοντέλα. Στα γραμμικά μοντέλα, δημοφιλείς επιλογές εκ των προτέρων κατανομής είναι αυτές που στηρίζονται στη συζυγή ανάλυση της Κανονικής-χ^2 κατανομής. Μεταξύ αυτών, προτιμότερη είναι η g-εκ των προτέρων κατανομή του Zellner επειδή οδηγεί σε περιθώριες πιθανοφάνειες αναλυτικής μορφής. Οι αλγόριθμοι ABC και ABC-MCMC κρίνονται για την ικανότητά τους στην εκτίμηση των παραμέτρων ενός κανονικού γραμμικού μοντέλου. Επίσης, παρουσιάζονται δύο παραδείγματα επιλογής μεταβλητών σε πολλαπλό γραμμικό μοντέλο χρησιμοποιώντας μία παραλλαγή του αλγορίθμου reversible jump MCMC και ενσωματώνοντας likelihood-free τεχνικές σε αυτόν.

Περιεχόμενα

Κεφάλαιο 1

Μπεϋζιανή Συμπερασματολογία και εισαγωγή στον αλγόριθμο ABC

1.1 Χαρακτηριστικά της μπεϋζιανής προσέγγισης

Ο σκοπός της Μπεϋζιανής προσέγγισης είναι η εξαγωγή συμπερασμάτων για μία πληθυσμιακή παράμετρο $\theta \in \Theta$, η οποία θεωρείται μία **τυχαία** ποσότητα/μεταβλητή, αξιοποιώντας τη γνώση μίας παρατηρούμενης τιμής από το δείγμα, $Y = y$. Αυτό επιτυγχάνεται με τη βοήθεια ενός πιθανοθεωρητικού μοντέλου $f(y|\theta)$ το οποίο καθορίζει πως κατανέμονται οι πιθανότητες στις διαφορετικές τιμές του Y, για μία δεδομένη τιμή του ϑ. Στη συνέχεια, με τη βοήθεια του μοντέλου $f(y|\theta)$ βγάζουμε συμπεράσματα για το μοντέλο $f(\theta|y)$, το οποίο αντιπροσωπεύει την κατανομή πιθανότητας της παραμέτρου ϑ δοθέντων των δεδομένων και αποτελεί το βασικότερο αντικείμενο ενδιαφέροντος.

Στο σημείο αυτό είναι απαραίτητο να γίνει μία αναφορά στη μαθηματική σημειογραφία η οποία θα χρησιμοποιηθεί παρακάτω. Πρώτα απ' όλα, ο συμβολισμός $f(\cdot|\cdot)$ δηλώνει μία υπό συνθήκη κατανομή πιθανότητας η οποία καθορίζεται κάθε φορά από τις αντίστοιχες τυχαίες μεταβλητές. Ομοίως, ο συμβολισμός $f(\cdot)$ δηλώνει την κατανομή πιθανότητας της αντίστοιχης τυχαίας μεταβλητής.

Η Μπεϋζιανή προσέγγιση, λοιπόν, χρησιμοποιεί:

1. **Πρότερη Πληροφορία:** Καθορίζουμε μία πρότερη κατανομή πιθανότητας, που συμβολίζεται με $f(\theta)$ και εκφράζει τις πρότερες πεποιθήσεις σχετικά με την κατανομή της παραμέτρου ϑ χωρίς την ύπαρξη κάποιας πληροφορίας για τα δεδομένα. Η πρότερη κατανομή μπορεί να αντανακλά την προσωπική άποψη ενός ερευνητή για το πείραμα το οποίο μελετάται.

2. **Συλλογή των δεδομένων:** Σε κάθε στατιστική μελέτη τα δεδομένα που συγκεντρώνονται μπορούν να γραφτούν στη μορφή ενός διανύσματος, έστω $y = (y_1, y_2, ..., y_n)$. Για μία στατιστική ανάλυση είναι συνήθης η υπόθεση ότι η αβεβαιότητα εκφράζεται μέσω της από κοινού συνάρτησης κατανομής των δεδομένων $f(y_1, y_2, ..., y_n)$ η οποία είναι αμετάβλητη ως προς τις μεταθέσεις των δεικτών.
   Μία ειδική περίπτωση που εμφανίζεται πολύ συχνά σε στατιστικές αναλύσεις είναι ότι τα δεδομένα αποτελούν ένα τυχαίο δείγμα, δηλαδή ότι είναι ανεξάρτητες και ισόνομα κατανεμημένες παρατηρήσεις δοθείσας της παραμέτρου ϑ. Έτσι, η από κοινού κατανομή των δεδομένων, έστω

$x = (y_1, y_2, ..., y_n)$, δίνεται από τη σχέση

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta). \tag{1.1}$$

3. **Διεξαγωγή Συμπερασμάτων:** Συνδυάζοντας την πρότερη πληροφορία, $f(\theta)$, μαζί με τα δεδομένα, $f(y|\theta)$, λαμβάνουμε την εκ των υστέρων κατανομή της παραμέτρου ϑ δοθέντος του y που συμβολίζεται με $f(\theta|y)$ και εμπεριέχει όλη την πληροφορία για την άγνωστη παράμετρο ϑ.

1.2 Μπεϋζιανή Συμπερασματολογία

1.2.1 Ο Κανόνας του Bayes

Η εξαγωγή συμπερασμάτων για την παράμετρο $\theta \in \Theta$ δοθέντος του y χρησιμοποιώντας ένα πιθανοθεωρητικό μοντέλο απαιτεί την χρήση της από κοινού συνάρτησης πιθανότητας του ϑ και του y. Η από κοινού συνάρτηση μάζας ή πυκνότητας πιθανότητας δίνεται από τη σχέση

$$f(\theta, y) = f(\theta) f(y|\theta)$$

όπου $f(\theta)$ είναι η πρότερη κατανομή και $f(y|\theta)$ είναι η από κοινού κατανομή των δεδομένων. Δεσμεύοντας ως προς τα δεδομένα y, που θεωρούνται γνωστά, και χρησιμοποιώντας το θεώρημα του Bayes καταλήγουμε στην εκ των υστέρων κατανομή, $f(\theta|y)$:

$$f(\theta|y) = \frac{f(\theta, y)}{f(y)} = \frac{f(\theta) f(y|\theta)}{f(y)}, \tag{1.2}$$

όπου $f(y) = \sum_\theta f(\theta) f(y|\theta)$ όταν το ϑ παίρνει διακριτές τιμές και $f(y) = \int_\theta f(\theta) f(y|\theta) d\theta$ όταν το ϑ παίρνει συνεχείς τιμές.

Μία ισοδύναμη μορφή της σχέσης (1.2) προκύπτει παραλείποντας τον παρανομαστή $f(y)$, ο οποίος δεν εξαρτάται από το ϑ και για σταθερή τιμή του y θεωρείται σταθερός με αποτέλεσμα να έχουμε

$$f(\theta|y) \propto f(\theta) f(y|\theta) \tag{1.3}$$

η οποία ονομάζεται μη κανονικοποιημένη εκ των υστέρων κατανομή.

1.2.2 Πιθανοφάνεια

Όταν η από κοινού κατανομή (1.1) θεωρείται μία συνάρτηση του ϑ για δοσμένο y καλείται συνάρτηση πιθανοφάνειας και συμβολίζεται με $L(\theta)$. Η πιθανοφάνεια συγκεντρώνει όλη την πληροφορία που φέρουν τα δεδομένα για την άγνωστη παράμετρο ϑ.

$$L(\theta) = f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Σε πολλές περιπτώσεις χρησιμοποιείται η λογαριθμοποιημένη πιθανοφάνεια για λόγους ευκολίας:

$$l(\theta) = \log \{f(y|\theta)\}$$

Επίσης, η Μπεϋζιανή συμπερασματολογία υπακούει στην **αρχή της πιθανοφάνειας** η οποία δηλώνει ότι αν δύο πειράματα παράγουν ανάλογες πιθανοφάνειες για ένα δοθέν δείγμα δεδομένων, τότε η συμπερασματολογία για την άγνωστη παράμετρο ϑ θα πρέπει να είναι ίδια και για τα δύο πειράματα.

1.2.3 Πρότερη Κατανομή

Η διαδικασία της Μπεϋζιανής συμπερασματολογίας περιλαμβάνει το πέρασμα από την πρότερη κατανομή, $f(\theta)$, στην ύστερη κατανομή, $f(\theta|y)$. Γι΄ αυτό, η επιλογή της πρότερης κατανομής αποτελεί κομβική ιδέα στη Μπεϋζιανή συμπερασματολογία και πρέπει να γίνεται πολύ προσεκτικά γιατί επηρεάζει την εκ των υστέρων κατανομή.

Υπάρχουν περιπτώσεις όπου η γνώση που έχουμε για την παράμετρο ϑ είναι επαρκής. Τότε, η καλύτερη προσέγγιση είναι να ενσωματώσουμε αυτή την γνώση στην πρότερη κατανομή (υποκειμενική Μπεϋζιανή προσέγγιση). Σε άλλες περιπτώσεις, δεν έχουμε καμία πληροφορία για την παράμετρο ϑ και ακολουθούμε μία αντικειμενική Μπεϋζιανή προσέγγιση. Διαλέγουμε δηλαδή, μία πρότερη κατανομή η οποία να ανταναχλά πλήρη άγνοια για την παράμετρο ϑ, όταν αυτό είναι δυνατό.

Όταν έχουμε μία ΄θολή ιδέα' για την πρότερη κατανομή είναι καλό να υπολογίζουμε κάποια χαρακτηριστικά της για να μπορέσουμε να κατανοήσουμε καλύτερα τη συνάρτηση. Για παράδειγμα, η κανονική κατανομή καθορίζεται από δύο παραμέτρους, που σημαίνει ότι μπορούμε να προσδιορίσουμε μία συγκεκριμένη πρότερη κατανομή εστιάζοντας σε δύο χαρακτηριστικά της κατανομής, όπως είναι η μέση τιμή και η διασπορά της.

Συζυγείς Πρότερες Κατανομές

Μία οικογένεια πρότερων κατανομών καλείται συζυγής όταν η εκ των υστέρων κατανομή ανήκει στην ίδια οικογένεια με την πρότερη. Οι συζυγείς πρότερες κατανομές είναι κομμάτι της **υποκειμενικής Μπεϋζιανής προσέγγισης.** Η επιλογή μιας τέτοιας κατανομής μπορεί να κάμψει πολλές δυσκολίες, όπως τον υπολογισμό της σταθεράς κανονικοποίησης στη σχέση (1.2), ο οποίος σε πολλές περιπτώσεις υπολογίζεται μόνο με αριθμητικές μεθόδους.

Έχει αποδειχθεί ότι όταν η κατανομή των δεδομένων ανήκει στην εκθετική οικογένεια κατανομών, μπορεί να βρεθεί μία συζυγής πρότερη κατανομή. Η κατανομή αυτή δηλαδή, μπορεί να γραφτεί στη μορφή

$$f(y_i|\theta) = h(y_i)g(\theta)\exp\{t(y_i)c(\theta)\}$$

για κάποιες συναρτήσεις h, g, t, c. Στην εκθετική οικογένεια κατανομών ανήκει η εκθετική και η Poisson κατανομή, η Γάμμα κατανομή με γνωστή την παράμετρο θέσης, η διωνυμική κατανομή και η κανονική κατανομή με γνωστή διασπορά.

Έτσι λοιπόν, για ένα τυχαίο δείγμα $y = (y_1, y_2, ..., y_n)$ η πιθανοφάνεια ως προς την άγνωστη παράμετρο ϑ είναι

$$\begin{aligned} f(y|\theta) &= \left(\prod_{i=1}^{n} h(y_i)\right) g(\theta)^n \exp\left\{\sum_{i=1}^{n} t(y_i)c(\theta)\right\} \\ &\propto g(\theta)^n \exp\left\{\sum_{i=1}^{n} t(y_i)c(\theta)\right\} \end{aligned}$$

Τότε, αν διαλέξουμε μία πρότερη κατανομή της μορφής

$$f(\theta) \propto g(\theta)^k \exp\{\nu c(\theta)\}$$

καταλήγουμε στην ύστερη κατανομή

$$\begin{aligned} f(\theta|y) &\propto f(\theta)f(y|\theta) \\ &\propto g(\theta)^k \exp\{\nu c(\theta)\} g(\theta)^n \exp\left\{\sum_{i=1}^{n} t(y_i)c(\theta)\right\} \\ &= g(\theta)^{n+k} \exp\left\{\sum_{i=1}^{n} t(y_i)c(\theta) + \nu c(\theta)\right\} \\ &= g(\theta)^{n+k} \exp\left\{\left(\sum_{i=1}^{n} t(y_i) + \nu\right) c(\theta)\right\} \end{aligned}$$

Συνεπώς, παρατηρούμε ότι η εκ των υστέρων κατανομή ανήκει στην ίδια οικογένεια με την πρότερη κατανομή έχοντας διαφορετικές παραμέτρους.

1.2.4 Παραδείγματα Συζυγών Πρότερων Κατανομών

1.Poisson κατανομή

Έστω τυχαίο δείγμα με ανεξάρτητες και ισόνομα κατανεμημένες παρατηρήσεις $x = (y_1, y_2, ..., y_n)$ από την κατανομή Poisson με παράμετρο ϑ. Τότε, έχουμε $y_i \sim Poisson(\theta)$ με συνάρτηση πυκνότητας πιθανότητας

$$f(y_i|\theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}, \quad \text{για } \; i = 1, 2, ..., n \text{ και } y_i = 0, 1, 2, 3, ...$$

Τότε, η συνάρτηση πιθανοφάνειας είναι

$$\begin{aligned} L(\theta) = f(y|\theta) &= \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \\ &\propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \\ &\propto e^{-n\theta} \exp\left\{\log\theta \sum_{i=1}^{n} y_i\right\} \end{aligned}$$

Άρα, η συζυγής εκ των προτέρων κατανομή είναι

$$f(\theta) \propto e^{-\beta\theta}\theta^{\alpha-1}$$

η οποία είναι η κατανομή Γάμμα μα παραμέτρους α και β. Τότε, από τη σχέση (1.3) έχουμε ότι η εκ των υστέρων κατανομή ισούται με

$$\begin{aligned} f(\theta|y) &\propto f(\theta)f(y|\theta) \\ &\propto e^{-\beta\theta}\theta^{\alpha-1} e^{-n\theta} \exp\left\{\log\theta \sum_{i=1}^{n} y_i\right\} \\ &\propto e^{-(n+\beta)\theta}\theta^{n\overline{y}+\alpha-1} \end{aligned}$$

Συνεπώς, η εκ των υστέρων κατανομή είναι $\theta|y \sim \Gamma\alpha\mu\mu\alpha(n\overline{y} + \alpha, n + \beta)$.

2.Κανονική κατανομή με γνωστή διασπορά

Έστω τυχαίο δείγμα με ανεξάρτητες και ισόνομα κατανεμημένες παρατηρήσεις $y = (y_1, y_2, ..., y_n)$ από την κανονική κατανομή με μέση τιμή ϑ και γνωστή διασπορά σ^2. Έχουμε, δηλαδή, $y_i \sim N(\theta, \sigma^2)$ και η συνάρτηση πυκνότητας πιθανότητας είναι

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\}, \quad \forall\ i = 1, ..., n.$$

Τότε, η πιθανοφάνεια είναι

$$\begin{aligned}
L(\theta) = f(y|\theta) &= \prod_{i=1}^{n} f(y_i|\theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\theta y_i + \theta^2)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(-2\theta\sum_{i=1}^{n} y_i + n\theta^2)\right\} = \exp\left\{-\frac{1}{2\sigma^2}(-2n\theta\overline{y} + n\theta^2)\right\} \\
&\propto \exp\left\{\frac{n\theta\overline{y}}{\sigma^2} - \frac{n\theta^2}{2\sigma^2}\right\}
\end{aligned}$$

Αν θεωρήσουμε για πρότερη κατανομή την

$$f(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right)$$

τότε έχουμε $\theta \sim N(\mu_0, \tau_0^2)$ με υπερπαραμέτρους μ_0 και τ_0^2. Τότε, από τη σχέση (1.3) έχουμε

$$\begin{aligned}
f(\theta|y) \propto f(\theta)f(y|\theta) &\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2 + \frac{n\theta\overline{y}}{\sigma^2} - \frac{n\theta^2}{2\sigma^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{(\theta - \mu_0)^2}{\tau_0^2} - \frac{2n\theta\overline{y}}{\sigma^2} + \frac{n\theta^2}{\sigma^2}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\theta^2 + \mu_0^2 - 2\theta\mu_0}{\tau_0^2} - \frac{2n\theta\overline{y}}{\sigma^2} + \frac{n\theta^2}{\sigma^2}\right)\right\} \\
&\propto \exp\left\{-\frac{\theta^2}{2\tau_0^2} + \frac{\theta\mu_0}{\tau_0^2} + \frac{n\theta\overline{y}}{\sigma^2} - \frac{n\theta^2}{2\sigma^2}\right\} \\
&= \exp\left\{-\left(\frac{1}{2\tau_0^2} + \frac{n}{2\sigma^2}\right)\theta^2 + \left(\frac{\mu_0}{\tau_0^2} + \frac{n\overline{y}}{\sigma^2}\right)\theta\right\}
\end{aligned}$$

Παρατηρούμε λοιπόν, ότι η εκ των προτέρων κατανομή εξαρτάται από τα δεδομένα μόνο από τον δειγματικό μέσο $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$ ο οποίος είναι επαρκής στατιστική συνάρτηση για την παράμετρο ϑ. Συνεπώς, $\theta|y \sim N(\mu_n, \tau_n^2)$ όπου

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\overline{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{και} \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

3. Διωνυμικό μοντέλο

Στο διωνυμικό μοντέλο, σκοπός είναι να εκτιμήσουμε ένα πληθυσμιακό ποσοστό από τα αποτελέσματα μιάς ακολουθίας από n ανεξάρτητες δοκιμές Bernoulli. Σε κάθε δοκιμή Bernoulli, τα δεδομένα $y_1, y_2, ..., y_n$ μπορούν να πάρουν μία από τις τιμές 1 ή 0, οι οποίες συμβατικά αντικατοπτρίζουν ¨επιτυχία' ή ¨αποτυχία' αντίστοιχα. Έτσι, στο διωνυμικό μοντέλο τα δεδομένα, έστω y, συμβολίζουν το συνολικό αριθμό επιτυχιών μέσα σε n δοκιμές και η άγνωστη παράμετρος ϑ το ποσοστό των επιτυχιών μέσα στον πληθυσμό που μελετάται. Η από κοινού κατανομή δίνεται ως

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

Θεωρώντας την παραπάνω σχέση ως συνάρτηση του ϑ, παίρνουμε την πιθανοφάνεια η οποία είναι της μορφής

$$L(\theta) \propto \theta^y(1-\theta)^{n-y}$$

Συνεπώς, μπορούμε να διαλέξουμε ως πρότερη κατανομή για το ϑ την

$$f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

η οποία είναι η κατανομή Beta με υπερπαραμέτρους α και β: $\theta \sim Beta(\alpha, \beta)$. Αυτή η επιλογή της πρότερης κατανομής υποδηλώνει ότι έχουμε εκ των προτέρων $\alpha - 1$ το πλήθος επιτυχιές και $\beta - 1$ αποτυχίες. Τότε, από τη σχέση (1.3) έχουμε

$$\begin{aligned} f(\theta|y) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

Άρα, η εκ των υστέρων κατανομή του $\theta|y$ είναι επίσης μία κατανομή Beta με παραμέτρους $y + \alpha$ και $n - y + \beta$: $\theta|y \sim Beta(y + \alpha, n - y + \beta)$

4. Κανονικό Μοντέλο με άγνωστες και τις δύο παραμέτρους

Μέχρι τώρα τα παραδείγματα που έχουν μελετηθεί περιλαμβάνουν μόνο μία άγνωστη παράμετρο. Όμως, τα περισσότερα στατιστικά προβλήματα συνδέονται με μοντέλα τα οποία περιλαμβάνουν περισσότερες από μία άγνωστες παραμέτρους. Σε αυτό το παράδειγμα λοιπόν, θα μελετήσουμε το κανονικό μοντέλο με άγνωστες και τις δύο παραμέτρους: τη μέση τιμή μ και τη διασπορά σ^2.

Έστω ότι $y = (y_1, y_2, ..., y_n)$ είναι ένα τυχαίο δείγμα από την κανονική κατανομή, $N(\mu, \sigma^2)$, με άγνωστες και τις δύο παραμέτρους. Θεωρούμε το διάνυσμα $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ των παραμέτρων για το οποίο επιθυμούμε να γίνει η αντίστοιχη συμπερασματολογία.

Η μέθοδος ανάλυσης ενός μοντέλου με περισσότερες από μία άγνωστες παραμετρους ακολουθεί ακριβώς την ίδια θεωρία με αυτή στην οπόια έχουμε αναφερθεί στις προηγούμενες ενότητες. Ειδικότερα, ο προσδιορισμός της εκ των υστέρων κατανομής $f(\theta|y)$ γίνεται χρησιμοποιώντας το θεώρημα του Bayes

$$f(\theta|y) = \frac{f(\theta, y)}{\int_\theta f(\theta)f(y|\theta)d\theta} = \frac{f(\theta)f(y|\theta)}{\int_\theta f(\theta)f(y|\theta)d\theta}$$

όπως πριν, με ιδιαίτερη προσοχή όμως στο γεγονός ότι εδώ το ϑ είναι ένα διάνυσμα.

Στη συνέχεια, θα θεωρήσουμε μία συζυγή εκ των προτέρων κατανομή για το δι-παραμετρικό κανονικό μοντέλο. Γι΄ αυτό, επιλέγουμε ως περιθώρια κατανομή του σ^2 την κλιμακωτή αντίστροφη -χ^2 και ως κατανομή του μ δοθέντος του σ^2 μία κανονική κατανομή με $\mu_0 > 0$ και $\kappa_0 > 0$:

$$\mu|\sigma^2 \sim N(\mu_0, \frac{\sigma^2}{\kappa_0}) \qquad \text{και} \qquad \sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2)$$

Έτσι, η από κοινού εκ των προτέρων κατανομή του μ και του σ^2 είναι :

$$\begin{aligned} f(\mu, \sigma^2) &= f(\mu|\sigma^2) f(\sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2/\kappa_0}} \exp\left\{-\frac{\kappa_0(\mu-\mu_0)^2}{2\sigma^2}\right\} \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\frac{\nu_0}{2})} \sigma_0^{\nu_0} (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_0}{2}+\frac{3}{2})} \exp\left\{-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2\right)\right\} \end{aligned} \tag{1.4}$$

Η παραπάνω κατανομή είναι η Κανονική Αντίστροφη-χ^2 και συμβολίζεται ως N-Inv-$\chi^2(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$.

Η από κοινού εκ των υστέρων κατανομή υπολογίζεται πολλαπλασιάζοντας την εκ των προτέρων κατανομή (1.4) με την κανονική πιθανοφάνεια:

$$\begin{aligned} f(\mu, \sigma^2|y) &\propto f(\mu, \sigma^2) f(y|\mu, \sigma^2) \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_0}{2}+\frac{3}{2})} \exp\left\{-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2\right)\right\} \times \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_0}{2}+\frac{3}{2}+\frac{n}{2})} \exp\left\{-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{n}(y_i-\overline{y}+\overline{y}-\mu)^2\right)\right\} \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_0+n}{2}+\frac{3}{2})} \exp\left\{-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2 + (n-1)s^2 + 2(\overline{y}-\mu)(n\overline{y}-n\overline{y}) + n(\overline{y}-\mu)^2\right)\right\} \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_0+n}{2}+\frac{3}{2})} \exp\left\{-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2 + (n-1)s^2 + n(\overline{y}-\mu)^2\right)\right\} \end{aligned} \tag{1.5}$$

όπου $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2$ η δειγματική διασπορά και $\overline{y}$ ο δειγματικός μέσος. Μπορεί να αποδειχθεί λοιπόν, ότι η εκ των υστέρων κατανομή $f(\mu, \sigma^2|y)$ έχει την ίδια μορφή με την εκ των προτέρων $f(\mu, \sigma^2)$, είναι δηλαδή μία Κανονική Αντίστροφη-χ^2 με τέσσερις παραμέτρους:

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\overline{y} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \sigma_n^2 &= \frac{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\overline{y}-\mu_0)^2}{\nu_0+n} \end{aligned} \tag{1.6}$$

Η περιθώρια εκ των υστέρων κατανομή του σ^2 προκύπτει ολοκληρώνοντας τη σχέση (1.5) ως προς μ:

$$\begin{aligned} f(\sigma^2|y) &= \int_{-\infty}^{+\infty} f(\mu, \sigma^2|y) d\mu \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_n}{2}+\frac{3}{2})} \exp\left\{-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right\} \int_{-\infty}^{+\infty} \exp\left\{-\frac{\kappa_n(\mu-\mu_n)^2}{2\sigma^2}\right\} d\mu \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_n}{2}+\frac{3}{2})} \exp\left\{-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right\} \sqrt{2\pi\sigma^2/\kappa_n} \\ &\propto \left(\sigma^2\right)^{-(\frac{\nu_n}{2}+1)} \exp\left\{-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right\} \end{aligned}$$

Άρα, η περιθώρια εκ των υστέρων κατανομή του σ^2 είναι μία κλιμακωτή αντίστροφη-χ^2:

$$\sigma^2|y \sim \text{SI-}\chi^2(\nu_n, \sigma_n^2)$$

Η περιθώρια εκ των υστέρων κατανομή του μ προκύπτει από τη σχέση (1.5) ολοκληρώνοντας ως προς σ^2 και κάνοντας αλλαγή μεταβλητής:

$$z := \frac{\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2}{2\sigma^2} := \frac{A}{2\sigma^2} \qquad dz = -\frac{A}{2\sigma^4} d\sigma^2$$

Άρα, έχουμε

$$\begin{aligned} f(\mu|y) &= \int_0^{+\infty} f(\mu, \sigma^2|y) d\sigma^2 \\ &\propto \int_0^{+\infty} \left(\sigma^2\right)^{-(\frac{\nu_n}{2}+\frac{3}{2})} \exp\left\{-\frac{1}{2\sigma^2}\left(\nu_n\sigma_n^2\kappa_n(\mu-\mu_n)^2\right)\right\} d\sigma^2 \\ &\propto \int_{+\infty}^0 \left(\frac{A}{2z}\right)^{-\frac{\nu_n+3}{2}} \exp\{-z\} \left(-\frac{2\sigma^4}{A}\right) dz \\ &\propto \int_0^{+\infty} \left(\frac{A}{2z}\right)^{-\frac{\nu_n+3}{2}} \frac{A^2/2z^2}{A} \exp\{-z\} dz \\ &\propto A^{-\frac{\nu_n+1}{2}} \int_0^{+\infty} \frac{\exp\{-z\}}{4z^3} dz \\ &\propto \left(\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2\right)^{-\frac{\nu_n+1}{2}} = \left[\nu_n\sigma_n^2\left(1 + \frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right)\right]^{-\frac{\nu_n+1}{2}} \\ &\propto \left(1 + \frac{\kappa_n(\mu-\mu_n)^2}{\nu_n\sigma_n^2}\right)^{-\frac{\nu_n+1}{2}} \end{aligned}$$

Παρατηρούμε λοιπόν, ότι η περιθώρια κατανομή του μ είναι η κατανομή t με ν_n βαθμούς ελευθερίας και παραμέτρους μ_n και σ_n^2:

$$\mu|y \sim t_{\nu_n}\left(\mu_n, \sigma_n^2\right)$$

1.2.5 Μη πληροφοριακές πρότερες κατανομές

Όταν δεν έχουμε διαθέσιμη πληροφορία για την μελέτη που διεξάγεται, είναι δύσκολη η κατασκευή μίας πρότερης κατανομής. Γι' αυτό, είναι απαραίτητη η ύπαρξη μίας πρότερης η οποία να επηρεάζει ελάχιστα την εκ των υστέρων κατανομή. Αυτές οι κατανομές χαρακτηρίζονται ως μη πληροφοριακές ή ως

ασαφείς εκ των προτέρων κατανομές με μεγάλη διασπορά. Η ιδέα για τη χρήση μιας μη πληροφοριακής πρότερης κατανομής βασίζεται και στο γεγονός ότι όσο περισσότερα δεδομένα έχουμε τόσο λιγότερη είναι η επίδραση της πρότερης στον καθορισμό της εκ των υστέρων κατανομής.

Καταχρηστικές εκ των προτέρων κατανομές (Improper prior distributions)

Μία καταχρηστική εκ των προτέρων κατανομή ονομάζεται μία μη αρνητική συνάρτηση f, ορισμένη στο Θ, για την οποία ισχύει

$$\int_{\Theta} f(\theta)\, d\theta = +\infty$$

Οι κατανομές αυτές δηλαδή, παραβιάζουν την υπόθεση ότι οι συναρτήσεις πιθανότητας ολοκληρώνουν στη μονάδα. Παρ΄ όλα αυτά, σε πολλές περιπτώσεις η χρήση μιας καταχρηστικής εκ των προτέρων κατανομής οδηγεί σε κατάλληλες εκ των υστέρων κατανομές, δοθέντων των δεδομένων.

Στο δεύτερο παράδειγμα του κεφαλαίου 1.2.4 μελετήσαμε ένα τυχαίο δείγμα με ανεξάρτητες και ισόνομα κατανεμημένες παρατηρήσεις $y = (y_1, y_2, ..., y_n)$ όπου $y_i \sim N(\theta, \sigma^2)$. Η παράμετρος προς εκτίμηση είναι η μέση τιμή θ, $\theta \in \mathbb{R}$ θεωρώντας ότι η διασπορά σ^2 είναι γνωστή. Τώρα, θα υποθέσουμε ότι η εκ των προτέρων κατανομή είναι ανάλογη με μία σταθερά c, $c \in \mathbb{R}$:

$$f(\theta) \propto c$$

Μία τέτοια κατανομή θεωρείται καταχρηστική γιατί το ολοκλήρωμα για $\theta \in \mathbb{R}$ είναι άπειρο. Ισοδύναμα, θα μπορούσαμε να θεωρήσουμε ότι $\tau_0 \to \infty$, όπου τ_0 η τυπική απόκλιση της εκ των προτέρων κατανομής στο Παράδειγμα 2. Όμως, υπολογίζοντας την εκ των υστέρων κατανομή έχουμε

$$f(\theta|y) \propto \exp\left\{-\frac{n}{2\sigma^2}\theta^2 + \frac{n\overline{y}}{\sigma^2}\theta\right\}$$

δηλαδή, $\theta|y \sim N(\overline{y}, \frac{\sigma^2}{n})$. Συνεπώς, παρατηρούμε ότι ενώ χρησιμοποιήσαμε μία καταχρηστική εκ των προτέρων κατανομή, τελικά καταλήξαμε σε μία καλώς ορισμένη εκ των υστέρων κατανομή.

Η εκ των προτέρων κατανομή του Jeffreys

Μία άλλη μη πληροφοριακή εκ των προτέρων κατανομή θεωρείται η εκ των προτέρων κατανομή του Jeffreys. Αν δεν υπάρχει εκ των προτέρων διαθέσιμη πληροφορία, η πρότερη κατανομή του Jeffreys προτιμάται στις περισσότερες περιπτώσεις επειδή είναι συνεπής στους 1-1 μετασχηματισμούς της παραμέτρου $\phi = h(\theta)$. Η συνέπεια στους 1-1 μετασχηματισμούς ισοδυναμεί με τη σχέση

$$f(\phi) = f(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right| \tag{1.7}$$

Η αρχή του Jeffreys οδηγεί στον προσδιορισμό της μη πληροφοριακής πρότερης κατανομής ως εξής

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

όπου $I(\theta)$ είναι η πληροφορία του Fisher και ορίζεται ως:

$$I(\theta) = \mathbb{E}\left(\left(\frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta}\right)^2\right)$$

Αν η λογαριθμοποιημένη πιθανοφάνεια $l(\theta)$ είναι δύο φορές παραγωγίσιμη ως προς ϑ, υπό ορισμένες συνθήκες, η πληροφορία του Fisher μπορεί να γραφτεί και ως $I(\theta) = -\mathbb{E}\left(\frac{d^2 l(\theta}{d\theta^2}\right)$.

Για να αποδειχθεί ότι η πρότερη κατανομή του Jeffreys είναι αμετάβλητη στους μετασχηματισμούς της άγνωστης παραμέτρου ϑ, πρέπει να υπολογίσουμε την $I(\phi)$:

$$I(\phi) = \mathbb{E}\left(\left(\frac{\mathrm{d}l(\phi)}{\mathrm{d}\phi}\right)^2\right) = \mathbb{E}\left(\left(\frac{\mathrm{d}l(\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2\right) = \left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 I(\theta) \tag{1.8}$$

Συνεπώς, λαμβάνοντας υπόψην τη σχέση (1.7), η σχέση (1.8) αποδεικνύει ότι η πρότερη κατανομή του Jeffreys είναι αμετάβλητη στους 1-1 μετασχηματισμούς της παραμέτρου ϑ.

1.3 Εισαγωγή στις προσεγγιστικές Μπεϋζιανές υπολογιστικές μεθόδους (ABC)

Όπως είδαμε στο κεφάλαιο 1.1, το σημαντικότερο εργαλείο στη Μπεϋζιανή συμπερασματολογία είναι η εκ των υστέρων κατανομή της παραμέτρου ϑ δοθέντων των δεδομένων y. Σύμφωνα με τη σχέση (1.2), η εκ των υστέρων κατανομή είναι ανάλογη της συνάρτησης πιθανοφάνειας, γεγονός το οποίο καθιστά κρίσιμο τον υπολογισμό της πιθανοφάνειας. Σε πολλές περιπτώσεις, άλλοτε η αδυναμία και άλλοτε η δυσκολία υπολογισμού της πιθανοφάνειας δημιουργεί εμπόδια στη Μπεϋζιανή συμπερασματολογία. Γι΄ αυτό το λόγο, τα μοντέλα τα οποία περιέχουν πιθανοφάνειες υπολογιστικά δύσκολες έχουν προσελκύσει το ενδιαφέρον των μελετητών τα τελευταία χρόνια.

Τα εμπόδια που παρουσιάζονται είναι δύο: (1) όταν η συνάρτηση πιθανοφάνειας δεν είναι διαθέσιμη σε κλειστή μορφή ως συνάρτηση της παραμέτρου ϑ και (2) όταν το υπολογιστικό κόστος για τον υπολογισμό της πιθανοφάνειας είναι αρκετά μεγάλο. Τα εμπόδια αυτά οδήγησαν στην εισαγωγή προσεγγιστικών Μπεϋζιανών υπολογιστικών μεθόδων(ABC methods) που να προσεγγίζουν την πιθανοφάνεια χωρίς να την υπολογίζουν αναλυτικά. Ήδη από το 1984 η μέθοδος ABC είχε αναφερθεί μέσω ενός φιλοσοφικού επιχειρήματος από τον Rubin.

Η πρώτη εφαρμογή πραγματοποιήθηκε από τους Tavare et al σε άρθρο πληθυσμιακής γενετικής, οι οποίοι εισήγαγαν τις μεθόδους ABC σαν μία τεχνική απόρριψης αποφεύγοντας τον υπολογισμό της συνάρτησης πιθανοφάνειας χρησιμοποιώντας προσομοιωμένες τιμές από την κατανομή ενδιαφέροντος. Σκοπός του άρθρου ήταν η εύρεση μεθόδων για την εκτίμηση του χρόνου συγκερασμού(coalescence time) ενός δείγματος αλληλουχιών DNA ενδοειδών. Η θεωρία του συγκερασμού(coalescent theory) είναι ένα μοντέλο το οποίο ερευνά κατά πόσο οι παραλλαγές γονιδίων σε έναν πληθυσμό μπορεί να προέρχονται από κοινό πρόγονο. Ο χρόνος συγκερασμού ορίζεται ως ο χρόνος από τον πιο πρόσφατο κοινό πρόγονο.

Οι μέθοδοι αυτές εκμεταλλεύονται προηγούμενη γνώση από τις μελέτες για τον ανθρώπινο πληθυσμό σε συνδυσμό με μοριακά δεδομένα. Συγκεκριμένα, στο άρθρο γίνεται η χρήση του μοντέλου του Kingman, Kingman's coalescent. Το μοντέλο αυτό είναι ένα πιθανοθεωρητικό μοντέλο για το γενεαλογικό δέντρο ενός τυχαίου δείγματος n γονιδίων που προέρχονται από έναν μεγάλο πληθυσμό. Υπάρχουν δύο σημαντικά μεγέθη που συνδέονται με ένα γενεαλογικό δέντρο. Το ύψος του δέντρου, T_n, το οποίο ορίζεται ως ο χρόνος μέχρι τον πιο πρόσφατο κοινό πρόγονο και το μήκος του δέντρου, L_n, το

οποίο ορίζεται ως το συνολικό μήκος όλων των κλαδιών του δέντρου. Τα μεγέθη αυτά εκφράζονται συναρτήσει του χρόνου W_j κατά τον οποίο το δείγμα έχει j διακριτούς προγόνους, $2 \leq j \leq n$. Ο χρόνος W_j μετριέται σε συνεχείς τιμές και ακολουθεί την εκθετική κατανομή με παράμετρο $\frac{j(j-1)}{2}$. Η περιγραφή αυτή παρέχει μία κλειστή προσέγγιση σε ένα εύρος μοντέλων πληθυσμιακής γενετικής στα οποία ο χρόνος εκφράζεται σε γενιές. Εξασφαλίζεται δηλαδή, ότι μία μονάδα χρόνου συγκερασμού ισοδυναμεί με N το πλήθος γενιές.

Το ύψος και το μήκος του δέντρου εκφράζονται από τις σχέσεις

$$T_n = \sum_{j=2}^{n} W_j \quad \text{και} \quad L_n = \sum_{j=2}^{n} jW_j$$

Οι χρόνοι κατά τους οποίους συμβαίνουν μεταλλάξεις γονιδίων μοντελοποιούνται στη θεωρία συγκερασμού υποθέτοντας ότι οι χρόνοι ακολουθούν κατανομή Poisson με σταθερό ρυθμό $\rho/2$. Αυτό σημαίνει ότι αν ένα κλαδί στο δέντρο έχει μήκος w, τότε ο αριθμός των μεταλλάξεων πάνω σε αυτό το κλαδί ακολουθεί κατανομή Poisson με μέση τιμή $w\rho/2$, ανεξάρτητα από τις μεταλλάξεις πάνω σε άλλα κλαδιά. Συνεπώς,

$$\rho = 2N\mu$$

όπου μ είναι ο ρυθμός μετάλλαξης ανά γονίδιο ανά γενιά.

Επίσης, στη μελέτη πρέπει να ληφθούν υπόψην οι θέσεις των μεταλλάξεων πάνω στο γενεαλογικό δέντρο καθώς και οι διαφορετικοί τύποι μεταλλάξεων. Όταν τα δεδομένα που διαθέτουμε είναι ακολουθίες DNA, μπορούμε να υποθέσουμε ότι δοθέντος του μήκους του δέντρου L_n, ο αριθμός S_n των διαφορετικών θέσεων όπου συμβαίνουν μεταλλάξεις στο δείγμα ακολουθεί την κατανομή Poisson με μέση τιμή $\rho L_n/2$. Συμβολίζουμε,

$$\mathbb{P}\left(S_n = k | L_n = l\right) = Po\left(k, \rho l/2\right),$$

όπου $Po\left(x, \lambda\right)$ χρησιμοποιείται για να υποδηλώσουμε την συνάρτηση μάζας πιθανότητας της κατανομής Poisson, $Po\left(x, \lambda\right) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ για $x = 0, 1, ...$ και $\lambda \leq 0$.

Ο σκοπός των Tavare et al ήταν να περιγράψουν όσο το δυνατόν καλύτερα το χρόνο συγκερασμού T_n υπό το πρίσμα των διαθέσιμων δεδομένων και των κατάλληλων υποθέσεων μοντελοποίησης. Γι΄ αυτό, επίκεντρο της μελέτης τους αποτέλεσε ο υπολογισμός της συνάρτησης $f_{T_n}\left(t|D\right), t > 0$, της κατανομής του T_n δοθέντος των δεδομένων D. Η μελέτη τους επικεντρώνεται στην περίπτωση όπου τα δεδομένα D είναι ακολουθίες DNA. Με αυτή την παραδοχή είναι εφικτή η αντικατάσταση των δεδομένων D από την στατιστική συνάρτηση S_n. Από τον ορισμό της δεσμευμένης πιθανότητας έχουμε,

$$f_{T_n}\left(t|D\right) = f_{T_n}\left(t\right)\frac{\mathbb{P}(D|T_n = t)}{\mathbb{P}(D)}$$

Είναι προφανές από την παραπάνω σχέση ότι ο όρος $\mathbb{P}(D)$ θεωρείται σταθερά ως προς τη μεταβλητή t και γι΄ αυτό μπορούμε ισοδύναμα να χρησιμοποιήσουμε την έκφραση

$$f_{T_n}\left(t|D\right) \propto f_{T_n}\left(t\right)\mathbb{P}(D|T_n = t)$$

Η παραπάνω σχέση δεν μπορεί να υπολογιστεί αναλυτικά γιατί η αναλυτική έκφραση του όρου $\mathbb{P}(D|T_n = t)$ είναι διαθέσιμη μόνο σε περιορισμένες περιπτώσεις, όπως όταν $n = 2$ ή $S_n = 0$. Έτσι,

οι Tavare et al προχώρησαν στη δημιουργία ενός αλγορίθμου ο οποίος προσεγγίζει τη συνάρτηση $\mathbb{P}(D|T_n = t)$ και περιγράφεται αναλυτικά στη συνέχεια προσαρμοσμένος στην ορολογία που έχει χρησιμοποιηθεί μέχρι τώρα.

Έστω ότι έχουμε την παράμετρο ενδιαφέροντος θ. Η ιδέα των Tavare et al στηρίζεται στην προσομοίωση τιμών θ^* της παραμέτρου θ από την εκ των προτέρων κατανομή $f(\theta)$ και η αποδοχή αυτής της τιμής εξαρτάται από το πόσο κοντά βρίσκονται οι προσομοιωμένες τιμές ενός δείγματος Z στις τιμές του πραγματικού δείγματος Y. Ο αλγόριθμος που δίνεται στη συνέχεια διατυπώνεται θεωρώντας ότι το σύνολο Y, το στήριγμα της κατανομής του y, είναι πεπερασμένο.

