ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ &
ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΕΠΙΛΟΓΗ ΟΙΚΟΓΕΝΕΙΑΣ ΜΕΤΑΣΧΗΜΑΤΙΣΜΩΝ ΣΕ ΜΠΕΫΖΙΑΝΑ ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ: ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΥΣΤΡΑΤΙΑΣ Η. ΧΑΡΙΤΙΔΟΥ

Διπλωματούχου Εφαρμοσμένων Μαθηματικών &
Φυσικών Επιστημών Ε.Μ.Π.

ΕΠΙΒΛΕΠΩΝ:

Δ. ΦΟΥΣΚΑΚΗΣ

Αν. Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, Οκτώβριος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ &
ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΕΠΙΛΟΓΗ ΟΙΚΟΓΕΝΕΙΑΣ ΜΕΤΑΣΧΗΜΑΤΙΣΜΩΝ ΣΕ ΜΠΕΫΖΙΑΝΑ ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ: ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΥΣΤΡΑΤΙΑΣ Η. ΧΑΡΙΤΙΔΟΥ

Διπλωματούχου Εφαρμοσμένων Μαθηματικών &
Φυσικών Επιστημών Ε.Μ.Π.

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 27η Οκτωβρίου 2016.

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

1. Δ. ΦΟΥΣΚΑΚΗΣ, Αν. Καθ. Ε.Μ.Π. (Επιβλέπων)
2. Ι. ΝΤΖΟΥΦΡΑΣ, Καθ. Ο.Π.Α.
3. Γ. ΚΟΚΟΛΑΚΗΣ, Ομ. Καθ. Ε.Μ.Π.

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

1. Δ. ΦΟΥΣΚΑΚΗΣ, Αν. Καθ. Ε.Μ.Π. (Επιβλέπων)
2. Ι. ΝΤΖΟΥΦΡΑΣ, Καθ. Ο.Π.Α.
3. Γ. ΚΟΚΟΛΑΚΗΣ, Ομ. Καθ. Ε.Μ.Π.
4. Μ. ΛΟΥΛΑΚΗΣ, Αν. Καθ. Ε.Μ.Π.
5. Π. ΤΣΙΑΜΥΡΤΖΗΣ, Αν. Καθ. Ο.Π.Α.
6. Θ. ΝΙΚΟΛΕΡΗΣ, Επ. Καθ. Ε.Κ.Π.Α.
7. Ν. ΔΕΜΙΡΗΣ, Επ. Καθ. Ο.Π.Α.

ΑΘΗΝΑ, Οκτώβριος 2016

. . .

ΕΥΣΤΡΑΤΙΑ Η. ΧΑΡΙΤΙΔΟΥ

Διδάκτωρ Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών Ε.Μ.Π.

Αφιερώνεται στη γιαγιά μου Στρατούλα

Ευχαριστίες  [1]

Η παρούσα διδακτορική διατριβή είναι το αποτέλεσμα δημιουργικής εργασίας και επίμονης μελέτης. Είμαι ευγνώμων σε πολλούς ανθρώπους για την υποστήριξη, την καθοδήγηση και την εμπιστοσύνη που επέδειξαν απέναντί μου σε αυτό το δύσκολο ακαδημαϊκό ταξίδι.

Το πρώτο μεγάλο ευχαριστώ απευθύνεται στον επιβλέποντά μου, Αναπληρωτή Καθηγητή Δημήτρη Φουσκάκη, για την εμπιστοσύνη και την καθοδήγησή του στα ερευνητικά μονοπάτια που διάνυσα. Η συνεργασία μας είχε ξεκινήσει ήδη πολύ ενθαρρυντικά στα προπτυχιακά μου χρόνια και συνεχίστηκε μετά την ολοκλήρωση των μεταπτυχιακών μου σπουδών στην Αγγλία και την επιστροφή μου στην Ελλάδα. Από την ανάθεση της παρούσας διατριβής μέχρι σήμερα είμαι ευτυχής να πω ότι μου έχει προσφέρει ανεκτίμητες συμβουλές και πολύπλευρες ακαδημαϊκές εμπειρίες οι οποίες εκτείνονται από προτάσεις για ερευνητικά έργα που καταθέσαμε (π.χ. ΠΕΒΕ 2010), διάφορα συνέδρια εντός και εκτός συνόρων στα οποία λάβαμε μέρος, την άκρως ενδιαφέρουσα ερευνητική μου επίσκεψη στη Santa Barbara της Καλιφόρνιας όπου βρισκόταν ως επισκέπτης καθηγητής, μέχρι εν τέλει την ίδια τη διαδικασία της διδακτορικής έρευνας. Τον ευχαριστώ από βάθους καρδιάς όχι μόνο για τη μετάδοση γνώσεων και τους νέους επιστημονικούς δρόμους που μου έδειξε, αλλά και για την ηθική υποστήριξη που μου προσέφερε.

Το ίδιο θερμά θα ήθελα να ευχαριστήσω και τον δεύτερο επιβλέποντά μου, Καθηγητή Γιάννη Ντζούφρα, μέλος της τριμελούς συμβουλευτικής επιτροπής μου, με τον οποίο συνεργαζόμασταν στενά ως μέλη της ίδιας ερευνητικής ομάδας καθ' όλη τη διάρκεια εκπόνησης της διατριβής μου. Η συνεργασία μας ήταν άψογη και η συμπαράστασή του ήταν καθοριστικής σημασίας για εμένα, ενώ ο ενθουσιώδης τρόπος με τον οποίο αντιμετωπίζει

[1] Εντολή `\Lisa` από το πακέτο `simpsons` του LaTeX.

την ερευνητική διαδικασία με ενθάρρυνε να συνεχίζω υπερπηδώντας τα όποια εμπόδια.

Ευχαριστώ ακόμη τον Ομότιμο Καθηγητή κ. Γεώργιο Κοκολάκη, μέλος της τριμελούς συμβουλευτικής επιτροπής μου, για τα εποικοδομητικά του σχόλια πάνω στην έρευνά μου και πάνω στην ακριβή μετάφραση των διάφορων αγγλικών στατιστικών όρων, καθώς και τους Αναπληρωτές Καθηγητές κ.κ. Μιχάλη Λουλάκη και Παναγιώτη Τσιαμυρτζή και τους Επίκουρους Καθηγητές κ.κ. Νικόλαο Δεμίρη και Θεόδωρο Νικολέρη που με τίμησαν ειλικρινά και δέχτηκαν να συμμετάσχουν στην επταμελή εξεταστική επιτροπή.