1. Για $i = 1, ..., N$:
2. Επανάλαβε μέχρι $Z = Y$
   α. Προσομοίωση τιμής θ^* από την εκ των προτέρων κατανομή $f(\theta)$
   β. Προσομοίωση τιμής Z από την πιθανοφάνεια $f(Y|\theta^*)$
3. Θέσε $\theta_i = \theta^*$

Αλγόριθμος 1.1. Likelihood-free rejection sampler από τους Tavare et al

Ο Αλγόριθμος 1.1 παράγει το διάνυσμα $(\theta_1, \theta_2, ..., \theta_N)$ το οποίο είναι ένα δείγμα ανεξάρτητων και ισόνομων μεταβλητών από την εκ των υστέρων κατανομή, όπως γίνεται φανερό και από την σχέση

$$\begin{aligned} f(\theta_i) &\propto \sum_{z \in Y} f(\theta_i) f(z|\theta_i) \mathbb{I}_y(z) = f(\theta_i) f(z|\theta_i) \\ &\propto f(\theta_i|y) \end{aligned}$$

Όμως, η μέθοδος αυτή παρουσιάζει ορισμένα μειονεκτήματα καθώς είναι υπολογιστικά δύσκολο να εφαρμοστεί σε σύνθετα μοντέλα. Και είναι κυρίως δύο οι λόγοι στους οποίους οφείλονται οι δυσκολίες αυτές. Πρώτον, όταν ο πληθάριθμος του συνόλου Y είναι υπερβολικά μεγάλος είναι ιδιαίτερα σπάνιο να ικανοποιηθεί η συνθήκη $Z = Y$. Δεύτερον, ο αλγόριθμος εφαρμόζεται όταν το σύνολο Y είναι διακριτό και πεπερασμένο και δεν μπορεί να υλοποιηθεί όταν τα δεδομένα προέρχονται από ένα συνεχές και μη πεπερασμένο σύνολο.

Έτσι, οι Pritchard et al το 1999 γενίκευσαν την παραπάνω ιδέα και για συνεχή σύνολα δεδομένων και εισήγαγαν τον πρώτο επίσημο αλγόριθμο ABC. Κι εδώ η ομοιότητα ανάμεσα στο προσομοιωμένο σύνολο δεδομένων Z και το παρατηρηθέν σύνολο δεδομένων Y είναι αναγκαία. Για να την εξετάσουμε ορίζουμε ένα μέτρο απόκλισης/ μία μετρική ανάμεσα σε αυτά τα δύο σύνολα, έστω $\rho(Z, Y)$. Για παράδειγμα, η ομοιότητα μπορεί να μετρηθεί χρησιμοποιώντας την Ευκλείδεια απόσταση μεταξύ αυτών των συνόλων. Όμως, λόγω του μεγάλου μεγέθους του δείγματος Y, είναι συνήθης η χρήση κατάλληλων στατιστικών περιγραφικών μέτρων. Ιδανικά το στατιστικό περιγραφικό μέτρο που επιλέγουμε, έστω η, πρέπει να είναι μία επαρκής στατιστική συνάρτηση για την άγνωστη παράμετρο ϑ, έτσι ώστε να μην συμβεί απώλεια σημαντικών πληροφοριών. Βέβαια, ειδικά σε μοντέλα τα οποία είναι αρκετά σύνθετα η εύρεση επαρκής στατιστικής συνάρτησης είναι σπάνιο φαινόμενο.

Η τιμή της μετρικής που έχουμε ορίσει θα καθορίσει το αν θα δεχθούμε ή θα απορρίψουμε την προσομοιωμένη τιμή θ^*. Δεχόμαστε την θ^* ως μία τιμή από την εκ των υστέρων κατανομή αν η τιμή της μετρικής $\rho(\eta(Z), \eta(Y))$ είναι μικρότερη από μία μικρή τιμή ε, τη λεγόμενη σταθερά tolerance threshold, όπου $\varepsilon > 0$. Για μικρές τιμές του ε, η εκ των υστέρων κατανομή $f(\theta|\rho(\eta(Z), \eta(Y)) \leq \varepsilon)$ προσεγγίζει

την εκ των υστέρων κατανομή $f(\theta|Y)$. Οι ιδέες αυτές λοιπόν, κατασκευάζουν τον **Αλγόριθμο 1.2** που ακολουθεί.

1. Για $i = 1, ..., N$:
2. Επανάλαβε μέχρι $\rho(\eta(Z), \eta(Y)) \leq \varepsilon$
   α. Προσομοίωση τιμής θ^* από την εκ των προτέρων κατανομή $f(\theta)$
   β. Προσομοίωση τιμής z από την πιθανοφάνεια $f(y|\theta^*)$
3. Θέσε $\theta_i = \theta^*$

Αλγόριθμος 1.2. Likelihood-free rejection sampler από τους Pritchard et al

Συνεπώς, οι προσομοιωμένες τιμές θ_i αντιστοιχούν σε ένα σύνολο δεδομένων Z το οποίο είναι 'παρόμοιο' με το αρχικό σύνολο δεδομένων Y και έχει τον ίδιο αριθμό παρατηρήσεων. Αποτέλεσμα αυτού είναι οι συναρτήσεις πιθανοφάνειας των δύο συνόλων δεδομένων να είναι ανάλογες και κατα συνέπεια ανάλογες να είναι και οι εκ των υστέρων κατανομές.

Χρησιμοποιώντας μία αρκετά αντιπροσωπευτική στατιστική συνάρτηση η συνδυασμένη με μία ικανοποιητικά μικρή τιμή της σταθεράς ε ο αλγόριθμος ABC είναι ικανός να παράξει μία αρκετά καλή εκτίμηση της εκ των υστέρων κατανομής, δηλαδή,

$$f_\varepsilon(\theta|y) = \int f_\varepsilon(\theta, z|y)\, dz \approx f(\theta|y)$$

Η προσομοίωση όμως τιμών από την εκ των προτέρων κατανομή είναι μη αποδοτική επειδή δεν λαμβάνονται υπόψη τα δεδομένα τη στιγμή της προσομοίωσης με αποτέλεσμα οι προτεινόμενες τιμές να οδηγούν εκ των υστέρων σε περιοχές χαμηλής πιθανότητας. Η δυσκολία αυτή κάμφθηκε χάρη στους Marjoram et al, 2003 οι οποίοι εισήγαγαν τον αλγόριθμο MCMC-ABC(**Αλγόριθμος 1.3**).

1. Αρχικοποίηση του αλγορίθμου με τιμές (θ^0, z^0) με τη βοήθεια του Αλγορίθμου 1.2
2. Για $i = 1, ..., N$:
3. Προσομοίωση τιμής θ^* από την κατανομή εισήγησης $q\left(\cdot|\theta^{(i-1)}\right)$
4. Προσομοίωση τιμής z^* από την πιθανοφάνεια $f(\cdot|\theta^*)$
5. Προσομοίωση τιμής u από την ομοιόμορφη κατανομή $u \sim U_{[0,1]}$
6. Αν $u \leq \dfrac{f(\theta^*)q\left(\theta^{(i-1)}|\theta^*\right)}{f(\theta^{(i-1)})q\left(\theta^*|\theta^{(i-1)}\right)}$ και $\rho\left(\eta(z^*), \eta(y)\right) \leq \varepsilon$
   τότε θέσε $\left(\theta^{(i)}, z^{(i)}\right) = (\theta^*, z^*)$
   αλλιώς θέσε $\left(\theta^{(i)}, z^{(i)}\right) = \left(\theta^{(i-1)}, z^{(i-1)}\right)$

Αλγόριθμος 1.3. Likelihood-free MCMC sampler από τους Marjoram et al

Η αρχικοποίηση του αλγορίθμου με τη χρήση του Likelihood-free rejection sampler (**Αλγόριθμος 1.2**) μπορεί να παραλειφθεί καθώς η Μαρκοβιανή αλυσίδα ξεχνάει την αρχική της κατάσταση. Το αποτέλεσμα είναι η απουσία του υπολογιστικού κόστους που προέρχεται από την αρχικοποίηση των τιμών. Όμως, σε αυτή την περίπτωση ο αλγόριθμος MCMC θα χρειαστεί περισσότερο χρόνο για να επιτευχθεί η σύγκλιση της αλυσίδας και επίσης, απαιτείται η παράλειψη των τιμών των πρώτων επαναλήψεων από το συνολικό αποτέλεσμα, διαδικασία με σημαντικό υπολογιστικό κόστος.

Γι΄ αυτό είναι πολύ σημαντική η επιλογή των σωστών παραμέτρων για τον αλγόριθμο. Ο αλγόριθμος ABC εξαρτάται από τρεις παραμέτρους, tuning parameters, την στατιστική συνάρτηση η, την σταθερά ε, tolerance threshold, και την απόσταση ρ. Οι παράμετροι αυτές επηρεάζουν καθοριστικά την πορεία και την επιτυχία του αλγορίθμου και γι΄ αυτό διαφοροποιούνται ανάλογα με το διαθέσιμο μοντέλο.

Στα παραδείγματα που ακολουθούν θα εξετάσουμε την αποτελεσματικότητα και την ακρίβεια των αλγορίθμων ABC σε μοντέλα των οποίων η πιθανοφάνεια είναι εύκολο να υπολογιστεί αναλυτικά. Έτσι, θα γίνει μία σύγκριση των αποτελεσμάτων από τους αλγορίθμους με τις ακριβείς εκ των υστέρων κατανομές των μοντέλων. Σε περίπτωση όπου η πιθανοφάνεια δεν είναι διαθέσιμη σε κλειστή μορφή, η λύση βρίσκεται σε αλγορίθμους MCMC, όπως είναι ο ψευδο-περιθωριακός αλγόριθμος Metropolis-Hastings, Lin et al., 2000, Beaumont, 2003, Andrieu and Roberts, 2009.

1.3.1 Εκτίμηση της εκ των υστέρων κατανομής χρησιμοποιώντας τον αλγόριθμο ABC στο Διωνυμικό μοντέλο

Στο Κεφάλαιο 1.2.4, Παράδειγμα 3, εκτιμήθηκε η εκ των υστέρων κατανομή του ποσοστού των επιτυχιών μέσα σε έναν πληθυσμό στα πλαίσια ενός διωνυμικού μοντέλου. Θα χρησιμοποιήσουμε την εκ των υστέρων κατανομή, η οποία είναι μία Beta κατανομή, για να εκτιμήσουμε την ακρίβεια των προσεγγιστικών εκ των υστέρων κατανομών που παράγονται από τον αλγόριθμο ABC.

Αρχικά, είναι απαραίτητος ο προσδιορισμός της κατάλληλης μετρικής για τη σύγκριση των προσομοιωμένων δεδομένων Z με τα παρατηρούμενα δεδομένα Y. Στο διωνυμικό μοντέλο τα δεδομένα Y συμβολίζουν το συνολικό αριθμό επιτυχιών μέσα σε n δοκιμές, δηλαδή παίρνουν διακριτές τιμές από το σύνολο $Y = \{0, 1, 2, ..., n\}$. Άρα, μπορούμε να ορίσουμε ως μέτρο απόκλισης την απόλυτη διαφορά του αριθμού των επιτυχιών στα δύο σύνολα διαιρούμενη με το πλήθος των δοκιμών n,

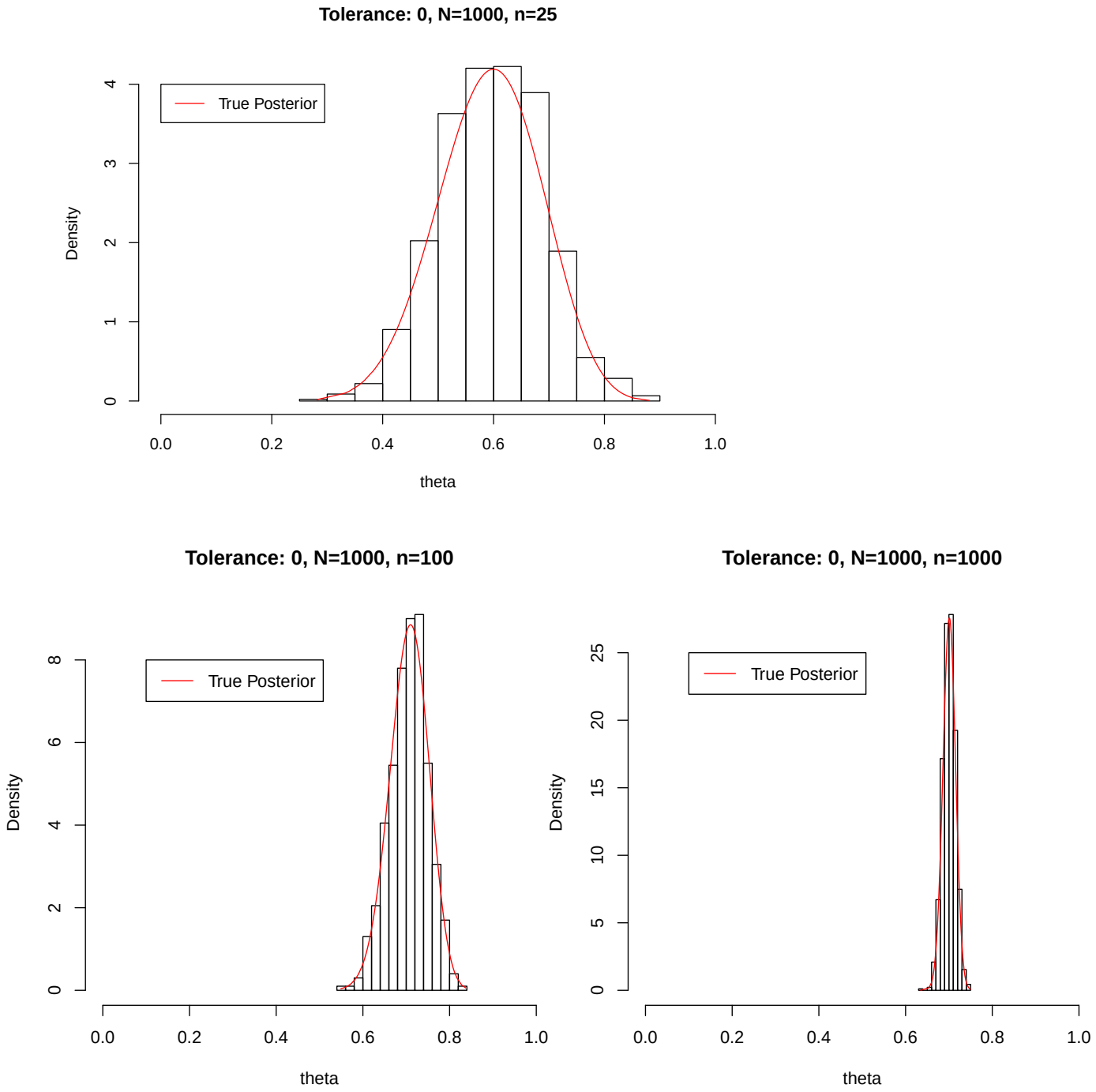
$$\rho(Z, Y) = \frac{1}{n}|Z - Y|$$

Το μέτρο απόκλισης $\rho(Z, Y)$ μετράει το βαθμό στον οποίο τα προσομοιωμένα δεδομένα Z ταιριάζουν με τα παρατηρούμενα δεδομένα Y. Έτσι, όταν $\rho(Z, Y) = 0$ ο αριθμός των επιτυχιών είναι ακριβώς ο ίδιος και στα δύο σύνολα.

Η προσομοίωση του μοντέλου γίνεται έχοντας τρία διαφορετικά σύνολα δεδομένων, με πιθανότητα επιτυχίας $p = 0.7$ στο καθένα. Το πρώτο σύνολο αποτελείται από $n = 25$, το δεύτερο από $n = 100$ και το τρίτο από $n = 1000$ ανεξάρτητες δοκιμές Bernoulli(p). Καθώς το πλήθος των δοκιμών αυξάνεται, η ποσότητα της πληροφορίας που σχετίζεται με την παράμετρο p αυξάνεται με αποτέλεσμα να οδηγούμαστε σε κατανομές που έχουν σημείο μεγίστου. Σε κάθε επανάληψη προσομοιώνουμε τιμή για την παράμετρο p από την κατανομή Beta$(1, 1)$, η οποία παράγει το προσομοιωμένο σύνολο δεδομένων z, σύμφωνα με τον ABC αλγόριθμο απόρριψης (**Αλγόριθμος 1.2**). Κρατάμε 1000 τιμές για την παράμετρο p, αυτές δηλαδή που παράγουν ένα προσομοιωμένο σύνολο δεδομένων z που είναι κοντά στο αρχικό y χρησιμοποιώντας το μέτρο απόκλισης $\rho(Z, Y)$ και τη σταθερά ε η οποία ισούται με μηδέν, $\varepsilon = 0$.

Στο **Σχήμα 1.1** παρουσιάζονται οι κατανομές των προσομοιωμένων τιμών για την παράμετρο p για κάθενα από τα τρία σύνολα δεδομένων.

Η πραγματική εκ των υστέρων κατανομή που επικαλύπτει κάθε ιστόγραμμα υποδεικνύει ότι καθώς ο αριθμός των δοκιμών αυξάνεται, η εκ των υστέρων κατανομή συγκεντρώνεται όλο και περισσότερο γύρω από την πραγματική τιμή της παραμέτρου p.



Σχήμα 1.1: Η προσαρμογή των εκ των υστέρων κατανομών σε τρία διαφορετικά σύνολα δεδομένων αποτελούμενα από $n = 25$(πάνω), $n = 100$ (αριστερά), $n = 1000$ (δεξιά) δοκιμές. Η κόκκινη γραμμή αντιπροσωπεύει την πραγματική εκ των υστέρων κατανομή.

Για κάθε δείγμα, η εκτίμηση της εκ των υστέρων που παράγεται από τον αλγόριθμο ABC προσεγγίζει με μεγάλη ακρίβεια την πραγματική εκ των υστέρων κατανομή. Η απλότητα του συγκεκριμένου παραδείγματος μας επιτρέπει να προσομοιώσουμε ένα μεγάλο αριθμό τιμών για την παράμετρο p με αμελητέο κόστος. Ο χρόνος που απαιτείται είναι 2, 5 και 120 δευτερόλεπτα για 25, 100 και 1000 δοκιμές αντίστοιχα χρησιμοποιωντας το περιβάλλον της R.

1.3.2 Εκτίμηση της εκ των υστέρων κατανομής χρησιμοποιώντας τον αλγόριθμο ABC στο Κανονικό μοντέλο με γνωστή διασπορά

Στο Παράδειγμα 2 του Κεφαλαίου 1.2.4 μελετήσαμε την περίπτωση όπου τα δεδομένα y προέρχονται από την κανονική κατανομή με άγνωστη μέση τιμή ϑ και γνωστή διασπορά σ^2. Χρησιμοποιώντας ως συζυγή εκ των προτέρων κατανομή για την παράμετρο ϑ την κανονική κατανομή με υπερπαραμέτρους μ_0 και τ_0^2, καταλήγουμε σε μία εκ των υστέρων κατανομή για την παράμετρο $\theta|y$, την κανονική κατανομή με παραμέτρους μ_n και τ_n^2. Η κλειστή μορφή της εκ των υστέρων κατανομής μας επιτρέπει να εξετάσουμε την ακρίβεια του αλγορίθμου ABC για τον υπολογισμό της εκ των υστέρων σε κανονικά δεδομένα.

Το κανονικό μοντέλο, όπως και οποιοδήποτε μοντέλο που είναι κατάλληλο για προσομοίωση συνεχών τιμών, παρουσιάζει περισσότερες δυσκολίες από το διωνυμικό μοντέλο. Όταν τα δεδομένα προέρχονται από σύνολο συνεχών τιμών, η πιθανότητα ταύτισης του συνόλου των πραγματικών δεδομένων Y με το σύνολο των προσομοιωμένων δεδομένων Z είναι μηδενική. Για να επιτευχθεί η μέγιστη δυνατή ακρίβεια, απαιτείται περισσότερη μελέτη γύρω από επιλεχθέν μέτρο απόκλισης, $\rho(\eta(Z), \eta(Y))$ και τη σταθερά ε, το tolerance threshold.

Εξετάζουμε τρία διαφορετικά μέτρα απόκλισης για τη σύγκριση του συνόλου των πραγματικών δεδομένων Y με το σύνολο των προσομοιωμένων δεδομένων Z, τα οποία είναι η απόλυτη διαφορά των μέσων, η σταθμισμένη ευκλείδεια απόσταση και η σταθμισμένη L_1 απόσταση διαιρούμενες με το πλήθος των δεδομένων. Αντίστοιχα, έχουμε

$$\rho_1 = |\overline{Y} - \overline{Z}|, \quad \rho_2 = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(\frac{y_i - z_i}{s}\right)^2}, \quad \rho_3 = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - z_i|}{s}$$

όπου s εκφράζει την τυπική απόκλιση των προσομοιωμένων δεδομένων z.

Πολύ σημαντική είναι και η επιλογή της σταθεράς ε. Η σταθερά ε είναι ουσιαστικά το επιθυμητό επίπεδο συμφωνίας των προσομοιωμένων και των πραγματικών δεδομένων. Η επιλογή της καθορίζεται σύμφωνα με μία σειρά διαδικασιών ετσι ώστε να εναρμονίζεται με τη φύση του κάθε μοντέλου και της αντίστοιχης απόστασης που έχει επιλεγθχεί. Συγκεκριμένα, αφού γίνει προσομοίωση ενός μεγάλου αριθμού τιμών για την παράμετρο ϑ, έστω M το πλήθος, πραγματοποείται ταξινόμηση των ζευγαριών $\{\theta_i, \rho_i\}$ ως προς την απόσταση ρ_i. Ρυθμίζοντας ένα φυσικό αριθμό κ, θεωρούμε ως ε την κ-οστή μικρότερη τιμή των ταξινομημένων αποστάσεων (Biau et al, 2015). Εναλλακτικά, κρατάμε N το πλήθος προσομοιωμένες τιμές, όπου $N = \alpha \times M$ έτσι ώστε να καθοριστεί το ε. Η μεταβλητή α συνηθίζεται να ισούται με 0.5 ή να κυμαίνεται ανάμεσα στην ακρίβεια και το σφάλμα Monte Carlo, (Fearnhead & Prangle, 2012).

Για να εξετάσουμε την ακρίβεια του αλγορίθμου ABC θεωρούμε ότι τα πραγματικά δεδομένα προέρχονται από την κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1, $y_i \sim N(0,1), \quad i = 1, .., n$. Σε κάθε επανάληψη προσομοιώνουμε τη μέση τιμή θ του μοντέλου από την κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 10 για να δηλώσουμε άγνοια για το που κυμαίνονται οι πραγματικές τιμές. Κάθε τιμή παράγει το προσομοιωμένο σύνολο δεδομένων z, σύμφωνα με τον αλγόριθμο απόρριψης ABC (**Αλγόριθμος 1.2).** Κρατάμε 1000 τιμές για την παράμετρο θ, αυτές δηλαδή που παράγουν ένα προσομοιωμένο σύνολο δεδομένων z που είναι κοντά στο αρχικό y χρησιμοποιώντας ένα από τα τρία μέτρα απόκλισης ρ_1, ρ_2, ρ_3 και την κατάλληλη σταθερά ε σε κάθε περίπτωση.

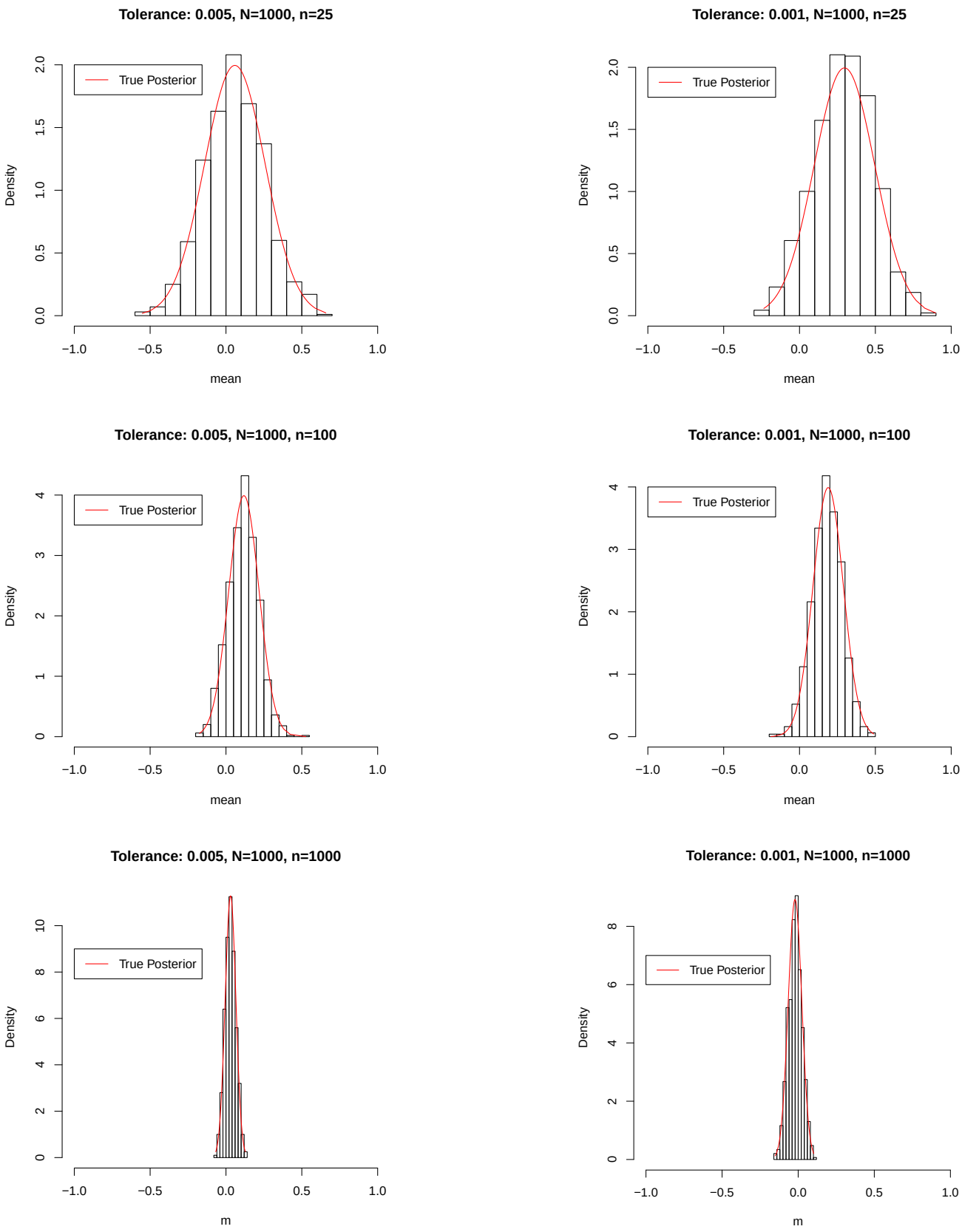
Α. Η απόλυτη διαφορά των μέσων ως μέτρο απόκλισης

Ξεκινώντας με την απόσταση ρ_1, ο αλγόριθμος ABC υλοποιείται σε τρεις διαφορετικές περιπτώσεις ανάλογα με το σύνολο των δεδομένων, το οποίο αποτελείται από $n = 25$, $n = 100$ και $n = 1000$ ανεξάρτητες τιμές από την τυποποιημένη κανονική κατανομή N(0,1) σε κάθε περίπτωση. Επιλέγουμε δύο τιμές για τη σταθερά ε, $\varepsilon_0 = 0.005$ και $\varepsilon_1 = 0.001$.

Στο **Σχήμα 1.2** παρουσιάζονται οι κατανομές των προσομοιωμένων τιμών για την παράμετρο θ για κάθενα από τα τρία σύνολα δεδομένων. Η πραγματική εκ των υστέρων κατανομή που επικαλύπτει κάθε ιστόγραμμα υποδεικνύει ότι καθώς το μέγεθος του δείγματος αυξάνεται, η εκ των υστέρων κατανομή συγκεντρώνεται όλο και περισσότερο γύρω από την πραγματική τιμή της παραμέτρου θ.

Για κάθε δείγμα, η εκτίμηση της εκ των υστέρων που παράγεται από τον αλγόριθμο ABC προσεγγίζει με μεγάλη ακρίβεια την πραγματική εκ των υστέρων κατανομή. Η απλότητα του συγκεκριμένου παραδείγματος μας επιτρέπει να προσομοιώσουμε ένα μεγάλο αριθμό τιμών για την παράμετρο θ με αμελητέο κόστος. Όταν η σταθερά ε έχει την τιμή 0.005, $\varepsilon_0 = 0.005$, ο χρόνος που απαιτείται είναι 1, 2 και 7 λεπτά για $n = 25$, $n = 100$ και $n = 1000$ δοκιμές αντίστοιχα χρησιμοποιωντας το περιβάλλον της R.

Από την άλλη, όταν η σταθερά ε μειώνεται και είναι ίση με 0.001, $\varepsilon_1 = 0.001$ ο χρόνος που απαιτείται για την ολοκλήρωση του αλγορίθμου αυξάνεται. Η ολοκλήρωση του αλγορίθμου πραγματοποιείται μέσα σε 6, 8 και 20 λεπτά για $n = 25$, $n = 100$ και $n = 1000$ δοκιμές. Τα διαγράμματα υποδεικνύουν ότι ο Αλγόριθμος 1.2 είναι ικανός να εκτιμήσει την άγνωστη μέση τιμή του μοντέλου με μεγάλη ακρίβεια. Ειδικότερα, καθώς το πλήθος των δεδομένων αυξάνεται ο αλγόριθμος γίνεται και πιο αποδοτικός.

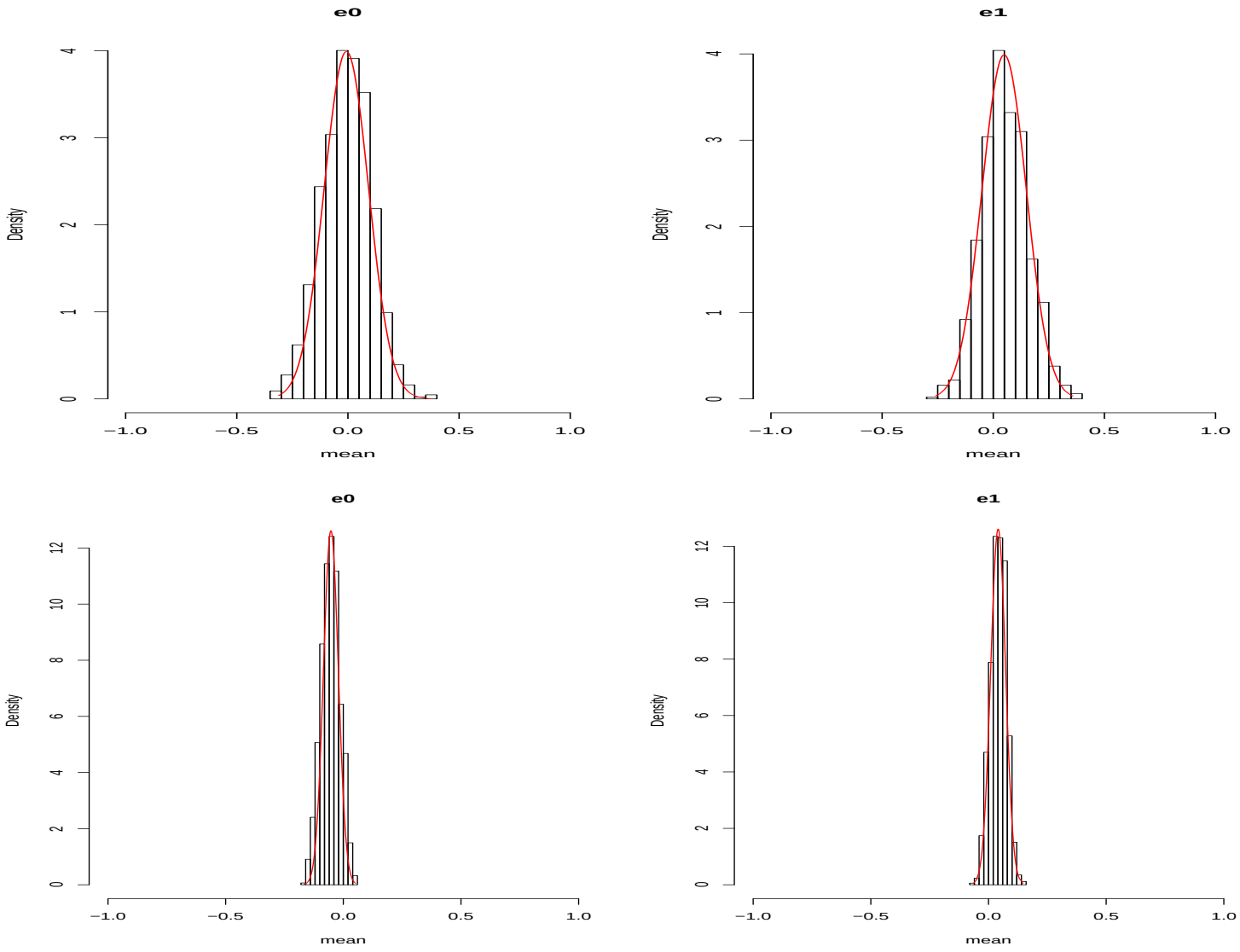


Σχήμα 1.2: Η προσαρμογή των εκ των υστέρων κατανομών σε τρία διαφορετικά σύνολα δεδομένων αποτελούμενα από $n = 25$ (πρώτη γραμμή), $n = 100$ (δεύτερη γραμμή), $n = 1000$ (τρίτη γραμμή) δοκιμές. Στην πρώτη στήλη η σταθερά ε ισούται με 0.005 ενώ στη δεύτερη ισούται με 0.001.

Β. Η σταθμισμένη Ευκλείδεια απόσταση ως μέτρο απόκλισης

Χρησιμοποιώντας τη δεύτερη απόσταση, ο αλγόριθμος ABC υλοποιείται σε ένα σύνολο δεδομένων από 100 ανεξάρτητες τιμές από την τυποποιημένη κανονική κατανομή $N(0,1)$ για δύο διαφορετικές τιμές της σταθεράς ε, με $\varepsilon_0 = 0.02$ και $\varepsilon_1 = 0.01$. Στη συνέχεια, υλοποιείται για ένα σύνολο δεδομένων που αποτελείται από 1000 ανεξάρτητες τιμές από την τυποποιημένη κανονική κατανομή $N(0,1)$ και παίρνουμε δύο τιμές για τη σταθερά ε, με $\varepsilon_0 = 0.002$ και $\varepsilon_1 = 0.001$.

Στο **Σχήμα 1.3** παρουσιάζονται τα ιστογράμματα των προσομοιωμένων τιμών για τη μέση τιμή θ μαζί με την εκ των υστέρων κατανομή. Χρησιμοποιώντας τον αλγόριθμο ABC στο σύνολο δεδομένων των 100 τιμών, ο χρόνος που απαιτείται είναι 4 και 55 λεπτά με τη σταθερά ε_0 και ε_1 αντίστοιχα. Στο σύνολο δεδομένων των 1000 τιμών, ο αλγόριθμος ABC υλοποιείται μέσα σε 6 και 70 λεπτά με τη σταθερά ε_0 και ε_1 αντίστοιχα.

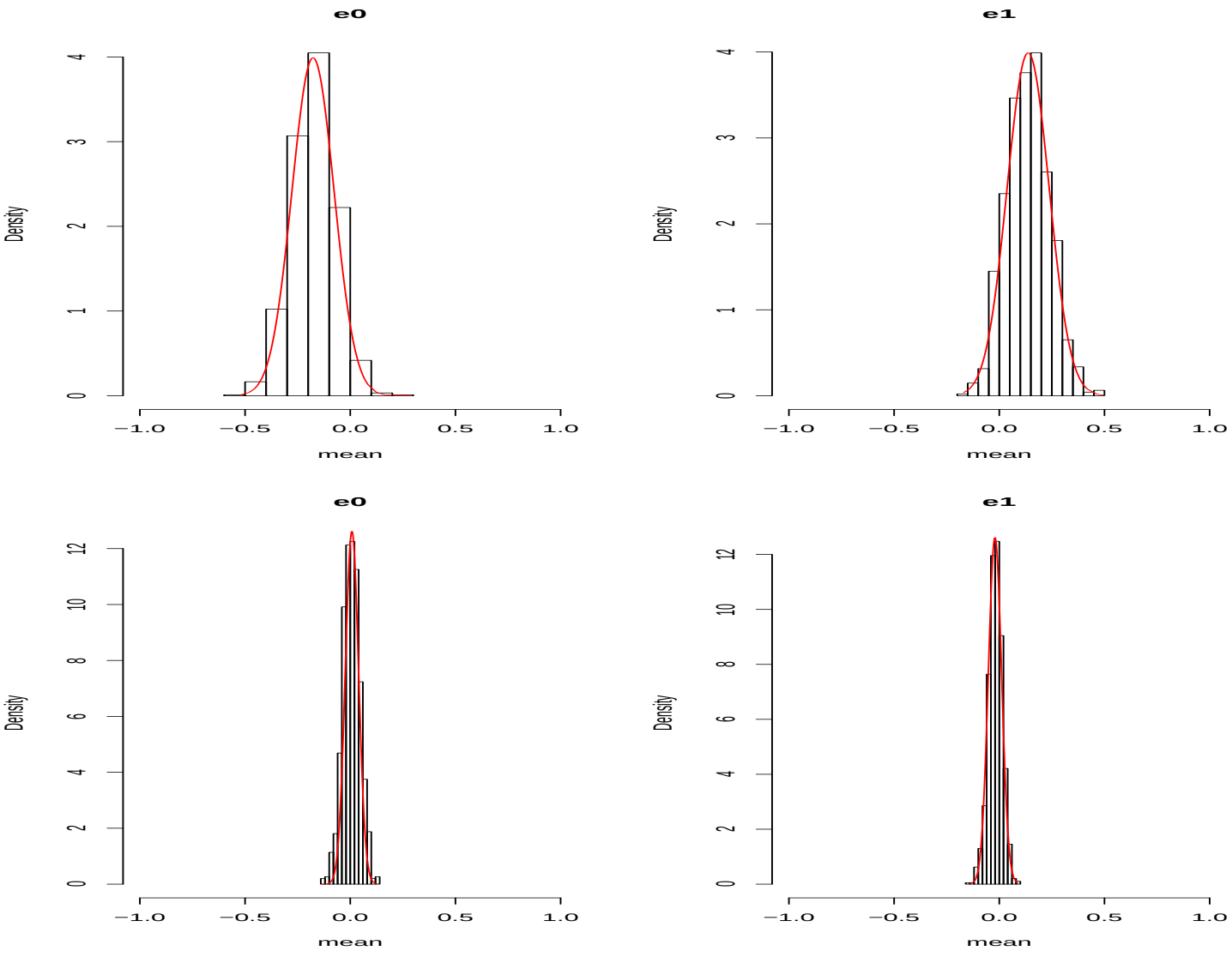


Σχήμα 1.3: Τα ιστογράμματα μαζί με τις εκ των υστέρων κατανομές για δύο διαφορετικά σύνολα δεδομένων αποτελούμενα από $n = 100$ (πρώτη γραμμή) και $n = 1000$ (δεύτερη γραμμή) δοκιμές και για δύο τιμές της σταθεράς ε.

Γ. Η σταθμισμένη L_1 απόσταση ως μέτρο απόκλισης

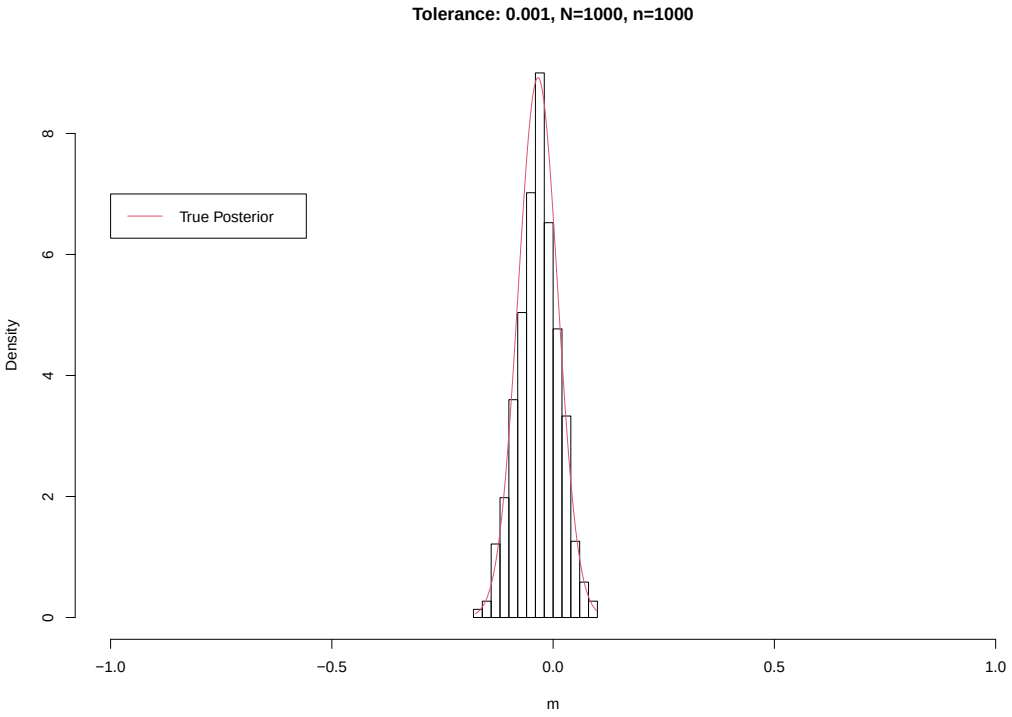
Κλείνοντας, ενδιαφέρον παρουσιάζουν και τα αποτελέσματα που παίρνουμε χρησιμοποιώντας τη τρίτη απόσταση. Κι εδώ, ο αλγόριθμος ABC υλοποιείται σε ένα σύνολο δεδομένων από 100 και 1000 ανεξάρτητες τιμές από την τυποποιημένη κανονική κατανομή $N(0,1)$. Στο σύνολο των 100 τιμών έχουμε δύο τιμές της σταθεράς ε, $\varepsilon_0 = 0.08$ και $\varepsilon_1 = 0.04$ και στο σύνολο των 1000 τιμών χρησιμοποιούμε τις τιμές $\varepsilon_0 = 0.05$ και $\varepsilon_1 = 0.03$.

Στο **Σχήμα 1.4** παρουσιάζονται τα ιστογράμματα των προσομοιωμένων τιμών για τη μέση τιμή θ μαζί με την εκ των υστέρων κατανομή. Χρησιμοποιώντας τον αλγόριθμο ABC στο σύνολο δεδομένων των 100 τιμών, ο χρόνος που απαιτείται είναι 9 και 80 λεπτά με τη σταθερά ε_0 και ε_1 αντίστοιχα. Στο σύνολο δεδομένων των 1000 τιμών, ο αλγόριθμος ABC υλοποιείται μέσα σε 7 και 22 λεπτά με τη σταθερά ε_0 και ε_1 αντίστοιχα.



Σχήμα 1.4: Οι εκ των υστέρων κατανομές για δύο διαφορετικά σύνολα δεδομένων αποτελούμενα $n = 100$ (πρώτη γραμμή), $n = 1000$ (δεύτερη γραμμή) δοκιμές και για δύο τιμές της σταθεράς ε.

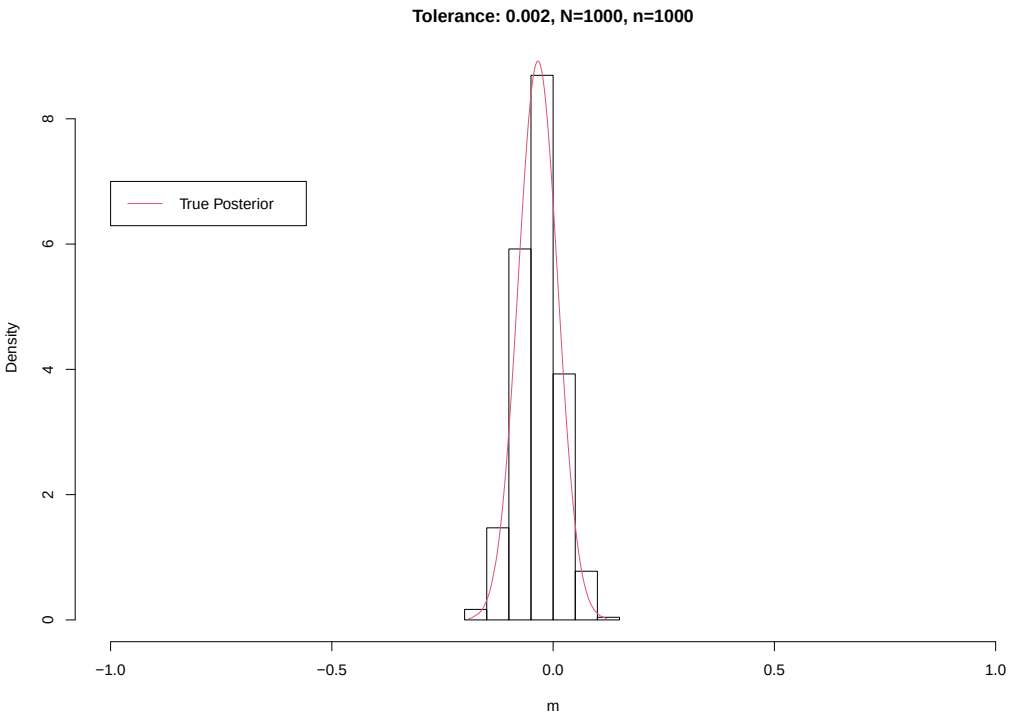
Στη συνέχεια, χρησιμοποιώντας το ίδιο σύνολο δεδομένων ο **Αλγόριθμος 1.2** υλοποιείται για τις τρεις διαφορετικές αποστάσεις που προαναφέρθηκαν. Στο **Σχήμα 1.4(α)** παρουσιάζεται η κατανομή των προσομοιωμένων τιμών για τη μέση τιμή χρησιμοποιώντας την πρώτη απόσταση, την απόλυτη διαφορά των μέσων, και η σταθερά ε ισούται με 0.001. Ο αλγόριθμος απαιτεί 10 λεπτά για την ολοκλήρωσή του. Η πραγματική εκ των υστέρων κατανομή που επικαλύπτει κάθε ιστόγραμμα υποδεικνύει ότι επιτυγχάνεται καλή προσαρμογή. Ο εκ των υστέρων μέσος ισούται με -0.0342 και η διασπορά με 0.045. Η μέση τιμή και η διασπορά των προσομοιωμένων τιμών είναι -0.0335 και 0.002 αντίστοιχα.



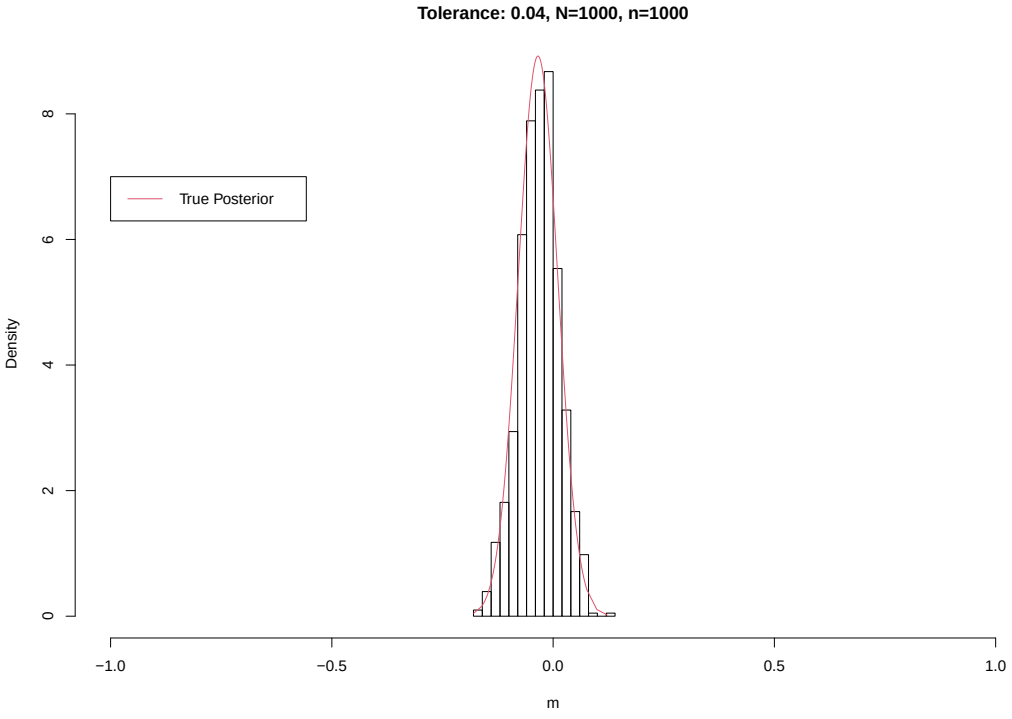
Σχήμα 1.4(α): Η προσαρμογή της εκ των υστέρων κατανομής χρησιμοποιώντας την πρώτη απόσταση. Η κόκκινη γραμμή αντιπροσωπεύει την πραγματική εκ των υστέρων κατανομή.

Στο **Σχήμα 1.4(β)** παρουσιάζεται η κατανομή των προσομοιωμένων τιμών για τη μέση τιμή χρησιμοποιώντας την δεύτερη απόσταση, την σταθμισμένη Ευκλείδεια απόσταση διαιρούμενη με το πλήθος των δεδομένων, και η σταθερά ε ισούται με 0.002. Ο αλγόριθμος απαιτεί 40 λεπτά για την ολοκλήρωσή του έτσι ώστε να προσομοιώσουμε 1000 αποδεκτές τιμές για τη μέση τιμή. Η πραγματική εκ των υστέρων κατανομή που επικαλύπτει κάθε ιστόγραμμα υποδεικνύει ότι επιτυγχάνεται καλή προσαρμογή. Ο εκ των υστέρων μέσος ισούται με -0.0342 και η διασπορά με 0.045. Η μέση τιμή και η διασπορά των προσομοιωμένων τιμών είναι -0.0343 και 0.002 αντίστοιχα.

Κλείνοντας, στο **Σχήμα 1.4(γ)** παρουσιάζεται η κατανομή των προσομοιωμένων τιμών για τη μέση τιμή χρησιμοποιώντας την τρίτη απόσταση, την σταθμισμένη L_1 απόσταση διαιρούμενη με το πλήθος των δεδομένων, και η σταθερά ε ισούται με 0.04. Ο αλγόριθμος απαιτεί 28 λεπτά για την ολοκλήρωσή του προκειμένου να προσομοιώσουμε 1000 αποδεκτές τιμές για τη μέση τιμή. Η πραγματική εκ των υστέρων κατανομή που επικαλύπτει κάθε ιστόγραμμα υποδεικνύει ότι επιτυγχάνεται καλή προσαρμογή. Ο εκ των υστέρων μέσος ισούται με -0.0342 και η διασπορά με 0.045. Η μέση τιμή και η διασπορά των προσομοιωμένων τιμών είναι -0.0313 και 0.002 αντίστοιχα.



Σχήμα 1.4(β): Η προσαρμογή των εκ των υστέρων κατανομών χρησιμοποιώντας την δεύτερη απόσταση. Η κόκκινη γραμμή αντιπροσωπεύει την πραγματική εκ των υστέρων κατανομή.



Σχήμα 1.4(γ): Η προσαρμογή των εκ των υστέρων κατανομών χρησιμοποιώντας την τρίτη απόσταση. Η κόκκινη γραμμή αντιπροσωπεύει την πραγματική εκ των υστέρων κατανομή.

1.3.3 Εκτίμηση της εκ των υστέρων κατανομής χρησιμοποιώντας τον αλγόριθμο ABC στο Κανονικό μοντέλο με άγνωστες και τις δύο παραμέτρους

Στο τέταρτο και τελευταίο παράδειγμα του Κεφαλαίου 1.2.4 μελετήσαμε την περίπτωση όπου τα δεδομένα y προέρχονται από την κανονική κατανομή, $N(\mu, \sigma^2)$ με άγνωστες και τις δύο παραμέτρους. Η συζυγής ανάλυση του παραπάνω μοντέλου στο προηγούμενο κεφάλαιο θα μας καθοδηγήσει για να υλοποιήσουμε μία συμπερασματολογία βασισμένη στον αλγόριθμο ABC.

Ο προσδιορισμός της κατάλληλης μετρικής για τη σύγκριση των προσομοιωμένων δεδομένων Z με τα παρατηρούμενα δεδομένα Y είναι πρωτεύουσας σημασίας για την επιτυχή ολοκλήρωση του αλγορίθμου. Η σύγκριση των δύο συνόλων πραγματοπείται χρησιμοποιώντας δύο διαφορετικά μέτρα απόκλισης, τη σταθμισμένη Ευκλείδεια απόσταση και τη σταθμισμένη L_1 απόσταση διαιρούμενες με το πλήθος των δεδομένων. Αντίστοιχα, έχουμε

$$\rho_1 = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(\frac{y_i - z_i}{s}\right)^2}, \quad \rho_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - z_i|}{s}$$

Σε κάθε επανάληψη προσομοιώνουμε τη διασπορά από την αντίστροφη γάμμα κατανομή,
$\sigma_*^2 \sim \text{Inv-Gamma}(3,3)$, η οποία είναι ισοδύναμη με την κλιμακωτή αντίστροφη-χ^2 κατανομή. Η μέση τιμή μ^* του μοντέλου εξαρτάται από τη διασπορά σ_*^2 και προσομοιώνεται από την κανονική κατανομή με μέση τιμή 5 και διασπορά ίση με την τιμή της προηγούμενης προσομοίωσης, $\mu^* \sim N(5, \sigma_*^2)$. Κάθε ζεύγος τιμών (μ^*, σ_*^2) παράγει το προσομοιωμένο σύνολο δεδομένων z, σύμφωνα με τον αλγόριθμο απόρριψης ABC (**Αλγόριθμος 1.2).** Κρατάμε 1000 τιμές για τις παραμέτρους μ^* και σ_*^2, αυτές δηλαδή που παράγουν ένα προσομοιωμένο σύνολο δεδομένων z που είναι κοντά στο αρχικό y χρησιμοποιώντας ένα από τα δύο μέτρα απόκλισης ρ_1, ρ_2 και την κατάλληλη σταθερά ε σε κάθε περίπτωση.

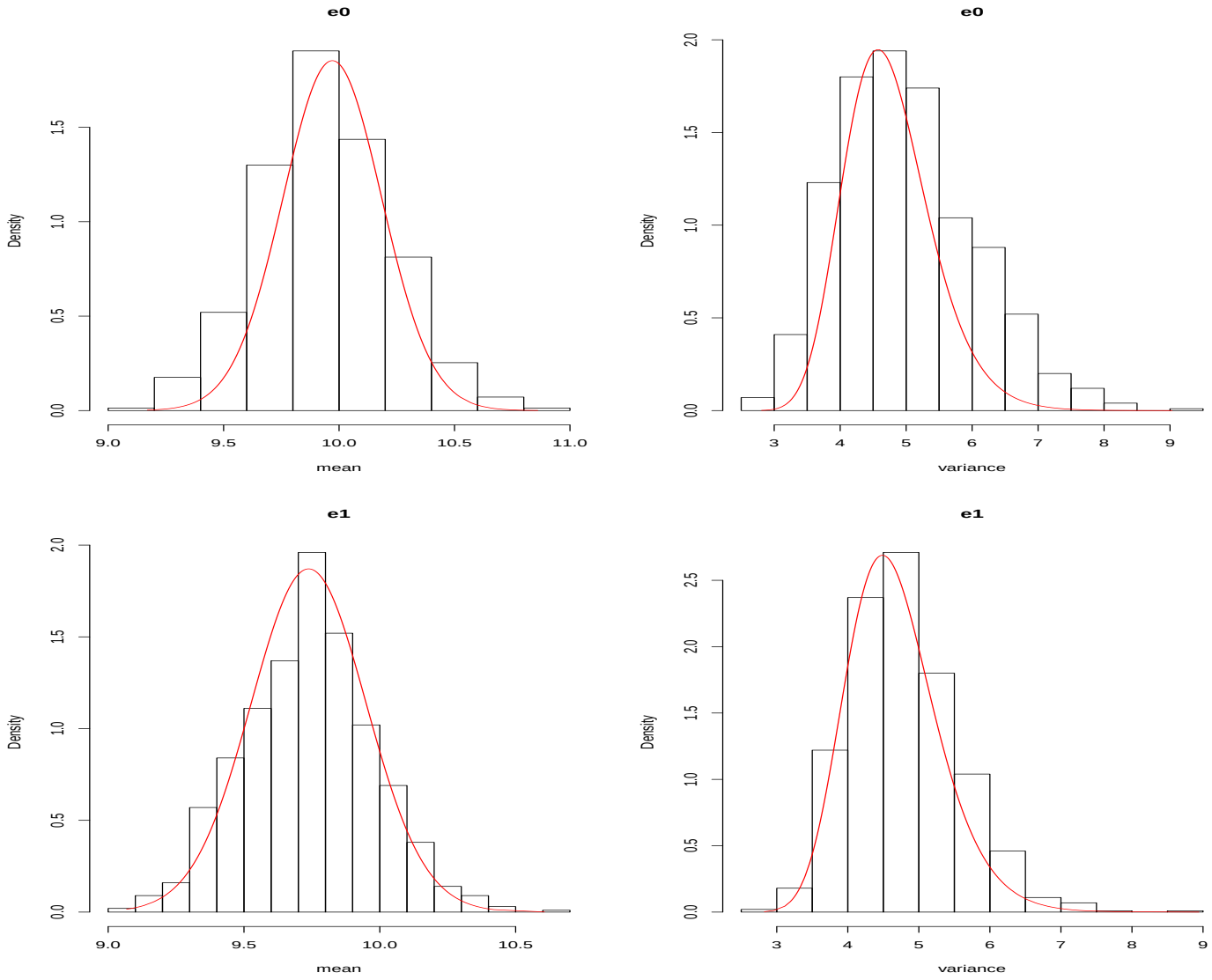
Α. Η σταθμισμένη Ευκλείδεια απόσταση ως μέτρο απόκλισης

Ξεκινώντας με την απόσταση ρ_1, ο αλγόριθμος ABC υλοποιείται σε δύο διαφορετικά σύνολα δεδομένων, στην πρώτη περίπτωση αποτελούμενο από $n = 100$ και στη δεύτερη από $n = 500$ ανεξάρτητες τιμές από την κανονική κατανομή N(10,4).

Αρχικά, μελετάμε τα αποτελέσματα που παίρνουμε στην πρώτη περίπτωση όπου το σύνολο των δεδομένων αποτελείται από 100 τιμές. Θα χρησιμοποιήσουμε δύο διαφορετικές τιμές για τη σταθερά ε, η πρώτη είναι η $\varepsilon_0 = 0.02$ και η δεύτερη είναι το μισό της πρώτης, δηλαδή $\varepsilon_1 = 0.01$. Με το πέρας του αλγορίθμου αποθηκεύονται 1000 προσομοιωμένες τιμές για τη μέση τιμή μ και τη διασπορά σ^2. Η απόσταση που χρησιμοποιούμε εκμεταλλεύεται όλη την πληροφορία από το πραγματικό και το προσομοιωμένο σύνολο με συνέπεια την αύξηση του χρόνου του αλγορίθμου.

Στο **Σχήμα 1.5** παρουσιάζονται τα ιστογράμματα των προσομοιωμένων τιμών για τη μέση τιμή μ και τη διασπορά σ^2 μαζί με τις εκ των υστέρων κατανομές. Τα ιστογράμματα των προσομοιωμένων τιμών μαζί με τις περιθώριες εκ των υστέρων κατανομές από τη συζυγή ανάλυση μας επιτρέπουν να εξετάσουμε την ακρίβεια του αλγορίθμου ABC.

Χρησιμοποιώντας τον αλγόριθμο απόρριψης ABC, (**Αλγόριθμος 1.2**), στο σύνολο δεδομένων των 100 τιμών και τη σταθερά $\varepsilon_0 = 0.02$ ο χρόνος που απαιτείται είναι 10 λεπτά ενώ χρησιμοποιώντας τη μικρότερη τιμή, $\varepsilon_1 = 0.01$ είναι 120 λεπτά. Η μικρότερη τιμή της σταθεράς απαιτεί τον δωδεκαπλάσιο χρόνο συγκριτικά με την μεγαλύτερη τιμή αλλά είναι εμφανές από το **Σχήμα 1.5** ότι τα αποτελέσματα που παίρνουμε από τον αλγόριθμο είναι πιο ικανοποιητικά.

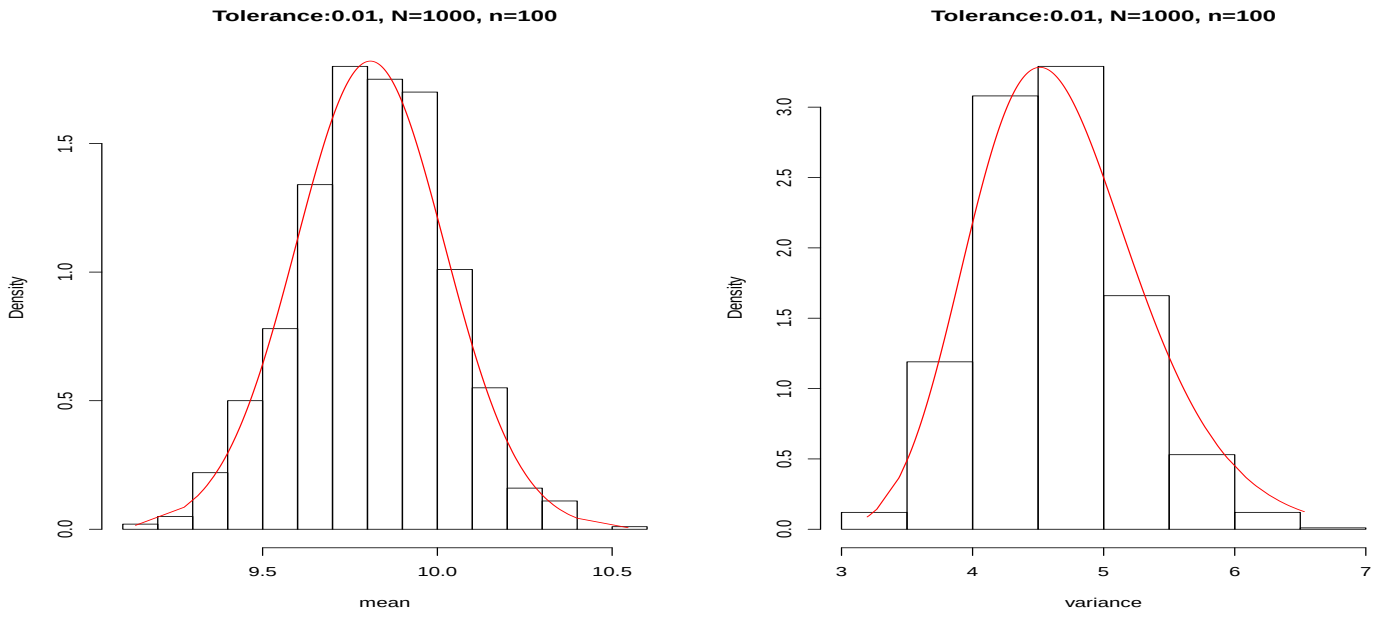


Σχήμα 1.5: Η προσαρμογή των εκ των υστέρων κατανομών σε σύνολο δεδομένων αποτελούμενο από $n = 100$ δοκιμές, χρησιμοποιώντας δύο διαφορετικές τιμές για τη σταθερά, την $\varepsilon_0 = 0.02$(πρώτη γραμμή) και την $\varepsilon_1 = 0.01$(δεύτερη γραμμή).

Όμως, το υπολογιστικό κόστος για την ολοκλήρωση του αλγορίθμου στη δεύτερη περίπτωση, όπου $\varepsilon_1 = 0.01$, είναι αξιοσημείωτα μεγάλο συγκριτικά με την πρώτη, όπου $\varepsilon_0 = 0.02$. Παράλληλα όμως η τιμή αυτή για τη σταθερά ε είναι απαραίτητη καθώς καθιστά τον αλγόριθμο ABC πιο ακριβή και αποτελεσματικό.

Για να μειωθεί το υπολογιστικό κόστος του αλγορίθμου και να διατηρήσουμε τη συγκεκριμένη τιμή για τη σταθερά ε χρησιμοποιούμε τον αλγόριθμο ABC-MCMC, (**Αλγόριθμος 1.3**).

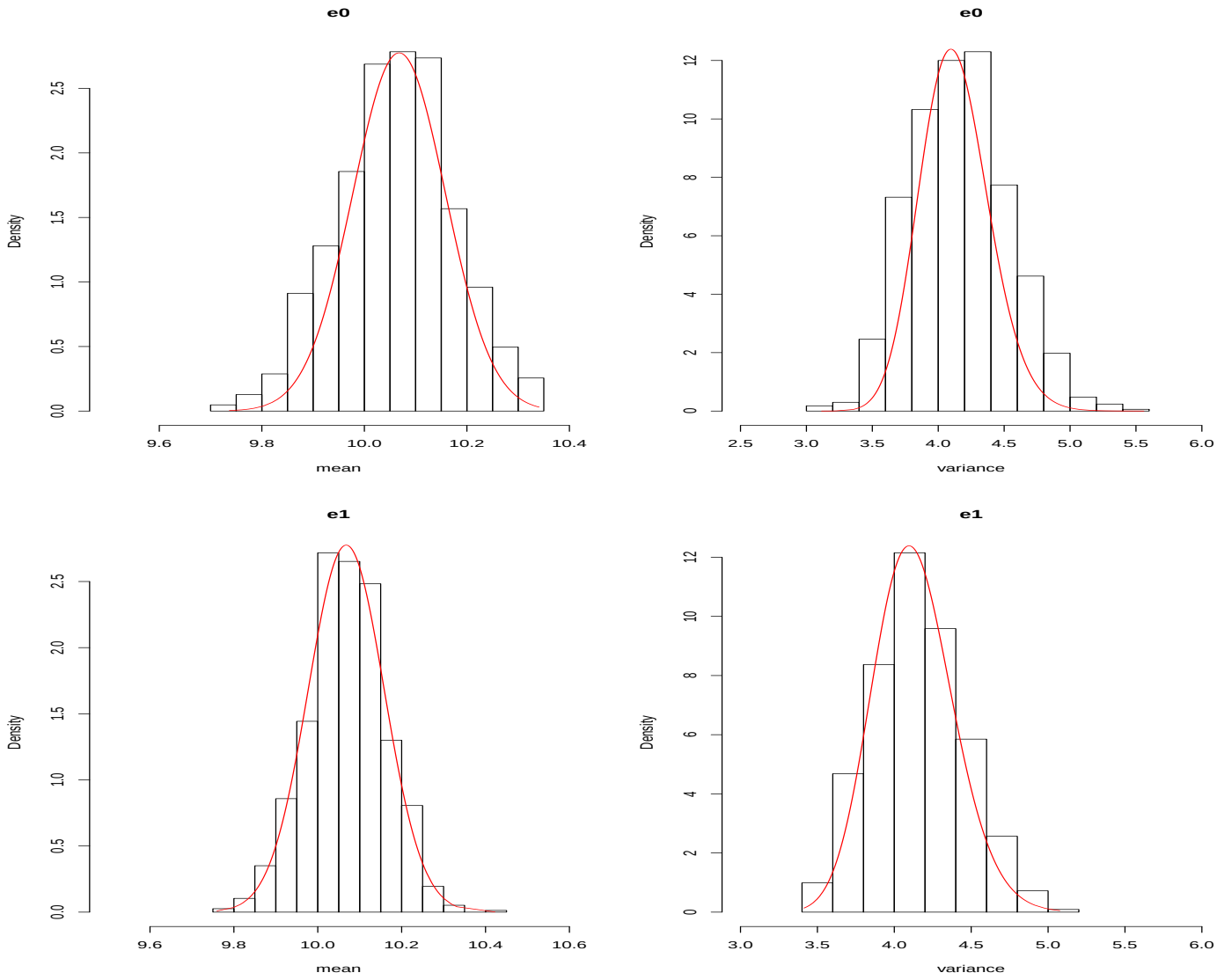
Στο **Σχήμα 1.6** παρουσιάζονται τα αποτελέσματα από τον αλγόριθμο ABC-MCMC. Η εκτίμηση της μέσης τιμής και της διασποράς του μοντέλου είναι πολύ ικανοποιητικές. Ο αλγόριθμος χρειάστηκε μόνο 3 λεπτά για να ολοκληρωθεί σε αντίθεση με τον αρχικό ο οποίος απαιτούσε 120 λεπτά.



Σχήμα 1.6: Η προσαρμογή των εκ των υστέρων κατανομών για τη μέση τιμή και τη διασπορά σε σύνολο δεδομένων αποτελούμενο από $n = 100$ δοκιμές χρησιμοποιώντας τη σταθερά $\varepsilon_1 = 0.01$ σύμφωνα με τον αλγόριθμο ABC-MCMC.

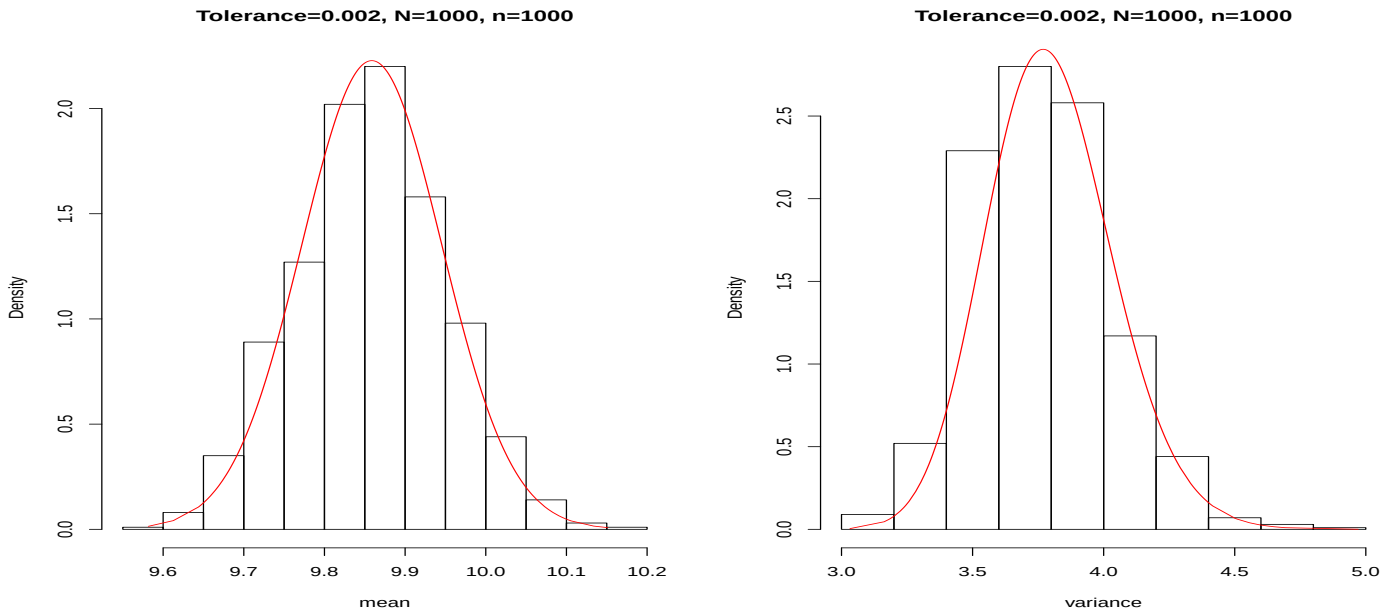
Στη συνέχεια, ο αλγόριθμος απόρριψης ABC, (**Αλγόριθμος 1.2**) υλοποιείται σε ένα σύνολο δεδομένων από 500 ανεξάρτητες τιμές από την κανονική κατανομή N(10,4). Κι εδώ θα χρησιμοποιήσουμε δύο διαφορετικές τιμές για τη σταθερά ε, η πρώτη είναι η $\varepsilon_0 = 0.004$ και η δεύτερη είναι το μισό της πρώτης, δηλαδή $\varepsilon_1 = 0.002$. Στο **Σχήμα 1.7** παρουσιάζονται τα ιστογράμματα των προσομοιωμένων τιμών για τη μέση τιμή μ και τη διασπορά σ^2 μαζί με τις εκ των υστέρων κατανομές.

Στο σύνολο δεδομένων των 500 τιμών και τη σταθερά $\varepsilon_0 = 0.004$ ο χρόνος που απαιτείται είναι 2 ώρες ενώ χρησιμοποιώντας τη μικρότερη τιμή, $\varepsilon_1 = 0.002$ είναι 45 ώρες. Ο συνολικός χρόνος που απαιτείται για τη τιμή ε_1 είναι αισθητά μεγαλύτερος αλλά τα αποτελέσματα που παίρνουμε είναι πιο ακριβή και η προσαρμογή αισθητά καλύτερη. Επίσης, παρατηρώντας το **Σχήμα 1.7** και συγκριτικά με το **Σχήμα 1.5** αντιλαμβανόμαστε ότι η αύξηση του συνόλου των δεδομένων επιφέρει και αύξηση της ακρίβειας του αλγορίθμου απόρριψης ABC, (**Αλγόριθμος 1.2**).



Σχήμα 1.7: Η προσαρμογή των εκ των υστέρων κατανομών σε σύνολο δεδομένων αποτελούμενο από $n = 500$ δοκιμές, χρησιμοποιώντας δύο διαφορετικές τιμές για τη σταθερά $\varepsilon_0 = 0.004$(πρώτη γραμμή) και $\varepsilon_1 = 0.002$(δεύτερη γραμμή).

Κι εδώ η χρήση του αλγορίθμου ABC-MCMC, (**Αλγόριθμος 1.3**), καθίσταται αναγκαία στη δεύτερη περίπτωση όπου ο αλγόριθμος απαιτεί 45 ώρες για τη διεξαγωγή αποτελεσμάτων. Στο **Σχήμα 1.8** παρουσιάζονται τα αποτελέσματα από τον αλγόριθμο ABC-MCMC. Η εκτίμηση της μέσης τιμής και της διασποράς του μοντέλου είναι πολύ ικανοποιητικές. Ο αλγόριθμος χρειάστηκε μόνο 4 λεπτά για να ολοκληρωθεί, σε αντίθεση με τον αρχικό ο οποίος απαιτούσε 45 ώρες. Ο αλγόριθμος ABC-MCMC είναι ικανός να κάνει καλή εκτίμηση των παραμέτρων και παράλληλα να επιτύχει μικρό υπολογιστικό κόστος.

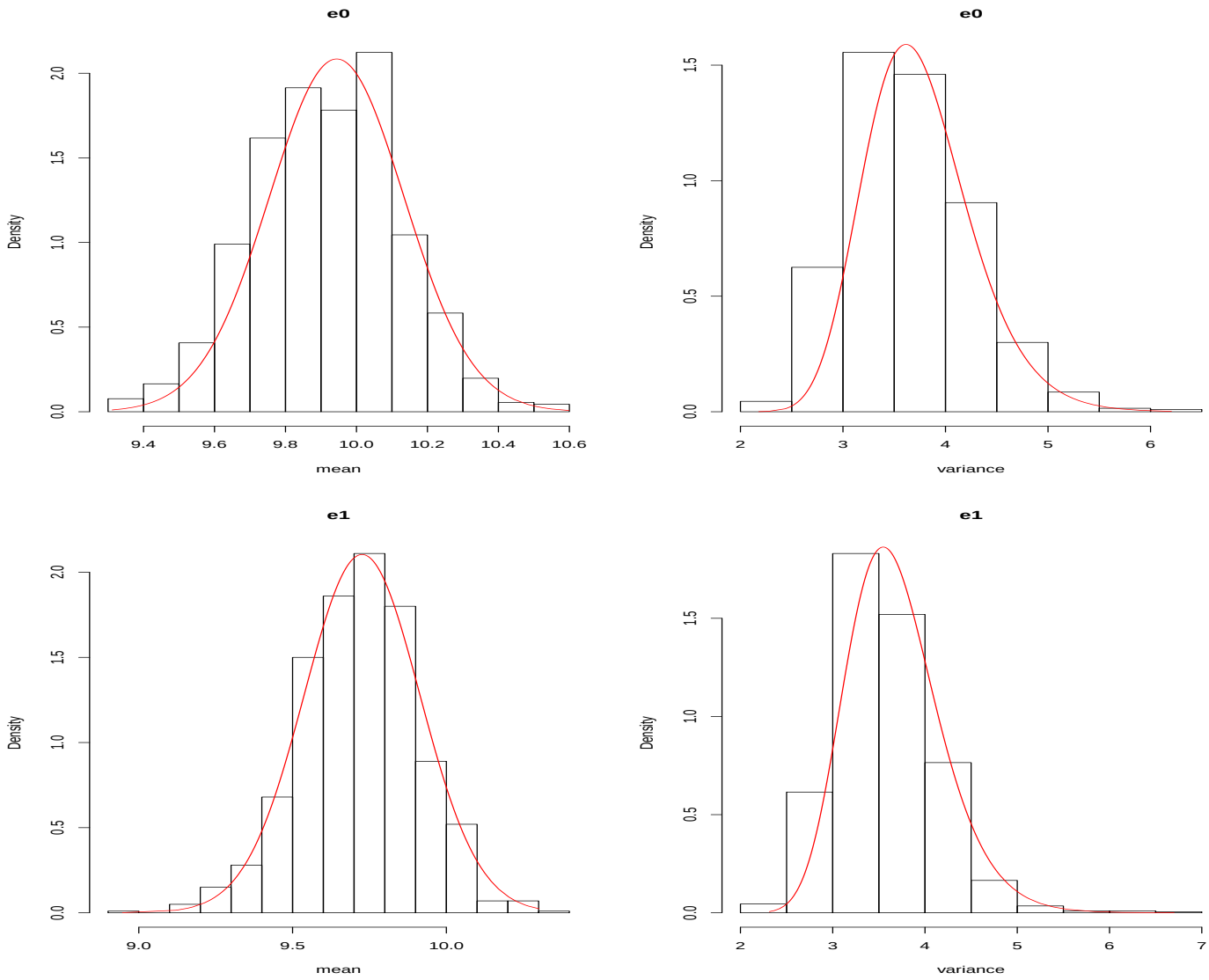


Σχήμα 1.8: Η προσαρμογή των εκ των υστέρων κατανομών για τη μέση τιμή και τη διασπορά σε σύνολο δεδομένων αποτελούμενο από $n = 500$ δοκιμές, χρησιμοποιώντας τη σταθερά $\varepsilon_1 = 0.002$ σύμφωνα με τον αλγόριθμο ABC-MCMC.

Β. Η σταθμισμένη L_1 απόσταση ως μέτρο απόκλισης

Στη συνέχεια, ενδιαφέρον παρουσιάζουν και τα αποτελέσματα που παίρνουμε χρησιμοποιώντας τη σταθμισμένη L_1 απόσταση διαιρούμενη με το πλήθος των δεδομένων. Ο αλγόριθμος απόρριψης ABC, (**Αλγόριθμος 1.2**), υλοποιείται σε ένα σύνολο δεδομένων από 100 και 500 ανεξάρτητες τιμές από την κανονική κατανομή $N(10, 4)$.

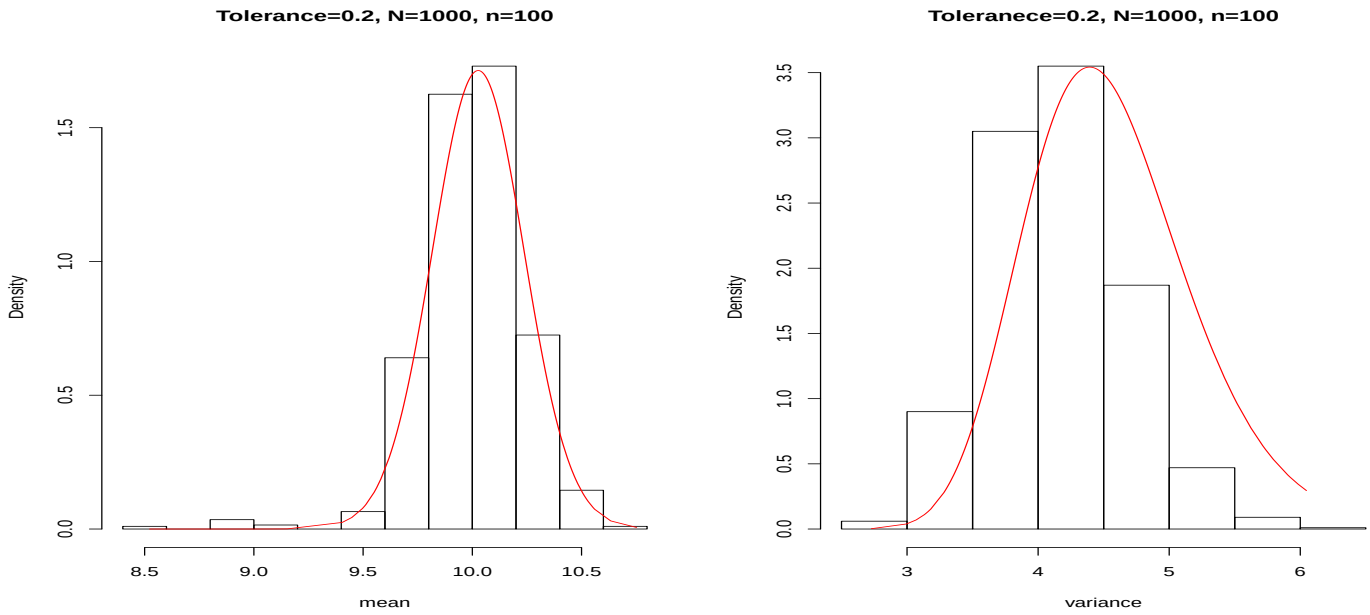
Ξεκινώντας με το σύνολο των 100 τιμών χρησιμοποιήσουμε δύο διαφορετικές τιμές για τη σταθερά ε, η πρώτη είναι η $\varepsilon_0 = 0.4$ και η δεύτερη είναι $\varepsilon_1 = 0.2$, δηλαδή το μισό της πρώτης. Στο **Σχήμα 1.9** παρουσιάζονται τα ιστογράμματα των προσομοιωμένων τιμών για τη μέση τιμή μ και τη διασπορά σ^2 μαζί με τις εκ των υστέρων κατανομές. Χρησιμοποιώντας τον αλγόριθμο απόρριψης ABC στο σύνολο δεδομένων των 100 τιμών, ο χρόνος που απαιτείται είναι 4 ώρες και 10 λεπτά με τη σταθερά $\varepsilon_0 = 0.4$ και 19 ώρες και 30 λεπτά με τη σταθερά $\varepsilon_1 = 0.2$ αντίστοιχα. Από τα διαγράμματα, είναι εμφανές ότι καθώς η σταθερά μειώνεται, τα αποτελέσματα που παίρνουμε από τον αλγόριθμο είναι πιο ικανοποιητικά.



Σχήμα 1.9: Η προσαρμογή των εκ των υστέρων κατανομών σε σύνολο δεδομένων αποτελούμενο από $n = 100$ δοκιμές, χρησιμοποιώντας δύο διαφορετικές τιμές για τη σταθερά $\varepsilon_0 = 0.4$(πρώτη γραμμή) και $\varepsilon_1 = 0.2$(δεύτερη γραμμή).

Η επιτυχία ενός αλγορίθμου όμως, κρίνεται τόσο από την ακρίβεια των αποτελεσμάτων όσο και από το μικρό υπολογιστικό κόστος. Ο αλγόριθμος απόρριψης ABC, (**Αλγόριθμος 1.2**), είναι ικανός να εκτιμήσει με ακρίβεια τις παραμέτρους του μοντέλου όμως υστερεί σε ταχύτητα. Ειδικά σε αυτή την περίπτωση όπου οι παράμετροι του μοντέλου είναι δύο υπάρχει ακόμα μεγαλύτερη καθυστέρηση στην υλοποίησή του φτάνοντας τις 19 ώρες και 30 λεπτά. Σε μία προσπάθεια να μειώσουμε το υπολογιστικό κόστος θα γίνει η χρήση του αλγορίθμου ABC-MCMC, (**Αλγόριθμος 1.3**).