Ευχαριστώ ιδιαίτερα και τον Σπύρο, αφενός για την αρχική παρότρυνση να ξεκινήσω αυτή την ακαδημαϊκή περιπέτεια, αλλά και για τη μετέπειτα ενθάρρυνση όταν κατά καιρούς την είχα ανάγκη.

Ακόμη, θα ήθελα να ευχαριστήσω τους φίλους, συναδέλφους και συνυποψήφιους (τότε) διδάκτορες Δημήτρη Στογιάννη, Δήμο Γκουνταρούλη, Νίκο Σταμάτη τους οποίους αναφέρω κατά σειρά πρώτης εμφάνισης στο γραφείο μας! Οι καθημερινές - ενίοτε επιστημονικές ενίοτε αστείες - συζητήσεις μας, οι κοινές μας έξοδοι, τα περιπατητικά διαλείμματα στην Πολυτεχνειούπολη και η συμπαράσταση εκ των έσω για κάθε μικρό ή μεγάλο εμπόδιο που προέκυπτε βοήθησαν σημαντικά.

Πολλά ευχαριστώ και στους υπέροχους φίλους μου Μαρίνα, Γεωργία, Ελισάβετ, Αλέξανδρο, Γιάννη, καθώς και στη Ρηνούλα οι οποίοι συνέβαλαν με τον τρόπο τους στην ολοκλήρωση αυτού του εγχειρήματος. Επίσης, είμαι ευγνώμων στη δασκάλα μουσικής μου, Χριστίνα, και στη συνακορντεονίστριά μου, Αλίνα, για τη μουσική πινελιά που συντρόφευε τις επιστημονικές μου αναζητήσεις.

Πέρα από τη βασική αφιέρωση που αναφέρεται σε προηγούμενη σελίδα, θα ήθελα να αφιερώσω τη διατριβή αυτή στους γονείς μου, Ηλία και Ντίνα, και στον αδελφό μου, Θοδωρή, για όλη την αγάπη που μου έχουν δείξει και οι οποίοι ανέμεναν με ανυπομονησία την ολοκλήρωση του πονήματος αυτού.

Αν και τελευταίο, το πιο ιδιαίτερο ευχαριστώ το κράτησα για τον Μιχάλη ο οποίος έδειξε αμέριστη συμπαράσταση, θετική διάθεση και αγάπη γεμίζοντάς με με αισιοδοξία και δύναμη κατά τη διάρκεια αυτής της πολυετούς προσπάθειας.

Η παρούσα διδακτορική διατριβή εκπονήθηκε μέσω χρηματοδότησης από τον Ειδικό Λογαριασμό Κονδυλίων Έρευνας (Ε.Λ.Κ.Ε.) του Εθνικού Μετσόβιου Πολυτεχνείου, εν είδει υποτροφίας, καθώς και μέσω μερικής χρηματοδότησης από το ερευνητικό πρόγραμμα Π.Ε.Β.Ε. 2010.

Περίληψη

Η παρούσα διδακτορική διατριβή αφορά τη διερεύνηση της Μπεϋζιανής επιλογής μετασχηματισμού σε στατιστικά μοντέλα και συνίσταται στα εξής βασικά σημεία.

(i) Το πρόβλημα της επιλογής μετασχηματισμού δεδομένων σε στατιστικά μοντέλα με σκοπό την κανονικότητα απασχολεί την ερευνητική κοινότητα εδώ και δεκαετίες καθώς συνδέεται άμεσα με τη διαδικασία βέλτιστης επιλογής μοντέλου και συνακόλουθα με την ανάδειξη μοντέλων υψηλότερης ακρίβειας καθώς και μοντέλων βελτιωμένης προβλεπτικής ικανότητας, μεταξύ άλλων. Η πιο γνωστή και ευρέως χρησιμοποιούμενη μονοπαραμετρική οικογένεια μετασχηματισμών (T) είναι η οικογένεια Box-Cox (Box & Cox 1964) με παράμετρο μετασχηματισμού λ_T, η οποία, για μια παρατήρηση y της τ.μ. Y, δίνεται από τον ακόλουθο τύπο:

$$y^{(\lambda_T)} = \begin{cases} \frac{y^{\lambda_T}-1}{\lambda_T}\,, & \lambda_T \neq 0 \\ \log(y)\,, & \lambda_T = 0 \end{cases} \qquad y > 0.$$

Με αφορμή την άνωθεν οικογένεια, διεξήχθει σε βάθος βιβλιογραφική ανασκόπηση όσον αφορά τους υπάρχοντες μετασχηματισμούς για τις τιμές των μεταβλητών σε Μπεϋζιανά αλλά και κλασικά στατιστικά μοντέλα και τις μεταξύ τους ομοιότητες και διαφορές, καθώς και τα προτερήματα και μειονεκτήματα αυτών στην πράξη. Στη συνέχεια, διερευνήθηκε το πρόβλημα επιλογής κατάλληλου μετασχηματισμού μονομεταβλητών δεδομένων, έστω $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, έτσι ώστε τα μετασχηματισμένα δεδομένα $\boldsymbol{y}^{(\lambda_T)} = \left(y_1^{(\lambda_T)}, \ldots, y_n^{(\lambda_T)}\right)^\top$ να προέρχονται από μια κανονική κατανομή με παραμέτρους (μ_T, σ_T^2) για συγκεκριμένη τιμή της παραμέτρου λ_T και για δεδομένη οικογένεια T. Το ενδιαφέρον εστιάστηκε στις κάτωθι μονοπαραμετρικές οικογένειες μετασχηματισμών: Box-Cox (BC), Modulus (Mod), Yeo-Johnson (YJ) και Dual. Το πρόβλημα μελετήθηκε υπό την

Μπεϋζιανή σκοπιά και ο παράγοντας Bayes (*Bayes factor*, BF) καθώς και οι εκ των υστέρων πιθανότητες των μοντέλων (*posterior model probabilities*, PMP) χρησιμοποιήθηκαν για την επιλογή, αξιολόγηση και σύγκριση των υπό μελέτη οικογενειών μετασχηματισμών. Σημειωτέον ότι η έννοια της επιλογής μετασχηματισμού συνίσταται κατά πρώτον στην επιλογή της βέλτιστης οικογένειας μετασχηματισμών T και κατά δεύτερον στην επιλογή της βέλτιστης τιμής της παραμέτρου λ_T δεδομένης της οικογένειας T από το πρώτο βήμα. Το κομμάτι της σύγκρισης και επιλογής μεταξύ οικογενειών μετασχηματισμών αποτελεί πρωτότυπη έρευνα και δεν απαντάται στη διεθνή βιβλιογραφία ως σήμερα.