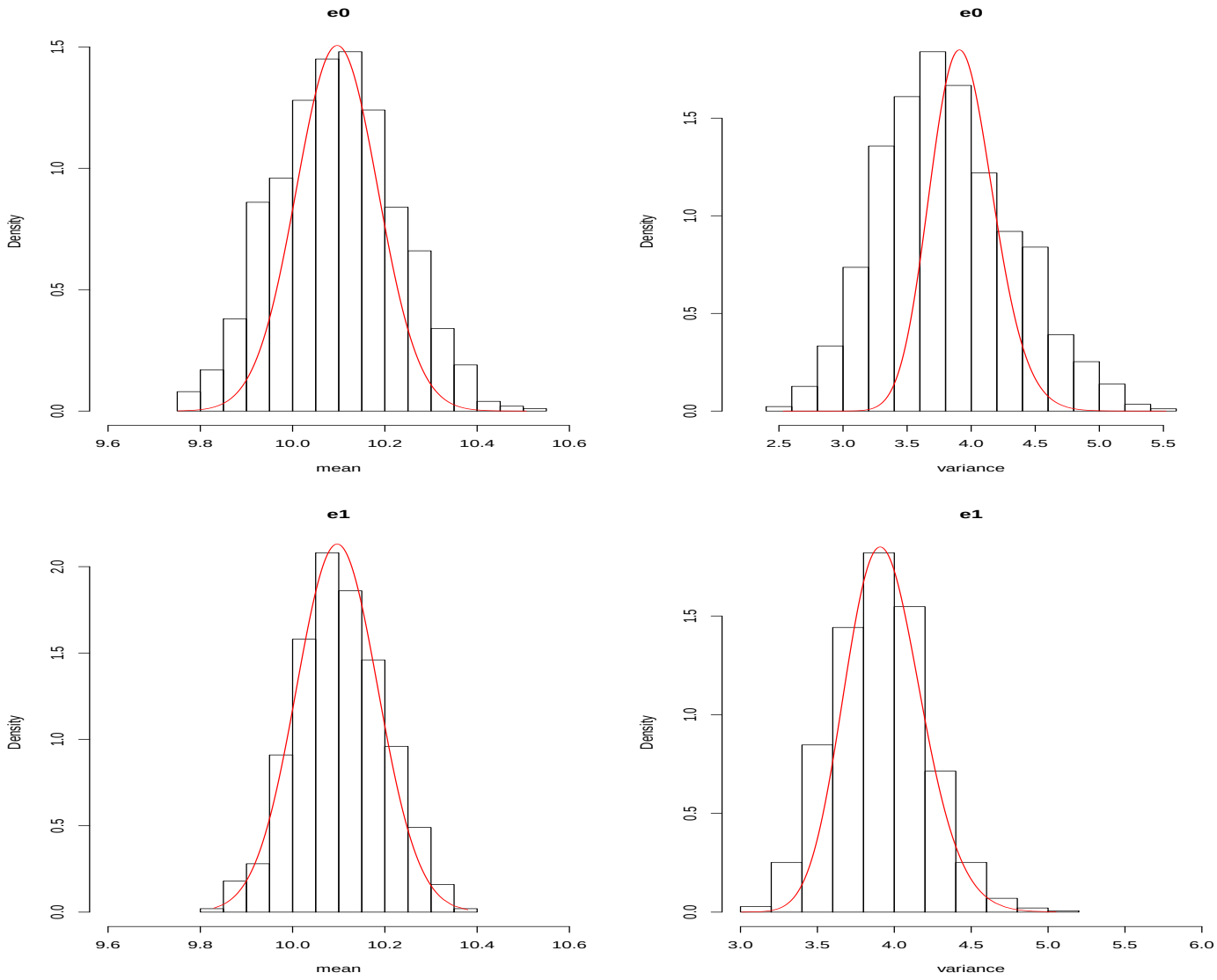
Στο **Σχήμα 1.10** παρουσιάζονται τα αποτελέσματα από τον αλγόριθμο ABC-MCMC. Η εκτίμηση της μέσης τιμής και της διασποράς του μοντέλου είναι πολύ ικανοποιητικές. Ο αλγόριθμος χρειάστηκε 1 ώρα και 40 λεπτά για να ολοκληρωθεί, σε αντίθεση με τον αρχικό ο οποίος απαιτούσε 19 ώρες και 30 λεπτά. Συμπεραίνουμε ότι και σε αυτή την περίπτωση ο αλγόριθμος ABC-MCMC είναι ικανός να κάνει καλή εκτίμηση των παραμέτρων και παράλληλα να επιτύχει μικρό υπολογιστικό κόστος.



Σχήμα 1.10: Η προσαρμογή των εκ των υστέρων κατανομών για τη μέση τιμή και τη διασπορά σε σύνολο δεδομένων αποτελούμενο από $n = 100$ δοκιμές, χρησιμοποιώντας τη σταθερά $\varepsilon_1 = 0.2$ σύμφωνα με τον αλγόριθμο ABC-MCMC.

Κλείνοντας, ο αλγόριθμος υλοποιείται σε ένα σύνολο δεδομένων από 500 ανεξάρτητες τιμές από την κανονική κατανομή N(10,4). Κι εδώ θα χρησιμοποιήσουμε δύο διαφορετικές τιμές για τη σταθερά ε, η πρώτη είναι η $\varepsilon_0 = 0.4$ και η δεύτερη είναι το μισό της πρώτης, δηλαδή $\varepsilon_1 = 0.2$. Στο **Σχήμα 1.11** παρουσιάζονται τα ιστογράμματα των προσομοιωμένων τιμών για τη μέση τιμή μ και τη διασπορά σ^2 μαζί με τις εκ των υστέρων κατανομές.

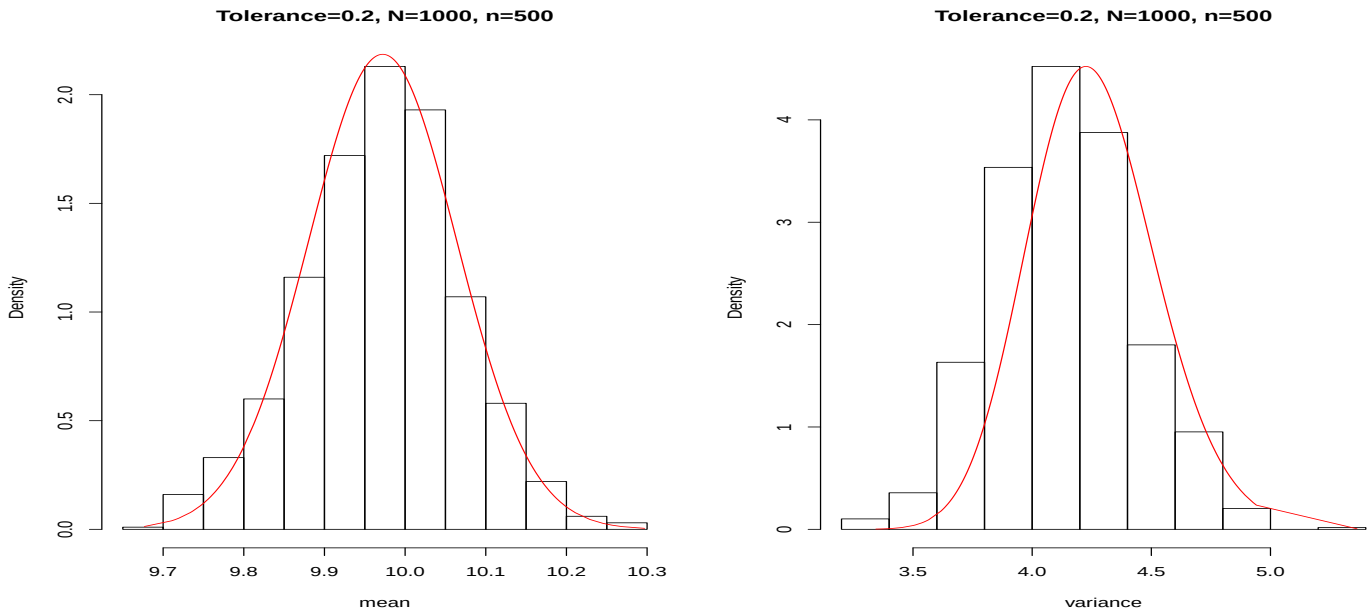
Χρησιμοποιώντας τον αλγόριθμο απόρριψης ABC, (**Αλγόριθμος 1.2**), στο σύνολο δεδομένων των 500 τιμών, ο χρόνος που απαιτείται είναι 30 λεπτά με τη σταθερά $\varepsilon_0 = 0.4$ και 18 ώρες με τη σταθερά $\varepsilon_1 = 0.2$. Από τα διαγράμματα, είναι εμφανές ότι στην πρώτη γραμμή, όπου έχουμε τη σταθερά $\varepsilon_0 = 0.4$ δεν επιτυγχάνεται καλή προσαρμογή ενώ στη δεύτερη γραμμή που η σταθερά μειώνεται, τα αποτελέσματα που παίρνουμε από τον αλγόριθμο είναι αισθητά πιο ικανοποιητικά.



Σχήμα 1.11: Η προσαρμογή των εκ των υστέρων κατανομών σε σύνολο δεδομένων αποτελούμενο από $n = 500$ δοκιμές, χρησιμοποιώντας δύο διαφορετικές τιμές για τη σταθερά $\varepsilon_0 = 0.4$(πρώτη γραμμή) και $\varepsilon_1 = 0.2$(δεύτερη γραμμή).

Είναι απαραίτητο να παρατηρήσουμε το **Σχήμα 1.9** και το **Σχήμα 1.11** και να γίνει η μεταξύ τους σύγκριση. Στην περίπτωση που έχουμε περισσότερα δεδομένα, όπως στο **Σχήμα 1.11**, τα αποτελέσματα που μας δίνει ο αλγόριθμος ABC είναι αδιαμφισβήτητα πιο ακριβή. Συνεπώς, η ακρίβεια του αλγορίθμου ABC θεμελιώνεται από τη στιγμή που χρησιμοποιούμε μεγάλο πλήθος δεδομένων, μειώνουμε τη σταθερά ε και χρησιμοποιούμε την κατάλληλη απόσταση για τη σύγκριση των πραγματικών και των προσομοιωμένων δεδομένων.

Επίσης, εύλογα εμφανίζεται κι εδώ η ανάγκη για μείωση του υπολογιστικού κόστους, όπως παρατηρήσαμε και στις προηγούμενες περιπτώσεις. Η μεγάλη χρονική διάρκεια του αλγορίθμου απόρριψης στην περίπτωση όπου $\varepsilon_1 = 0.2$ μας οδηγεί στον αλγόριθμο ABC-MCMC.



Σχήμα 1.12: Η προσαρμογή των εκ των υστέρων κατανομών για τη μέση τιμή και τη διασπορά σε σύνολο δεδομένων αποτελούμενο από $n = 500$ δοκιμές, χρησιμοποιώντας τη σταθερά $\varepsilon_1 = 0.2$ σύμφωνα με τον αλγόριθμο ABC-MCMC.

Στο **Σχήμα 1.12** παρουσιάζονται τα αποτελέσματα από τον αλγόριθμο ABC-MCMC. Η εκτίμηση της μέσης τιμής και της διασποράς του μοντέλου είναι πολύ ικανοποιητικές. Ο αλγόριθμος χρειάστηκε μόνο 4 λεπτά για να ολοκληρωθεί, σε αντίθεση με τον αρχικό ο οποίος απαιτούσε 18 ώρες. Ο αλγόριθμος ABC-MCMC είναι αδιαμφισβήτητα αποδοτικότερος συγκριτικά με τον αλγόριθμο απόρριψης ABC και η ανωτερότητά του γίνεται ακόμα πιο έκδηλη σε μοντέλα με παραπάνω από μία παραμέτρους.

Κεφάλαιο 2

Μπεϋζιανές Μέθοδοι Επιλογής Μοντέλου και ο αλγόριθμος ABC

2.1 Σύγκριση μοντέλων και έλεγχος υποθέσεων

Υπό τη Μπεϋζιανή προσέγγιση, η σύγκριση μεταξύ των υποψήφιων μοντέλων και η απόφαση για το ποιο μοντέλο είναι ικανό να περιγράψει καλύτερα τα δεδομένα καθορίζεται από τη σύγκριση των εκ των υστέρων πιθανοτήτων των μοντέλων. Έστω ότι τα δεδομένα τα οποία έχουμε παρατηρήσει $y = (y_1, y_2, ..., y_n)$ έχουν προέλθει κάτω από τις συνθήκες που έχουν δημιουργήσει ένα από τα δύο ακόλουθα μοντέλα, M_0 και M_1, με πιθανότητα $f(y|M_0)$ και $f(y|M_1)$ αντίστοιχα. Σκοπός μας είναι να πραγματοποιήσουμε τον ακόλουθο έλεγχο υποθέσεων

$$H_0 : M = M_0$$
$$H_1 : M = M_1$$

δοθέντων των εκ των προτέρων πιθανοτήτων $\pi(M_0)$ και $\pi(M_1) = 1-\pi(M_0)$ για το μοντέλο M_0 και M_1 αντίστοιχα. Η κατανομή των δεδομένων σε συνδυασμό με την εκ των προτέρων πιθανότητα παράγουν την εκ των υστέρων πιθανότητα $f(M_0|y)$ και $f(M_1|y)$ για τα δύο μοντέλα. Έτσι, ο έλεγχος υποθέσεων για ποιό από τα δύο μοντέλα είναι το καταλληλότερο περνάει στη σύγκριση των εκ των υστέρων πιθανοτήτων $f(M_0|y)$ και $f(M_1|y)$. Αν $f(M_0|y) > f(M_1|y)$ τότε έχουμε ενδείξεις να απορρίψουμε το μοντέλο M_1. Από το θεώρημα του Bayes και το Θεώρημα Ολικής Πιθανότητας έχουμε:

$$\begin{aligned} f(M_i|y) &= \frac{f(M_i \cap y)}{f(y)} \\ &= \frac{f(y|M_i)\pi(M_i)}{f(y|M_0)\pi(M_0) + f(y|M_1)\pi(M_1)} \qquad i = 0, 1 \end{aligned} \tag{2.1}$$

Για λόγους ευκολίας, ο υπολογισμός της κάθε εκ των υστέρων πιθανότητας ξεχωριστά μπορεί να αποφευχθεί υπολογίζοντας μία ισοδύναμη απλούστερη μορφή, η οποία καλείται posterior odds. Posterior odds ονομάζεται ο λόγος των εκ των υστέρων κατανομών και ισούται με το κλάσμα $\frac{f(M_1|y)}{f(M_0|y)}$ και χρησιμοποιώντας τη σχέση (2.1), ισούται με

$$\frac{f(M_1|y)}{f(M_0|y)} = \frac{f(y|M_1)}{f(y|M_0)} \frac{\pi(M_1)}{\pi(M_0)} \tag{2.2}$$

Το κλάσμα $B_{10} := \dfrac{f(y|M_1)}{f(y|M_0)}$ που εμφανίζεται στο δεξί μέλος της σχέσης (2.2) καλείται **παράγοντας του Bayes (Bayes Factor).** Συνεπώς, αντί να ελέγχουμε αν $f(M_0|y) > f(M_1|y)$, σύγκριση η οποία περιλαμβάνει αρκετούς υπολογισμούς, εξετάζουμε ισοδύναμα αν το κλάσμα $\dfrac{f(M_1|y)}{f(M_0|y)}$ είναι μεγαλύτερο ή μικρότερο της μονάδας(απόρριψη της υπόθεσης H_0 ή απόρριψη της υπόθεσης H_1 αντίστοιχα).

Η ευελιξία στη χρήση του παράγοντα του Bayes παρουσιάζεται στο γεγονός ότι ο υπολογισμός του παρανομαστή στη σχέση (2.1) μπορεί να αποφευχθεί. Επίσης, αν οι εκ των προτέρων κατανομές των μοντέλων είναι ίσες, $\pi(M_1) = \pi(M_0)$, ο παράγοντας του Bayes ισούται με το λόγο των εκ των υστέρων κατανομών.

Ο παράγοντας του Bayes συγκρίνει πόσο πιθανότερη είναι εκ των υστέρων η H_0 από την H_1 με το πόσο πιθανότερη ήταν εκ των προτέρων. Η πρώτη αναφορά στον παράγοντα του Bayes έγινε από τον Jeffreys, ο οποίος καθόρισε και μία κλίμακα αξιολόγησης της τιμής του και αργότερα τροποποιήθηκε από τους Kass and Raftery. Στον πίνακα 2.1 παρουσιάζεται μία ερμηνεία των τιμών του παράγοντα του Bayes σύμφωνα με τους Kass and Raftery.

$2\log(B_{10})$	B_{10}	Ένδειξη κατά της H_0
0-2	1-3	Πολύ αδύναμη
2-6	3-20	Θετική
6-10	20-150	Ισχυρή
> 10	> 150	Πολύ ισχυρή

Πίνακας 2.1 Ερμηνεία του παράγοντα του Bayes του μοντέλου M_1 έναντι του M_0.

Στην πιο απλή περίπτωση όπου τα δύο μοντέλα περιλαμβάνουν μόνο μία άγνωστη παράμετρο ο παράγοντας του Bayes

$$B_{10} := \frac{f(y|M_1)}{f(y|M_0)} \tag{2.3}$$

είναι ο λόγος των πιθανοφανειών, όπως είναι φανερό και από τη σχέση (2.3). Όμως, σε πολλές περιπτώσεις τα υπό σύγκριση μοντέλα περιέχουν πάνω από μία άγνωστες παραμέτρους. Έτσι, ο υπολογισμός του παράγοντα του Bayes παρουσιάζει ορισμένες δυσκολίες καθώς οι όροι του κλάσματος υπολογίζονται ολοκληρώνοντας ως προς τον παραμετρικό χώρο,

$$f(y|M_i) = \int f(y|\theta_{ip}, M_i)\pi(\theta_{ip}|M_i)d\theta_{ip} \tag{2.4}$$

όπου p είναι η διάσταση του διανύσματος των παραμέτρων και $\pi(\theta_{ip}|M_i)$ είναι η εκ των προτέρων κατανομή της παραμέτρου θ_{ip} υπό το μοντέλο M_i. Η ποσότητα $f(y|M_i)$ καλείται περιθώρια πιθανοφάνεια. Αντιπροσωπεύει δηλαδή, την πιθανότητα, ή την πυκνότητα πιθανότητας σε περίπτωση που έχουμε συνεχή δεδομένα, να παρατηρήσουμε τα πραγματικά δεδομένα πριν γίνει διαθέσιμη οποιαδήποτε πληροφορία, υποθέτωντας ότι το μοντέλο M_i είναι αυτό από το οποίο έχουν προέλθει τα πραγματικά δεδομένα.

Αξιοσημείωτο είναι το γεγονός ότι για τον υπολογισμό του παράγοντα του Bayes, (2.3), όλες οι σταθερές που εμφανίζονται στην περιθώρια πιθανοφάνεια $f(y|M_i)$ πρέπει να παραμείνουν. Το αποτέλεσμα

είναι ο υπολογισμός του ολοκληρώματος στη σχέση (2.4) να είναι ανέφικτο να υπολογιστεί αναλυτικά και έτσι, να είναι απαραίτητη η χρήση αριθμητικών μεθόδων. Ο ακριβής υπολογισμός του ολοκληρώματος (2.4) είναι εφικτός όταν χρησιμοποιούμε συζυγείς εκ των προτέρων κατανομές στην ανάλυσή μας.

Από την άλλη, οι περισσότερες αριθμητικές μέθοδοι είναι μη αποτελεσματικές εξαιτίας του μεγάλου μεγέθους των δειγμάτων και της μεγάλης διάστασης της παραμέτρου προς εκτίμηση. Γι΄ αυτό, οι μέθοδοι Monte Carlo και οι ασυμπτωτικές προσεγγίσεις αποτελούν ιδιαίτερα χρήσιμα εργαλεία. Ένας τρόπος ασυμπτωτικής προσέγγισης του παράγοντα του Bayes είναι η χρήση του κριτηρίου BIC (Bayesian Information Criterior), το οποίο ορίζεται ως:

$$S = \log f\left(y|\widehat{\theta_{1p}}, M_1\right) - \log f\left(y|\widehat{\theta_{2p}}, M_2\right) - \frac{1}{2}(p_{M_1} - p_{M_2})\log n \tag{2.5}$$

όπου $\widehat{\theta_{ip}}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας υπό το μοντέλο M_i, p_{M_i} είναι η διάσταση του διανύσματος θ_i και n είναι το μέγεθος του δείγματος. Καθώς $n \to +\infty$ η ποσότητα αυτή, η οποία καλείται κριτήριο Schwarz, ικανοποιεί

$$\frac{S - \log B_{12}}{\log B_{12}} \to 0 \tag{2.6}$$

και έτσι μπορεί να θεωρηθεί μία προσέγγιση του λογαρίθμου του παράγοντα Bayes. Έχει καθιερωθεί ότι μείον δύο φορές το κριτήριο Schwarz διαμορφώνει το κριτήριο BIC.

Οι Kass & Wasserman (1995) μελέτησαν υπό ποιες προϋποθέσεις η σχέση (2.6) ευσταθεί και εξασφάλισαν μία προσαρμοσμένη μορφή της προσέγγισης όπου ήταν απαραίτητο. Απέδειξαν ότι για ένα μεγάλο εύρος εκ των προτέρων κατανομών η σχέση (2.5) παρέχει μία χρήσιμη προσέγγιση και υποστήριξαν ότι δεν είναι απαραίτητο μεγάλο μέγεθος δείγματος για να έχουμε επαρκή αποτελέσματα. Γενικά, αν και η σχέση (2.5) θεωρείται η πιο απλή και όχι πάντα η καλύτερη προσέγγιση του παράγοντα του Bayes, είναι αρκετά ικανή να εκτιμήσει την ένδειξη υπέρ της αρχικής υπόθεσης σε περιπτώσεις όπου οι εκ των προτέρων κατανομές δύσκολα υπολογίζονται.

2.2 Μέθοδοι ABC για Μπεϋζιανή επιλογή μοντέλου

Εκτός από την εκτίμηση παραμέτρων ο αλγόριθμος ABC μπορεί να χρησιμοποιηθεί για τον υπολογισμό των εκ των υστέρων πιθανοτήτων των υποψήφιων μοντέλων. Σε τέτοιες εφαρμογές, μία επιλογή είναι να χρησιμοποιήσουμε τον αλγόριθμο rejection-sampling με έναν ιεραρχικό τρόπο. Πρώτα, επιλέγεται ένα μοντέλο από την εκ των προτέρων κατανομή των μοντέλων και ύστερα, δοθέντος του μοντέλου που επιλέγθηκε, γίνεται προσομοίωση των παραμέτρων του μοντέλου από την εκ των προτέρων κατανομή. Η προσομοίωση των δεδομένων πραγματοποιείται όπως στο single-model ABC. Η σχετική συχνότητα αποδοχής των διαφορετικών μοντέλων προσεγγίζει την εκ των υστέρων κατανομή αυτών των μοντελων. Κατ΄ επέκταση, ο λόγος των εκ των υστέρων κατανομών δύο μοντέλων, που σχετίζεται με τον παράγοντα του Bayes, προσεγγίζεται από το λόγο των σχετικών πιθανοτήτων των μοντέλων. Στην πράξη, όπως θα συζητηθεί ακολούθως, η προσέγγιση αυτή μπορεί να γίνει ιδιαίτερα ευαίσθητη ως προς την επιλογή των εκ των προτέρων κατανομών των παραμέτρων και των στατιστικών συναρτήσεων.

Για την κατασκευή του αλγορίθμου, εισάγουμε μία νέα παράμετρο για το συμβολισμό του μοντέλου, έστω $m \in \{1, ..., M\}$, όπου M είναι ο αριθμός των υπό μελέτη μοντέλων. Συνεπώς, το διάνυσμα

των παραμέτρων εξαρτάται και από το εκάστοτε μοντέλο υποψήφιο προς επιλογή και γι΄ αυτό συμβολίζουμε $\theta_m = \left(\theta_m^{(1)}, ..., \theta_m^{(k_m)}\right)$, όπου $m = 1, 2, ..., M$ και k_m ο αριθμός των παραμέτρων στο μοντέλο m.

Ξεκινώντας, χρησιμοποιούμε μία εκ των προτέρων κατανομή, η οποία εκφράζει την πιθανότητα επιλογής του μοντέλου m, έστω $\pi(m)$. Για το μοντέλο που επιλέγεται, γίνεται προσομοίωση τιμών για τις άγνωστες παραμέτρους και η διαδικασία συνεχίζεται όπως και στον Αλγόριθμο 1.2 στο Κεφάλαιο 1.3.

1. Για $i = 1, ..., N$:
2. Επανάλαβε μέχρι $\rho(\eta(Z), \eta(Y)) \leq \varepsilon_0$
   α. Επιλογή μοντέλου m με πιθανότητα $\pi(m)$
   β. Προσομοίωση τιμής θ^* από την εκ των προτέρων κατανομή $f_m(\theta^*)$
   γ. Προσομοίωση τιμής z από την πιθανοφάνεια $f_m(z|\theta^*)$
3. Θέσε $\theta_i = \theta^*$ και $m_i = m$

Αλγόριθμος 2.1. Likelihood-free model choice sampler(ABC-MC)

Σκοπός της υλοποίησης του παραπάνω αλγορίθμου είναι η επιλογή του καταλληλότερου μοντέλου υπολογίζοντας προσεγγιστικά την περιθώρια εκ των υστέρων κατανομή του κάθε μοντέλου m, $f(m|y)$. Η εκ των υστέρων πιθανότητα $f(m|y)$ εκτιμάται από τη συχνότητα του πλήθους των αποδοχών του μοντέλου m,

$$f(m|y) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{m_i = m} \tag{2.7}$$

Στη συνέχεια χρησιμοποιούμε γνωστές κατανομές για να διασφαλίσουμε την ακρίβεια και την αποτελεσματικότητα της μεθόδου. Ξεκινάμε με το διωνυμικό μοντέλο, προχωράμε με το κανονικό μοντέλο με γνωστή διασπορά και κλείνουμε με το κανονικό μοντέλο με άγνωστες και τις δύο παραμέτρους.

2.2.1 Διωνυμικό μοντέλο

Στο διωνυμικό μοντέλο θα πραγματοποιήσουμε τη σύγκριση των μοντέλων M_0 και M_1 όπου,

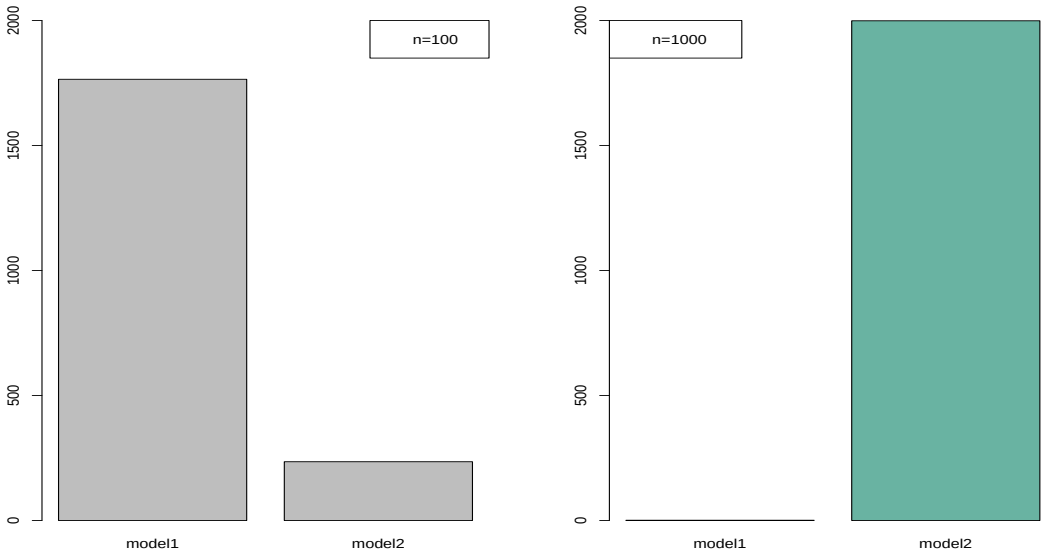
$$\begin{aligned} M_0 &: \ y_i \sim B(p_0), \ \text{ με } \ p_0 = 0.6 \\ M_1 &: \ y_i \sim B(p_0), \ \ i = 1, ..., n \ \text{ όπου } \ p_0 \sim Beta(1, 1) \end{aligned}$$

Ξεκινώντας, θεωρούμε ότι είναι ισοπίθανο να διαλέξουμε είτε το πρώτο είτε το δεύτερο μοντέλο και δίνουμε εκ των προτέρων πιθανότητα και στα δύο μοντέλα ίση με 0.5, $\pi(M_0) = \pi(M_1) = 0.5$. Προσομοιώνουμε ένα σύνολο τιμών y από την κατανομή Bernoulli με πιθανότητα επιτυχίας $p = 0.7$ και χρησιμοποιώντας τον Αλγόριθμο 2.1 θα το συγκρίνουμε σε κάθε επανάληψη με το σύνολο τιμών z που παράγει είτε το μοντέλο M_0 είτε το M_1. Η σύγκριση των συνόλων υλοποιείται χρησιμοποιώντας την απόσταση

$$\rho(Z, Y) = \frac{1}{n}|Z - Y|$$

η οποία αναφερθηκε στο Κεφάλαιο 1. Με το πέρας του αλγορίθμου έχουμε 2000 προσομοιωμένες τιμές για την παράμετρο p_0 καθώς και το μοντέλο από το οποίο κρατήθηκε.

Ο αλγόριθμος εκτελείται δύο φορές, την πρώτη προσομοιώνουμε $n = 100$ τιμές για να δημιουργήσουμε το σύνολο y, ενώ τη δεύτερη προσομοιώνουμε $n = 1000$ ανεξάρτητες τιμές από την κατανομή Bernoulli. Θα παρατηρήσουμε ότι καθώς το πλήθος των τιμών αυξάνεται, η ακρίβεια του αλγορίθμου επίσης αυξάνεται και τα αποτελέσματα διαφοροποιούνται δραματικά.



Σχήμα 2.1: Η σύγκριση του μοντέλου M_0, (model1), με το μοντέλο M_1, (model2), για διαφορετικό πλήθος προσομοιωμένων δεδομένων.

Στην πρώτη δοκιμή, παρατηρούμε οτι το πλήθος των φορών που αποδεχόμαστε το πρώτο μοντέλο M_0 ισούται με 1765 και αντίστοιχα οι φορές που αποδεχόμαστε το δεύτερο μοντέλο M_1 είναι 235, **Σχήμα 2.1**. Υπολογίζοντας τον παράγοντα του Bayes χρησιμοποιώντας την προσέγγιση (2.7) για την πιθανότητα $f(M_i|y)$ βρίσκουμε να ισούται με $\widehat{B_{10}} = \dfrac{235}{1765} = \dfrac{1}{7.51}$. Ο αλγόριθμος κρατάει περισσότερες τιμές του μοντέλου M_0 και δεν έχουμε επαρκείς ενδείξεις για να το απορρίψουμε.

Στο συγκεκριμένο παράδειγμα, το διωνυμικό μοντέλο μας επιτρέπει να υπολογίσουμε και αναλυτικά τον παράγοντα του Bayes από τη σχέση (2.3). Πρώτα, υπολογίζουμε την πιθανοφάνεια του μοντέλου M_1, για το οποίο ισχύει ότι $p_0 \sim Beta(1,1)$ ή ισοδύναμα $p_0 \sim U(0,1)$.

$$\begin{aligned} f(y|M_1) &= \int_0^1 f(y|p_0) f(p_0|M_1) dp_0 \qquad (2.8) \\ &= \int_0^1 \binom{n}{y} p_0^y (1-p_0)^{n-y} \cdot \frac{p_0^{1-1}(1-p_0)^{1-1}}{B(1,1)} dp_0 \\ &= \binom{n}{y} \int_0^1 p_0^y (1-p_0)^{n-y} dp_0 \\ &= \binom{n}{y} B(y+1, n-y+1) \end{aligned}$$

Στον αλγόριθμο η προσομοίωση του συνόλου τιμών y από την κατανομή Bernoulli δίνει $\sum_{i=1}^{n=100} y_i = 62$. Συνεπώς, ο παράγοντας του Bayes ισούται με,

$$B_{10} = \frac{f(y|M_1)}{f(y|M_0)} = \frac{\binom{100}{62} B(63,39)}{\binom{100}{62} 0.6^{62} \cdot 0.4^{38}} = \frac{1}{7.61}$$

Παρατηρούμε λοιπόν, ότι η προσεγγιστική τιμή υπολογισμού του παράγοντα του Bayes, η οποία ισούται με $\widehat{B_{10}} = \frac{1}{7.51}$, βρίσκεται πολύ κοντά στην πραγματική τιμή του.

Συνεχίζοντας, αυξάνουμε το πλήθος των τιμών που προσομοιώνουμε και επιλέγουμε $n = 1000$ ανεξάρτητες τιμές από την κατανομή Bernoulli. Το πρώτο μοντέλο M_0 δεν επιλέγεται ούτε μία φορά και όλες οι τιμές της παραμέτρου ρ_0 προέρχονται από το δεύτερο μοντέλο M_1, **Σχήμα 2.1**. Υπολογίζοντας τον παράγοντα του Bayes χρησιμοποιώντας την προσέγγιση (2.7) για την πιθανότητα $f(M_i|y)$ βρίσκουμε να ισούται με $\widehat{B_{10}} \to +\infty$. Η τιμή αυτή προφανώς αντιστοιχεί σε μία πολύ ισχυρή ένδειξη στο να απορρίψουμε το μοντέλο M_0. Χρησιμοποιώντας τη σχέση (2.8) και την πληροφορία ότι ο αλγόριθμος δίνει $\sum_{i=1}^{n=1000} y_i = 708$ υπολογίζουμε αναλυτικά τον παράγοντα του Bayes, ο οποίος ισούται με,

$$B_{10} = \frac{f(y|M_1)}{f(y|M_0)} = \frac{\binom{1000}{708} B(709,293)}{\binom{1000}{708} 0.6^{708} \cdot 0.4^{292}} = 3,461,298,929$$

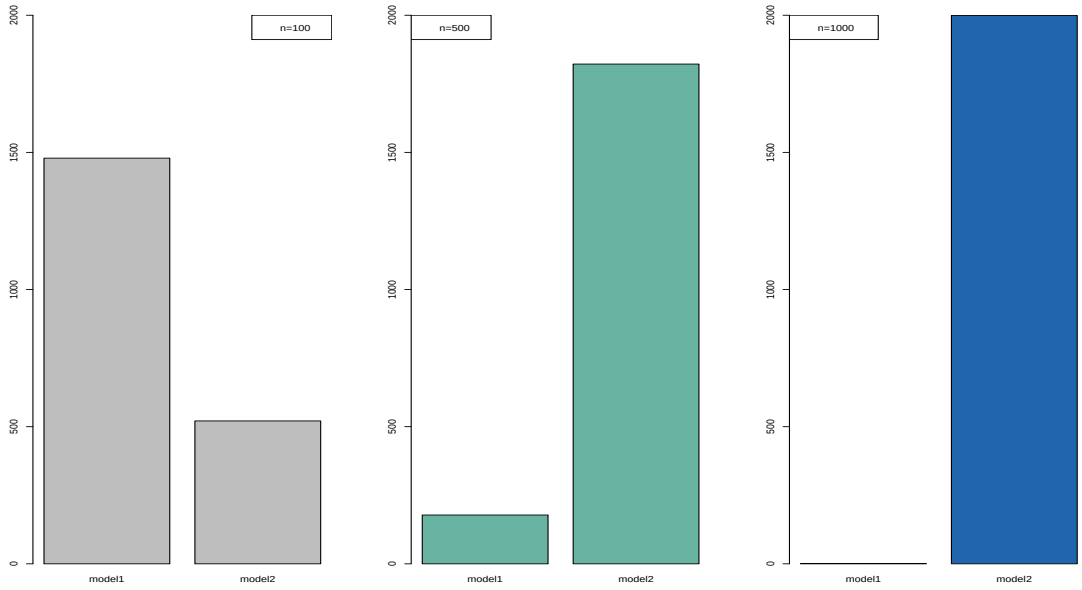
2.2.2 Κανονικό μοντέλο με γνωστή διασπορά

Στην περίπτωση όπου τα δεδομένα προέρχονται από την κανονική κατανομή θα πραγματοποιήσουμε τη σύγκριση των μοντέλων,

$$\begin{aligned} M_0 &: \ y_i \sim N(\theta_0, 1), \quad \theta_0 = 0.2 \\ M_1 &: \ y_i \sim N(\theta_0, 1), \quad i = 1, ..., n \ \ \text{όπου} \ \ \theta_0 \sim N(0, 10) \end{aligned}$$

Ξεκινώντας, θεωρούμε ότι είναι ισοπίθανο να διαλέξουμε είτε το πρώτο είτε το δεύτερο μοντέλο και δίνουμε εκ των προτέρων πιθανότητα και στα δύο μοντέλα ίση με 0.5, $\pi(M_0) = \pi(M_1) = 0.5$. Προσομοιώνουμε το σύνολο δεδομένων, y, που αποτελείται από n ανεξάρτητες και ισόνομες τιμές από την κανονική κατανομή με μέση τιμή μηδέν και τυπική απόκλιση ίση με 1, $y_i \sim N(0,1)$, $i = 1, ..., n$. Για να μειώσουμε το υπολογιστικό κόστος χρησιμοποιούμε την πρώτη μετρική του αντίστοιχου παραδείγματος στο Κεφάλαιο 1, $\rho_1 = |\overline{z} - \overline{y}|$, και τη σταθερά, (tolerance threshold), $\varepsilon_0 = 0.01$. Με το πέρας του αλγορίθμου έχουμε 2000 προσομοιωμένες τιμές για την παράμετρο θ_0 καθώς και το μοντέλο από το οποίο κρατήθηκε.

Ο αλγόριθμος υλοποιείται για τρία διαφορετικά σύνολα δεδομένων, για $n = 100$, $n = 500$ και $n = 1000$. Στην πρώτη περίπτωση, όπου $n = 100$, παρατηρούμε οτι το πλήθος των φορών που αποδεχόμαστε το πρώτο μοντέλο M_0 ισούται με 1479 και αντίστοιχα οι φορές που αποδεχόμαστε το δεύτερο μοντέλο M_1 είναι 521, **Σχήμα 2.2**. Υπολογίζοντας τον παράγοντα του Bayes χρησιμοποιώντας την προσέγγιση (2.7) για την πιθανότητα $f(M_i|y)$ βρίσκουμε να ισούται με $\widehat{B_{10}} = \frac{521}{1479} = 0.352$. Ο αλγόριθμος κρατάει περισσότερες τιμές του μοντέλου M_0 και δεν έχουμε επαρκείς ενδείξεις για να το απορρίψουμε.



Σχήμα 2.2: Η σύγκριση του μοντέλου M_0, (model1), με το μοντέλο M_1, (model2), για διαφορετικό πλήθος προσομοιωμένων δεδομένων.

Στο συγκεκριμένο παράδειγμα, το κανονικό μοντέλο μας επιτρέπει να υπολογίσουμε και αναλυτικά τον παράγοντα του Bayes από τη σχέση (2.3). Πρώτα, υπολογίζουμε την πιθανοφάνεια του μοντέλου M_1,

$$
\begin{aligned}
f(y|M_1) &= \int_{\theta_0 \in \mathbb{R}} f(y|\theta_0 \neq 0.2) f(\theta_0|M_1) d\theta_0 \qquad (2.9)\\
&= \int_{-\infty}^{+\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i-\theta_0)^2\right\} \cdot \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left\{-\frac{\theta_0^2}{2\tau_0^2}\right\} d\theta_0 \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right\} \int_{-\infty}^{+\infty} \exp\left\{\frac{n\theta_0\overline{y}}{\sigma^2} - \frac{n\theta_0^2}{2\sigma^2} - \frac{\theta_0^2}{2\tau_0^2}\right\} d\theta_0 \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right\} \exp\left\{-\frac{\mu_0^2}{2\tau_0^2}\right\} \times \\
&\times \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}(\tau_n^2)^{-1}\theta_0^2 + \mu_n(\tau_n^2)^{-1}\theta_0\right\} d\theta_0 \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \frac{\sqrt{\tau_n^2}}{\sqrt{\tau_0^2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right\} \exp\left\{-\frac{\mu_0^2}{2\tau_0^2} + \frac{\mu_n^2(\tau_n^2)^{-1}}{2}\right\}
\end{aligned}
$$

Στο παράδειγμα, ο μέσος όρος των προσομοιωμένων δεδομένων y ισούται με $\overline{y} = -0.0667$ και έτσι, ο εκ των υστέρων μέσος και η εκ των υστέρων διασπορά ισούνται με,

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\overline{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{100 \cdot (-0.0667)}{\frac{1}{10} + 100} = \frac{-66.74}{1001} = -0.067$$

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{\frac{1}{10} + 100} = \frac{10}{1001}$$

Συνεπώς, έχοντας την πληροφορία ότι $\sum_{i=1}^{n} y_i^2 = 99.41$ και χρησιμοποιώντας τη σχέση (2.9), ο παράγοντας του Bayes υπολογίζεται και ισούται με,

$$\begin{aligned} B_{10} = \frac{f(y|M_1)}{f(y|M_0)} &= \frac{\sqrt{2\pi}^{-100}\sqrt{1001}^{-1}\exp\left\{-\frac{99.41}{2}+\frac{0.067^2\cdot 1001}{20}\right\}}{\left(\sqrt{2\pi}\right)^{-100}\exp\left\{-\sum_{i=1}^{n}\frac{(y_i-0.2)^2}{2}\right\}} \\ &= \frac{\sqrt{1001}^{-1}\exp\left\{-\frac{99.41}{2}+\frac{0.067^2\cdot 1001}{20}\right\}}{\exp\left\{-\frac{99.41}{2}+20\overline{y}-2\right\}} = \sqrt{1001}^{-1}\exp\left\{0.067^2\cdot 50.05-20\overline{y}+2\right\} \\ &= 1.111 \end{aligned}$$

Στη δεύτερη περίπτωση, όπου $n = 500$, παρατηρούμε οτι το πλήθος των φορών που αποδεχόμαστε το πρώτο μοντέλο M_0 ισούται με 178 και αντίστοιχα οι φορές που αποδεχόμαστε το δεύτερο μοντέλο M_1 είναι 1822, **Σχήμα 2.2**. Υπολογίζοντας τον παράγοντα του Bayes χρησιμοποιώντας την προσέγγιση (2.7) για την πιθανότητα $f(M_i|y)$ βρίσκουμε να ισούται με $\widehat{B_{10}} = \dfrac{1822}{178} = 10.24$. Σύμφωνα με τον πίνακα 2.1, η τιμή αυτή αντιστοιχεί σε μία θετική ένδειξη στο να απορρίψουμε το μοντέλο M_0.

Επίσης, μπορούμε να υπολογίσουμε αναλυτικά τον παράγοντα του Bayes από τη σχέση (2.3). Ο μέσος όρος των προσομοιωμένων δεδομένων ισούται με $\overline{y} = 0.02$, άρα έχουμε τον εκ των υστέρων μέσο και την εκ των υστέρων διασπορά να ισούνται με,

$$\mu_n = \frac{500\cdot 0.02}{\frac{1}{10}+500} = \frac{10}{5001} = 0.002 \quad \text{και} \quad \tau_n^2 = \frac{1}{\frac{1}{10}+500} = \frac{10}{5001}$$

Συνεπώς, έχοντας την πληροφορία ότι $\sum_{i=1}^{n} y_i^2 = 539.1281$ και χρησιμοποιώντας τη σχέση (2.9), ο παράγοντας του Bayes υπολογίζεται και ισούται με,

$$\begin{aligned} B_{10} = \frac{f(y|M_1)}{f(y|M_0)} &= \frac{\sqrt{2\pi}^{-500}\sqrt{5001}^{-1}\exp\left\{\frac{0.002^2\cdot 5001}{20}\right\}}{\left(\sqrt{2\pi}\right)^{-500}\exp\left\{-\sum_{i=1}^{n}\frac{(y_i-0.2)^2}{2}\right\}} \\ &= \frac{\sqrt{5001}^{-1}\exp\left\{\frac{0.002^2\cdot 5001}{20}\right\}}{\exp\left\{100\overline{y}-10\right\}} = \sqrt{5001}^{-1}\exp\left\{0.002^2\cdot 250.05-100\cdot 0.02+10\right\} \\ &= 42.2 \end{aligned}$$

Κλείνοντας, στην τρίτη περίπτωση όπου $n = 1000$, όλες οι τιμές που έχουν γίνει δεκτές προέρχονται από το δεύτερο μοντέλο, M_1, **Σχήμα 2.2**. Ο παράγοντας του Bayes τείνει στο άπειρο, $\widehat{B_{10}} \to +\infty$ και συνεπώς, υπάρχει πολύ ισχυρή ένδειξη στο να απορρίψουμε το μοντέλο M_0. Αναλυτικά, έχουμε

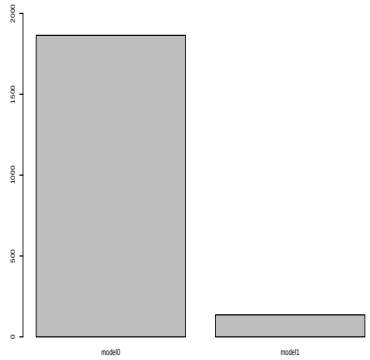
$$\begin{aligned} B_{10} &= \frac{f(y|M_1)}{f(y|M_0)} \\ &= \frac{\sqrt{10001}^{-1}\exp\left\{\frac{0.008^2\cdot 10001}{20}\right\}}{\exp\left\{200\overline{y}-20\right\}} \\ &= 24,612,771 \end{aligned}$$

Παρατηρούμε λοιπόν ότι υπάρχουν σημαντικές αποκλίσεις μεταξύ των προσεγγιστικών τιμών του παράγοντα του Bayes και των πραγματικών του τιμών. Στην πρώτη περίπτωση, όπου $n = 100$, η προσεγγιστική τιμή $\widehat{B_{10}} = 0.352$ μας οδηγεί στο συμπέρασμα να απορρίψουμε το μοντέλο M_1 ενώ η πραγματική του τιμή δεν μας επιτρέπει να καταλήξουμε σε κάποιο συμπέρασμα, $B_{10} = 1.111$.

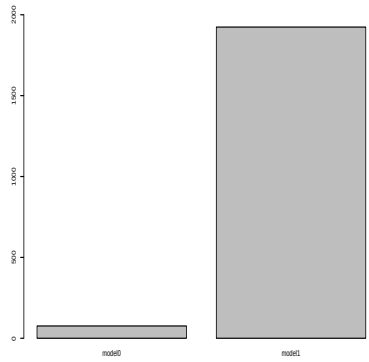
Στην δεύτερη περίπτωση, όπου $n = 500$, η προσεγγιστική τιμή $\widehat{B_{10}} = 10.24$ μας παρέχει μία θετική ένδειξη στο να απορρίψουμε το μοντέλο M_0 ενώ η πραγματική του τιμή υποδεικνύει μία ισχυρή ένδειξη ως προς την απόρριψή του, $B_{10} = 42.2$. Μόνο στην τρίτη περίπτωση όπου διαθέτουμε αρκετά μεγάλο δείγμα, $n = 1000$, ο αλγόριθμος ABC-MC δίνει ακριβή αποτελέσματα.

Συνεπώς, παρουσιάζεται η ανάγκη βελτίωσης των παραμέτρων του αλγορίθμου. Η πρώτη βελτίωση που μπορεί να υλοποιηθεί είναι η μείωση της σταθεράς ε, από τη τιμή $\varepsilon_0 = 0.01$ που χρησιμοποιήθηκε να μειωθεί στη τιμή $\varepsilon_1 = 0.005$. Όμως, συμπεραίνουμε ότι η μείωση της σταθεράς ε δεν βελτιώνει την απόδοση του αλγορίθμου σε αυτή την περίπτωση. Η αλλαγή του μέτρου απόκλισης και της στατιστικής συνάρτησης θα οδηγήσει στην καλύτερη απόδοση του αλγορίθμου. Χρησιμοποιούμε την σταθμισμένη Ευκλείδεια απόσταση διαιρούμενη με το πλήθος των δεδομένων.

Ο αλγόριθμος υλοποιείται σε δύο διαφορετικά σύνολα δεδομένων από $n = 100$ ανεξάρτητες και ισόνομες τιμές από την κανονική κατανομή. Το πρώτο σύνολο έχει μέση τιμή $\overline{y}_1 = 0.0917$ και η σύγκριση του αρχικού συνόλου με το προσομοιωμένο πραγματοποιείται αφού πρώτα γίνει διάταξη των στοιχείων τους. Η σταθερά ε ισούται με 0.01. Υπολογίζοντας την πραγματική τιμή του παράγοντα Bayes από τη σχέση (2.9) το αποτέλεσμα είναι 0.277. Επίσης, χρησιμοποιώντας τον αλγόριθμο ABC-MC, αποθηκεύονται 1864 τιμές από το μοντέλο M_0 και 136 από το μοντέλο M_1 με αποτέλεσμα η προσεγγιστική τιμή του παράγοντα Bayes να ισούται με 0.0729.



Η προσεγγιστική τιμή του παράγοντα Bayes βρίσκεται πολύ κοντά στην πραγματική του και η απόδοση του αλγορίθμου έχει βελτιωθεί αισθητά.



Στο δεύτερο σύνολο δεδομένων, ο μέσος όρος ισούται με $\overline{y}_2 = -0.11$ και σύμφωνα με τη σχέση (2.9) η πραγματική τιμή του παράγοντα Bayes είναι 141.1775. Ο αλγόριθμος ABC-MC υλοποείται με τις ίδιες παραμέτρους που χρησιμοποιήθηκαν και στο πρώτο σύνολο δεδομένων. Το αποτέλεσμα είναι να αποθηκευτούν 76 τιμές από το μοντέλο M_0 και 1924 από το μοντέλο M_1. Σύμφωνα με τη σχέση (2.7), η προσεγγιστική τιμή του παράγοντα Bayes ισούται με 25.31579. Σε αυτό το παράδειγμα, και οι δύο τιμές του παράγοντα Bayes, τόσο η ρεαλιστική όσο και η προσεγγιστική, μας οδηγούν σε μία ισχυρή ένδειξη να απορρίψουμε το μοντέλο M_0. Όμως, η προσεγγιστική τιμή δεν είναι ικανή να εκτιμήσει με ακρίβεια την ρεαλιστική.

2.2.3 Κανονικό μοντέλο με άγνωστες και τις δύο παραμέτρους

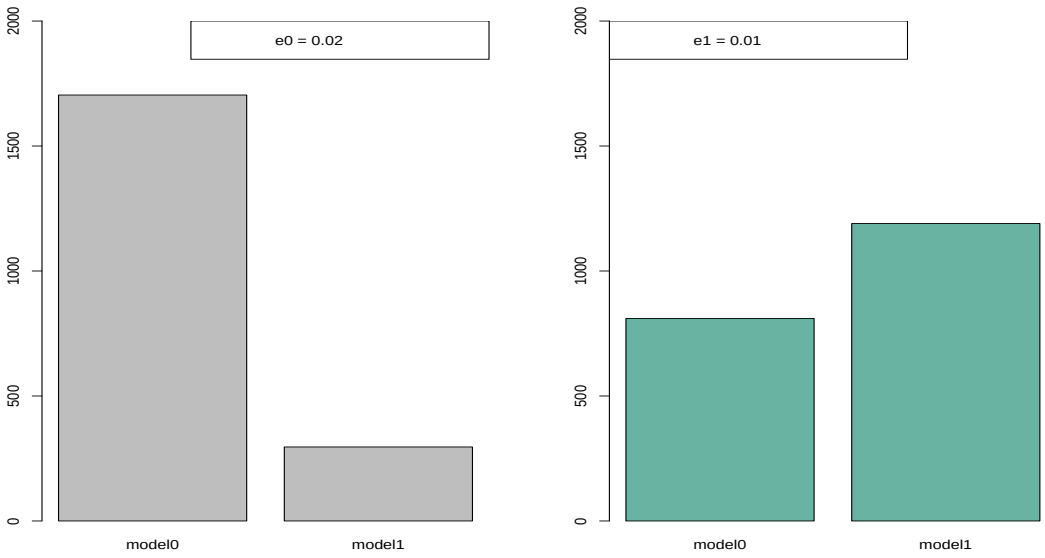
Στην περίπτωση όπου τα δεδομένα προέρχονται από την κανονική κατανομή με άγνωστες και τις δύο παραμέτρους θα πραγματοποιήσουμε τη σύγκριση των μοντέλων,

$$M_0: \ y_i \sim N(\mu, \sigma^2), \ \ \mu = 9.5 \ \ \text{και} \ \ \sigma^2 = 3, \ \ i = 1, ..., n$$
$$M_1: \ y_i \sim N(\mu, \sigma^2), \ \ \text{όπου} \ \ \mu|\sigma^2 \sim N(\mu_0, \frac{\sigma^2}{\kappa_0}) \ \ \text{και} \ \ \sigma^2 \sim \text{SI-}\chi^2\left(\nu_0, \sigma_0^2\right)$$

όπου δίνονται οι ακόλουθες τιμές στις υπερπαραμέτρους, $\mu_0 = 5$, $\kappa_0 = 1$, $\nu_0 = 6$ και $\sigma_0^2 = 1$. Ξεκινώντας, θεωρούμε ότι είναι ισοπίθανο να διαλέξουμε είτε το πρώτο είτε το δεύτερο μοντέλο και δίνουμε εκ των προτέρων πιθανότητα και στα δύο μοντέλα ίση με 0.5, $\pi(M_0) = \pi(M_1) = 0.5$. Προσομοιώνουμε το σύνολο δεδομένων, y, που αποτελείται από $n = 100$ ανεξάρτητες και ισόνομες τιμές από την κανονική κατανομή με μέση τιμή δέκα και τυπική απόκλιση ίση με 2, $y_i \sim N(10, 4)$, $i = 1, ..., n$. Ξρησιμοποιούμε την πρώτη μετρική του αντίστοιχου παραδείγματος στο Κεφάλαιο 1,

$$\rho_1 = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(\frac{y_i - z_i}{s}\right)^2}$$

Ο αλγόριθμος εκτελείται δύο φορές για δύο διαφορετικές τιμές της σταθεράς ε. Στην πρώτη δοκιμή δίνεται η τιμή $\varepsilon_0 = 0.02$ και στην επόμενη η τιμή της σταθεράς μειώνεται στο μισό, $\varepsilon_1 = 0.01$. Με το πέρας του αλγορίθμου έχουμε 2000 προσομοιωμένες τιμές για το διάνυσμα των παραμέτρων $\overrightarrow{\theta} = (\mu, \sigma^2)$ καθώς και το μοντέλο από το οποίο κρατήθηκε.



Σχήμα 2.3: Η σύγκριση του μοντέλου M_0 με το μοντέλο M_1 για διαφορετικές τιμές της σταθεράς ε.

Στην πρώτη περίπτωση, όπου $\varepsilon_0 = 0.02$, παρατηρούμε οτι το πλήθος των φορών που αποδεχόμαστε το πρώτο μοντέλο M_0 ισούται με 1704 και αντίστοιχα οι φορές που αποδεχόμαστε το δεύτερο μοντέλο M_1 είναι 296, **Σχήμα 2.3**. Υπολογίζοντας τον παράγοντα του Bayes χρησιμοποιώντας την προσέγγιση (2.7) για την πιθανότητα $f(M_i|y)$ βρίσκουμε να ισούται με $\widehat{B_{10}} = \frac{296}{1704} = 0.174$. Ο αλγόριθμος κρατάει

περισσότερες τιμές του μοντέλου M_0 και δεν έχουμε επαρκείς ενδείξεις για να το απορρίψουμε.

Στη δεύτερη περίπτωση, όπου $\varepsilon_0 = 0.01$, παρατηρούμε οτι το πλήθος των φορών που αποδεχόμαστε το πρώτο μοντέλο M_0 ισούται με 810 και αντίστοιχα οι φορές που αποδεχόμαστε το δεύτερο μοντέλο M_1 είναι 1190, **Σχήμα 2.3**. Υπολογίζοντας τον παράγοντα του Bayes χρησιμοποιώντας την προσέγγιση (2.7) για την πιθανότητα $f(M_i|y)$ βρίσκουμε να ισούται με $\widehat{B_{10}} = \dfrac{1190}{810} = 1.47$. Σε αυτή την περίπτωση, η εικόνα φαίνεται να διαφοροποιείται συγκριτικά με την πρώτη αλλά σύμφωνα με τον Πίνακα 2.1, δεν έχουμε επαρκείς ενδείξεις για να απορρίψουμε το μοντέλο M_0.

Για τον υπολογισμό της πραγματικής τιμής του παράγοντα Bayes από τη σχέση (2.3), απαιτείται πρώτα ο υπολογισμός της πιθανοφάνειας του κανονικού μοντέλου με παραμέτρους $\overrightarrow{\theta} = (\mu, \sigma^2) \in \Omega = (\mathbb{R}, \mathbb{R}^+)$.

$$
\begin{aligned}
f(y|M_1) &= \int_{\theta\in\Omega} f(y|\theta \neq (9.5,3)) f(\theta|M_1) d\theta \qquad (2.10)\\
&= \int_{\theta\in\Omega} f(y|\theta \neq (9.5,3)) f(\mu|\sigma^2) f(\sigma^2) d\theta \\
&= \int_{\theta\in\Omega} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i-\mu)^2\right\} \times \\
&\times \frac{1}{\sqrt{2\pi\sigma^2/\kappa_0}} \exp\left\{-\frac{\kappa_0(\mu-\mu_0)^2+\nu_0\sigma_0^2}{2\sigma^2}\right\} \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\frac{\nu_0}{2})} \sigma_0^{\nu_0} (\sigma^2)^{-(\frac{\nu_0}{2}+1)} d\theta \\
&= \frac{\sqrt{\kappa_0}\,(\nu_0/2)^{\nu_0/2}}{\left(\sqrt{2\pi}\right)^n \Gamma(\frac{\nu_0}{2})} \sigma_0^{\nu_0} \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_0}{2}+1)}}{\sqrt{2\pi\sigma^2}\sqrt{\sigma^2}^n} \exp\left\{-\frac{\kappa_0(\mu-\mu_0)^2+\nu_0\sigma_0^2+\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right\} d\theta \\
&= \Lambda \cdot \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_n}{2}+1)}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\kappa_0(\mu-\mu_0)^2+\nu_0\sigma_0^2+\sum_{i=1}^n (y_i-\overline{y}+\overline{y}-\mu)^2}{2\sigma^2}\right\} d\theta \\
&= \Lambda \cdot \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_n}{2}+1)}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\kappa_0(\mu-\mu_0)^2+\nu_0\sigma_0^2+\sum_{i=1}^n (y_i-\overline{y})^2+n(\overline{y}-\mu)^2}{2\sigma^2}\right\} d\theta \\
&= \Lambda \cdot \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_n}{2}+1)}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\kappa_0(\mu-\mu_0)^2+\nu_0\sigma_0^2+(n-1)s^2+n(\overline{y}-\mu)^2}{2\sigma^2}\right\} d\theta \\
&= \Lambda \cdot \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_n}{2}+1)}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\nu_0\sigma_0^2+(n-1)s^2+n\overline{y}^2+\kappa_n\mu^2-2\mu\kappa_n\mu_n+\kappa_0\mu_0^2}{2\sigma^2}\right\} d\theta \\
&= \Lambda \cdot \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_n}{2}+1)}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\nu_0\sigma_0^2+(n-1)s^2+n\overline{y}^2+\kappa_n(\mu-\mu_n)^2-\kappa_n\mu_n^2+\kappa_0\mu_0^2}{2\sigma^2}\right\} d\theta \\
&= \Lambda \cdot \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_n}{2}+1)}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\nu_0\sigma_0^2+(n-1)s^2+\frac{\kappa_0 n}{\kappa_n}(\overline{y}-\mu_0)^2+\kappa_n(\mu-\mu_n)^2}{2\sigma^2}\right\} d\theta \\
&= \Lambda \cdot \int_{\theta\in\Omega} \frac{(\sigma^2)^{-(\frac{\nu_n}{2}+1)}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\nu_n\sigma_n^2+\kappa_n(\mu-\mu_n)^2}{2\sigma^2}\right\} d\theta \\
&= \frac{\sqrt{\kappa_0}\left(\frac{\nu_0}{2}\cdot\sigma_0^2\right)^{\nu_0/2}}{\sqrt{\kappa_n}\left(\sqrt{2\pi}\right)^n} \cdot \frac{\Gamma\left(\frac{\nu_n}{2}\right)\sigma_n^{-\nu_n}}{\Gamma(\frac{\nu_0}{2})\left(\frac{\nu_n}{2}\right)^{\nu_n/2}}
\end{aligned}
$$

όπου $\Lambda = \frac{\sqrt{\kappa_0}\left(\frac{\nu_0}{2}\right)^{\nu_0/2}}{\left(\sqrt{2\pi}\right)^n \Gamma(\frac{\nu_0}{2})}\sigma_0^{\nu_0}$ και οι παράμετροι μ_n, κ_n, ν_n και σ_n^2 δίνονται από τις σχέσεις (1.6) στο Κεφάλαιο 1.

Στο παράδειγμα ο μέσος όρος των προσομοιωμένων δεδομένων ισούται με $\overline{y} = 10.20$ και η δειγματική διασπορά ισούται με $s^2 = 4.13$, άρα μπορούμε να υπολογίσουμε τις εκ των υστέρων παραμέτρους:

$$\kappa_n = \kappa_0 + n = 101$$
$$\nu_n = \nu_0 + n = 106$$
$$\sigma_n^2 = \frac{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\overline{y}-\mu_0)^2}{\nu_0+n} = \frac{6 + 99\cdot 4.13 + 0.99\cdot(10.20-5)^2}{106} = 4.17$$

Χρησιμοποιώντας τη σχέση (2.10) υπολογίζεται ο παράγοντας του Bayes,

$$\begin{aligned} B_{10} &= \frac{f(y|M_1)}{f(y|M_0)} \\ &= \frac{3^3\Gamma(53)\sqrt{4.17}^{-106}}{\sqrt{2\pi}^{-100}\Gamma(3)\sqrt{101}\cdot 53^{53}} \cdot \frac{1}{\left(\sqrt{2\pi\cdot 3}\right)^{-100}\exp\left\{-\sum_{i=1}^{100}\frac{(y_i-9.5)^2}{6}\right\}} \\ &= \frac{\sqrt{3}^{106}\Gamma(53)\sqrt{4.17}^{-106}\cdot 53^{-53}}{\Gamma(3)\sqrt{101}\exp\left\{-\frac{464.38}{6}\right\}} \\ &= 17.78 \end{aligned}$$

Η πραγματική τιμή του παράγοντα Bayes ισούται με 17.78, ενώ αντίθετα στις περιπτώσεις όπου $\varepsilon = 0.02$ και $\varepsilon = 0.01$ οι προσεγγιστικές τιμές είναι 0.174 και 1.47 αντίστοιχα, γεγονός που καθιστά τον αλγόριθμο ABC-MC ακατάλληλο για επιλογή μοντέλου.

2.3 Επιλογή στατιστικών περιγραφικών μέτρων

Οι Robert et al. (2011) υπογραμμίζουν τους κινδύνους που ελλοχεύουν όταν γίνεται χρήση του αλγορίθμου ABC για επιλογή μοντέλου. **Συγκεκριμένα, τονίζουν ότι ακόμα και όταν μία στατιστική συνάρτηση είναι επαρκής για κάθε μοντέλο ενδιαφέροντος, η ίδια στατιστική συνάρτηση δεν είναι επαρκής για επιλογή μοντέλου**. Ο ισχυρισμός αυτός αποδεικνύεται άμεσα.

Θεωρούμε ότι η στατιστική συνάρτηση $S(y)$ είναι επαρκής για την παράμετρο δύο μοντέλων με πιθανοφάνειες, $f\left(y|m_i,\theta_i\right)$, για $i = 1, 2$. Τότε, από το παραγοντικό κριτήριο του Neyman η πιθανοφάνεια μπορεί να γραφτεί στη μορφή $g\left(S(y)|m_i,\theta_i\right)h(y|m_i)$ για $i = 1, 2$. Υποθέτωντας ότι τα μοντέλα 1 και 2 είναι εκ των προτέρων ισοπίθανα, η πραγματική τιμή του παράγοντα του Bayes, B_{12}, είναι ο λόγος των περιθωριακών πιθανοφανειών

$$\begin{aligned} B_{12} &= \frac{\int_{\theta_1} f\left(y|m_1,\theta_1\right)\pi\left(\theta_1|m_1\right)d\theta_1}{\int_{\theta_2} f\left(y|m_2,\theta_2\right)\pi\left(\theta_2|m_2\right)d\theta_2} \\ &= \frac{h\left(y|m_1\right)\int_{\theta_1} g\left(S(y)|m_1,\theta_1\right)\pi\left(\theta_1|m_1\right)d\theta_1}{h\left(y|m_2\right)\int_{\theta_2} g\left(S(y)|m_2,\theta_2\right)\pi\left(\theta_2|m_2\right)d\theta_2} \\ &= \frac{h\left(y|m_1\right)}{h\left(y|m_2\right)}B_{12}^{S(y)} \end{aligned}$$

όπου $B_{12}^{S(y)}$ θα ήταν ο παράγοντας του Bayes ο οποίος λαμβάνεται όταν η επιλογή μοντέλου βασίζεται στο $S(y)$ και όχι στα δεδομένα y. Ο λόγος $\dfrac{h\left(y|m_1\right)}{h\left(y|m_2\right)}$ ισούται με 1 μόνο σε πολύ ιδιαίτερες περιπτώσεις, όπως για παράδειγμα σε εμφωλευμένα μοντέλα που ανήκουν στην εκθετική οικογένεια κατανομών ή σε μοντέλα που γίνεται χρήση όλης της πληροφορίας των δεδομένων. Συνεπώς, σε ένα γενικότερο πλαίσιο, η προσέγγιση του παράγοντα του Bayes από τον αλγόριθμο ABC δεν είναι ικανοποιητική.

Σε πολλές περιπτώσεις στατιστικά περιγραφικά μέτρα και στατιστικές συναρτήσεις που είναι επαρκείς για επιλογή παραμέτρων δεν είναι απαραίτητα επαρκείς για την επιλογή μοντέλου. Όμως, η συνθήκη της επάρκειας παρέχει έγκυρα αποτελέσματα όταν ο αλγόριθμος πρόκειται να εφαρμοστεί σε εμφωλευμένα μοντέλα που ανήκουν στην εκθετική οικογένεια κατανομών (Didelot et al., 2010). Μερικά από τα πιο χαρακτηριστικά μοντέλα τα οποία ανήκουν σε μία τετοια οικογένεια και παράλληλα η πιθανοφάνειά τους δεν υπολογίζεται αναλυτικά είναι το μοντέλο των τυχαίων πεδίων του Gibbs, (Grelaud et al., 2009), τα εκθετικά μοντέλα τυχαίων γραφημάτων και τα μοντέλα autologistic, (Drovandi and Pettitt, 2011a). Αποδεικνύεται ότι το διάνυσμα των επαρκών στατιστικών συναρτήσεων του πλήρους μοντέλου είναι επαρκές για την επιλογή μοντέλου και για την επιλογή των παραμέτρων του κάθε μοντέλου. Αντίστοιχο παράδειγμα παρουσιάζεται στο Κεφάλαιο 3.

Εναλλακτικά, δύο μη εμφωλευμένα μοντέλα τα οποία ανήκουν στην εκθετική οικογένεια κατανομών μπορούν να ενσωματωθούν σε ένα άλλο μοντέλο εκθετικής οικογένειας το οποίο αντιλαμβάνεται τα αρχικά ως ειδικές περιπτώσεις. Το αποτέλεσμα αυτής της προσέγγισης είναι η επαρκής στατιστική συνάρτηση για το συνδυαστικό μοντέλο να καθίσταται επαρκής για τη σύγκριση των εμφωλευμένων σε αυτό μοντέλων (Didelot et al., 2010). Στη διαδικασία εκτίμησης των παραμέτρων κάθε μοντέλου θα υπάρξει ένας πλεονασμός στατιστικών περιγραφικών μέτρων, γεγονός το οποίο θα πρέπει να αντιμετωπίσουμε προκειμένου να εξασφαλίσουμε από κοινού επάρκεια.

2.4 Συμπεράσματα

Στα δύο τελευταία παραδείγματα, τόσο στο στο Κεφάλαιο 2.2.2 στο κανονικό μοντέλο με μία άγνωστη παράμετρο όσο και στο Κεφάλαιο 2.2.3 στο κανονικό με δύο άγνωστες παραμέτρους, ο αλγόριθμος ABC-MC δεν ήταν ικανός να εκτιμήσει σωστά την πραγματική τιμή του παράγοντα Bayes. Η μείωση της σταθεράς ε δεν στάθηκε ικανοποιητική λύση στη βελτίωση της απόδοσης του αλγορίθμου αλλά ούτε και η αλλαγή του μέτρου απόκλισης διαφοροποίησε σημαντικά τα αποτελέσματα.

Είναι ιδιαίτερα χρήσιμο να παρατηρήσουμε ότι η περιθώρια πιθανοφάνεια του μοντέλου M_1 στο Κεφάλαιο 2.2.2, (2.9), είναι ισοδύναμη με την πυκνότητα της n-διάστατης κανονικής κατανομής $N_n\left(M,\Sigma\right)$ με δύο παραμέτρους. Το διάνυσμα - στήλη $M=\left(\mu_0,...,\mu_0\right)$ διάστασης $n\times 1$, όπου $\mu_0=0$ ο εκ των προτέρων μέσος και ο πίνακας Σ διάστασης $n\times n$ με στοιχεία :

- $\Sigma\left[i,i\right]=\sigma_0^2+\sigma^2, \quad i=1,...,n$
- $\Sigma\left[i,j\right]=\sigma_0^2, \quad \text{για } i\neq j$

όπου σ_0^2 η εκ των προτέρων διασπορά της άγνωστης μέσης τιμής θ_0 με $\sigma_0^2=10$ και σ^2 η γνωστή πληθυσμιακή διασπορά με $\sigma^2=1$.

Επίσης, η περιθώρια πιθανοφάνεια του μοντέλου M_1 στο Κεφάλαιο 2.2.3, (2.10), είναι ισοδύναμη με την πυκνότητα της n-διάστατης κατανομής $T_n\left(M,\Sigma,df\right)$ με τρεις παραμέτρους. Το διάνυσμα - στήλη

$M = (\mu_0, ..., \mu_0)$ διάστασης $n \times 1$, όπου $\mu_0 = 5$ ο εκ των προτέρων μέσος της άγνωστης μέσης τιμής μ. Ο πίνακας Σ διάστασης $n \times n$, όπου $\Sigma = \sigma_0^2 A$ με τον πίνακα A να περιέχει τα στοιχεία :

- $A[i,i] = 1 + \dfrac{1}{\kappa_0}, \quad i = 1, ..., n$
- $A[i,j] = \dfrac{1}{\kappa_0}, \quad \text{για } i \neq j$

όπου σ_0^2 η εκ των προτέρων παράμετρος κλίμακας της άγνωστης διασποράς σ^2, $\sigma_0^2 = 1$ και $\kappa_0 = 1$. Η παράμετρος df ισούται με την παράμετρο ν_0, $\text{df} = \nu_0 = 6$.

Το μοντέλο που χρησιμοποιήθηκε σε κάθε ένα από τα προαναφερθέντα παραδείγματα είναι της μορφής $y_i \sim N(\mu, \sigma^2)$, δηλαδή είναι ειδική περίπτωση του κανονικού γραμμικού μοντέλου στο οποίο υπάρχει μόνο ο σταθερός όρος στο μ. Το γεγονός αυτό μας δίνει την αφορμή να επιστήσουμε την προσοχή στα κανονικά γραμμικά μοντέλα και να προχωρήσουμε στην δημιουργία ενός αλγορίθμου ABC ικανού για την επιλογή μοντέλων αυτής της μορφής. Ο αλγόριθμος αυτός είναι ένας κατάλληλα διαμορφωμένος αγόριθμος ABC-MCMC, η ανάλυση του οποίου ακολουθεί στο Κεφάλαιο 3.

Κεφάλαιο 3

Επιλογή μεταβλητών υπό την Μπεϋζιανή προσέγγιση

3.1 Κανονικό γραμμικό μοντέλο

Πολλές μελέτες επικεντρώνονται στην εύρεση οποιασδήποτε μορφής συσχέτισης μεταξύ δύο ή περισσότερων μεταβλητών. Το ενδιαφέρον προσανατολίζεται στο κατα πόσο μία ποσότητα, έστω y, εξαρτάται και διαμορφώνεται από μία άλλη ποσότητα, έστω x, η οποία συνήθως έχει τη μορφή διανύσματος. Έτσι, μελετάμε την υπό συνθήκη κατανομή του y δοθέντος του x, η οποία συμβολίζεται ως $f(y|\theta, x)$, σε ένα μοντέλο που περιέχει n παρατηρήσεις της μορφής $(x, y)_i$.

Υποθέτουμε ότι Y είναι η μεταβλητή ενδιαφέροντος, η οποία καλείται μεταβλητή απόκρισης, και οι μεταβλητές $x_1, x_2, ..., x_p$ αποτελούν το σύνολο των επεξηγηματικών μεταβλητών. Κάθε επεξηγηματική μεταβλητή είναι ένα n-διάστατο διάνυσμα, οι συντεταγμένες του οποίου αντιστοιχούν στις n παρατηρήσεις. Η κατανομή του Y δοθέντων των μεταβλητών $x_1, x_2, ..., x_p$ μελετάται έχοντας ένα σύνολο n παρατηρήσεων, $i = 1, 2..., n$ όπου το y_i εξαρτάται από τις μεταβλητές $x_{i1}, x_{i2}, ..., x_{ip}$.

Στο κανονικό γραμμικό μοντέλο η κατανομή του Y δοθέντων των επεξηγηματικών μεταβλητών x_1, $x_2, ..., x_p$ είναι η κανονική κατανομή με τη μέση τιμή να είναι μία γραμμική συνάρτηση των επεξηγηματικών μεταβλητών. Σε δείγμα n ανεξάρτητων παρατηρήσεων το μοντέλο είναι της μορφής,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \text{όπου} \quad \varepsilon_i \sim N_n(0, \sigma^2 I_n) \tag{3.1}$$

Στην παραπάνω σχέση, η μεταβλητή απόκρισης θεωρείται συνεχής ενώ οι επεξηγηματικές μεταβλητές είναι είτε διακριτές είτε συνεχείς. Οι όροι ε_i είναι τα τυχαία σφάλματα, τα όποια είναι ανεξάρτητα και έχουν ίσες διασπορές, και οι όροι $\beta_j, j = 0, 1, 2, ..., p$ είναι οι παράμετροι του μοντέλου. Ισοδύναμα, η σχέση (3.1), όπως παρουσιάστηκε από τους McCullagh, Nelder (1989) γράφεται,

$$y_i|x_{i1}, x_{i2}, ..., x_{ip} \sim N(\mu_i, \sigma^2) \quad \text{όπου} \quad \mu_i = \mathbb{E}(y_i|\beta, x_{ij}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.2}$$

Θεωρούμε ότι οι παρατηρήσεις y_i είναι υπό συνθήκη ανεξάρτητες δοθέντων των θ, x_{ij} όπου $\theta = (\beta_0, ..., \beta_p, \sigma^2)$ είναι το διάνυσμα με τις παραμέτρους προς εκτίμηση. Επίσης, οι υπό συνθήκη διασπορές είναι ίσες, $Var(y_i|\theta, X) = \sigma^2$ για κάθε $i = 1, 2, ..., n$. Ισοδύναμα οι σχέσεις (3.1) και (3.2)

γράφονται σε μορφή διανυσμάτων,

$$Y \sim N_n(X\beta, \sigma^2 I_n) \tag{3.3}$$

όπου $Y = (y_1, ..., y_n)^t$, $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$, I_n ο ταυτοτικός πίνακας διάστασης $n \times n$ και X ο πίνακας των επεξηγηματικών μεταβλητών διάστασης $n \times (p+1)$, ο οποίος καλείται πίνακας σχεδιασμού,

$$X = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}$$

Σε πρώτη φάση, ο σκοπός μας είναι η εκτίμηση των παραμέτρων υπο την Μπεϋζιανή σκοπιά, δηλαδή του διανύσματος $\theta = (\beta_0, ..., \beta_p, \sigma^2)$. Ξεκινώντας, ο υπολογισμός της πιθανοφάνειας θα μας οδηγήσει στην κατάλληλη επιλογή των εκ των προτέρων κατανομών για το διάνυσμα $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$ και τη διασπορά σ^2. Έχοντας $Y \sim N_n(X\beta, \sigma^2 I_n)$, η συνάρτηση πιθανοφάνειας για το διάνυσμα $Y = (y_1, ..., y_n)^t$ ισούται με

$$\begin{aligned} f(y|\beta, \sigma^2, X) &= (2\pi)^{-n/2} \det\left(\sigma^2 I_n\right)^{-1/2} \exp\left\{-\frac{1}{2}(y - X\beta)^t (\sigma^2 I_n)^{-1}(y - X\beta)\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^t (y - X\beta)\right\} \end{aligned} \tag{3.4}$$

Παρατηρούμε ότι είναι της μορφής του κανονικού μοντέλου με άγνωστη μέση τιμή και διασπορά, το οποίο αναλύθηκε στο Κεφάλαιο 1.2.4. Η πιο χρήσιμη επιλογή εκ των προτέρων κατανομής για το διάνυσμα $\theta = (\beta_0, ..., \beta_p, \sigma^2)$ είναι η πολυδιάστατη κανονική αντίστροφη-χ^2. Ειδικότερα, το διάνυσμα $\beta = (\beta_0, \beta_1, ..., \beta_p)$ ακολουθεί την πολυδιάστατη κανονική κατανομή και η διασπορά σ^2 ακολουθεί την κλιμακωτή αντίστροφη-χ^2. Έχουμε λοιπόν,

$$\beta|\sigma^2 \sim N_q(\tilde{\beta}, \sigma^2\Sigma) \qquad \text{και} \qquad \sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \tag{3.5}$$

όπου $q = p + 1$, ο εκ των προτέρων μέσος $\tilde{\beta}$ είναι ένα διάνυσμα στήλη διάστασης $q \times 1$, ο πίνακας Σ είναι διάστασης $q \times q$ και οι τιμές ν_0, σ_0^2 θεωρούνται σταθερές, θετικές υπερ-παράμετροι του μοντέλου. Οι εκ των προτέρων κατανομές 3.5 καθορίζονται από τις υπερ-παραμέτρους $\tilde{\beta}, \Sigma, \nu_0, \sigma_0^2$, οι τιμές των οποίων πρέπει να επιλεγούν προσεχτικά όπως θα συζητηθεί στο Κεφάλαιο 3.3.