Στα πλαίσια της Μπεϋζιανής μοντελοποίησης, έχει εξέχουσα σημασία η κατασκευή πρότερων κατανομών όσον αφορά τον δείκτη του μοντελοχώρου $T \in \mathcal{T}$ καθώς και το διάνυσμα παραμέτρων $\boldsymbol{\theta}_T = (\mu_T, \sigma_T^2, \lambda_T)^\top$ για κάθε οικογένεια T. Ο μοντελοχώρος του προβλήματος ορίζεται ως $\mathcal{T} = \{\text{Id, Log, BC, Mod, YJ, Dual}\}$ στον οποίο, πέρα από τις προαναφερθείσες παραμετρικές οικογένειες μετασχηματισμών, προστέθηκαν και ο Ταυτοτικός (*Identical, Id*) και ο Λογαριθμικός (Log) μετασχηματισμός. Επιλέχθηκε μια διακριτή ομοιόμορφη πρότερη κατανομή στον χώρο $\mathcal{T}$ που εκφράζει πρότερη άγνοια: $\pi(T) = |\mathcal{T}|^{-1}$. Η πρότερη κατανομή του διανύσματος παραμέτρων $\boldsymbol{\theta}_T$ είναι ιεραρχικά δομημένη:

$$\pi(\boldsymbol{\theta}_T|T) = \pi\left(\mu_T, \sigma_T^2 \middle| \lambda_T, T\right) \pi(\lambda_T|T),$$

ενώ για τις παραμέτρους (μ_T, σ_T^2) χρησιμοποιήθηκε η γνωστή συζυγής πρότερη κατανομή normal-inverse-gamma (NIG).

Για την παράμετρο μετασχηματισμού λ_T επιλέχτηκαν και εφαρμόστηκαν πρότερες κατανομές με βάση κάθε οικογένεια μετασχηματισμού οι οποίες βασίζονται στην έννοια της πρότερης κατανομής δύναμης (*power prior*) των Ibrahim & Chen (2000) και στη χρήση φανταστικών δεδομένων $\boldsymbol{y}^*$ προερχόμενων από μοντέλο που στηρίζει την αρχή της φειδωλότητας, με σκοπό την επίτευξη συμβατότητας των πρότερων κατανομών μεταξύ των διαφορετικών οικογενειών μετασχηματισμών. Η επίτευξη συμβατότητας κρίνεται ως επιτακτική λόγω της διαφορετικής ερμηνείας της παραμέτρου λ_T ανάλογα με την οικογένεια στην οποία αναφέρεται. Παράλληλα, κατασκευάστηκε μια κανονική πρότερη κατανομή μοναδιαίας πληροφορίας με μέσο $\mu_{\lambda_T} = 1$ και διασπορά $\sigma_{\lambda_T}^2$ βασιζόμενη στην παρατηρούμενη πληροφορία κατά Fisher εισάγοντας και εδώ τη χρήση φανταστικών δε-

δομένων $\boldsymbol{y}^*$.

Ως βέλτιστη οικογένεια $T \in \mathcal{T}$ για ένα σύνολο παρατηρήσεων $\boldsymbol{y}$ θεωρείται αυτή με την υψηλότερη εκ των υστέρων πιθανότητα:

$$\pi\left(T|\ \boldsymbol{y}, \boldsymbol{y}^*\right) = \frac{f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right)\pi(T)}{\sum_{T\in\mathcal{T}} f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right)\pi(T)}$$

όπου $f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right)$ είναι η περιθώρια πιθανοφάνεια (*marginal likelihood*) του $\boldsymbol{y}$ υπό την οικογένεια T. Προς διευκόλυνση, η περιθώρια πιθανοφάνεια μπορεί να γραφτεί ως εξής:

$$f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right) = \int f\left(\boldsymbol{y}|\lambda_T, T\right)\pi\left(\lambda_T|\boldsymbol{y}^*, T\right)\mathrm{d}\lambda_T$$

όπου $f\left(\boldsymbol{y}|\lambda_T, T\right)$ η πιθανοφάνεια του $\boldsymbol{y}$ υπό την οικογένεια T περιθωριοποιημένη ως προς μ_T και σ_T^2 και $\pi\left(\lambda_T|\boldsymbol{y}^*, T\right)$ η πρότερη κατανομή του λ_T υπό την T. Όσον αφορά στο πρόβλημα της εκτίμησης της περιθώριας πιθανοφάνειας των δεδομένων, εφαρμόστηκαν τρεις διαφορετικές μέθοδοι: ο εκτιμητής του Chib, η μέθοδος Laplace-Metropolis και μια προσεγγιστική αριθμητική μέθοδος βασισμένη στον κανόνα τετραγωνισμού των Gauss-Kronrod (*Gauss-Kronrod quadrature*).

Η σύγκριση της απόδοσης των προαναφερθέντων οικογενειών μετασχηματισμών έγινε για τυχαία προσομοιωμένα δείγματα από μια πληθώρα στατιστικών κατανομών, καθώς και για ένα σύνολο δεδομένων από το πεδίο του ελέγχου ποιότητας. Ιδιαίτερη έμφαση δόθηκε σε δεδομένα με υψηλό βαθμό ασυμμετρίας (όπως δεδομένα από μια γάμμα κατανομή), καθώς επίσης και σε δεδομένα προερχόμενα από κατανομή με παχιές ουρές (Student), καθώς τα τελευταία διακρίνονται στη βιβλιογραφία για τον υψηλό βαθμό δυσκολίας εύρεσης κατάλληλου μετασχηματισμού προς την κανονικότητα. Σύνολα δεδομένων με έντονη ασυμμετρία φαίνεται να αντιμετωπίζονται ικανοποιητικά μέσω του μετασχηματισμού Box-Cox, ενώ σημαντική μείωση του βαθμού ασυμμετρίας μιας κατανομής τονώνει σχετικά τον ρόλο του μετασχηματισμού YJ. Συμμετρικές κατανομές με παχιές ουρές (όπως οι κατανομές Student και διπλή εκθετική) συνδέονται με την οικογένεια Modulus. Τα παραγόμενα αποτελέσματα δείχνουν ότι η κυριαρχία του μοντέλου Box-Cox στη βιβλιογραφία μέχρι στιγμής δεν είναι πάντα ακριβής και η επιλογή από ένα ευρύτερο σύνολο οικογενειών μετασχηματισμών θα έπρεπε να γίνει κοινή πρακτική.