Η από κοινού εκ των προτέρων κατανομή του διανύσματος β και του σ^2 είναι:

$$\begin{aligned} f(\beta, \sigma^2) &= f(\beta|\sigma^2) f(\sigma^2) \\ &= (2\pi)^{-q/2} (\det\left(\sigma^2\Sigma\right))^{-1/2} \exp\left\{-\frac{1}{2}(\beta - \tilde{\beta})^t \left(\sigma^2\Sigma\right)^{-1} (\beta - \tilde{\beta})\right\} \\ &\quad \times \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\frac{\nu_0}{2})} \sigma_0^{\nu_0} (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \\ &\propto (\sigma^2)^{-\frac{q}{2}-\frac{\nu_0}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \tilde{\beta})^t \Sigma^{-1} (\beta - \tilde{\beta}) - \frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \end{aligned} \tag{3.6}$$

Ο υπολογισμός της εκ των υστέρων κατανομής των παραμέτρων, η οποία προκύπτει από το γινόμενο της πιθανοφάνειας (3.4) και της εκ των προτέρων κατανομής (3.6), απαιτεί αρκετές πράξεις με τη

βοήθεια προτάσεων από τη Γραμμική Άλγεβρα πάνω στις τετραγωνικές μορφές στο $\mathbb{R}^{\nu}$.

$$
\begin{aligned}
f(\beta,\sigma^2|y,X) &\propto f(y|\beta,\sigma^2,X)f(\beta,\sigma^2) \\
&\propto (\sigma^2)^{-(\frac{n+q+\nu_0}{2}+1)}\exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \\
&\quad\times\exp\left\{-\frac{1}{2\sigma^2}(\beta-\tilde{\beta})^t\Sigma^{-1}(\beta-\tilde{\beta})-\frac{1}{2\sigma^2}(y-X\beta)^t(y-X\beta)\right\} \\
&\propto (\sigma^2)^{-(\frac{n+q+\nu_0}{2}+1)}\exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} \\
&\quad\times\exp\left\{-\frac{1}{2\sigma^2}\left[(\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)-h^t\tilde{\Sigma}h+y^ty+\tilde{\beta}^t\Sigma^{-1}\tilde{\beta}\right]\right\} \\
&\propto (\sigma^2)^{-(\frac{n+\nu_0}{2}+1)}\exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+y^ty+\tilde{\beta}^t\Sigma^{-1}\tilde{\beta}-h^t\tilde{\Sigma}h\right]\right\} \\
&\quad\times(\sigma^2)^{-q/2}\exp\left\{-\frac{1}{2\sigma^2}(\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)\right\} \\
&= \ \mathrm{N}_q\text{-Inv-}\chi^2\left(\tilde{m},\sigma^2\tilde{\Sigma};\tilde{\nu_n},\tilde{\sigma_n}^2\right)
\end{aligned}
\tag{3.7}
$$

όπου $\tilde{\Sigma}=(X^TX+\Sigma^{-1})^{-1}$ και $h=X^Ty+\Sigma^{-1}\tilde{\beta}$. Αποδεικνύεται ότι η παραπάνω κατανομή είναι μία πολυδιάστατη κανονική αντίστροφη-χ^2 με τέσσερις παραμέτρους:

$$
\begin{aligned}
\tilde{m} &= \tilde{\Sigma}h=\tilde{\Sigma}(X^Ty+\Sigma^{-1}\tilde{\beta}) \\
\tilde{\Sigma} &= (X^TX+\Sigma^{-1})^{-1} \\
\tilde{\nu_n} &= n+\nu_0 \\
\tilde{\sigma_n}^2 &= \frac{\nu_0\sigma_0^2+y^ty+\tilde{\beta}^t\Sigma^{-1}\tilde{\beta}-h^t\tilde{\Sigma}h}{n+\nu_0}
\end{aligned}
$$

Η περιθώρια εκ των υστέρων κατανομή του σ^2 προκύπτει ολοκληρώνοντας τη σχέση (3.7) ως προς β :

$$
\begin{aligned}
f(\sigma^2|y,X) &= \int_{\mathbb{R}^{q\times 1}} f(\beta,\sigma^2|y,X)d\beta \\
&\propto (\sigma^2)^{-(\frac{n+\nu_0}{2}+1)}\exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+y^ty+\tilde{\beta}^t\Sigma^{-1}\tilde{\beta}-h^t\tilde{\Sigma}h\right]\right\} \\
&\quad\times\int_{\mathbb{R}^{q\times 1}}(\sigma^2)^{-q/2}\exp\left\{-\frac{1}{2\sigma^2}(\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)\right\}d\beta \\
&\propto (\sigma^2)^{-(\frac{n+\nu_0}{2}+1)}\exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+y^ty+\tilde{\beta}^t\Sigma^{-1}\tilde{\beta}-h^t\tilde{\Sigma}h\right]\right\}\times(2\pi)^{q/2}(\det\tilde{\Sigma})^{1/2} \\
&\propto (\sigma^2)^{-(\frac{n+\nu_0}{2}+1)}\exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+y^ty+\tilde{\beta}^t\Sigma^{-1}\tilde{\beta}-h^t\tilde{\Sigma}h\right]\right\}
\end{aligned}
$$

Άρα, η περιθώρια εκ των υστέρων κατανομή του σ^2 είναι μία κλιμακωτή αντίστροφη-χ^2:

$$
\sigma^2|y,X\sim\text{SI-}\chi^2(\tilde{\nu_n},\tilde{\sigma_n}^2) \tag{3.8}
$$

Η περιθώρια εκ των υστέρων κατανομή του β προκύπτει από τη σχέση (3.7) ολοκληρώνοντας ως προς σ^2. Για να διευκολύνουμε τη διαδικασία, θέτουμε

$$
\tilde{\kappa_n}^2=\frac{\tilde{\nu_n}\tilde{\sigma_n}^2+(\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)}{\tilde{\nu_n}+q}
$$

και θα εμφανίσουμε στους υπολογισμούς το ολοκλήρωμα της συνάρτησης πυκνότητας της κλιμακωτής αντίστροφης-χ^2 κατανομής.

$$
\begin{aligned}
f(\beta|y,X) &= \int_0^{+\infty} f(\beta,\sigma^2|y,X)d\sigma^2 \\
&\propto \int_0^{+\infty} (\sigma^2)^{-(\frac{q+\tilde{\nu_n}}{2}+1)} \exp\left\{-\frac{\tilde{\nu_n}{\tilde{\sigma_n}}^2}{2\sigma^2} - \frac{1}{2\sigma^2}(\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)\right\} d\sigma^2 \\
&\propto \int_0^{+\infty} (\sigma^2)^{-(\frac{q+\tilde{\nu_n}}{2}+1)} \exp\left\{-\frac{\tilde{\nu_n}{\tilde{\sigma_n}}^2 + (\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)}{2\sigma^2}\right\} d\sigma^2 \\
&\propto \int_0^{+\infty} (\sigma^2)^{-(\frac{q+\tilde{\nu_n}}{2}+1)} \exp\left\{-\frac{(\tilde{\nu_n}+q){\tilde{\kappa_n}}^2}{2\sigma^2}\right\} d\sigma^2 \\
&\propto \frac{\Gamma\left(\frac{q+\tilde{\nu_n}}{2}\right)}{\left(\frac{q+\tilde{\nu_n}}{2}\right)^{\frac{q+\tilde{\nu_n}}{2}} {\tilde{\kappa_n}}^{(q+\tilde{\nu_n})}} \int_0^{+\infty} \frac{\left(\frac{q+\tilde{\nu_n}}{2}\right)^{\frac{q+\tilde{\nu_n}}{2}} {\tilde{\kappa_n}}^{(q+\tilde{\nu_n})}}{\Gamma\left(\frac{q+\tilde{\nu_n}}{2}\right)} (\sigma^2)^{-(\frac{q+\tilde{\nu_n}}{2}+1)} \exp\left\{-\frac{(\tilde{\nu_n}+q){\tilde{\kappa_n}}^2}{2\sigma^2}\right\} d\sigma^2 \\
&\propto \frac{\Gamma\left(\frac{q+\tilde{\nu_n}}{2}\right)}{\left(\frac{q+\tilde{\nu_n}}{2}\right)^{\frac{q+\tilde{\nu_n}}{2}} {\tilde{\kappa_n}}^{(q+\tilde{\nu_n})}} = \Gamma\left(\frac{q+\tilde{\nu_n}}{2}\right)\left(\frac{q+\tilde{\nu_n}}{2}{\tilde{\kappa_n}}^2\right)^{-\frac{q+\tilde{\nu_n}}{2}} \\
&\propto \left[\frac{\tilde{\nu_n}{\tilde{\sigma_n}}^2 + (\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)}{2}\right]^{-\frac{q+\tilde{\nu_n}}{2}} \\
&\propto \left[\frac{\tilde{\nu_n}{\tilde{\sigma_n}}^2}{2}\left(1+\frac{(\beta-\tilde{\Sigma}h)^t\tilde{\Sigma}^{-1}(\beta-\tilde{\Sigma}h)}{\tilde{\nu_n}{\tilde{\sigma_n}}^2}\right)\right]^{-\frac{q+\tilde{\nu_n}}{2}} \\
&\propto \left(1+\frac{1}{\tilde{\nu_n}}(\beta-\tilde{\Sigma}h)^t({\tilde{\sigma_n}}^2\tilde{\Sigma})^{-1}(\beta-\tilde{\Sigma}h)\right)^{-\frac{q+\tilde{\nu_n}}{2}}
\end{aligned}
$$

Παρατηρούμε λοιπόν, ότι η περιθώρια εκ των υστέρων κατανομή του β είναι η πολυδιάστατη κατανομή t, διάστασης q, με $\tilde{\nu_n}$ βαθμούς ελευθερίας. Συνεπώς, η περιθώρια εκ των υστέρων κατανομή των συντελεστών του γραμμικού μοντέλου είναι

$$\beta|y,X \sim t_q\left(\tilde{m}, {\tilde{\sigma_n}}^2\tilde{\Sigma}, \tilde{\nu_n}\right) \tag{3.9}$$

Από την τελευταία σχέση έπεται ότι οι $q = p+1$ περιθωριακές κατανομές $\beta_j|y,X$ είναι μονοδιάστατες κατανομές t με $\tilde{\nu_n}$ βαθμούς ελευθερίας,

$$\beta_j|y,X \sim t_1\left(\tilde{m}_j, {\tilde{\sigma_n}}^2\tilde{\Sigma_{jj}}, \tilde{\nu_n}\right), \quad j = 0,1,...,p$$

3.2 Επιλογή μεταβλητών για το κανονικό γραμμικό μοντέλο

Το πρόβλημα της επιλογής μεταβλητών προκύπτει όταν υπάρξει η ανάγκη μέτρησης / μοντελοποίησης της σχέσης ανάμεσα στο Y και σε ένα υποσύνολο των $X_1, X_2, ..., X_p$ αλλά υπάρχει αβεβαιότητα για το ποιο θα είναι το υποσύνολο που θα χρησιμοποιηθεί. Το πρόβλημα αυτό έχει ιδιαίτερο ενδιαφέρον ειδικά όταν το πλήθος p των επεξηγηματικών μεταβλητών είναι μεγάλο και το σύνολο $X_1, X_2, ..., X_p$

περιέχει πολλές περιττές μεταβλητές που δεν επηρεάζουν τη μεταβλητή Y. Το πρόβλημα της επιλογής μεταβλητών θεωρείται επίσης μία ειδική περίπτωση του προβλήματος επιλογής μοντέλου, όπου κάθε μοντέλο υπό εξετάση αντιστοιχεί σε ένα ξεχωριστό υποσύνολο του συνόλου των επεξηγηματικών μεταβλητών $X_1, X_2, ..., X_p$.

Το πρόβλημα αυτό είναι πιο γνωστό στο πλαίσιο της πολλαπλής παλινδρόμησης όπου το ενδιαφέρον περιορίζεται σε κανονικά γραμμικά μοντέλα. Πολλές από τις θεμελιώδεις εξελίξεις στην επιλογή μεταβλητών προκύπτουν μέσα στο πλαίσιο του γραμμικού μοντέλου, γιατί η αναλυτική του ικανότητα διευκολύνει σε μεγάλο βαθμό τη διεξαγωγή συμπερασμάτων και την υπολογιστική μείωση.

Υπό την Μπεϋζιανή προσέγγιση, θεωρούμε ότι το κανονικό γραμμικό μοντέλο χρησιμοποιείται για να συσχετίσει την μεταβλητή Y με τις επεξηγηματικές μεταβλητές $X_1, X_2, ..., X_p$ μέσα από τη σχέση (3.3). Το πρόβλημα της επιλογής μεταβλητών προκύπτει όταν υπάρχει κάποιο άγνωστο υποσύνολο επεξηγηματικών μεταβλητών, οι συντελεστές των οποίων είναι αρκετά μικροί, που θα ήταν προτιμότερο να τις αγνοήσουμε. Θα ήταν αρκετά βολικό να προσδιορίσουμε κάθε ένα από αυτά τα 2^q πιθανά μοντέλα, $q = p + 1$, χρησιμοποιώντας το διάνυσμα

$$\gamma = (\gamma_0, \gamma_1, ..., \gamma_p)^{'}$$

όπου $\gamma_j = 0$ ή $\gamma_j = 1$ ανάλογα με το αν το β_j είναι μικρό ή μεγάλο αντίστοιχα. Χρησιμοποιώντας αυτό το συμβολισμό, μπορεί να ελεγχθεί η ύπαρξη ή όχι της υποψήφιας επεξηγηματικής μεταβλητής στο τελικό μοντέλο, όντας παρούσα όταν $\gamma_j = 1$ ή απούσα όταν $\gamma_j = 0$. Η μεταβλητή $q_\gamma \equiv \gamma^{'} 1$ υποδηλώνει το πλήθος των επεξηγηματικών μεταβλητών στο εκάστοτε υποσύνολο.

Οι George, McCulloch (1993) εισήγαγαν την ένδειξη γ_j χωρίς να ενσωματωθεί στη γραμμική σχέση (3.2). Σε αυτή την περίπτωση, το διάνυσμα γ εμπεριέχεται στο μοντέλο μέσω της ακόλουθης ιεραρχικής δομής,

$$\begin{gathered} y|\beta, \sigma^2, X \sim f\left(y|\beta, \sigma^2, X\right) \\ \beta, \sigma^2|\gamma \sim f\left(\beta, \sigma^2|\gamma\right) \\ \gamma \sim \pi\left(\gamma\right) \end{gathered}$$

όπου η $f\left(y|\beta, \sigma^2, X\right)$ είναι η συνάρτηση πιθανοφάνειας, η $f\left(\beta, \sigma^2|\gamma\right)$ είναι η εκ των προτέρων κατανομή των παραμέτρων και η $\pi\left(\gamma\right)$ είναι η εκ των προτέρων κατανομή του διανύσματος γ.

Για τον προσδιορισμό της εκ των προτέρων κατανομής $\pi\left(\gamma\right)$, σε πολλές εφαρμογές Μπεϋζιανής επιλογής μεταβλητών χρησιμοποιείται μία μορφή ανεξάρτητης εκ των προτέρων κατανομής

$$\pi(\gamma) = \prod_{i=0}^{p} w_i^{\gamma_i} \left(1 - w_i\right)^{1-\gamma_i} \tag{3.10}$$

η οποία είναι εύκολο να προσδιοριστεί, να μειώσει ουσιαστικά υπολογιστικές απαιτήσεις και συχνά να οδηγήσει σε λογικά αποτελέσματα (Clyde, Desimone, Parmigiani (1996), Raftery, Madigan, Hoeting (1997)). Υπό τη σκοπιά αυτής της εκ των προτέρων κατανομής, κάθε X_i εισάγεται στο μοντέλο ανεξάρτητα από τις άλλες μεταβλητές με πιθανότητα $\pi(\gamma_i = 1) = 1 - \pi(\gamma_i = 0) = w_i$. Μικρότερες τιμές του w_i χρησιμοποιούνται για να δώσουν μικρότερη βαρύτητα σε μεταβλητές X_i που έχουν αμελητέα επιρροή στο μοντέλο.

Μία χρήσιμη βελτίωση της σχέσης (3.10) προκύπτει θέτοντας $w_i = w$, καταλήγοντας στη σχέση

$$\pi(\gamma) = w^{q_\gamma} (1-w)^{q-q_\gamma} \tag{3.11}$$

όπου εδώ η υπερπαράμετρος w δηλώνει το εκ των προτέρων αναμενόμενο πλήθος των επεξηγηματικών μεταβλητών στο μοντέλο. Πιο συγκεκριμένα, θέτοντας $w = 1/2$, οδηγούμαστε στη γνωστή ομοιόμορφη εκ των προτέρων κατανομή

$$\pi(\gamma) \equiv 1/2^q$$

η οποία χρησιμοποιείται για να εκφράσουμε άγνοια. Όμως, η επιλογή αυτής της εκ των προτέρων κατανομής δίνει μεγαλύτερη βαρύτητα σε μοντέλα μεγέθους $q_\gamma = q/2$ επειδή είναι περισσότερα στο πλήθος.
Για να δώσουμε αυξημένη βαρύτητα σε parsimonious μοντέλα, δηλαδή μοντέλα που περιέχουν τον ελάχιστο δυνατό αριθμό παραμέτρων και ταυτόχρονα έχουν καλή προβλεπτική ικανότητα, μπορούμε, για παράδειγμα, να χρησιμοποιήσουμε μικρή τιμή για το w. Εναλλακτικά, μπορούμε να βάλουμε μία εκ των προτέρων κατανομή στο w. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε την κατανομή Beta $w \sim \text{Beta}(\alpha, \beta)$ και η σχέση (3.11) γίνεται

$$\pi(\gamma) = \frac{B(\alpha + q_\gamma, \beta + q - q_\gamma)}{B(\alpha, \beta)}$$

3.3 Προσδιορισμός των παραμέτρων των εκ των προτέρων κατανομών του μοντέλου

Σε αυτό το κεφάλαιο εξετάζουμε τις επιλογές που έχουμε για τις παραμέτρους των εκ των προτέρων κατανομών σε ένα πρόβλημα επιλογής μεταβλητών για το κανονικό γραμμικό μοντέλο. Σε προβλήματα επιλογής μεταβλητών, στόχος μας είναι να αγνοήσουμε μόνο αυτά τα X_i για τα οποία $\beta_i = 0$ στη σχέση (3.3). Συνεπώς, το ζητούμενο είναι να επιλέξουμε ένα υπο-μοντέλο της μορφής

$$Y|\beta_\gamma, \sigma^2, \gamma \sim N_n(X_\gamma \beta_\gamma, \sigma^2 I_n)$$

όπου ο X_γ είναι ο πίνακας διάστασης $n \times q_\gamma$ του οποίου οι στήλες αντιστοιχούν στο υποσύνολο του $X_1, X_2, ..., X_p$ σε αντιστοιχία με το διάνυσμα γ, το β_γ είναι το διάνυσμα $q_\gamma \times 1$ των παραμέτρων / συντελεστών του μοντέλου και το σ^2 η άγνωστη διασπορά των τυχαίων σφαλμάτων.

Όπως είδαμε και στο Κεφάλαιο 3.1, η πιο χρήσιμη επιλογή εκ των προτέρων κατανομής για το μοντέλο αυτό είναι η πολυδιάστατη κανονική αντίστροφη-χ^2, η οποία αποτελείται από την πολυδιάστατη κανονική κατανομή διάστασης q_γ για το διάνυσμα β_γ,

$$\beta_\gamma|\sigma^2, \gamma \sim N_{q_\gamma}\left(\tilde{\beta}_\gamma, \sigma^2 \Sigma_\gamma\right) \tag{3.12}$$

σε συνδυασμό με την κλιμακωτή αντίστροφη-χ^2 για την παράμετρο σ^2,

$$\sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \tag{3.13}$$

η οποία είναι ισοδύναμη με την $\text{IG}\left(\alpha = \frac{\nu_0}{2}, \beta = \frac{\nu_0}{2}\sigma_0^2\right)$. Αυτή η επιλογή για την εκ των προτέρων κατανομή μας οδηγεί σε εκ των υστέρων κατανομή που ανήκει στην ίδια οικογένεια (όπως αποδείχθηκε

στο Κεφάλαιο 3.1) και γι΄ αυτό καθίσταται ως η πιο χρήσιμη. Επίσης, ολοκληρώνοντας την πολυδιάστατη κανονική αντίστροφη-χ^2 ως προς σ^2 παίρνουμε την κατανομή του β_γ δεδομένου του διανύσματος γ σε αναλυτική μορφή,

$$\beta_\gamma|\gamma \sim t_{q_\gamma}\left(\nu_0, \tilde{\beta_\gamma}, \sigma_0^2\Sigma_\gamma\right) \tag{3.14}$$

η οποία είναι η πολυδιάστατη κατανομή Student-t διάστασης q_γ κεντραρισμένη στο $\tilde{\beta_\gamma}$ με ν_0 βαθμούς ελευθερίας και παράμετρο κλίμακας $\sigma_0^2\Sigma_\gamma$.

Οι εκ των προτέρων κατανομές (3.12) και (3.13) καθορίζονται από τις υπερπαραμέτρους $\tilde{\beta_\gamma}, \Sigma_\gamma, \nu_0, \sigma_0^2$, οι οποίες πρέπει να προσδιοριστούν για να προχωρήσουμε σε εφαρμογές. Από τη μία πλευρά, ο προσδιορισμός των παραμέτρων μπορεί να γίνει από υποκειμενική σκοπιά και να οδηγήσει σε επαρκή αποτελέσματα ειδικά σε μικρότερα προβλήματα και με την κατάλληλη επιστημονική πληροφόρηση (Garthwaite and Dickey, 1996). Από την άλλη, ιδιαίτερο ενδιαφέρον έχει η περίπτωση όπου μία τέτοια πληροφόρηση δεν είναι διαθέσιμη και ο στόχος μας είναι να εκχωρήσουμε τιμές στις παραμέτρους οι οποίες ελαχιστοποιούν την επιρροή της εκ των προτέρων κατανομής.

Ξεκινώντας με την επιλογή των παραμέτρων σ_0^2 και ν_0, το σ_0^2 μπορεί να θεωρηθεί ως μία εκ των προτέρων εκτίμηση του σ^2 και το ν_0 ως το εκ των προτέρων μέγεθος δείγματος που σχετίζεται με αυτή την εκτίμηση (Chipman, George, McCulloch, 2001). Χρησιμοποιώντας τα δεδομένα και θεωρώντας το s_Y^2, τη δειγματική διασπορά του Y, ως το ανώτερο όριο εκτίμησης για το σ^2, μία συνηθισμένη επιλογή είναι να διαλέξουμε μια μικρή τιμή για το ν_0 και το σ_0^2 κοντά στο s_Y^2. Εναλλακτικά, η διαδικασία επιλογής τιμών για αυτές τις παραμέτρους μπορεί να αποφευχθεί, χρησιμοποιώντας την εκ των προτέρων κατανομή $f\left(\sigma^2\right) \propto 1/\sigma^2$, το όριο της (3.13) καθώς $\nu_0 \to 0$. Παρά το γεγονός ότι είναι μία καταχρηστική κατανομή, μπορεί να χρησιμοποιηθεί γιατί οδηγεί σε κατάλληλη περιθώρια κατανομή $f\left(Y|\gamma\right)$ όταν συνδυάζεται με την (3.12).

Στη συνέχεια, το ενδιαφέρον στρέφεται στην εκ των προτέρων κατανομή των συντελεστών του μοντέλου, όπου η πιο συνηθισμένη επιλογή για την εκ των προτέρων μέση τιμή $\tilde{\beta_\gamma}$ είναι $\tilde{\beta_\gamma} = 0$, μία επιλογή που αντανακλά ουδετερότητα μεταξύ θετικών και αρνητικών τιμών. Για τον πίνακα διακύμανσης - συνδιακύμανσης Σ_γ συνηθίζεται η επιλογή $\Sigma_\gamma = cV_\gamma$, όπου c είναι μία σταθερά και V_γ είναι πίνακας προκαθορισμένης μορφής όπως $V_\gamma = \left(X_\gamma^t X_\gamma\right)^{-1}$ ή $V_\gamma = I_{q_\gamma}$, ο ταυτοτικός πίνακας διάστασης $q_\gamma \times q_\gamma$. Η επιλογή $V_\gamma = \left(X_\gamma^t X_\gamma\right)^{-1}$ ενισχύει τη δομή διακύμανσης της πιθανοφάνειας και αντιστοιχεί στη g-εκ των προτέρων κατανομή του Zellner (1986). Αντίθετα, η επιλογή $V_\gamma = I_{q_\gamma}$ καθιστά τις συνιστώσες του β_γ υπο συνθήκη ανεξάρτητες, προκαλώντας εξασθένηση στη δομή διακύμανσης της πιθανοφάνειας.

Έχοντας καθορίσει τον πίνακα V_γ, η επόμενη ενέργεια έιναι να δώσουμε μία τιμή στη σταθερά c τόσο μεγάλη έτσι ώστε η εκ των προτέρων κατανομή να είναι σχετικά επίπεδη στην περιοχή όλων των πιθανών τιμών του β_γ με αποτέλεσμα να μειώνεται η εκ των προτέρων επιρροή (Edwards, Lindman and Savage, 1963). Παράλληλα όμως, είναι σημαντικό να αποφύγουμε αρκετά μεγάλες τιμές για τη σταθερά c. Χρησιμοποιώντας μεγάλες τιμές για το c, η εκ των προτέρων θα δώσει μεγαλύτερη βαρύτητα στο μοντέλο που περιέχει μόνο το σταθερό όρο (null model) καθώς $c \to +\infty$, οδηγούμαστε δηλαδή στο παράδοξο των Bartlett-Lindley (Bartlett, 1957). Για πρακτικούς λόγους, μία λύση είναι να επιλέξουμε το c με τέτοιο τρόπο έτσι ώστε η (3.14) να εκχωρεί ουσιαστική πιθανότητα σε όλο το εύρος των πιθανών τιμών του β_γ. Οι Raftery, Madigan and Hoeting, (1997), πραγματοποίησαν διαδικασία προσομοίωσης για να αξιολογήσουν την επίδραση διάφορων τιμών του c, με πίνακα $V_\gamma = \left(X_\gamma^t X_\gamma\right)^{-1}$, $f\left(\sigma^2\right) \propto 1/\sigma^2$ και $\pi\left(\gamma\right) = 2^{-p}$, πάνω στην εκ των υστέρων κατανομή του πραγματικού μοντέλου. Κατέληξαν στο να προτείνουν την επιλογή $c = \max\left\{p^2, n\right\}$.

3.4 Η g εκ των προτέρων κατανομή του Zellner

Η g εκ των προτέρων κατανομή του Zellner (1986) ορίζεται ως Κανονική Αντίστροφη χ^2 κατανομή ή ισοδύναμα Κανονική Αντίστροφη Γάμμα κατανομή, με μέση τιμή κεντραρισμένη στο μηδέν, δηλαδή $\tilde{\beta}_\gamma = 0$ και πίνακα ίσο με $V_\gamma = \left(X_\gamma^t X_\gamma\right)^{-1}$ όπου X_γ είναι ο πίνακας σχεδιασμού του μοντέλου M_γ. Ο πίνακας V_γ καθορίζει τον εκ των προτέρων συσχετισμό των συντελεστών β_γ, μέσω του πίνακα X_γ που περιέχει τις επιδράσεις του διανύσματος των συντελεστών που είναι διαφορετικές από το μηδέν. Συγκεκριμένα, για ένα μοντέλο M_γ με p_γ επεξηγηματικές μεταβλητές, ο Zellner θέτοντας $g = c$, εισήγαγε στην περίπτωση της κανονικής παλινδρόμησης την g εκ των προτέρων κατανομή για τους συντελεστές υπό τη μορφή

$$\beta_\gamma|\sigma^2, \gamma \sim N_{q_\gamma}\left(0, c \cdot \sigma^2 (X_\gamma^t X_\gamma)^{-1}\right)$$

με εκ των προτέρων κατανομή $f(\sigma^2) \propto \dfrac{1}{\sigma^2}$ κοινή για όλα τα μοντέλα $M_\gamma \in M$ (prior του Jeffreys).

Το σημαντικότερο πλεονέκτημα της g εκ των προτέρων κατανομής του Zellner είναι ο αναλυτικός υπολογισμός του παράγοντα Bayes καθώς και ο υπολογισμός της περιθώριας πιθανοφάνειας του μοντέλου, χρησιμοποιώντας τις προαναφερθείσες εκ των προτέρων κατανομές για τις παραμέτρους του μοντέλου. Η περιθώρια πιθανοφάνεια του μοντέλου δίνεται σε κλειστή μορφή και είναι η ακόλουθη:

$$f(Y|M_\gamma) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}^{n-1}\sqrt{n}} \|Y - \overline{Y}\|^{-(n-1)} \frac{(1+c)^{\frac{n-q_\gamma}{2}}}{\left[1 + c\left(1 - R_\gamma^2\right)\right]^{\frac{n-1}{2}}} \tag{3.15}$$

όπου $\|\cdot\|$ η Ευκλείδεια νόρμα και R_γ^2 είναι ο συντελεστής προσδιορισμού του μοντέλου παλινδρόμησης M_γ (Fully Bayes factor with a generalized g-prior, Yuzo Maruyama et al.).

Δύο κλειστές μορφές του παράγοντα του Bayes είναι ιδιαίτερα χρήσιμες: η κλειστή μορφή του παράγοντα Bayes που σχετίζεται με τη σύγκριση του μοντέλου M_γ ($\beta_\gamma \in \mathbb{R}^{q_\gamma}$ με α) το μηδενικό μοντέλο, δηλαδή το μοντέλο που περιέχει μόνο το σταθερό όρο (null model) και β) με το πλήρες μοντέλο, δηλαδή το μοντέλο που περιέχει όλες τις επεξηγηματικές μεταβλητές (full model). Με τη βοήθεια αυτών μπορούμε να καταλήξουμε στον υπολογισμό του παράγοντα Bayes που σχετίζεται με τη σύγκριση δύο οποιωνδήποτε μοντέλων M_γ και $M_{\gamma'}$.

Α. Ο παράγοντας Bayes για σύγκριση με το μηδενικό μοντέλο

Από τον τύπο της περιθώριας πιθανοφάνειας στη σχέση (3.15) μπορούμε να παρατηρήσουμε την ανεξαρτησία του μηδενικού μοντέλου M_N, null model, $(H_0 : \gamma = (1, 0, 0, ..., 0))$ από την παράμετρο c καθώς έχουμε ότι $R_\gamma^2 = 0$ και $q_\gamma = p_\gamma + 1 = 1$. Βάσει αυτών των ισοτήτων και του κλειστού τύπου της περιθώριας πιθανοφάνειας, (3.15), ο παράγοντας Bayes για τη σύγκριση του μοντέλου M_γ με το μηδενικό μοντέλο M_N δίνεται ως:

$$BF_{M_\gamma M_N} = (1+c)^{\frac{n-q_\gamma}{2}} \left[1 + c(1 - R_\gamma^2)\right]^{-\frac{n-1}{2}} \tag{3.16}$$

Β. Ο παράγοντας Bayes για σύγκριση με το πλήρες μοντέλο

Για τη σύγκριση του μοντέλου M_γ, για το οποίο X_γ ο πίνακας σχεδιασμού του, με το πλήρες μοντέλο M_F απαιτείται κατάλληλη τροποποίηση. Ο πίνακας σχεδιασμού του πλήρους μοντέλου γράφεται ως

$X = [\mathbb{I}, X_\gamma, X_{-\gamma}]$, έτσι ώστε το πλήρες μοντέλο M_F να μπορεί να γραφεί στη μορφή:

$$M_F : \mu_F = \mathbb{I}\beta_0 + X_\gamma\beta_\gamma + X_{-\gamma}\beta_{-\gamma}$$

όπου ο $X_{-\gamma}$ αναφέρεται στις στήλες του πίνακα σχεδιασμού X που δεν περιλαμβάνονται στο μοντέλο M_γ. Το μοντέλο M_γ αντιστοιχεί στην υπόθεση $H_0 : \beta - \gamma = 0$ ενώ το πλήρες μοντέλο M_F στην εναλλακτική $H_1 : \beta_{-\gamma} \in \mathbb{R}^{q-q_\gamma}$. Προκειμένου να συγκρίνουμε τα δύο αυτά μοντέλα με τη βοήθεια του παράγοντα Bayes, για τα οποία το β_γ είναι κοινό, χρησιμοποιούμε τια ακόλουθες εκ των προτέρων κατανομές:

$$M_\gamma : f\left(\beta_\gamma, \sigma^2\right) \propto \frac{1}{\sigma^2}$$

$$M_F : f\left(\beta_\gamma, \sigma^2\right) \propto \frac{1}{\sigma^2} \quad \text{και} \quad \beta_{-\gamma}|\sigma^2 \sim N\left(0, c\sigma^2(X_{-\gamma}^t X_{-\gamma})^{-1}\right)$$

με αποτέλεσμα ο παράγοντας Bayes να δίνεται στην ακόλουθη κλειστή μορφή

$$BF_{M_\gamma M_F} = (1+c)^{-(\frac{n-q}{2})}\left[1 + c \cdot \frac{1 - R_F^2}{1 - R_\gamma^2}\right]^{\frac{n-q_\gamma}{2}} \tag{3.17}$$

όπου R_F^2 και R_γ^2 οι συντελεστές προσδιορισμού για το πλήρες μοντέλο M_F και το μοντέλο M_γ αντίστοιχα.

Αξίζει να σημειωθεί ότι η σύγκριση ενός μοντέλου με το μηδενικό ή το πλήρες μοντέλο μπορεί να χρησιμοποιηθεί προκειμένου να συγκρίνουμε δύο οποιαδήποτε μοντέλα μεταξύ τους χρησιμοποιώντας το μηδενικό ή το πλήρες μοντέλο αντίστοιχα ως μοντέλο αναφοράς. Στην περίπτωση αυτή ο παράγοντας Bayes των δύο συγκρινόμενων μοντέλων θα δίνεται ως το πηλίκο των παραγόντων Bayes του κάθε μοντέλου ξεχωριστά σε σχέση με το μοντέλο αναφοράς.

Έτσι, για παράδειγμα χρησιμοποιώντας ως μοντέλο αναφοράς το μηδενικό μοντέλο, ο παράγοντας Bayes για τα συγκρινόμενα μοντέλα M_γ και $M_{\gamma'}$ θα δίνεται ως:

$$BF_{M_\gamma M_{\gamma'}} = \frac{BF_{M_\gamma M_N}}{BF_{M_{\gamma'} M_N}}$$

Κατά ανάλογο τρόπο λαμβάνουμε τον παράγοντα Bayes εάν χρησιμοποιήσουμε ως μοντέλο αναφοράς το πλήρες μοντέλο ή και οποιοδήποτε άλλο.

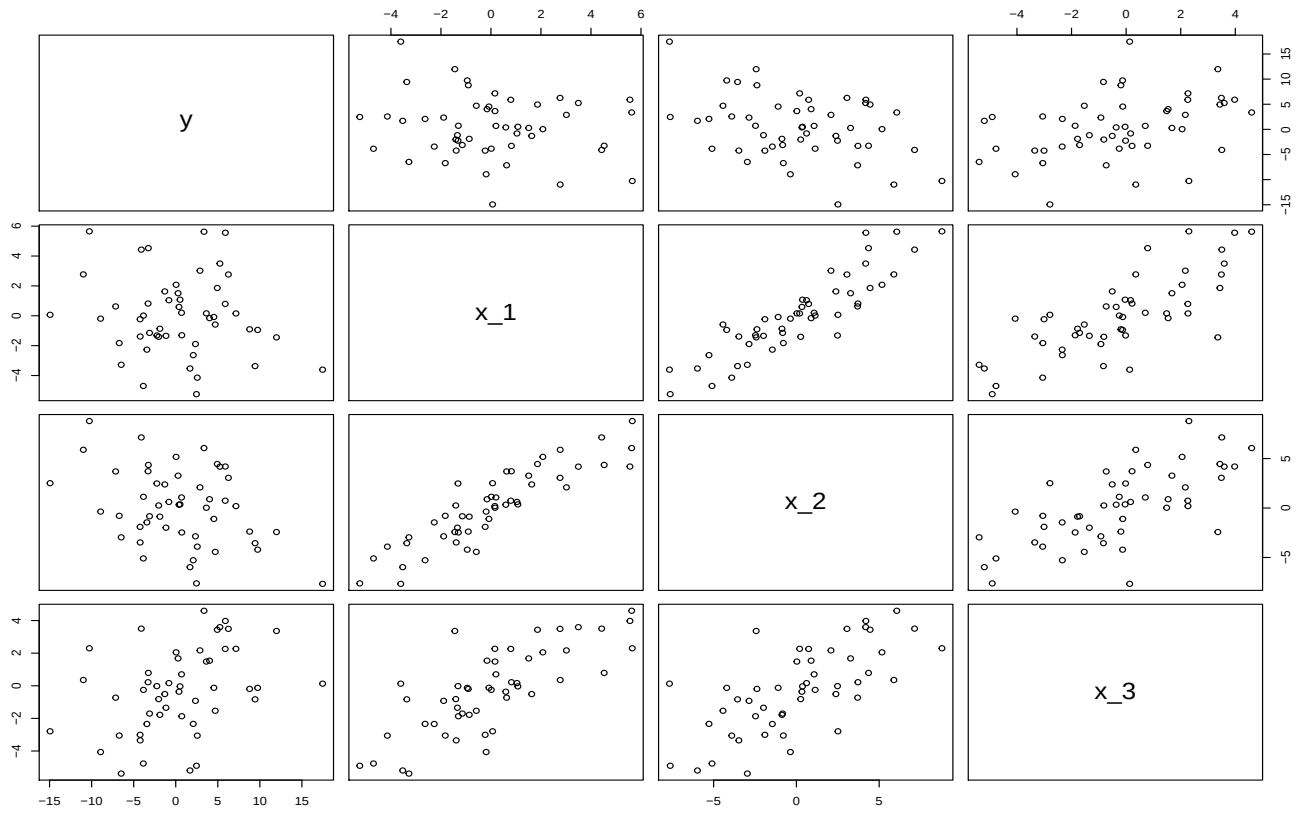
3.5 Ο αλγόριθμος ABC rejection sampler στο πολλαπλό γραμμικό μοντέλο

Στο παράδειγμα που ακολουθεί θα εξετάσουμε την απόδοση του αλγορίθμου ABC likelihood-free rejection sampler στην εκτίμηση των παραμέτρων ένος πολλαπλού γραμμικού μοντέλου.

Θεωρούμε δείγμα μεγέθους $n = 50$ παρατηρήσεων / προσομοιωμένων τιμών με τη βοήθεια του οποίου θα κρίνουμε την ικανότητα του αλγορίθμου στην εκτίμηση των παραμέτρων ενός πολλαπλού γραμμικού μοντέλου. Υποθέτουμε ότι έχουμε $p = 3$ επεξηγηματικές μεταβλητές, X_1, X_2, X_3, και έχουμε το πολλαπλό γραμμικό μοντέλο της μορφής,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Θέτουμε $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = -2$, $\beta_3 = 3$ και $\sigma^2 = 3$ και χρησιμοποιώντας το πακέτο bayesianRegression, (Παπασταμούλης, 2019, Ο.Π.Α), και την εντολή simulate_data προσομοιώνουμε 50 τιμές για την μεταβλητή απόκρισης $y_i, \quad i = 1, ..., 50$, με τις αντίστοιχες επεξηγηματικές μεταβλητές x_{i1}, x_{i2}, x_{i3}. Στο **Σχήμα 3.1** μπορούμε να δούμε τη γραμμική συσχέτιση ανάμεσα στις μεταβλητές.



Σχήμα 3.1: Η γραμμική συσχέτιση ανάμεσα στις μεταβλητές

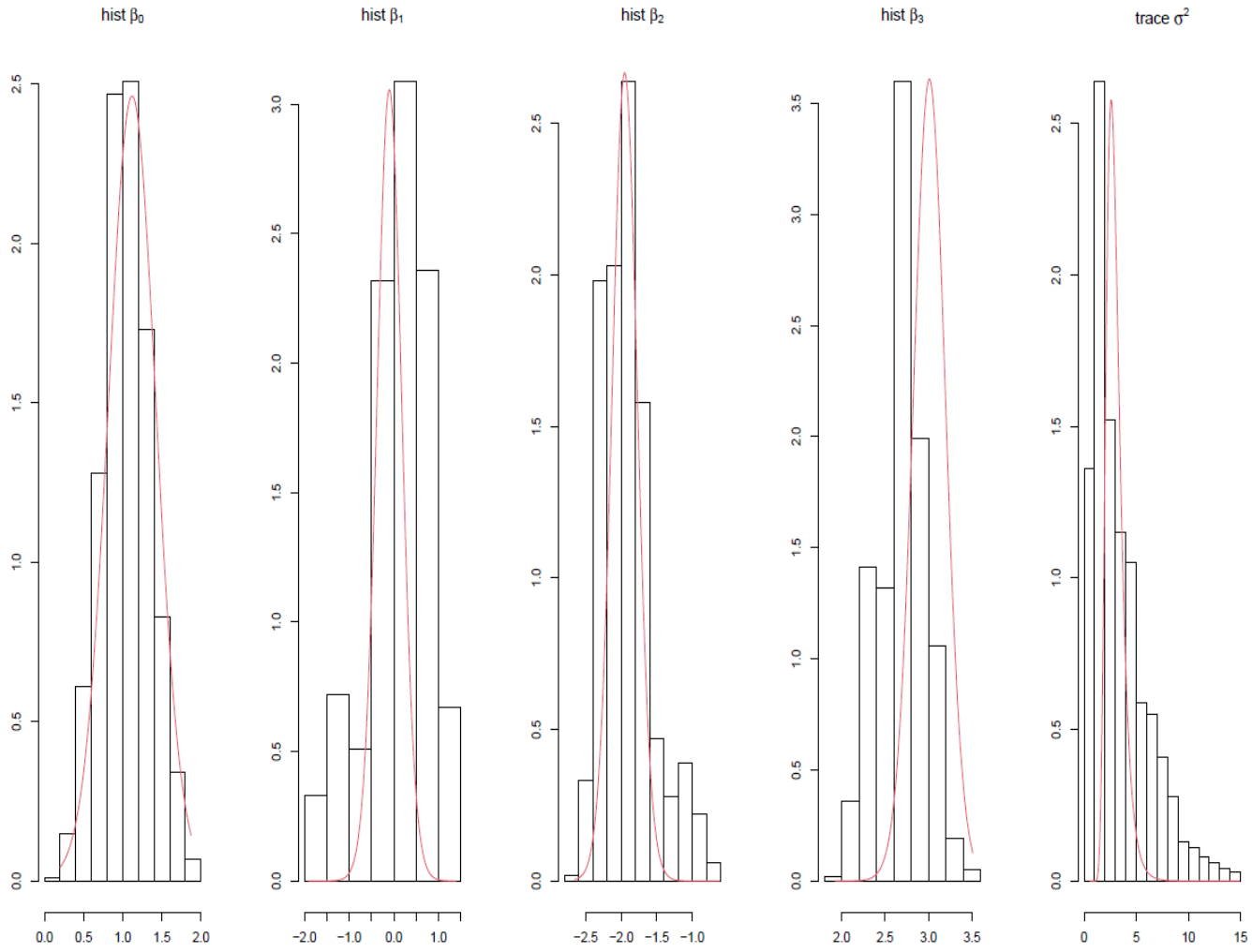
Το **Σχήμα 3.1** μας δίνει μία πρώτη εικόνα για τα δεδομένα του παραδείγματος. Μπορούμε να πούμε ότι η μεταβλητή Y δεν φαίνεται να συνδέεται γραμμικά με τη μεταβλητή X_1, αντίθετα φαίνεται να υπάρχει μία αρνητική και μία θετική γραμμική συσχέτιση με τη μεταβλητή X_2 και X_3 αντίστοιχα.

Για να κατασκευάσουμε τον αλγόριθμο πρέπει πρώτα να προσδιορίσουμε τις τιμές των υπερπαραμέτρων των εκ των προτέρων κατανομών. Όπως είδαμε και στο **Κεφάλαιο 3.1** επιλέγουμε

$$\beta|\sigma^2 \sim N_4(\tilde{\beta}, \sigma^2\Sigma) \qquad \text{και} \qquad \sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2)$$

Θεωρούμε ότι $\tilde{\beta} = (0,0,0,0)^t$ και στον πίνακα Σ δίνεται η g-εκ των προτέρων κατανομή του Zellner, $\Sigma = n\,(X^tX)^{-1}$ όπου X ο πίνακας σχεδιασμού. Οι υπερπαράμετροι ν_0 και σ_0^2 επιλέγονται με τέτοιο τρόπο ώστε η εκ των προτέρων διασπορά του σ^2 να είναι μεγάλη και έτσι θέτουμε $\nu_0 = 0.2$ και $\sigma_0^2 = 1$.

Η απόσταση που θα χρησιμοποιήσουμε είναι ίδια με αυτή που χρησιμοποιήθηκε και στα παραδείγματα 2 και 3 του Κεφαλαίου 1.3.1, η σταθμισμένη Ευκλείδεια απόσταση διαιρούμενη με το πλήθος των δεδομένων, και θέτουμε $\varepsilon_0 = 0.02$. Χρησιμοποιώντας τον αλγόριθμο ABC likelihood-free rejection sampler, θα πάρουμε $N = 1000$ προσομοιωμένες τιμές για τις παραμέτρους $\beta_0, \beta_1, \beta_2, \beta_3$ και σ^2. Τα αποτέλεσματα που παίρνουμε με την ολοκλήρωση του αλγορίθμου συνοψίζονται στο **Σχήμα 3.2**.



Σχήμα 3.2: Η εκτίμηση των παραμέτρων $\beta_0, \beta_1, \beta_2, \beta_3$ και σ^2 του πολλαπλού γραμμικού μοντέλου από τον αλγόριθμο ABC rejection sampler

Κατασκευάζοντας τα ιστογράμματα των προσομοιωμένων τιμών μαζί με τις πραγματικές εκ των υστέρων κατανομές των παραμέτρων (**Σχήμα 3.2**) μπορούμε να κρίνουμε την ικανότητα του αλγορίθμου ως

προς την εκτίμηση αυτών. Από τα διαγράμματα είναι εμφανές ότι δεν έχουμε καλή προσαρμογή των δεδομένων στις πραγματικές εκ των υστέρων κατανομές. Γι΄ αυτό θα προχωρήσουμε στην υλοποίηση του αλγορίθμου ABC MCMC για την εκτίμηση των παραμέτρων του πολλαπλού γραμμικού μοντέλου.

3.6 Ο αλγόριθμος ABC MCMC στο πολλαπλό γραμμικό μοντέλο

Όπως παρατηρήσαμε στο προηγούμενο κεφάλαιο, ο αλγόριθμος ABC likelihood-free rejection sampler **δε** δίνει ικανοποιητικά αποτελέσματα στην εκτίμηση των παραμέτρων ενός πολλαπλού γραμμικού μοντέλου. Γι΄ αυτό, καθίσταται αναγκαία η χρήση του αλγορίθμου ABC MCMC. Οι τιμές των υπερπαραμέτρων των εκ των προτέρων κατανομών παραμένουν ίδιες αλλά η απόσταση που θα χρησιμοποιήσουμε για τη σύγκριση των προσομοιωμένων δεδομένων Z και των πραγματικών δεδομένων Y διαφοροποιείται.

Αρχικά, δίνονται αρχικές τιμές στις παραμέτρους από τους εκτιμητές μέγιστης πιθανοφάνειας. Στη συνέχεια, σημαντικό βήμα στον αλγόριθμο είναι η σωστή επιλογή κατανομών εισήγησης / πρότασης. Για τις παραμέτρους $\beta_j,\ \ j = 0, 1, 2, 3$ προτείνουμε τιμές, β_j^*, από την κανονική κατανομή με μέση τιμή $\beta_j^{(t-1)}$, δηλαδή την κατάσταση στην οποία βρισκόταν η αλυσίδα στο προηγούμενο βήμα και κατάλληλη τυπική απόκλιση. Η επιλογή της τιμής για την τυπική απόκλιση πρέπει να γίνεται με τέτοιο τρόπο έτσι ώστε να γίνεται καλή εξερεύνηση του χώρου και να επιτυγχάνεται γρήγορα η σύγκλιση της αλυσίδας. Σε αυτό το παράδειγμα θα δοκιμάσουμε δύο τιμές για την τυπική απόκλιση, αρχικά είναι ίση με 0.25 και στη συνέχεια είναι 0.1. Για την παράμετρο σ^2 προτείνουμε τιμές σ_*^2 από την λογαριθμική κανονική κατανομή με μέση τιμή $\log \sigma_{(t-1)}^2$ και τυπική απόκλιση ίση με 0.25 στην πρώτη δοκιμή και 0.1 στην δεύτερη.

Στον αλγόριθμο ABC MCMC έχουμε δύο κριτήρια τα οποία πρέπει να ικανοποιούνται για να δεχθούμε τις προτεινόμενες τιμές. Πρώτα, τα προσομοιωμένα δεδομένα Z πρέπει να βρίσκονται ΄κοντά΄ στα πραγματικά δεδομένα Y και για το σκοπό αυτό χρησιμοποιούμε κατάλληλη απόσταση και κατάλληλο ε_0.

Υπολογίζουμε τον πίνακα στήλη $A = X^tY$ και παίρνουμε το διάνυσμα $c = (\alpha_1/n, \alpha_2/n, \alpha_3/n, \alpha_4/n, s_Y^2)$ όπου s_Y^2 η δειγματική διασπορά του Y. Αντίστοιχα, υπολογίζουμε $\Delta = X^tZ$ και παίρνουμε το διάνυσμα $d = (\delta_1/n, \delta_2/n, \delta_3/n, \delta_4/n, s_Z^2)$, όπου s_Z^2 η δειγματική διασπορά του Z. Η απόσταση που χρησιμοποιούμε είναι η ευκλείδεια απόσταση αυτών των δύο διανυσμάτων,

$$\rho_1 = \sqrt{\sum_{i=1}^{5} (c_i - d_i)^2} \tag{3.18}$$

η οποία είναι επαρκής στατιστική συνάρτηση. Οι τιμές που δίνουμε στο ε_0 είναι δύο, αρχικά ισούται με 0.75 αλλά όπως θα διαπιστώσουμε δεν είναι ικανοποιητικό οπότε το μειώνουμε στο 0.5.

Το δεύτερο κριτήριο που πρέπει να ικανοποιείται προέρχεται από τον αλγόριθμο MCMC. Η νέα τιμή γίνεται αποδεκτή με πιθανότητα αποδοχής,

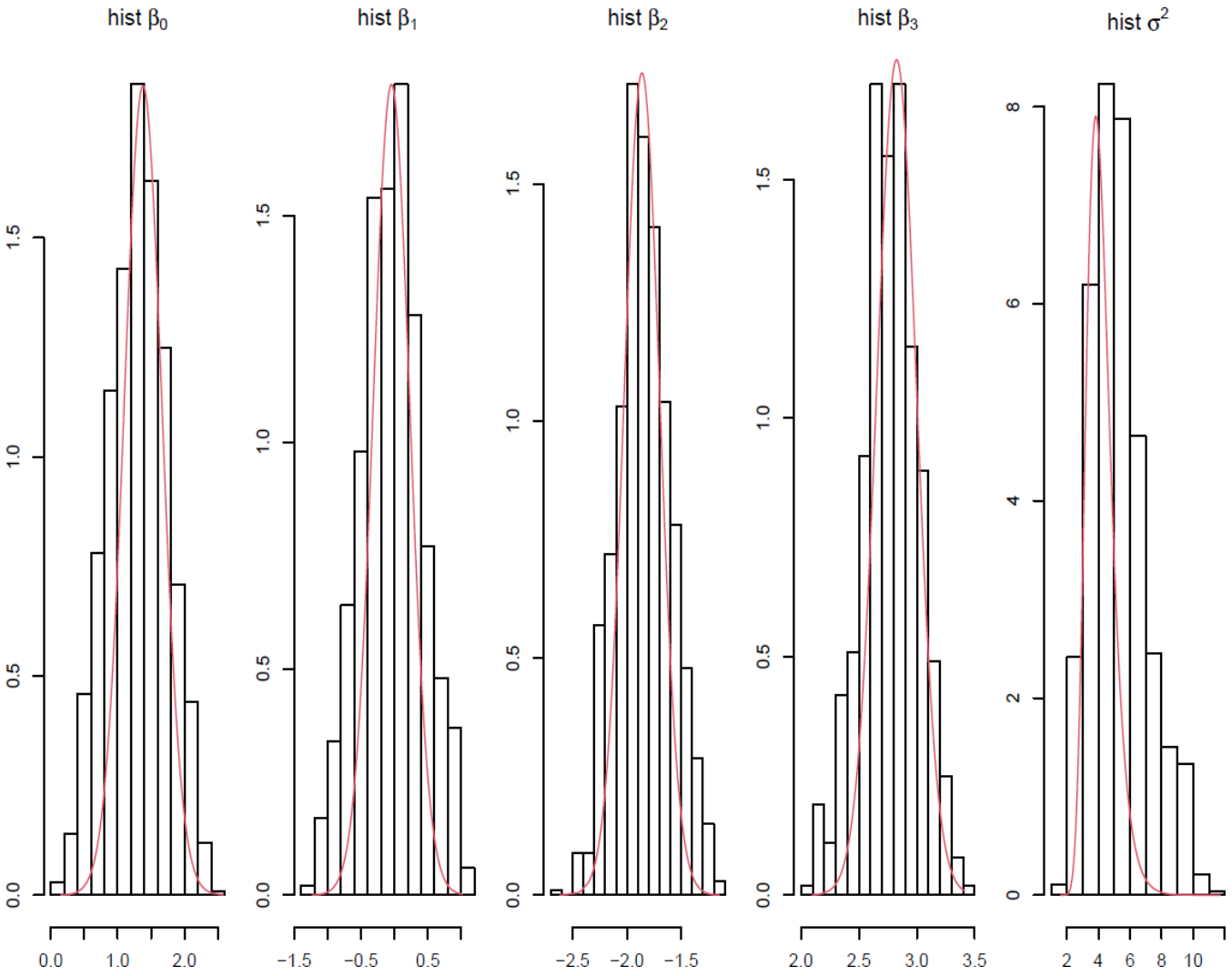
$$\alpha_{MH} = \min\left\{1, \frac{f(\theta^*)}{f(\theta^{(t-1)})} \cdot \frac{q(\theta^{(t-1)}|\theta^*)}{q(\theta^*|\theta^{(t-1)})}\right\}$$

όπου $\theta^* = (\beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*, \sigma_*^2)$ το προτεινόμενο διάνυσμα των παραμέτρων, $\theta^{(t-1)}$ η προηγούμενη κατάσταση στην οποία βρισκόταν η αλυσίδα, f η εκ των προτέρων κατανομή των παραμέτρων, η πολυδιάστατη κανονική αντίστροφη-χ^2 και q η κατανομή εισήγησης. Βάζουμε λογαριθμική κλίμακα για

περισσότερη ευκολία.

Χρησιμοποιούμε 10,000,000 επαναλήψεις και αποθηκεύουμε τιμές ανά 10,000 με αποτέλεσμα να πάρουμε 1,000 προσομοιωμένες τιμές για τις παραμέτρους με την ολοκλήρωση του αλγορίθμου.

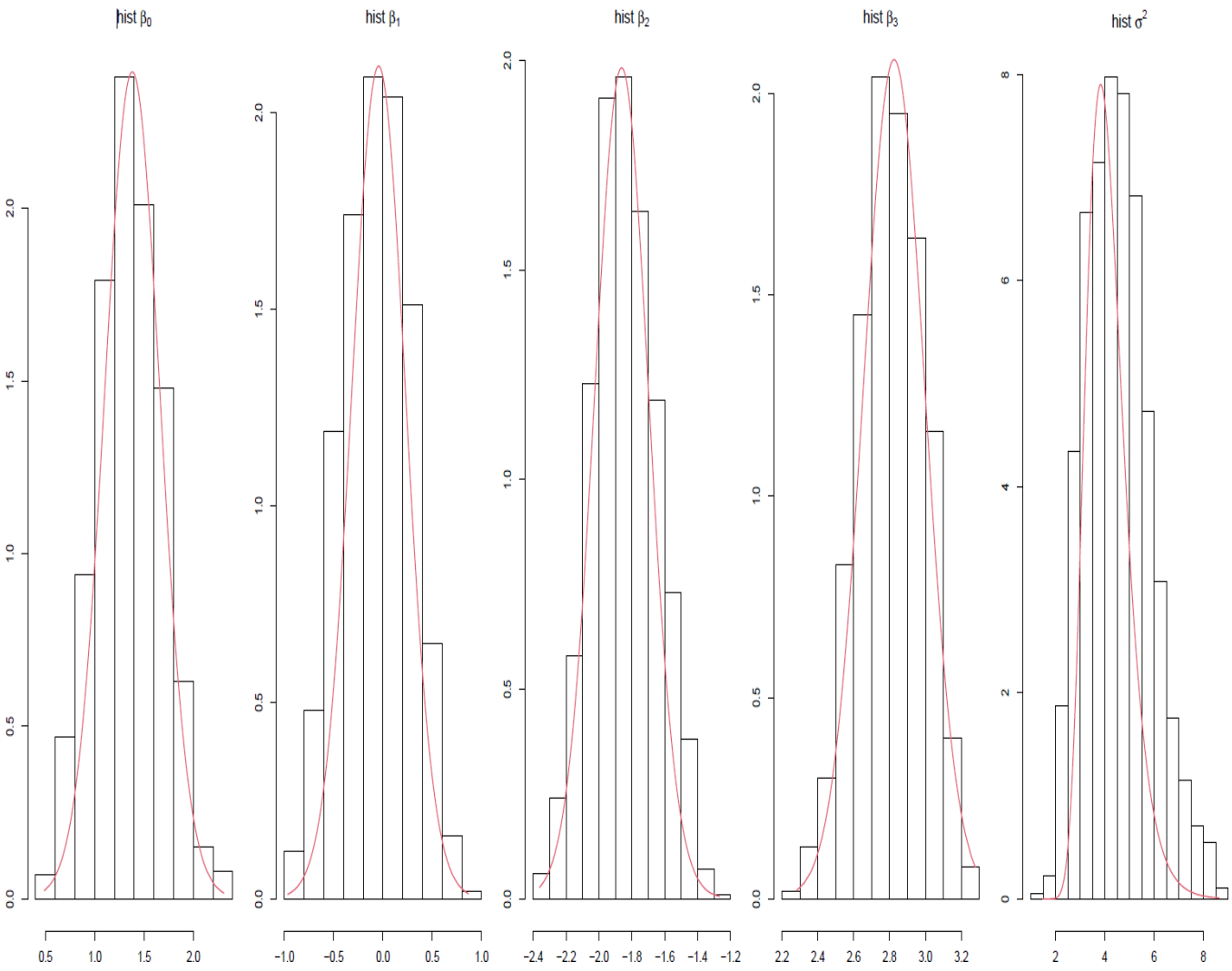
Αρχικά, τρέχουμε τον αλγόριθμο έχοντας τυπική απόκλιση ίση με 0.25 για την κατανομή εισήγησης και $\varepsilon_0 = 0.75$ και τα αποτελέσματα παρουσιάζονται στο **Σχήμα 3.3**. Παρατηρούμε ότι υπάρχει σημαντική



Σχήμα 3.3: Η εκτίμηση των παραμέτρων $\beta_0, \beta_1, \beta_2, \beta_3$ και σ^2 του πολλαπλού γραμμικού μοντέλου από τον αλγόριθμο ABC likelihood-free MCMC sampler με τυπική απόκλιση ίση με 0.25 και $\varepsilon_0 = 0.75$

βελτίωση στα αποτελέσματα συγκριτικά με το **Σχήμα 3.2**. Σε μία προσπάθεια να εξετάσουμε αν ο αλγόριθμος μπορεί να κάνει ακόμα καλύτερη εκτίμηση των παραμέτρων μειώνουμε την τυπική απόκλιση στο 0.1 και το ε_0 στο 0.5.

Κατασκευάζοντας τα ιστογράμματα των προσομοιωμένων τιμών μαζί με τις πραγματικές εκ των υστέρων κατανομές των παραμέτρων στο **Σχήμα 3.4** μπορούμε να πούμε ότι επιτυγχάνεται καλύτερη προσαρμογή μειώντας τη τιμή της τυπικής απόκλισης και του ε_0. Το ποσοστό αποδοχής είναι ίσο με 12%.



Σχήμα 3.4: Η εκτίμηση των παραμέτρων $\beta_0, \beta_1, \beta_2, \beta_3$ και σ^2 του πολλαπλού γραμμικού μοντέλου από τον αλγόριθμο ABC likelihood-free MCMC sampler με τυπική απόκλιση ίση με 0.1 και $\varepsilon_0 = 0.5$.

3.7 Επιλογή μοντέλου

Θεωρούμε το γενικότερο πρόβλημα επιλογής ενός μοντέλου $m \in \mathbb{M}$, όπου $\mathbb{M}$ συμβολίζει το σύνολο των δυνατών μοντέλων. Το καλύτερο από όλα τα δυνατά μοντέλα θεωρείται αυτό με τη μεγαλύτερη εκ των υστέρων πιθανότητα $f(m|y)$. Η επιλογή αυτή αντιστοιχεί στην κορυφή της εκ των υστέρων κατανομής των μοντέλων, Maximum a posteriori, (MAP). Κάθε μοντέλο καθορίζει την κατανομή των δεδομένων Y, $f(y|m, \beta_m)$, όπου β_m είναι το άγνωστο διάνυσμα των παραμέτρων του μοντέλου m. Γνωρίζουμε ότι $\beta_m \in \mathbb{B}_m$ όπου $\mathbb{B}_m$ είναι το σύνολο όλων των πιθανών τιμών για τους συντελεστές του μοντέλου m. Αν $\pi(m)$ είναι η εκ των προτέρων πιθανότητα του μοντέλου m, τότε η εκ των υστέρων πιθανότητα δίνεται από τη σχέση,

$$f(m|y) = \frac{\pi(m) f(y|m)}{\sum_{m \in \mathbb{M}} \pi(m) f(y|m)}, \qquad m \in \mathbb{M} \tag{3.19}$$

όπου $f(y|m)$ η περιθωριακή πιθανοφάνεια που ισούται με $f(y|m) = \int f(y|m, \beta_m) f(\beta_m|m) d\beta_m$ και $f(\beta_m|m)$ είναι η δεσμευμένη εκ των προτέρων κατανομή του β_m.

Εναλλακτικά, μπορούμε να επιλέξουμε το μοντέλο που περιέχει όλες τις επεξηγηματικές μεταβλητές των οποίων η πιθανότητα ένταξης είναι τουλάχιστον ίση με 0.5. Ως πιθανότητα ένταξης της μεταβλητής j ορίζεται η περιθωριακή πιθανότητα του ενδεχομένου το μοντέλο k να περιέχει τη j επεξηγηματική μεταβλητή, $k = 1, 2, ..., M$. Έστω λοιπόν οι δίτιμες μεταβλητές,

$$\gamma_{kj} = \begin{cases} 1, & \text{αν το μοντέλο } k \text{ περιέχει τη μεταβλητή } j \\ 0, & \text{διαφορετικά} \end{cases}$$

για $k = 1, 2, ..., M$ και $j = 0, 1, ..., p$. Τότε, η πιθανότητα ένταξης της μεταβλητής j είναι ίση με,

$$\gamma_{\cdot j} = \sum_{k=1}^{M} f(M_k|y)\gamma_{kj}, \qquad j = 0, ..., p \tag{3.20}$$

Το ολοκλήρωμα που εμφανίζεται στον όρο $f(y|m)$ της σχέσης (3.19) υπολογίζεται αναλυτικά μόνο σε συγκεκριμένες περιορισμένες περιπτώσεις. Επίσης, το μέγεθος του συνόλου των πιθανών μοντέλων $\mathbb{M}$ μπορεί να γίνει αρκετά μεγάλο σε σημείο που ο υπολογισμός του $f(y|m)$ για όλα τα μοντέλα m να καθίσταται αδύνατος. Το πλήθος των πιθανών μοντέλων αυξάνεται εκθετικά καθώς σε ένα γραμμικό μοντέλο με p επεξηγηματικές μεταβλητές, το πλήθος των δυνατών μοντέλων είναι $2^{p+1} - 1$.

Συνεπώς, οι μέθοδοι MCMC, οι οποίες παράγουν παρατηρήσεις / δεδομένα από την από κοινού εκ των υστέρων κατανομή $f(m, \beta_m|y)$ του διανύσματος (m, β_m) είναι ιδιαίτερα χρήσιμες για την εκτίμηση των κατανομών $f(m|y)$ και $f(\beta_m|m, y)$. Ο διανυσματικός χώρος του διανύσματος (m, β_m) είναι $B = \cup_{m\in\mathbb{M}} \{m\} \times \mathbb{B}_m$. Θα χρησιμοποιήσουμε τον αλγόριθμο reversible jump MCMC για την επιλογή μεταβλητών στο πολλαπλό γραμμικό μοντέλο.

3.8 Ο αλγόριθμος RJMCMC στην επιλογή μεταβλητών για το πολλαπλό γραμμικό μοντέλο

Ο αλγόριθμος reversible jump MCMC, (Green, 1995), παράγει τιμές από την από κοινού εκ των υστέρων κατανομή $f(m, \beta_m|y)$ και εξερευνά και το χώρο των παραμέτρων και των μοντέλων, επιτρέποντας τη δειγματοληψία / προσομοίωση τιμών από μοντέλα τα οποία έχουν διαφορετικές διαστάσεις, (Han and Carlin, 2001).

Θεωρώντας ότι η τωρινή κατάσταση του αλγορίθμου είναι (m, β_m), όπου β_m είναι το διάνυσμα των παραμέτρων του μοντέλου m διάστασης $\dim(\beta_m)$, ο αλγόριθμος προχωράει προτείνοντας νέες τιμές $(m', \beta'_{m'})$, διάστασης $\dim(\beta'_{m'})$, η οποία μπορεί να διαφέρει από τη προηγούμενη διάσταση $\dim(\beta_m)$. Εξαιτίας αυτής της αλλαγής στη διάσταση του διανύσματος των παραμέτρων, η σύγκλιση του αλγορίθμου εξασφαλίζεται με την προϋπόθεση ότι υπάρχει αντιστρεψιμότητα και αντιστοίχιση διαστάσεων, (Hartman and Hart, 2009).

Η ικανοποίηση αυτών των προϋποθέσεων επιτυγχάνεται εισάγωντας μία βοηθητική τυχαία μεταβλητή $u \sim q\left(u|\beta_m, m, m'\right)$. Η μεταβλητή αυτή συνδέεται με κάθε μοντέλο m, $m \in \mathbb{M}$, έτσι ώστε η διάσταση του διανύσματος (β_m, u) να παραμένει σταθερή για όλα τα μοντέλα, να έχουμε δηλαδή αντιστοίχιση διαστάσεων. Επίσης, κάθε διάνυσμα (β_m, u) συνδέεται με το $\left(\beta'_{m'}, u'\right)$ μέσω μίας αντιστρέψιμης συνάστησης g έτσι ώστε

$$\left(\beta'_{m'}, u'\right) = g_{m,m'}\left(\beta_m, u\right) \tag{3.21}$$

Η σχέση αυτή ικανοποιεί την αντιστρεψιμότητα με τέτοιο τρόπο έτσι ώστε ο αλγόριθμος να μπορεί να κινείται και ανάποδα, από την προτεινόμενη τιμή στην τωρινή κατάσταση, να ισχύει δηλαδή $g_{m',m} = g^{-1}_{m,m'}$. Λόγω της σχέσης (3.21), επιτυγχάνεται αντιστοίχιση διαστάσεων, δηλαδή όταν προτείνεται μία κίνηση από το μοντέλο m στο $m^{'}$, όπου $\dim(\beta_m) \neq \dim(\beta^{'}_{m'})$, ισχύει η ακόλουθη ισότητα,

$$\dim(\beta_m) + \dim(u) = \dim(\beta^{'}_{m'}) + \dim(u^{'}) \tag{3.22}$$

Τέλος, η προτεινόμενη τιμή είτε γίνεται αποδεκτή είτε απορρίπτεται υπολογίζοντας την πιθανότητα αποδοχής. Ο υπολογισμός είναι παρόμοιος με αυτόν που απαιτείται για την πιθανότητα αποδοχής του αλγορίθμου Metropolis - Hastings με ένα επιπλέον χαρακτηριστικό. Υπάρχει προσαρμογή του αλγορίθμου στην αλλαγή της διάστασης πολλαπλασιάζοντας με την Ιακωβιανή ορίζουσα $J = \left|\frac{\mathrm{d}g\left(\beta_m, u\right)}{\mathrm{d}\left(\beta_m, u\right)}\right|$. Θεωρώντας ότι η Μαρκοβιανή αλυσίδα βρίσκεται στην κατάσταση (m, β_m), τότε ο αλγόριθμος reversible jump MCMC ακολουθεί τα παρακάτω βήματα:

1. Προτείνεται το νέο μοντέλο $m^{'}$ με πιθανότητα $j(m^{'}, m)$
2. Προσομοίωση τιμής u από την κατανομή εισήγησης $q\left(u|\beta_m, m, m^{'}\right)$
3. Θέσε $\left(\beta^{'}_{m'}, u^{'}\right) = g_{m,m'}\left(\beta_m, u\right)$ όπου g είναι η αντιστρέψιμη συνάρτηση
4. Αποδοχή της νέας τιμής του μοντέλου $m^{'}$ με πιθανότητα

$$\alpha_{m\to m'} = \min\left\{1, \frac{f(y|m^{'}, \beta^{'}_{m'})f(\beta^{'}_{m'}|m^{'})\pi(m^{'})j(m^{'}, m)q(u^{'}|\beta^{'}_{m'}, m^{'}, m)}{f(y|m, \beta_m)f(\beta_m|m)\pi(m)j(m, m^{'})q(u|\beta_m, m, m^{'})} \times \left|\frac{\mathrm{d}g\left(\beta_m, u\right)}{\mathrm{d}\left(\beta_m, u\right)}\right|\right\}$$

Αλγόριθμος 3.1. Ο αλγόριθμος reversible jump MCMC

Υπάρχουν πολλές παραλλαγές του αλγορίθμου reversible jump MCMC που μπορούν να εφαρμοστούν σε προβλήματα επιλογής μοντέλου. Μία πιο απλή παραλλαγή συγκεκριμένα προκύπτει αν όλες οι παράμετροι του προτεινόμενου μοντέλου προσομοιωθούν απευθείας από μια κατανομή εισήγησης. Σε αυτή την περίπτωση έχουμε $\left(\beta^{'}_{m'}, u^{'}\right) = (u, \beta_m)$ με $\dim(\beta_m) = \dim(u^{'})$ και $\dim(\beta^{'}_{m'}) = \dim(u)$, και η Ιακωβιανή ορίζουσα στην πιθανότητα αποδοχής ισούται με 1.

Αξίζει να σημειωθεί ότι ένα χαρακτηριστικό του αλγορίθμου reversible jump που φανερώνει την ευελιξία του είναι η δυνατότητα κίνησης ανάμεσα σε μοντέλα διαφορετικής διάστασης. Στην απλή παραλλαγή που συζητήθηκε παραπάνω, αν το μοντέλο m είναι εμφωλευμένο στο μοντέλο $m^{'}$, τότε με την κατάλληλη κατανομή εισήγησης και κατάλληλη συνάρτηση $g_{m,m'}$, όπως είναι η ταυτοτική συνάρτηση, έχουμε $\dim(u^{'}) = 0$ και $\beta^{'}_{m'} = g_{m,m'}\left(\beta_m, u\right)$. Συνεπώς, όταν η αντίστροφη κίνηση προταθεί, οι παράμετροι του μοντέλου προτείνονται ντετερμινιστικά, (Dellaportas and Forster, 1999).

3.9 Ο αλγόριθμος ABC RJMCMC

Ο αλγόριθμος reversible jump MCMC μπορεί να διαμορφωθεί κατάλληλα σε περιπτώσεις όπου η πιθανοφάνεια δεν είναι διαθέσιμη σε κλειστή μορφή ως συνάρτηση των παραμέτρων. Στο πρόβλημα επιλογής μεταβλητών σε ένα πολλαπλό γραμμικό μοντέλο ο αλγόριθμος θα τροποποιηθεί κατάλληλα έτσι ώστε, **α)** η ανανέωση του μοντέλου να βασίζεται σε κίνηση τύπου γέννησης - θανάτου και **β)** η

ανανέωση των παραμέτρων να γίνεται σύμφωνα με τον αλγόριθμο Likelihood-free MCMC sampler.

Ας υποθέσουμε ότι έχουμε p επεξηγηματικές μεταβλητές, n παρατηρήσεις και το πολλαπλό γραμμικό μοντέλο της μορφής,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \quad \text{όπου } \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n$$

Οι παράμετροι του μοντέλου είναι $q = p + 1$ στο πλήθος μαζί με το σταθερό όρο β_0. Χρησιμοποιώντας το διάνυσμα $m \in \{0, 1\}^q - \{0, ..., 0\}$ συμβολίζουμε το μοντέλο, όπου $m_j = 1$ ή $m_j = 0$ αν το β_j εμπεριέχεται ή όχι στο μοντέλο αντίστοιχα. Το πλήθος των μεταβλητών που εμπεριέχονται ή αποκλείονται από το μοντέλο m εκφράζονται από τους συμβολισμούς

$$n_1(m) = \sum_{j=1}^{q} m_j \quad \text{και} \quad n_0(m) = q - n_1(m)$$

αντίστοιχα. Επίσης, κάνουμε την εκ των προτέρων υπόθεση ότι ο συνολικός αριθμός όλων των πιθανών μοντέλων είναι $M = 2^q - 1$ και η εκ των προτέρων πιθανότητα επιλογής μοντέλου m είναι $\pi(m) = \dfrac{1}{M}$, δηλαδή επιλέγουμε τη διακριτή ομοιόμορφη κατανομή.

Ξεκινώντας με τον αλγόριθμο, επιλέγουμε πρώτα το μοντέλο / γίνεται ανανέωση μοντέλου. Το μοντέλο ανανεώνεται εκτελώντας είτε γέννηση μίας νέας μεταβλητής είτε θάνατο μίας προϋπάρχουσας μεταβλητής. Η γέννηση μίας νέας μεταβλητής σημαίνει κίνηση σε μοντέλο μεγαλύτερης διάστασης από το τωρινό μοντέλο και ο θάνατος μιας προυπάρχουσας μεταβλητής σημαίνει κίνηση σε μοντέλο μικρότερης διάστασης από το τωρινό. Η πιθανότητα γέννησης είναι,

$$P_m(\text{birth}) = \begin{cases} 1, & n_1(m) = 1 \\ \frac{1}{2}, & 1 < n_1(m) < q \\ 0, & n_1(m) = q \end{cases} \tag{3.23}$$

και συνεπώς, η πιθανότητα θανάτου είναι $P_m(\text{death}) = 1 - P_m(\text{birth})$.

Ας υποθέσουμε ότι προτείνεται μία κίνηση **γέννησης** μίας νέας μεταβλητής και από το μοντέλο m, τωρινή κατάσταση, με παραμέτρους $\theta_m = (\beta_m, \sigma^2)$ προτείνεται το νέο μοντέλο $m^{'}$ με παραμέτρους $\theta^{'}_{m^{'}} = \left(\beta^{'}_{m^{'}}, \sigma^2\right)$. Τα βήματα που ακολουθούν είναι τα εξής :

1. Τυχαία επιλέγουμε ένα δείκτη h στο μοντέλο m, έτσι ώστε $m_h = 0$. Το νέο μοντέλο που προτείνεται $m^{'}$ είναι αυτό στο οποίο $m^{'}_j = m_j$ για κάθε $j \neq h$ και $m^{'}_h = 1$.

2. Προστίθεται μία νέα παράμετρος και η τιμή u αυτής προσδιορίζεται μέσω μιάς κατανομής εισήγησης, $u \sim q(u)$

3. Το διάνυσμα των νέων παραμέτρων $\theta^{'}_{m^{'}}$ προσδιορίζεται μέσω της ταυτοτικής συνάρτησης $g_{m,m^{'}}$, $\theta^{'}_{m^{'}} = g_{m,m^{'}}(\theta_m, u)$, και έχουμε

$$\beta^{'}_j = \begin{cases} \beta_j, & j < h \\ u, & j = h \\ \beta_{j-1}, & j > h \end{cases} \quad \text{και} \quad (\sigma^{'})^2 = \sigma^2 \tag{3.24}$$

Χρησιμοποιώντας την ταυτοτική συνάρτηση επιλέγουμε την απλή παραλλαγή του αλγορίθμου reversible jump που συζητήθηκε στο προηγούμενο κεφάλαιο και η Ιακωβιανή ορίζουσα ισούται με 1.

4. Η προτεινόμενη τιμή γίνεται αποδεκτή με πιθανότητα αποδοχής,

$$\alpha_{m\to m'} = \min\left\{1, \frac{f(\beta'_{m'}|m')\pi(m')j(m',m)}{f(\beta_m|m)\pi(m)j(m,m')} \times \left|\frac{\mathrm{d}g\left(\beta_m,u\right)}{\mathrm{d}\left(\beta_m,u\right)}\right|\right\} \tag{3.25}$$

$$\text{όπου εδώ } \pi(m) = \pi(m') = \frac{1}{M}, \quad \frac{j(m',m)}{j(m,m')} = \frac{P_{m'}(\text{ death}) \cdot \frac{1}{n_1(m')}}{P_m(\text{ birth}) \cdot \frac{1}{n_0(m)} \cdot q(u)} \text{ και } \left|\frac{\mathrm{d}g\left(\beta_m,u\right)}{\mathrm{d}\left(\beta_m,u\right)}\right| = 1$$

Ας υποθέσουμε τώρα ότι προτείνεται μία κίνηση **θανάτου** μίας προϋπάρχουσας μεταβλητής από το αρχικό μοντέλο m με παραμέτρους $\theta_m = (\beta_m, \sigma^2)$. Το νέο μοντέλο m' με παραμέτρους $\theta'_{m'} = \left(\beta'_{m'}, \sigma^2\right)$ περιέχει μία μεταβλητή λιγότερη από το αρχικό με αποτέλεσμα να είναι μικρότερης διάστασης. Τα βήματα που ακολουθούν είναι τα εξής :

1. Τυχαία επιλέγουμε ένα δείκτη h στο μοντέλο m, έτσι ώστε $m_h = 1$. Το νέο μοντέλο που προτείνεται m' είναι αυτό στο οποίο $m'_j = m_j$ για κάθε $j \neq h$ και $m'_h = 0$.

2. Λόγω της αντιστρεψιμότητας της συνάρτησης (3.23), οι παράμετροι του νέου μοντέλου προτείνονται ντετερμινιστικά. Συγκεκριμένα, οι παράμετροι του νέου μοντέλου παραμένουν ίδιες με αυτές του αρχικού εξαιρώντας αυτή που αντιστοιχεί στη μεταβλητή που διαγράφεται στο **βήμα 1**. Η τυχαία βοηθητική μεταβλητή u ισούται με το β_h και έχουμε

$$\left(\theta'_{m'}, u\right) = g^{-1}_{m,m'}\left(\theta_m\right) \tag{3.26}$$

   Δηλαδή, $\beta'_j = \{\beta_j : m_j = 1, j \neq h\}$, $(\sigma')^2 = \sigma^2$ και $u = \beta_h$.

3. Η προτεινόμενη τιμή γίνεται αποδεκτή με πιθανότητα αποδοχής που ισούται με την αντίστροφη τιμή της πιθανότητας (3.25), $\alpha^{-1}_{m'\to m}$.

Ο αλγόριθμος RJMCMC με κίνηση τύπου γέννησης - θανάτου συνδυάζεται με τις συνθήκες και τις προϋποθέσεις του αλγορίθμου ABC δημιουργώντας τον ακόλουθο αλγόριθμο, Αλγόριθμος 3.2.

---

1. Γίνεται αρχικοποίηση του (m, θ_m)
2. Για $i = 1, ..., N$:
   - α. Ανανέωση μοντέλου m μέσω κίνησης γέννησης-θανάτου
     - (i) Γίνεται κίνηση γέννησης ή θανάτου με πιθανότητα $P_m(\text{birth})$ ή $1 - P_m(\text{birth})$ αντίστοιχα
     - (ii) Αν πραγματοποιηθεί γέννηση μίας νέας μεταβλητής, προσομοίωσε $u \sim q(u)$ και υπολόγισε το $\theta'_{m'}$ μέσω της σχέσης (3.24) αλλιώς, υπολόγισε το $\theta'_{m'}$ μέσω της σχέσης (3.26)
     - (iii) Προσομοίωση δεδομένων z από την πιθανοφάνεια $f(\cdot|m', \theta'_{m'})$
     - (iv) Προσομοίωση τιμής $w \sim \mathbb{U}_{[0,1]}$
     - (v) Αν $w \leq \alpha_{m\to m'}$, για κίνηση γέννησης, ή $w \leq \alpha^{-1}_{m'\to m}$, για κίνηση θανάτου, και $\rho(\eta(Z), \eta(Y)) \leq \varepsilon$ τότε θέσε $(m^t, \theta^t_m) = \left(m', \theta'_{m'}\right)$
   - β. Ανανέωση των παραμέτρων
     - (i) Υλοποίηση του Αλγορίθμου (1.3) Likelihood-free MCMC sampler από το Κεφάλαιο 1.3

Αλγόριθμος 3.2. Ο αλγόριθμος ABC RJMCMC για την επιλογή μεταβλητών σε ένα πολλαπλό γραμμικό μοντέλο

Στα παραδείγματα που ακολουθούν εξετάζουμε την αποδοτικότητα και την ακρίβεια του αλγορίθμου ABC RJMCMC για την επιλογή μεταβλητών σε ένα μοντέλο της μορφής (3.1), αποτελούμενο από τρεις επεξηγηματικές μεταβλητές στο πρώτο παράδειγμα και από έξι επεξηγηματικές μεταβλητές στο δεύτερο.

3.9.1 Παράδειγμα επιλογής μεταβλητών σε πολλαπλό γραμμικό μοντέλο με 3 επεξηγηματικές μεταβλητές

Θεωρούμε δείγμα μεγέθους $n = 100$ παρατηρήσεων / προσομοιωμένων τιμών με τη βοήθεια του οποίου θα γίνει η επιλογή του καλύτερου μοντέλου εξετάζοντας ποιες μεταβλητές είναι απαραίτητες γι΄ αυτό και ποιες είναι περιττές. Θεωρούμε ότι έχουμε $p = 3$ επεξηγηματικές μεταβλητές, X_1, X_2, X_3, και έχουμε το πολλαπλό γραμμικό μοντέλο του **Κεφαλαίου 3.4**,

$$Y = 1 + 0 \cdot X_1 - 2X_2 + 3X_3 + \varepsilon$$

όπου $\varepsilon \sim N(0, \sigma^2 \mathbb{I}_n)$, με $\sigma^2 = 3$. Οι σχέσεις των μεταβλητών απεικονίζονται στο Σχήμα 3.1.

Ξεκινώντας τον αλγόριθμο, γίνεται αρχικοποίηση του (m, θ_m). Θεωρούμε ότι $m = (1, 1, 1, 1)$, ξεκινάμε δηλαδή έχοντας όλες τις επεξηγηματικές μεταβλητές, και στο διάνυσμα των παραμέτρων θ_m δίνονται οι εκτιμητές μέγιστης πιθανοφάνειας. Γίνεται κίνηση γέννησης ή θανάτου με πιθανότητα $P_m(\text{birth})$ ή $1 - P_m(\text{birth})$ αντίστοιχα που υπολογίζεται από τη σχέση (3.23).

Στην περίπτωση της γέννησης μίας νέας μεταβλητής, πραγματοποιείται προσομοίωση της τιμής αυτής από την κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 2, $u \sim N(0, 4)$. Το προτεινόμενο μοντέλο $m^{'}$ είναι μεγαλύτερης διάστασης από το αρχικό m. Τα προσομοιωμένα δεδομένα z προέρχονται από την κανονική κατανομή, $z \sim N_n\left(X\beta^{'}_{m^{'}}, \sigma^2 \mathbb{I}_n\right)$ όπου ο πίνακας σχεδιασμού X είναι διάστασης $n \times n_1(m^{'})$.

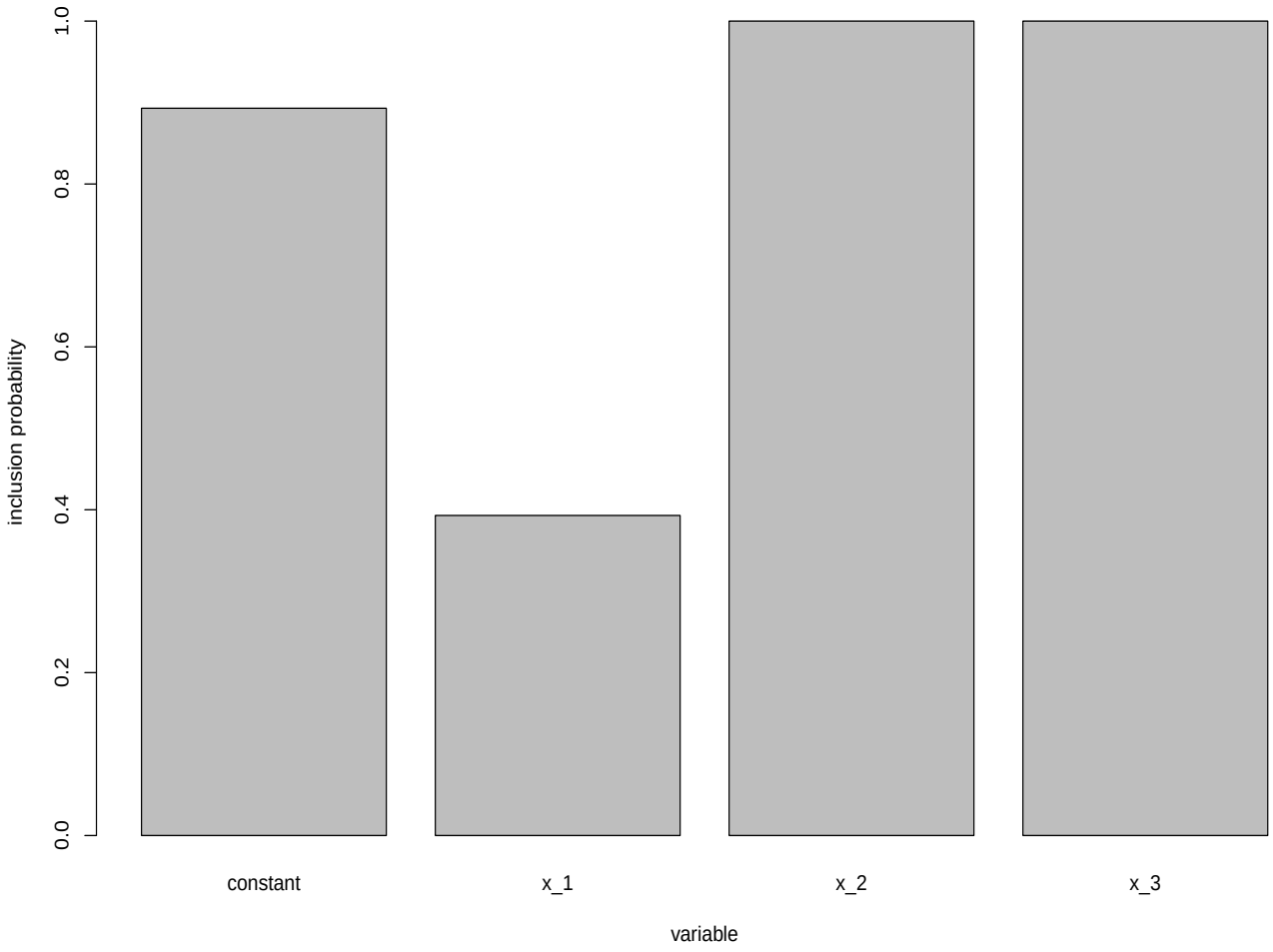
Εξετάζουμε αν τα προσομοιωμένα δεδομένα z είναι ΄κοντά΄ στα πραγματικά δεδομένα y χρησιμοποιώντας την απόσταση (3.18) και $\varepsilon_0 = 0.5$. Αν είναι κοντά τότε προχωράμε στον υπολογισμό της πιθανότητας (3.25) όπου στο διάνυσμα $\beta_m|m$ δίνεται η εκ των προτέρων κατανομή του Zellner που συζητήθηκε στο **Κεφάλαιο 3.3**, $f\left(\beta_m|m\right) = N_{n_1(m)}(\tilde{\beta}, \sigma^2\Sigma)$, όπου $\tilde{\beta}$ είναι το μηδενικό διάνυσμα-στήλη διάστασης $n_1(m)$ και ο πίνακας $\Sigma = n(X^tX)^{-1}$, με πίνακα σχεδιασμού X διάστασης $n \times n_1(m)$. Προχωράμε στην εξέταση της επόμενης συνθήκης και αν $w \leq \alpha_{m \to m'}$ τότε ανανεώνουμε το μοντέλο και το αντίστοιχο διάνυσμα των παραμέτρων. Σε αντίθετη περίπτωση κρατάμε το υπάρχον μοντέλο και τις παραμέτρους του.

Έχοντας ολοκληρώσει το πρώτο σκέλος του δεύτερου βήματος στον Αλγόριθμο 3.2, περνάμε στο δεύτερο σκέλος και γίνεται η ανανέωση των παραμέτρων χρησιμοποιώντας τον αλγόριθμο ABC-MCMC. Για την ανανέωση των δεικτών του διανύσματος β_m χρησιμοποιούμε την κανονική κατανομή για κατανομή εισήγησης, $\beta_j \sim N\left(\beta_j^{(t-1)}, 0.1^2\right)$. Η κατανομή εισήγησης που χρησιμοποιείται για την ανανέωση του σ^2 είναι η κανονικοποιημένη λογαριθμική, (Log-Normal), $\log \sigma^2 \sim \text{Log-Norm}\left((\sigma^2)^{(t-1)}, 0.1^2\right)$.