(ii) Η μεθοδολογία που αναπτύχθηκε στην Παράγραφο (i) μπορεί κάλλιστα να χρησιμοποιηθεί σαν ένας Μπεϋζιανός έλεγχος κανονικότητας, πολύ πιο ευέλικτος από τους

κλασικούς ελέγχους που απαντώνται στη βιβλιογραφία (π.χ. Shapiro-Wilk, Kolmogorov-Smirnov κ.α.). Πάνω σε αυτό, η βασική ιδέα συνοψίζεται στα εξής: ξεκινώντας από την παραδοχή ότι ένα διάνυσμα παρατηρήσεων $\boldsymbol{y}$ προέρχεται από μια τ.μ. $Y \sim N(\mu, \sigma^2)$, αναμένουμε η προαναφερθείσα μέθοδος να αναδείξει ως βέλτιστο μετασχηματισμό προς την κανονικότητα τον Ταυτοτικό. Με Μπεϋζιανούς όρους, αναμένεται το μοντέλο του Ταυτοτικού μετασχηματισμού να έχει πολύ υψηλότερη εκ των υστέρων πιθανότητα σε σχέση με τα υπόλοιπα πέντε μοντέλα υπό σύγκριση.

(iii) Επιπλέον, εξετάστηκε και το θέμα της επέκτασης της μεθοδολογίας που περιγράφηκε στην Παράγραφο (i) από μονομεταβλητά στατιστικά μοντέλα σε μοντέλα παλινδρόμησης με επεξηγηματικές μεταβλητές με στόχο την ταυτόχρονη αντιμετώπιση της επιλογής μετασχηματισμού δεδομένων της μεταβλητής απόκρισης και της επιλογής επεξηγηματικών μεταβλητών. Εδώ, το πλήρες μοντέλο θεωρήθηκε το κανονικό μοντέλο παλινδρόμησης $\boldsymbol{y}^{(\lambda_T)}|\mathbf{X}, \lambda_T, T \sim N_n\left(\boldsymbol{y}^{(\lambda_T)}|\mathbf{X}\boldsymbol{\beta}_T, \sigma_T^2\mathbf{I}_n\right)$ με p συμμεταβλητές και άγνωστη διασπορά σ_T^2, όπου $\mathbf{X}$ είναι ο $n \times (p+1)$ πίνακας σχεδιασμού, $\boldsymbol{\beta}_T$ είναι το διάνυσμα των άγνωστων συντελεστών και $\mathbf{I}_n$ είναι ο $n \times n$ ταυτοτικός πίνακας.

Λόγω ύπαρξης επεξηγηματικών μεταβλητών στο μοντέλο, η χρήση των φανταστικών δεδομένων $\boldsymbol{y}^*$ για την προσέγγιση της πρότερης κατανομής δύναμης για την παράμετρο λ_T περιπλέκεται, καθώς είναι δύσκολο να προσδιοριστεί πλήρως το μοντέλο αναφοράς από το οποίο θα γεννηθούν τα δεδομένα αυτά. Ειδικότερα, το μοντέλο αναφοράς θα μπορούσε να θεωρηθεί και πάλι το κανονικό μοντέλο, αλλά ο προσδιορισμός της μέσης τιμής και της διασποράς του μοντέλου αυτού είναι αρκετά δύσκολος. Στα μονομεταβλητά προβλήματα ο προσδιορισμός των εν λόγω παραμέτρων βασίστηκε στην τυποποίηση των δεδομένων, διαδικασία που περιπλέκεται παρουσία επεξηγηματικών μεταβλητών καθώς η ταυτόχρονη γέννηση τιμών $\boldsymbol{y}^*$ και $\mathbf{X}^*$ από το μοντέλο αναφοράς δεν είναι ξεκάθαρη. Εναλλακτικές πρότερες κατανομές χρειάζεται να αναζητηθούν για την παράμετρο λ_T οι οποίες να μη βασίζονται σε φανταστικά δεδομένα και να είναι "αντικειμενικές". Η χρήση μη γνήσιων (*improper*) κατανομών είναι συχνά απαραίτητη στο πρόβλημα της επιλογής μετασχηματισμού καθώς δε μπορεί εύκολα να υπάρξει υποκειμενική πρότερη πληροφορία για την παράμετρο μετασχηματισμού, ενώ μια πρότερη κατανομή μεγάλης διασποράς (*diffuse prior*) ενδεχομένως να ενεργοποιούσε το παράδοξο του Lindley. Όμως, ο κλα-

σικός παράγοντας Bayes δεν επιτρέπει εν γένει τη χρήση μη γνήσιων πρότερων κατανομών. Σαν λύση στο ανωτέρω ζήτημα, για τη σύγκριση μεταξύ των οικογενειών μετασχηματισμών με χρήση μη γνήσιων πρότερων κατανομών, επιστρατεύθηκαν εναλλακτικές μορφές παραγόντων Bayes, όπως ο ενδογενής παράγοντας Bayes (*intrinsic Bayes factor*, IBF) και ο κλασματικός παράγοντας Bayes (*fractional Bayes factor*, FBF). Στην ουσία, τα φανταστικά δεδομένα αντικαθίστανται από τα δεδομένα εκπαίδευσης του ενδογενούς παράγοντα Bayes και από την κλασματική παράμετρο εκπαίδευσης b του κλασματικού παράγοντα Bayes.

Σε αυτό το σημείο, μια δεύτερη βιβλιογραφική αναζήτηση ήταν αναγκαία ώστε να επιλεγούν οι εναλλακτικές μορφές παραγόντων Bayes οι οποίες θα ήταν οι πιο κατάλληλες για τη σύγκριση πολλαπλών μη εμφωλευμένων μοντέλων μετασχηματισμών. Εν τέλει, εφαρμόστηκαν ο διάμεσος ενδογενής παράγοντας Bayes και ο κλασματικός παράγοντας Bayes με μια χαμηλή και μια υψηλή τιμή για την κλασματική παράμετρο b.

Αναφορικά με τον μοντελοχώρο, ο οποίος πλέον ορίζεται από τις οικογένειες μετασχηματισμών αλλά και από τα δυνατά υποσύνολα επεξηγηματικών μεταβλητών, εξετάστηκαν διάφοροι συνδυασμοί πρότερων κατανομών, μεταξύ των οποίων η διακριτή ομοιόμορφη κατανομή και η βήτα-διωνυμική κατανομή. Ο συνδυασμός που επικράτησε, με βάση τα αποτελέσματα, είναι αυτός με την διακριτή ομοιόμορφη πρότερη πιθανότητα στον χώρο των μεταβλητών και τη διαφορετική πρότερη πιθανότητα στον χώρο των οικογενειών μετασχηματισμών με βάση το αν αυτές είναι παραμετρικές ή όχι.

Η μεθοδολογία που αναπτύχθηκε παρουσιάστηκε μέσα από τρεις εφαρμογές: ένα προσομοιωμένο σύνολο δεδομένων από ένα κανονικό μοντέλο πολλαπλής γραμμικής παλινδρόμησης, το γνωστό σύνολο δεδομένων Hald που συνδέεται με σοβαρά θέματα πολυσυγγραμικότητας και πολύ μικρό μέγεθος δείγματος και, τέλος, το πραγματικό σύνολο δεδομένων Highway με πλήθος επεξηγηματικών μεταβλητών.

Στο πρώτο παράδειγμα, όλες οι προσεγγίσεις ανέδειξαν πρώτο τον σωστό μετασχηματισμό, αλλά όσον αφορά την επιλογή μεταβλητών ο ενδογενής παράγοντας Bayes επέδειξε την καλύτερη συμπεριφορά ακολουθούμενος στενά από τον κλασματικό παράγοντα Bayes με τη χαμηλή τιμή της κλασματικής παραμέτρου b. Στα δεδομένα Hald, που είναι πιο ιδιόρρυθμα, η κατάσταση αντιστρέφεται με τον κλασματικό παράγοντα Bayes

με χαμηλή τιμή της κλασματικής παραμέτρου b να έχει σχετική υπεροχή, αν και είναι σημαντικό ότι πολυσυγγραμικά μοντέλα δεν προτάθηκαν ως βέλτιστα από καμιά προσέγγιση. Στο παράδειγμα των δεδομένων Highway με το αυξημένο πλήθος επεξηγηματικών μεταβλητών, όλες οι προσεγγίσεις αναδεικνύουν την ίδια οικογένεια μετασχηματισμών σαν βέλτιστη, αν και όχι απαραίτητα το ίδιο σύνολο επεξηγηματικών μεταβλητών. Σε καμία εκ των τριών εφαρμογών δεν προτιμήθηκε η υψηλή τιμή που δοκιμάστηκε για την κλασματική παράμετρο b, καθώς ο αντίστοιχος κλασματικός παράγοντας Bayes τείνει να αναδεικνύει κάθε άλλο παρά φειδωλά μοντέλα και συνεπώς δεν μπορεί να διαχειριστεί σωστά ούτε το πρόβλημα της πολυσυγγραμικότητας.

Συμπερασματικά, η καταλληλότητα και η αξιοπιστία των διαθέσιμων εναλλακτικών παραγόντων Bayes διαφέρουν ανάλογα με το εκάστοτε ερευνητικό πλαίσιο. Ο αριθμητικός ενδογενής παράγοντας Bayes είναι συχνά υπολογιστικά ασύμφορος, ενώ είναι ακατάλληλος όταν έχουμε πολλαπλά μη εμφωλευμένα μοντέλα υπό σύγκριση, όπως στην περίπτωση του προβλήματος επιλογής μετασχηματισμού. Ο γεωμετρικός ενδογενής παράγοντας Bayes είναι σημαντικά βελτιωμένος σε σχέση με τον αριθμητικό στην περίπτωση πολλών ακραίων τιμών με βάση τα δείγματα εκπαίδευσης και ο διάμεσος ενδογενής παράγοντας Bayes δείχνει να είναι ο πιο ευέλικτος σε περίπλοκες περιπτώσεις υπό σύγκριση μοντέλων. Όλες οι ανωτέρω μορφές αντιμετωπίζουν προβλήματα, παρόλα αυτά, λόγω του τρόπου επιλογής των δειγμάτων εκπαίδευσης και, συνακόλουθα, όταν συμπεριλαμβάνονται ποιοτικές συμμεταβλητές στα μοντέλα. Ο κλασματικός παράγοντας Bayes έχει σημαντικά μειωμένους χρόνους επεξεργασίας και διεκπεραίωσης ακόμα και όταν συμπεριλαμβάνει σημαντικό πλήθος επεξηγηματικών μεταβλητών, ενώ υπερτερεί σε σχέση με τον ενδογενή παράγοντα Bayes και στο ότι δε χρησιμοποιεί δείγματα εκπαίδευσης. Εντούτοις, η επιλογή της βέλτιστης τιμής της εμπλεκόμενης κλασματικής παραμέτρου b απαιτεί διερεύνηση σε κάθε περίπτωση.

Λέξεις - κλειδιά: Αλγόριθμοι MCMC, Πρότερη κατανομή δύναμης, Ενδογενής παράγοντας Bayes, Εκ των υστέρων πιθανότητες μοντέλων, Επιλογή μεταβλητών, Επιλογή οικογένειας μετασχηματισμών, Κλασματικός παράγοντας Bayes, Μπεϋζιανή επιλογή μοντέλου, Πρότερη κατανομή μοναδιαίας πληροφορίας, Συμβατότητα πρότερων κατανομών, Φανταστικά δεδομένα.

Summary

The present thesis delves into the investigation of Bayesian transformation selection in statistical models and consists of the following key points.