Στην περίπτωση που πραγματοποιείται ο θάνατος μιας προϋπάρχουσας μεταβλητής τα βήματα είναι αντίστοιχα. Οι διαφοροποιήσεις που παρουσιάζονται συγκριτικά με την κίνηση γέννησης είναι δύο. Πρώτον, το μοντέλο που προκύπτει $m^{'}$ είναι μικρότερης διάστασης από το αρχικό m. Δεύτερον, η πιθανότητα αποδοχής του προτεινόμενου διανύσματος των παραμέτρων ισούται με την αντίστροφη τιμή

της πιθανότητας (3.25).

Πραγματοποιούνται 1,000,000 επαναλήψεις και κρατάμε τα 1,000 επικρατέστερα μοντέλα μαζί με τις προσομοιωμένες τιμές των παραμέτρων τους. Ο χρόνος που απαιτείται για την ολοκλήρωση του αλγορίθμου είναι στις 4 ώρες. Η πιθανότητα να δεχτούμε το μοντέλο που προτείνεται σε κάθε βήμα ισούται με 0.035% και η πιθανότητα να δεχτούμε τις αντίστοιχες τιμές που προτείνονται ισούται με 0.096%. Το **Σχήμα 3.5** παρουσιάζει τις πιθανότητες ένταξης για την κάθε μεταβλητή. Όπως παρα-



Σχήμα 3.5: Η πιθανότητα ένταξης κάθε μεταβλητής στο πολλαπλό γραμμικό μοντέλο σύμφωνα με τον αλγόριθμο ABC RJMCMC, Αλγόριθμος 3.2.

τηρούμε και από το σχήμα **Σχήμα 3.5** έχουμε τις ακόλουθες πιθανότητες ένταξης κάθε μεταβλητής: $\widehat{P}(constant \in M) = 0.893$, $\widehat{P}(X_1 \in M) = 0.393$, $\widehat{P}(X_2 \in M) = 1.00$ και τέλος, $\widehat{P}(X_3 \in M) = 1.00$. Συνεπώς, το βέλτιστο μοντέλο κρίνεται να είναι αυτό που περιλαμβάνει τον σταθερό όρο, την μεταβλητή X_2 και τη μεταβλητή X_3. Η μεταβλητή X_1 αποκλείεται γιατί είναι $\widehat{P}(X_1 \in M) = 0.393 < 0.5$.

Χρησιμοποιώντας τα αποτελέσματα που παρουσιάζονται στον Πίνακα 3.1 μπορούμε να υπολογίσουμε τις **πραγματικές πιθανότητες** ένταξης των μεταβλητών. Με τη βοήθεια της σχέσης (3.20) και της εκ των υστέρων πιθανότητας κάθε μοντέλου υπολογίζουμε την πιθανότητα ένταξης κάθε μετα-

βλητής : $P(constant \in M) = 0.9996$, $P(X_1 \in M) = 0.1906$, $P(X_2 \in M) = 1.00$ και τέλος, $P(X_3 \in M) = 1.00$.

j	m_j	$f(y\|m_j)$	$f(m_j\|y)$
1	1000	-339.6144	$5.25 \cdot 10^{-57}$
2	0100	-339.4601	$6.13 \cdot 10^{-57}$
3	1100	-341.7216	$6.39 \cdot 10^{-58}$
4	0010	-332.9197	$4.25 \cdot 10^{-54}$
5	1010	-334.9362	$5.65 \cdot 10^{-55}$
6	0110	-321.5753	$3.59 \cdot 10^{-49}$
7	1110	-323.4031	$5.77 \cdot 10^{-50}$
8	0001	-326.9451	$1.67 \cdot 10^{-51}$
9	1001	-329.2245	$1.71 \cdot 10^{-52}$
10	0101	-296.6984	$2.28 \cdot 10^{-38}$
11	1101	-297.4871	$1.04 \cdot 10^{-38}$
12	0011	-223.2192	$1.86 \cdot 10^{-6}$
13	**1011**	**-210.1212**	**0.909**
14	0111	-225.5171	$1.87 \cdot 10^{-7}$
15	1111	-212.4272	$9.06 \cdot 10^{-2}$

Πίνακας 3.1 Η περιθωριακή πιθανοφάνεια και η αντίστοιχη εκ των υστέρων πιθανότητα για κάθε μοντέλο

Ο Πίνακας 3.1. περιέχει την περιθωριακή πιθανοφάνεια και την εκ των υστέρων πιθανότητα για κάθε μοντέλο, οι οποίες υπολογίζονται με τη βοήθεια της συνάρτησης variable_selection της βιβλιοθήκης bayesianRegression, (Παπασταμούλης, 2019, Ο.Π.Α), στην R και της εντολής summary. Το μοντέλο 1011, δηλαδή το μοντέλο με τον σταθερό όρο και τις μεταβλητές X_2, X_3 είναι αυτό με τη μεγαλύτερη περιθωριακή πιθανοφάνεια και τη μεγαλύτερη εκ των υστέρων πιθανότητα.

3.9.2 Παράδειγμα επιλογής μεταβλητών σε πολλαπλό γραμμικό μοντέλο με 6 επεξηγηματικές μεταβλητές

Στο παράδειγμα αυτό διπλασιάζουμε το πλήθος των παρατηρήσεων n καθώς και το πλήθος των επεξηγηματικών μεταβλητών για να κρίνουμε την ικανότητα του αλγορίθμου στην επιλογή του βέλτιστου μοντέλου. Θεωρούμε δείγμα μεγέθους $n = 200$ παρατηρήσεων / προσομοιωμένων τιμών χρησιμοποιώντας τη συνάρτηση simulate_data από τη βιβλιοθήκη bayesianRegression της R. Επιλέγουμε το πλήθος των επεξηγηματικών μεταβλητών να ισούται με $p = 6$ και έχουμε το πολλαπλό γραμμικό μοντέλο,

$$Y = 1 + 0 \cdot X_1 + 0 \cdot X_2 + 0 \cdot X_3 - 2X_4 + 0 \cdot X_5 + 3X_6 + \varepsilon$$

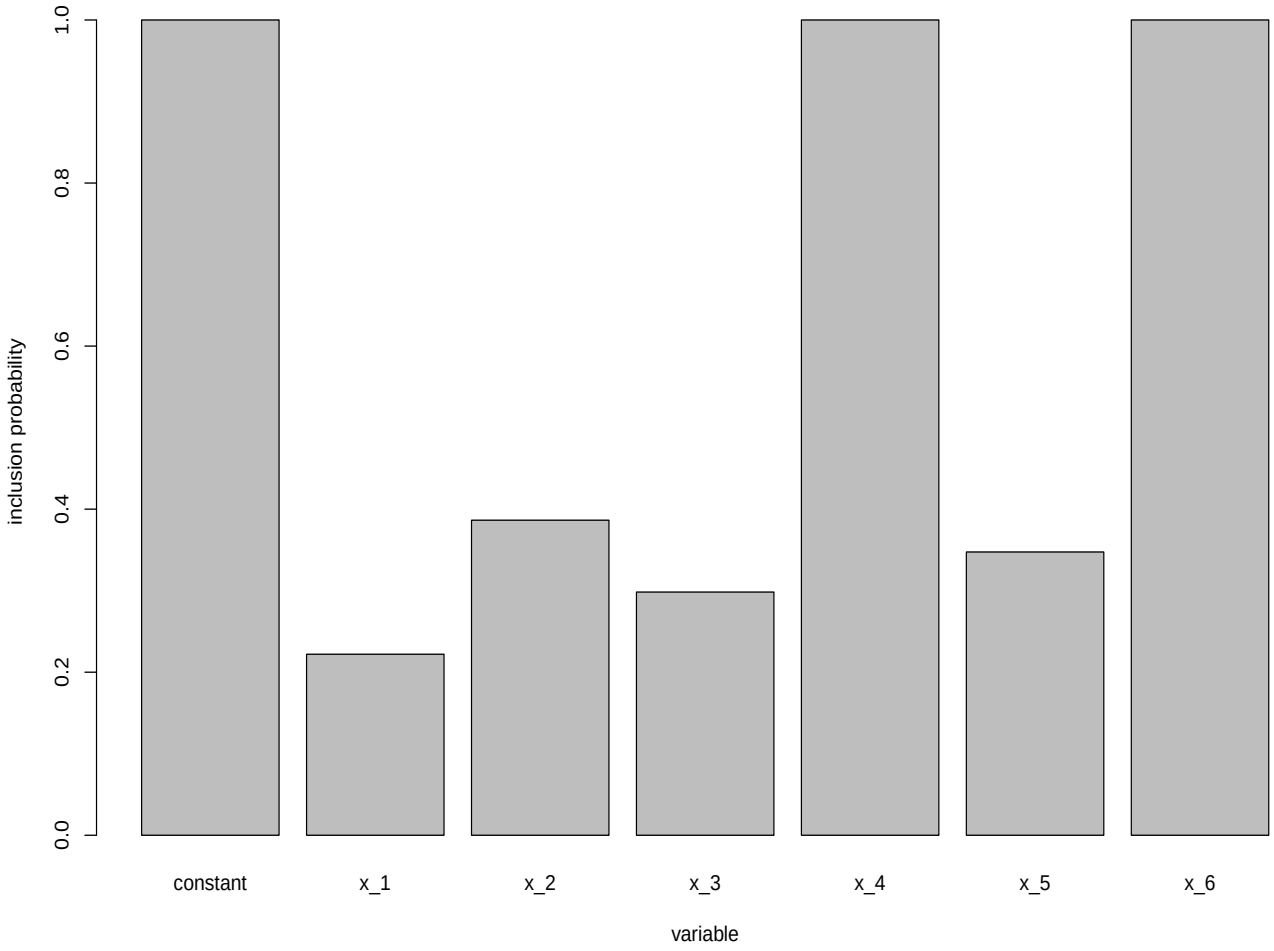
όπου $\varepsilon \sim N(0, \sigma^2 \mathbb{I}_n)$, με $\sigma^2 = 3$. Σε αυτό το παράδειγμα τα υποψήφια προς επιλογή μοντέλα είναι 127 στο πλήθος και καθίσταται αδύνατος ο υπολογισμός της εκ των υστέρων πιθανότητας (3.19) για κάθε ένα από αυτά. Χρησιμοποιώντας τον **Αλγόριθμο 3.2** θα αποφασίσουμε ποιο είναι το βέλτιστο μοντέλο εξετάζοντας ποιες μεταβλητές είναι απαραίτητες γι΄ αυτό και ποιες είναι περιττές.

Ξεκινώντας τον αλγόριθμο, γίνεται αρχικοποίηση του (m, θ_m). Θεωρούμε ότι $m = (1, 1, 1, 1)$, ξεκινάμε δηλαδή έχοντας όλες τις επεξηγηματικές μεταβλητές, και στο διάνυσμα των παραμέτρων β_m δίνονται οι εκτιμητές μέγιστης πιθανοφάνειας. Γίνεται κίνηση γέννησης ή θανάτου με πιθανότητα $P_m(\text{birth})$ ή

$1 - P_m(\text{birth})$ αντίστοιχα που υπολογίζεται από τη σχέση (3.23).

Στην περίπτωση της γέννησης μίας νέας μεταβλητής, πραγματοποιείται προσομοίωση της τιμής αυτής από την κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 2, $u \sim N(0, 4)$. Εξετάζουμε αν τα προσομοιωμένα δεδομένα z είναι 'κοντά' στα πραγματικά δεδομένα y χρησιμοποιώντας την απόσταση (3.18) και $\varepsilon_0 = 0.75$. Στο δεύτερο σκέλος του αλγορίθμου 3.2 γίνεται η ανανέωση των παραμέτρων χρησιμοποιώντας τον αλγόριθμο ABC-MCMC. Για την ανανέωση των δεικτών του διανύσματος β_m χρησιμοποιούμε την κανονική κατανομή για κατανομή εισήγησης, $\beta_j \sim N\left(\beta_j^{(t-1)}, 0.3^2\right)$. Η κατανομή εισήγησης που χρησιμοποιείται για την ανανέωση του σ^2 είναι η κανονικοποιημένη λογαριθμική, (Log-Normal), $\log \sigma^2 \sim \text{Log-Norm}\left((\sigma^2)^{(t-1)}, 0.3^2\right)$.

Πραγματοποιούνται 1,000,000 επαναλήψεις και κρατάμε 1,000 τιμές. Ο χρόνος που απαιτείται για την ολοκλήρωση του αλγορίθμου είναι στις 16 ώρες. Η πιθανότητα να δεχτούμε το μοντέλο που προτείνεται σε κάθε βήμα του αλγορίθμου ισούται με 0.058% και η πιθανότητα να δεχτούμε τις αντίστοιχες τιμές που προτείνονται ισούται με 0.072%.



Σχήμα 3.6: Η πιθανότητα ένταξης κάθε μεταβλητής στο πολλαπλό γραμμικό μοντέλο σύμφωνα με τον αλγόριθμο ABC RJMCMC, Αλγόριθμος 3.2.

Όπως παρατηρούμε από το **Σχήμα 3.6** οι πιθανότητες ένταξης κάθε μεταβλητής είναι οι ακόλουθες: $\widehat{P}(constant \in M) = 1.00$, $\widehat{P}(X_1 \in M) = 0.222$, $\widehat{P}(X_2 \in M) = 0.386$, $\widehat{P}(X_3 \in M) = 0.298$, $\widehat{P}(X_4 \in M) = 1.00$, $\widehat{P}(X_5 \in M) = 0.347$, $\widehat{P}(X_6 \in M) = 1.00$. Συνεπώς, το βέλτιστο μοντέλο κρίνεται να είναι αυτό που περιλαμβάνει τον σταθερό όρο, την μεταβλητή X_4 και τη μεταβλητή X_6. Οι υπόλοιπες αποκλείονται από το μοντέλο γιατί οι πιθανότητες ένταξής τους σε αυτό είναι μικρότερες από 0.5.

Αντίστοιχα με το προηγούμενο παράδειγμα, χρησιμοποιώντας τα αποτελέσματα που παρουσιάζονται στον Πίνακα 3.2 μπορούμε να υπολογίσουμε τις πραγματικές πιθανότητες ένταξης των μεταβλητών. Με τη βοήθεια της σχέσης (3.20) και της εκ των υστέρων πιθανότητας κάθε μοντέλου υπολογίζουμε την **πραγματική πιθανότητα** ένταξης κάθε μεταβλητής και παρατηρούμε ότι οι προσεγγιστικές τιμές βρίσκονται πολύ κοντά με τις πραγματικές. Οι πραγματικές τιμές είναι οι ακόλουθες: $P(constant \in M) = 1.00$, $P(X_1 \in M) = 0.12$, $P(X_2 \in M) = 0.25$, $P(X_3 \in M) = 0.17$, $P(X_4 \in M) = 1.00$, $P(X_5 \in M) = 0.19$, $P(X_6 \in M) = 1.00$.

j	m_j	$f(y\|m_j)$	$f(m_j\|y)$
...	...	...	...
16	0000100	-599.0592	$6.08 \cdot 10^{-86}$
17	1000100	-595.7317	$1.69 \cdot 10^{-84}$
...	...	...	...
64	0000001	-517.7236	$1.28 \cdot 10^{-50}$
65	1000001	-506.2627	$1.22 \cdot 10^{-45}$
...	...	...	...
80	0000101	-432.0890	$1.99 \cdot 10^{-13}$
81	**1000101**	**-403.1700**	**0.720**
82	0100101	-434.4989	$1.78 \cdot 10^{-14}$
83	1100101	-405.7768	$5.31 \cdot 10^{-2}$
84	0010101	-434.0216	$2.88 \cdot 10^{-14}$
85	1010101	-405.3734	$7.95 \cdot 10^{-2}$
86	0110101	-436.6686	$2.04 \cdot 10^{-15}$
87	1110101	-407.9445	$6.08 \cdot 10^{-3}$
88	0001101	-434.6113	$1.59 \cdot 10^{-14}$
89	1001101	-405.6985	$5.75 \cdot 10^{-2}$
90	0101101	-437.1222	$1.29 \cdot 10^{-15}$
91	1101101	-408.3452	$4.07 \cdot 10^{-3}$
...	...	...	...
127	1111111	-412.9723	$3.98 \cdot 10^{-5}$

Πίνακας 3.2 Η περιθωριακή πιθανοφάνεια και η αντίστοιχη εκ των υστέρων πιθανότητα για κάθε μοντέλο

Ενδεικτικά, εμφανίζονται στον Πίνακα 3.2 η περιθωριακή πιθανοφάνεια και η εκ των υστέρων πιθανότητα για μερικά από τα μοντέλα με τη βοήθεια της συνάρτησης variable_selection της βιβλιοθήκης bayesianRegression στην R και της εντολής summary. Το μοντέλο 1000101, δηλαδή το μοντέλο με τον σταθερό όρο και τις μεταβλητές X_4, X_6 είναι αυτό με τη μεγαλύτερη περιθωριακή πιθανοφάνεια και τη μεγαλύτερη εκ των υστέρων πιθανότητα.

Κεφάλαιο 4

Συμπεράσματα

4.1 Συμπεράσματα

Η παρούσα διπλωματική εργασία επικεντρώνεται στη μελέτη των αλγορίθμων ABC, ξεκινώντας από την αρχική μορφή που εισήγαγαν οι Tavare et al, 1997, προχωρώντας στην αυθεντικότερη μορφή από τους Pritchard et al, 1999 και καταλήγοντας σε βελτιωμένες μορφές και προεκτάσεις αυτών.

Ο αλγόριθμος απόρριψης (**Αλγόριθμος 1.2**) κρίνεται ιδιαίτερα αποτελεσματικός σε περιπτώσεις όπου οι προσομοιωμένες τιμές του δείγματος προέρχονται από ένα πεπερασμένο σύνολο, όπως εξετάστηκε στο διωνυμικό μοντέλο. Στο διωνυμικό μοντέλο απαιτείται μικρό χρονικό διάστημα για να υλοποιηθεί ο αλγόριθμος ανεξερτήτως του μεγέθους του δείγματος και δεν είναι απαραίτητη η διερεύνηση γύρω από τις τιμές της σταθεράς ε λόγω των διακριτών τιμών του συνόλου δεδομένων.

Αντίθετα, στο κανονικό μοντέλο με άγνωστη μέση τιμή απαιτείται περισσότερη διερεύνηση τόσο γύρω από τις τιμές της σταθεράς ε όσο και ως προς το μέτρο απόκλισης. Αυτό οφείλεται στο γεγονός ότι οι προσομοιωμένες τιμές του δείγματος αυτή τη φορά προέρχονται από το συνεχές σύνολο των πραγματικών αριθμών. Συνεπώς, αυξάνεται και ο απαιτούμενος χρόνος ολοκλήρωσης του αλγορίθμου. Ο **Αλγόριθμος 1.2** εξακολουθεί να ανταποκρίνεται αποτελεσματικά και στο κανονικό μοντέλο με άγνωστη μέση τιμή, όμως στην περίπτωση που οι άγνωστες παράμετροι είναι δύο, η μέση τιμή και η διασπορά, κρίνεται καταλληλότερος ο αλγόριθμος ABC-MCMC, (**Αλγόριθμος 1.3**).

Παρατηρούμε ότι σε κάθε περίπτωση ο **Αλγόριθμος 1.2** και ο **Αλγόριθμος 1.3** είναι ικανοί να εκτιμήσουν με μεγάλη ακρίβεια τις άγνωστες παραμέτρους. Συγκεκριμένα, οι αλγόριθμοι γίνονται ακόμα πιο αποδοτικοί όταν η σταθερά ε τείνει στο μηδέν ή όταν το μέγεθος του δείγματος τείνει στο άπειρο.

Παρ΄ όλα αυτά, τα αποτελέσματα δεν είναι τόσο ενθαρρυντικά όσον αφορά στην ικανότητα του αλγορίθμου ABC-MC στην επιλογή του καταλληλότερου μοντέλου. Ο **Αλγόριθμος 2.1** είναι ικανός να επιλέξει το καταλληλότερο μοντέλο μεταξύ δύο διωνυμικών μοντέλων ανεξαρτήτως του μεγέθους του δείγματος. Όμως, τόσο η σύγκριση μεταξύ δύο κανονικών μοντέλων με διαφορετική μέση τιμή όσο και η σύγκριση μεταξύ δύο κανονικών μοντέλων με διαφορετική μέση τιμή και διασπορά δεν ολοκληρώνονται με επιτυχία. Τα δύο αυτά μοντέλα είναι ειδική περίπτωση του κανονικού γραμμικού μοντέλου και απαιτείται κατάλληλη τροποποίηση του αλγορίθμου με ενσωμάτωση των μεθόδων MCMC.

Πρέπει να σημειωθεί ότι οι αλγόριθμοι ABC-MC κρίνονται ανεπαρκείς για την επιλογή του καταλ-

ληλότερου μοντέλου σε πιο σύνθετα στοχαστικά μοντέλα. Σύμφωνα με τους Robert et al, 2011, η ακαταλληλότητα των αλγορίθμων οφείλεται στο γεγονός ότι μία στατιστική συνάρτηση $\eta(y)$ μπορεί να είναι επαρκής για δύο μοντέλα $f_1(y|\theta_1)$ και $f_2(y|\theta_2)$ ξεχωριστά, αυτό όμως **δεν** σημαίνει ότι η σ.σ. $\eta(y)$ είναι απαραίτητα επαρκής για το συνδυασμό $\{m, f_m(y|\theta_m)\}$.

Οι Didelot et al, 2010 παρατήρησαν ότι η συνθήκη της επάρκειας παρέχει έγκυρα αποτελέσματα όταν ο αλγόριθμος εφαρμοστεί σε εμφωλευμένα μοντέλα που ανήκουν στην εκθετική οικογένεια κατανομών. Συνεπώς, το ενδιαφέρον μας επικεντρώνεται στα κανονικά γραμμικά μοντέλα. Πρώτα απ΄ όλα, ο αλγόριθμος ABC-MCMC είναι ικανός να εκτιμήσει με μεγάλη ακρίβεια τις παραμέτρους ενός κανονικού γραμμικού μοντέλου υπό κατάλληλες προϋποθέσεις. Η επιλογή της επαρκούς στατιστικής συνάρτησης (3.18) παίζει καθοριστικό ρόλο. Επίσης, η μείωση της σταθεράς ε καθώς και η μείωση στη τυπική απόκλιση της κατανομής πρότασης q συμβάλλουν στην αποδοτικότητα του αλγορίθμου, **Σχήμα 3.4**.

Στη συνέχεια, ο καταλληλότερος αλγόριθμος για την επιλογή μεταβλητών και κατ΄ επέκταση για την επιλογή μοντέλου σε ένα πολλαπλό γραμμικό μοντέλο είναι μία παραλλαγή του αλγορίθμου reversible jump MCMC, ο αλγόριθμος ABC-RJMCMC, **Αλγόριθμος 3.2**. Σύμφωνα με το **Κεφάλαιο 3.7**, προκειμένου να ενταγχθεί μια μεταβλητή στο μοντέλο πρέπει η πιθανότητα ένταξής της να είναι τουλάχιστον ίση με 0.5. Στα παραδείγματα των **Κεφαλαίων 3.9.1** και **Κεφαλαίων 3.9.2** ο αλγόριθμος υπολογίζει τις προσεγγιστικές πιθανότητες ένταξης των μεταβλητών σε ένα πολλαπλό γραμμικό μοντέλο. Υπολογίζοντας και τις πραγματικές πιθανότητες από τη σχέση (3.20) παρατηρούμε ότι οι προσεγγιστικές τιμές βρίσκονται πολύ κοντά στις πραγματικές. Συνεπώς, ο **Αλγόριθμος 3.2** κρίνεται πολύ ικανοποιητικός για την επιλογή μεταβλητών σε ένα πολλαπλό γραμμικό μοντέλο και κατ΄ επέκταση σε ελέγχους υποθέσεων για κανονικά δεδομένα.

Παράρτημα

Παράρτημα Α

Οι κυριότερες διακριτές και συνεχείς μονοδιάστατες κατανομές

Α.1 Συνεχείς κατανομές

Ομοιόμορφη

Η ομοιόμορφη κατανομή χρησιμοποιείται για την αναπαράσταση μίας μεταβλητής θ για την οποία είναι γνωστό ότι παίρνει τιμές σε ένα διάστημα με κάτω άκρο το α και άνω άκρο το β. Είναι εξίσου πιθανό η μεταβλητή θ να βρεθεί οπουδήποτε εντός του διαστήματος (α, β). Συμβολίζουμε ως $\theta \sim U(\alpha, \beta)$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{1}{\beta - \alpha}$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{\alpha + \beta}{2} \qquad Var(\theta) = \frac{(\beta - \alpha)^2}{12}$$

Κανονική

Η κανονική ή Gaussian κατανομή είναι πανταχού παρούσα στην στατιστική συμπερασματολογία, είτε ύπο την κλασική είτε υπό την Μπεϋζιανή προσέγγιση. Χαρακτηρίζεται από δύο παραμέτρους, την παράμετρο θέσης μ και την παράμετρο κλίμακας σ. Συμβολίζουμε ως $\theta \sim N(\mu, \sigma^2)$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta - \mu)^2}{2\sigma^2}\right\}$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \mu \qquad Var(\theta) = \sigma^2$$

Η διασπορά είναι αυστηρά θετική. Μία μη πληροφοριακή κατανομή αποκτάται καθώς η διασπορά $\sigma \to \infty$. Αν z είναι μία τυχαία μεταβλητή η οποία ακολουθεί την τυποποιημένη κανονική κατανομή, $N(0, 1)$, τότε η τυχαία μεταβλητή $\theta = \mu + \sigma z$ ακολουθεί την κατανομή $N(\mu, \sigma^2)$.

Κανονικοποιημένη Λογαριθμική

Έστω θ μία τυχαία μεταβλητή, η οποία είναι υποχρεωτικά θετική και $\log\theta \sim N(\mu, \sigma^2)$, τότε θεωρούμε ότι η μεταβλητή θ ακολουθεί την κανονικοποιημένη λογαριθμική κατανομή. Χαρακτηρίζεται από δύο παραμέτρους, την παράμετρο θέσης μ και την παράμετρο λογαριθμικής κλίμακας σ, όπου $\sigma > 0$. Συμβολίζουμε ως $\theta \sim \text{lognormal}(\mu, \sigma^2)$ με συνάρτηση πυκνότητας πιθανότητας ίση με,

$$f(\theta) = \frac{1}{\sqrt{2\pi}\sigma\theta} \exp\left\{-\frac{(\log\theta - \mu)^2}{2\sigma^2}\right\}$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \exp\left\{\mu + \frac{1}{2}\sigma^2\right\} \qquad Var(\theta) = \left(e^{\sigma^2} - 1\right)\exp\left\{2\mu + \sigma^2\right\}$$

Στον αλγόριθμο ABC MCMC χρησιμοποιείται ως κατανομή εισήγησης για την προσομοίωση τιμών για τη διασπορά του τυχαίου σφάλματος σε ένα πολλαπλό γραμμικό μοντέλο.

Γάμμα

Η Γάμμα κατανομή χρησιμοποιείται συχνά ως συζυγή εκ των προτέρων κατανομή μεταβλητών οι οποίες κατανέμονται ασύμμετρα και είναι αυστηρά θετικές. Χαρακτηρίζεται από δύο θετικές παραμέτρους, την παράμετρο θέσης α και την παράμετρο κλίμακας β. Συμβολίζουμε ως $\theta \sim Gamma(\alpha, \beta)$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}, \quad \theta > 0$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{\alpha}{\beta} \qquad Var(\theta) = \frac{\alpha}{\beta^2}$$

Αντίστροφη Γάμμα

Αν η μεταβλητή θ^{-1} ακολουθεί Γάμμα κατανομή με παραμέτρους α και β, τότε η μεταβλητή θ ακολουθεί την αντίστροφη-γάμμα κατανομή. Χρησιμοποιείται ως συζυγή εκ των προτέρων κατανομή για τη διασπορά μιάς κανονικής κατανομής. Συμβολίζουμε ως $\theta \sim Inv - gamma(\alpha, \beta)$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta}, \quad \alpha > 0,\ \beta > 0,\ \theta > 0$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{\beta}{\alpha - 1}, \ \ \alpha > 1 \qquad Var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \ \ \alpha > 2$$

χ^2

Αν $Z_1, Z_2, ..., Z_\nu$ είναι ν ανεξάρτητες τυποποιημένες κανονικές τυχαίες μεταβλητές, $Z_i \sim N(0,1)$ για $i = 1, 2, ..., \nu$ τότε το άθροισμα

$$Z_1^2 + Z_2^2 + ... + Z_\nu^2$$

λέμε ότι ακολουθεί την κατανομή χ^2 με ν βαθμούς ελευθερίας, $\nu > 0$. Η κατανομή χ^2 είναι μία ειδική περίπτωση της κατανομής γάμμα, με $\alpha = \nu/2$ και $\beta = 1/2$. Συμβολίζουμε ως $\theta \sim \chi^2_\nu$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{\nu/2-1}e^{-\theta/2}, \quad \theta > 0$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \nu \qquad Var(\theta) = 2\nu$$

Αντίστροφη-χ^2

Η αντίστροφη-χ^2 με ν βαθμούς ελευθερίας, $\nu > 0$, είναι μία ειδική περίπτωση της αντίστροφης-γάμμα κατανομής, με $\alpha = \nu/2$ και $\beta = 1/2$. Συμβολίζουμε ως $\theta \sim$ Inv-χ^2_ν με συνάρτηση πυκνότητας πιθανότητας

$$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{-(\nu/2+1)}e^{-1/(2\theta)}, \quad \theta > 0$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{1}{\nu-2}, \quad \nu > 2 \quad \text{και} \quad Var(\theta) = \frac{2}{(\nu-2)^2(\nu-4)}, \quad \nu > 4$$

Κλιμακωτή αντίστροφη-χ^2

Ορίζουμε επίσης, την κλιμακωτή αντίστροφη-χ^2, η οποία χρησιμοποιείται για την εκτίμηση της άγνωστης διασποράς σε κανονικά μοντέλα. Επίπλέον, αν η μεταβλητή X ακολουθεί την κατανομή χ^2_ν, τότε η μεταβλητή θ που προκύπτει από το μετασχηματισμό $\theta = \frac{\nu s^2}{X}$ ακολουθεί την κλιμακωτή αντίστροφη-$\chi^2(\nu, s^2)$ με ν βαθμούς ελευθερίας και παράμετρο κλίμακας s. Είναι ισοδύναμη με την αντίστροφη-γάμμα κατανομή με $\alpha = \frac{\nu}{2}$ και $\beta = \frac{\nu}{2}s^2$. Δίνεται η συνάρτηση πυκνότητας πιθανότητας από το τύπο

$$f(\theta) = \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma(\frac{\nu}{2})}s^\nu\theta^{-(\nu/2+1)}e^{-\nu s^2/(2\theta)}, \quad \theta > 0$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{\nu}{\nu-2}s^2, \quad \nu > 2 \quad \text{και} \quad Var(\theta) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)}s^4, \quad \nu > 4$$

Εκθετική

Η εκθετική κατανομή είναι μία ειδική περίπτωση της κατανομής γάμμα με $\alpha = 1$. Συμβολίζουμε ως $\theta \sim Exp(\lambda)$ με παράμετρο $\lambda > 0$ και συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \lambda e^{-\lambda\theta}, \quad \theta > 0$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{1}{\lambda} \quad \text{και} \quad Var(\theta) = \frac{1}{\lambda^2}$$

Student-t

Η κατανομή t ή αλλιώς Student-t εμφανίζεται ως η περιθώρια εκ των υστέρων κατανομή για την άγνωστη μέση τιμή σε ένα κανονικό μοντέλο με άγνωστες και τις δύο παραμέτρους. Χαρακτηρίζεται από τρεις παραμέτρους, τους βαθμούς ελευθερίας $\nu > 0$, την παράμετρο θέσης μ και την παράμετρο κλίμακας $\sigma > 0$. Στην ειδική περίπτωση όπου $\nu \to +\infty$ η κατανομή t προσεγγίζει την κανονική κατανομή $N(\mu, \sigma^2)$. Συμβολίζουμε ως $\theta \sim t_\nu(\mu, \sigma^2)$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi\sigma^2}} \left(1 + \frac{(\theta-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \mu, \quad \nu > 1 \quad \text{και} \quad Var(\theta) = \frac{\nu}{\nu-2}\sigma^2, \quad \nu > 2$$

Beta

Η κατανομή Beta είναι η συζυγής εκ των προτέρων κατανομή για την πιθανότητα επιτυχίας στο διωνυμικό μοντέλο. Χαρακτηρίζεται από δύο θετικές παραμέτρους $\alpha > 0$ και $\beta > 0$ και στην περίπτωση όπου $\alpha = \beta = 1$ είναι ισοδύναμη με την τυποποιημένη ομοιόμορφη κατανομή $U(0, 1)$. Μπορεί να χρησιμοποιηθεί και ως μη πληροφοριακή κατανομή για $\alpha = \beta = \frac{1}{2}$ ή $\alpha = \beta = 0$. Συμβολίζουμε ως $\theta \sim Beta(\alpha, \beta)$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0, 1]$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta} \quad \text{και} \quad Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Α.2 Διακριτές κατανομές

Poisson

Η κατανομή Poisson εκφράζει την πιθανότητα ενός δεδομένου αριθμού γεγονότων που συμβαίνουν σε ένα σταθερό διάστημα χρόνου. Χαρακτηρίζεται από μία θετική παράμετρο $\lambda > 0$, η οποία αντιπροσωπεύει το ρυθμό εμφάνισης ενός γεγονότος. Συμβολίζουμε ως $\theta \sim Poisson(\lambda)$ με συνάρτηση μάζας πιθανότητας,

$$f(\theta) = \frac{\lambda^{\theta}}{\theta!} e^{-\lambda}, \quad \theta = 0, 1, 2, ...$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \lambda \quad \text{και} \quad Var(\theta) = \lambda$$

Διωνυμική

Η διωνυμική κατανομή χρησιμοποιείται για την αναπαράσταση του αριθμού των επιτυχιών σε μία ακολουθία από n ανεξάρτητες και ισόνομα κατανεμημένες δοκιμές Bernoulli με πιθανότητα επιτυχίας ρ σε κάθε δοκιμή. Στην ειδική περίπτωση όπου $n = 1$, η διωνυμική καλείται κατανομή Bernoulli. Συμβολίζουμε με $\theta \sim Bin(n, \rho)$ με συνάρτηση μάζας πιθανότητας,

$$f(\theta) = \binom{n}{\theta} \rho^{\theta} (1 - \rho)^{n-\theta}, \quad \theta = 0, 1, 2, ..., n$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = n\rho \quad \text{και} \quad Var(\theta) = n\rho(1 - \rho)$$

Αρνητική Διωνυμική

Η αρνητική διωνυμική κατανομή χρησιμοποιείται για την αναπαράσταση του αριθμού των αποτυχιών Bernoulli μέχρι να επιτευχθούν α το πλήθος επιτυχίες, με πιθανότητα επιτυχίας ρ. Συμβολίζουμε με $\theta \sim \text{Neg-Bin}(\alpha, \rho)$ με συνάρτηση μάζας πιθανότητας,

$$f(\theta) = \binom{\theta + \alpha - 1}{\alpha - 1} \rho^{\alpha} (1 - \rho)^{\theta}, \quad \theta = 0, 1, 2, ...$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \frac{\alpha(1 - \rho)}{\rho} \quad \text{και} \quad Var(\theta) = \frac{\alpha(1 - \rho)}{\rho^2}$$

Παράρτημα Β

Οι κυριότερες πολυδιάστατες κατανομές

Πολυμεταβλητή κανονική κατανομή

Η πολυμεταβλητή κανονική κατανομή ενός διανύσματος $\theta = (\theta_1, \theta_2, ..., \theta_k)^T$ διάστασης k γράφεται με το συμβολισμό

$$\theta \sim N_k(\mu, \Sigma)$$

όπου η μέση τιμή ισούται με $\mu = E(\theta) = (E(\theta_1), E(\theta_2), ..., E(\theta_k))$ και ο πίνακας Σ είναι ο $k \times k$ συμμετρικός και θετικά ορισμένος πίνακας διασποράς $\Sigma_{ij} := E[(\theta_i - \mu_i)(\theta_j - \mu_j)] = Cov[\theta_i, \theta_j]$
Η συνάρτηση πυκνότητας πιθανότητας είναι

$$f(\theta) = (2\pi)^{-k/2} (\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)\right\}$$

Οι περιθώριες κατανομές των τυχαίων μεταβλητών θ_i είναι μονοδιάστατες κανονικές κατανομές, $\theta_i \sim N(\mu_i, \Sigma_{ii})$, για κάθε $i = 1, ..., k$ όπου $\mu_i = E(\theta_i)$ και Σ_{ii} το στοιχείο του πίνακα διασποράς Σ.

Πολυμεταβλητή κατανομή t

Η πολυμεταβλητή κατανομή t εμφανίζεται ως η περιθώρια εκ των υστέρων κατανομή για το διάνυσμα $\beta = (\beta_0, \beta_1, ..., \beta_p)$ των παραμέτρων ενός κανονικού γραμμικού μοντέλου. Χαρακτηρίζεται από τέσσερις παραμέτρους, τους βαθμούς ελευθερίας $\nu > 0$, την διάσταση $q = p + 1$, την παράμετρο θέσης $\mu = (\mu_1, \mu_2, ..., \mu_q)$ και τον συμμετρικό και θετικά ορισμένο πίνακα διασποράς $\Sigma > 0$, διάστασης $q \times q$. Συμβολίζουμε ως $\theta \sim t_\nu(\mu, \Sigma)$ με συνάρτηση πυκνότητας πιθανότητας,

$$f(\theta) = \frac{\Gamma(\frac{\nu+q}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{q/2}} (\det \Sigma)^{-1/2} \left(1 + \frac{(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)}{\nu}\right)^{-(\nu+q)/2}$$

Η μέση τιμή και η διασπορά της μεταβλητής θ είναι αντίστοιχα

$$\mathbb{E}(\theta) = \mu, \quad \nu > 1 \quad \text{και} \quad Var(\theta) = \frac{\nu}{\nu - 2}\Sigma, \quad \nu > 2$$

Παράρτημα Γ

Ο κώδικας στο περιβάλλον της R

Στα πλαίσια της διπλωματικής εργασίας όλοι οι κώδικες για την υλοποίηση των αλγορίθμων δημιουργήθηκαν στο περιβάλλον της R και είναι διαθέσιμοι στον παρακάτω σύνδεσμο:

https://github.com/elpidapatta/approximate-bayesian-computation

Βιβλιογραφία

[1] Bartlett, M. (1957) *A comment on D. V. Lindley's Statistical Paradox. Biometrika 44*, 533-534.

[2] Biau G, Cerou F, Guyader A (2015) *New insights into approximate Bayesian computation*, Ann. Inst. H. Poincare Probab. Statist. 51, 376–403.

[3] Chipman H, George E, McCulloch R (2001) *The practical implementation of Bayesian model selection*, IMS Lecture Notes-Monograph Series Volume 38

[4] Dellaportas, P. and Foster, J.J. (1999) *Markov Chain Monte Carlo Model determination for hierarchical and graphical log-linear Models, Biometrika* 86, 615-634

[5] Dellaportas, P. Forster, J.J. and Ntzoufras, I. (2000) *On Bayesian model and variable selection using MCMC. Statist. Comput. To appear.*

[6] Drovandi C (2012) *Bayesian Algorithms with Applications*, Ph.D. dissertation.

[7] Edwards, W Lindman H. and Savage, L.J. (1963) *Bayesian statistical inference for psychological research, Psychological Review* 70 193-242.

[8] Fearnhead P. and D. Prangle (2012) *Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation (with discussion)*, Journal of the Royal Statistical Society, Series B 74, 419–474.

[9] Garthwaite, P. H. and Dickey, J.M. (1996) *Quantifying and using expert opinion for variable-selection problems in regression (with discussion). Chemomet. Intel. Lab.Syst.* 35, 1-34.

[10] Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2014) *Bayesian Data Analysis*, 3rd ed. Chapman & Hall/ CRC Press, Texts in Statistical Science

[11] George, E.I. and McCulloch, R.E. (1993) *'Variable selection via Gibbs sampling', Journal of the american statistical association* 88 (423), 881-889.

[12] Green, P.J. (1995) *'Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination', Biometrika* 82 (4), 711-732.

[13] Han, C. and Carlin, B. (2001), *"Markov chain Monte Carlo methods for computing Bayes factors: a comparative review"* , Journal of the American Statistical Association, 96, 1122–1133.

[14] Hartman, B.M, and Hart, J.D. (2009) *'Using reversible jump MCMC to account for model uncertainty', Actuarial research clearing house.*

[15] Jeffreys, H. (1961) *'Theory of probability', New York: Oxford University Press*

[16] Kass, R.E. and Raftery, A.E. (1995), *'Bayes factors', Journal of the American Statistical Association* 90 (430), 773-795

[17] Kass, R.E. and Wasserman, L. (1995), *'Bayes factors', Journal of the American Statistical Association* 52 (2), 93-100.

[18] Marin J, Pudlo P, Robert C, Ryder R (2011) *Approximate Bayesian Computation*, Statistics and Computing, 22:1167–1180

[19] Marjoram P, Molitor J, Plagnol V, Tavare S (2003) *Markov chain Monte Carlo without likelihoods, Proceedings of the National Academy of Sciences* 100(26):15,324–15,328

[20] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models, 2nd Ed. Chapman and Hall, New York*.

[21] Papastamoulis P, (2019) *Methods of Bayesian Inference*

[22] Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) *Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution* 16:1791–1798

[23] Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) *Bayesian model averaging for linear regression models. J. Amer. Statist. Asso* 92, 179-191

[24] Robert C, Cornuet JM, Marin JM, Pillai N (2011) *Lack of confidence in ABC model choice.* Tech. Rep. 1102.4432, arXiv.org

[25] Rubin D (1984) *Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics* 12(4):1151–1172

[26] Tavare S, Balding D, Griffith R, Donnelly P (1997) *Inferring coalescence times from DNA sequence data, Genetics* 145(2):505–518

[27] Zellner, A. (1986) *On assessing prior distributions and Bayesian regression analysis with g-prior distributions, in Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti, North-Holland/Elsevier*, 233-243.

[28] Panagiotis Papastamoulis (2019) *bayesianRegression: Bayesian Analysis of the Conjugate Normal Regression Model. R package version 1.0* (available from author)