(i) The problem of transformation selection in statistical models aiming for normality attracts the interest of the research community for decades, since it is directly linked to the general process of model selection and hence to the demarcation of higher accuracy models and improved predictive power models, among others. The most known and widely used uniparametric family of transformations (T) is the Box-Cox family (Box & Cox 1964) with a transformation parameter λ_T, which, for an observation y of the random variable (r.v.) Y, is given by the following formula:

$$y^{(\lambda_T)} = \begin{cases} \frac{y^{\lambda_T}-1}{\lambda_T}\,, & \lambda_T \neq 0 \\ \log(y)\,, & \lambda_T = 0 \end{cases} \qquad y > 0.$$

Considering the above family as a starting point, an in-depth literature review was conducted regarding existing variable transformations for Bayesian and frequentist statistical models, exploring the similarities and differences among them as well as the advantages and disadvantages of these transformations in practice. The next step was the investigation of the problem of selecting a suitable transformation in the case of univariate data $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, so that the transformed data $\boldsymbol{y}^{(\lambda_T)} = \left(y_1^{(\lambda_T)}, \ldots, y_n^{(\lambda_T)}\right)^\top$ approximately follow a normal distribution with parameters (μ_T, σ_T^2) for a specific value of the parameter λ_T and for a given family T. Focus was on the following uniparametric transformation families: Box-Cox (BC), Modulus (Mod), Yeo-Johnson (YJ) and Dual. The problem was treated under a Bayesian perspective, with the Bayes factor (BF) and the posterior model probabilities (PMP) being employed for the selection, evaluation and comparison of the

transformation families under study. Note that the concept of transformation selection consists firstly in selecting the optimal family of transformations T and then in selecting the optimal value of the parameter λ_T given the family T from the first step. The part of the comparison and selection among various transformation families constitutes original research since it is not found in the literature to date.

Within the context of Bayesian modeling, it is of utmost importance to construct prior distributions regarding the model space index $T \in \mathcal{T}$ as well as the parameter vector $\boldsymbol{\theta}_T = (\mu_T, \sigma_T^2, \lambda_T)^\top$ for every transformation family T. The model space of the problem is defined as $\mathcal{T} = \{\text{Id, Log, BC, Mod, YJ, Dual}\}$ which, apart from the aforementioned parametric transformation families, includes also the Identity (Id) and the Logarithmic (Log) transformations. A discrete uniform prior has been chosen in the space $\mathcal{T}$ expressing a priori ignorance about the optimal family anticipated: $\pi(T) = |\mathcal{T}|^{-1}$. The prior distribution of the parameter vector $\boldsymbol{\theta}_T$ is hierarchically structured:

$$\pi(\boldsymbol{\theta}_T|T) = \pi\left(\mu_T, \sigma_T^2 \middle| \lambda_T, T\right) \pi(\lambda_T|T),$$

while for the parameters (μ_T, σ_T^2) we used the well known normal-inverse-gamma (NIG) conjugate prior.

For the transformation parameter λ_T, prior distributions for each transformation family T have been developed. These priors are based on the concept of the power-prior distribution of Ibrahim & Chen (2000) and on the use of imaginary data $\boldsymbol{y}^*$ derived by a reference model that supports the parsimony principle, in order to achieve compatibility of the prior distributions among the different families of transformations. Achieving compatibility is considered urgent because of the different interpretation of the parameter λ_T depending on the family T in question. Furthermore, a normal prior distribution of unit-information has been employed with mean $\mu_{\lambda_T} = 1$ and variance $\sigma_{\lambda_T}^2$ based on the observed Fisher information, introducing here as well the device of imaginary data $\boldsymbol{y}^*$.

The optimal family $T \in \mathcal{T}$ for a set of observations $\boldsymbol{y}$ is considered the one with the highest posterior probability:

$$\pi\left(T|\, \boldsymbol{y}, \boldsymbol{y}^*\right) = \frac{f\left(\boldsymbol{y}|\, \boldsymbol{y}^*, T\right) \pi(T)}{\sum_{T \in \mathcal{T}} f\left(\boldsymbol{y}|\, \boldsymbol{y}^*, T\right) \pi(T)}$$

where $f\left(\boldsymbol{y}|\, \boldsymbol{y}^*, T\right)$ is the marginal likelihood of $\boldsymbol{y}$ under the family T. For convenience,

the marginal likelihood can be expressed as:

$$f\left(\boldsymbol{y}|\ \boldsymbol{y}^{*}, T\right)=\int f\left(\boldsymbol{y}|\lambda_{T}, T\right)\pi\left(\lambda_{T}|\boldsymbol{y}^{*}, T\right)\mathrm{d}\lambda_{T}$$

where $f\left(\boldsymbol{y}|\lambda_{T}, T\right)$ is the likelihood of $\boldsymbol{y}$ under the family T marginalized as to μ_{T} and σ_{T}^{2} and $\pi\left(\lambda_{T}|\boldsymbol{y}^{*}, T\right)$ is the prior distribution of λ_{T} under T. In terms of the estimation of the marginal likelihood of the data, three different approaches were employed: Chib's estimator, the Laplace-Metropolis method, as well as a numerical approximation method based on the Gauss-Kronrod quadrature.

The comparison of the above transformations families in terms of performance was carried out for randomly generated samples from a variety of statistical distributions as well as for a dataset from the quality control area. Particular emphasis was given to data with a high degree of skewness (such as gamma datasets) as well as data coming from distributions with fat tails (such as the Student ditribution), since fat tails are often mentioned in the literature as being associated with a high difficulty of finding a suitable transformation towards normality. Highly skewed data are sufficiently treated by the Box-Cox family, whereas considerable drops in the density skewness result in boosting to some extent the role of the YJ family in transforming the data. Heavy-tailed symmetric distributions (such as the student and the double exponential) are associated with the Modulus family. Overall, empirical evidence entails that the predominance of the Box-Cox transformation in the literature so far is not always accurate and the selection from a wider set of transformations should become common practice.

(ii) The methodology developed in Paragraph (i) may well be used as a Bayesian normality test much more flexible than the frequentist tests encountered in the literature (e.g. Shapiro-Wilk, Kolmogorov-Smirnov etc). Elaborating a bit more on this, the basic idea is summarized as follows: starting from the assumption that a vector of observations $\boldsymbol{y}$ comes from a r.v. $Y\sim N(\mu, \sigma^{2})$, we anticipate the aforementioned method to highlight the Identity transformation as the best transformation towards normality of $\boldsymbol{y}$. In Bayesian terms, it is expected that the model of the Identity transformation has a much higher posterior probability in relation to the remaining five models under comparison.

(iii) What is more, the methodology described in Paragraph (i) was expanded to include not only univariate models but also regression models with explanatory variables, aiming

to treat transformation selection of the response variable as well as variable selection. In this context, the full model was taken to be the normal regression model $\boldsymbol{y}^{(\lambda_T)}|\mathbf{X}, \lambda_T, T \sim N_n\left(\boldsymbol{y}^{(\lambda_T)}|\mathbf{X}\boldsymbol{\beta}_T, \sigma_T^2\mathbf{I}_n\right)$ with p covariates and unknown variance σ_T^2, where $\mathbf{X}$ is the $n \times (p+1)$ design matrix, $\boldsymbol{\beta}_T$ is the vector of unknown regression coefficients and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Due to the presence of explanatory variables, defining the model from which to generate the imaginary data $\boldsymbol{y}^*$ required by the power prior approach for the parameter λ_T gets very complicated. This model could again be assumed normal but it is hard to choose values for the respective mean and variance. In univariate problems, the latter parameters were easily chosen after standardizing the data, but here the simultaneous generation of values for $\boldsymbol{y}^*$ and $\mathbf{X}^*$ is quite unclear. So, it is crucial to search for alternative priors for λ_T which would not rely on imaginary data and would be objective. The use of improper priors is often indispensable when it comes to transformation selection since no subjective prior information is easily available on the transformation parameter, whereas a diffuse prior could potentially activate Lindley's paradox. However, the Bayes factor does not allow the use of improper priors. As a solution, for the comparison among the various transformation families using improper priors, we employed alternative Bayes factor forms, such as the intrinsic Bayes factor (IBF) and the fractional Bayes factor (FBF). In fact, the imaginary data are replaced by the training data in the intrinsic Bayes factor case and by the fractional parameter b in the fractional Bayes factor case.

At this point, a second literature review was necessary so as to choose alternative Bayes factor forms that would be most proper in order to compare multiple non-nested model structures. Finally, the median intrinsic Bayes factor and the fractional Bayes factor were utilized, the latter corresponding to a low and a high value for the fractional parameter b.

Regarding the model space, which has also expanded to include all transformation families under consideration combined with all the potential variable subsets, various combinations of priors were investigated, including the discrete uniform prior and the beta-binomial prior. The best combination seems to be the one with the discrete uniform prior on the variable space and the family prior that differentiates itself based on the parametric or the non parametric nature of the families.

The methodology developed was illustrated via three applications: a simulated dataset from a normal regression model, the known Hald dataset linked with multicollinearity issues and very small sample size, and the Highway dataset with a considerable number of explanatory variables.

In the first of the three examples, all the approaches selected the correct transformation as optimal, but regarding the variable selection the intrinsic Bayes factor showed the best behaviour followed closely by the fractional Bayes factor with the low value for the fractional parameter b. In the - more peculiar - Hald data, the situation was reversed with the low-b fractional Bayes factor showing better results, although it is important to note that models suffering with multicollinearity were not demarcated as optimal by the IBF either. In the Highway example, where a larger set of potential covariates are considered, all approaches gave the same transformation family as optimal, even though not with the same set of covariates. In none of the three illustrative examples was the high value for b preferred, since the associated FBF did not lead to the selection of parsimonious models nor multicollinearity-free models.

In conclusion, the suitability and the reliability of the available alternative Bayes factors differ depending on the research context. The arithmetic form of the intrinsic Bayes factor is often computationally inefficient and is certainly unsuitable when dealing with multiple non nested models, as in the transformation selection problem. The geometric form of the intrinsic Bayes factor is much improved compared to the arithmetic form in the case of outlying training samples. The median intrinsic Bayes factor proves to be the most efficient form when it comes to complex cases of comparing models. However, all the above forms suffer a number of issues, especially when having to appropriately select the training samples and also when having to include qualitative explanatory variables in the models. The fractional Bayes factor is associated with much lower processing times even when the number of the explanatory variables is considerable, while it also outcompetes the intrinsic Bayes factor in that it does not use training samples for its calculation. Nevertheless, selecting the optimal value of the included fractional parameter b is not straighforward and requires a new investigation for every research task.

Key words: Bayesian model selection, Fractional Bayes factor, Imaginary data, Intrinsic

Bayes factor, MCMC, Posterior model probabilities, Power prior, Prior compatibility, Transformation family selection, Unit-information prior, Variable selection.

Δημοσιεύσεις και Ανακοινώσεις σε Συνέδρια

Τα αποτελέσματα της παρούσας διατριβής έχουν δημοσιευθεί σε διεθνή επιστημονικά περιοδικά με κριτές καθώς και σε πρακτικά συνεδρίων μετά από κρίση. Επιπλέον, η μεθοδολογία που αναπτύχθηκε έχει παρουσιαστεί σε 7 συνέδρια μέχρι στιγμής, τρία εκ των οποίων έλαβαν χώρα εντός Ελλάδος (Καλαμάτα, Πειραιάς, Αθήνα) και τα υπόλοιπα τέσσερα διεξήχθησαν σε άλλες Ευρωπαϊκές χώρες (Ηνωμένο Βασίλειο, Ιταλία, Ισπανία).

Δημοσιεύσεις μετά από κρίση

• Charitidou, E., Fouskakis, D. & Ntzoufras, I. (2015), 'Bayesian transformation family selection: moving towards a transformed Gaussian universe', *The Canadian Journal of Statistics*, **43**, 600-623.

• Charitidou, E., Fouskakis, D. & Ntzoufras, I. (2014), 'On Bayesian transformation selection: problem formulation and preliminary results', in Lanzarone, E. & Ieva, F. (eds.), *The Contribution of Young Researchers to Bayesian Statistics*, Springer Proceedings in Mathematics & Statistics, **63**, Switzerland: Springer International Publishing, pp. 11-14.

• Charitidou, E., Fouskakis, D. & Ntzoufras, I. (2013), 'On Bayesian transformation selection: problem formulation and preliminary results', *Πρακτικά του 26ου Πανελληνίου Συνεδρίου Στατιστικής*, Αθήνα: Ε.Σ.Ι., σελ. 253-260.

Άρθρα υποβληθέντα σε περιοδικά με κριτές

• Charitidou, E., Fouskakis, D. & Ntzoufras, I. (2016), 'Objective Bayesian transformation and variable selection using default Bayes factors'.

Άρθρα υπό προετοιμασία

• Charitidou, E., Fouskakis, D. & Ntzoufras, I. (2016), 'A Bayesian normality test based on transformation selection: an alternative to classical approaches'.

Ανακοινώσεις σε συνέδρια

• Charitidou E., Fouskakis D. & Ntzoufras I. (2016). 'Bayesian Transformation and Variable Selection Using Different Forms of Bayes Factors'. ISBA 2016, 13-17 Ιουνίου, Σαρδηνία, Ιταλία.

• Charitidou E., Fouskakis D. & Ntzoufras I. (2015). 'On Bayesian Transformation Selection: Comparison based on different forms of Bayes factors'. 3rd Meeting on Statistics, 24-26 Ιουνίου, Αθήνα.

• Charitidou E., Fouskakis D. & Ntzoufras I. (2015). 'On Bayesian Transformation Selection: comparison of results based on different forms of Bayes factors'. 11^{th} International Workshop on Objective Bayes Methodology (O-Bayes15), 1-5 Ιουνίου, Βαλένθια, Ισπανία.

• Charitidou E., Fouskakis D. & Ntzoufras I. (2013). 'On Bayesian Transformation Selection: problem formulation and preliminary results'. Bayesian Young Statisticians Meeting (BAYSM-2013), 5-6 Ιουνίου, Μιλάνο, Ιταλία.

• Charitidou E., Fouskakis D. & Ntzoufras I. (2013). 'Bayesian Transformation Selection: theoretical approach and applications'. 26 Πανελλήνιο Συνέδριο Στατιστικής, Ελληνικό Στατιστικό Ινστιτούτο (Ε.Σ.Ι.-2013), 8-11 Μαΐου, Πανεπιστήμιο Πειραιώς, Πειραιάς.

• Charitidou E., Fouskakis D. & Ntzoufras I. (2013). 'On Bayesian Transformation Selection: problem formulation and preliminary results'. 36^{th} Research Students Conference in Probability, Statistics and Social Statistics (RSC-2013), 25-28 Μαρτίου, Λάνκαστερ, Ηνωμένο Βασίλειο.

• Charitidou E., Fouskakis D. & Ntzoufras I. (2012). 'On Bayesian Transformation Selection in Univariate Problems'. Greek Stochastics δ meeting, 25-28 Αυγούστου, Καλαμάτα.

Περιεχόμενα

Κατάλογος Διαγραμμάτων

Κατάλογος Πινάκων

Κατάλογος Συντμήσεων

AIBF	:	Arithmetic intrinsic Bayes factor
ANOVA	:	Analysis of variance
BF	:	Bayes factor
BIC	:	Bayes information criterion
BMA	:	Bayesian model averaging
BC	:	Box-Cox
df	:	degrees of freedom
FBF	:	Fractional Bayes factor
FIP	:	Fractional intrinsic prior
GIBF	:	Geometric intrinsic Bayes factor
IBF	:	Intrinsic Bayes factor
i.i.d.	:	Independent and identically distributed
KL	:	Kullback-Leibler
MH	:	Metropolis-Hastings
MIBF	:	Median intrinsic Bayes factor
Mod	:	Modulus
MCMC	:	Markov chain Monte Carlo
MLE	:	Maximum likelihood estimation
PBF	:	Posterior Bayes factor
PMP	:	Posterior model probability
RMIBF	:	Ratio of medians intrinsic Bayes factor
RSS	:	Royal Statistical Society
SBC	:	Schwarz Bayesian criterion
τ.μ.	:	Τυχαία μεταβλητή
YJ	:	Yeo-Johnson

Εισαγωγή

Βασικός στόχος της παρούσας διδακτορικής διατριβής είναι να εντρυφήσει στην επιλογή συγκεκριμένων δομικών χαρακτηριστικών των στατιστικών μοντέλων. Πολύ συχνά, η επιλογή μοντέλου είναι άμεσα συνυφασμένη με την επιλογή του κατάλληλου μετασχηματισμού αναφορικά με τις τιμές των μεταβλητών που εμπλέκονται σε αυτό έχοντας συνήθως ως γνώμονα θεωρητικές προϋποθέσεις, κριτήρια καλής προσαρμογής ή και δείκτες προβλεπτικής ικανότητας του μοντέλου.

Συγκεκριμένα, μια θεμελιώδης προϋπόθεση του γραμμικού μοντέλου είναι η κανονικότητα σχετικά με τα σφάλματα του μοντέλου. Επιπλέον, στα πλαίσια της Μπεϋζιανής ανάλυσης, η κανονικότητα επιτρέπει τη χρήση συζυγών μορφών πρότερων κατανομών διευκολύνοντας σημαντικά την υπολογιστική διαδικασία. Στη βάση της Μπεϋζιανής επιλογής μετασχηματισμού δεδομένων, θεωρήσαμε τέσσερις μονοπαραμετρικές οικογένειες μετασχηματισμών. Οι οικογένειες αυτές είναι οι εξής: Box-Cox, Modulus, Yeo & Johnson και Dual.

Σε πρώτη φάση, το πρόβλημα της επιλογής οικογένειας μετασχηματισμού εξετάστηκε πλήρως από Μπεϋζιανή σκοπιά σε μονομεταβλητά προβλήματα. Οι ανωτέρω οικογένειες μετασχηματισμών λαμβάνονται υπόψη με σκοπό να προσεγγίσουμε την κανονικότητα όσον αφορά την κατανομή ενός συνόλου δεδομένων. Αλγόριθμοι Markov Chain Monte Carlo (MCMC) κατασκευάστηκαν ώστε να προσομοιώσουμε δείγμα από την εκ των υστέρων κατανομή της παραμέτρου μετασχηματισμού λ_T η οποία συνδέεται με κάθε οικογένεια μετασχηματισμού T. Διερευνήσαμε διαφορετικές προσεγγίσεις με στόχο την κατασκευή συμβατών πρότερων κατανομών για την παράμετρο λ_T μεταξύ των οικογενειών, κάνοντας χρήση της πρότερης κατανομής δύναμης και εναλλακτικά μιας κανονικής εκ των προτέρων κατανομής μοναδιαίας πληροφορίας. Η επιλογή και ανάδειξη

της βέλτιστης οικογένειας βασίζεται στον υπολογισμό των εκ των υστέρων πιθανοτήτων των οικογενειών-μοντέλων, ενώ παράλληλα ο βαθμός καλής προσαρμογής των μοντέλων εκτιμάται μέσα από μέτρα σχετικά με τη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) στη βάση των εκ των υστέρων τεταρτημορίων. Χρησιμοποιώντας προσομοιωμένα δεδομένα, καθώς και ένα παράδειγμα από τον χώρο του ελέγχου ποιότητας στη βιομηχανία, αναδεικνύεται η αποτελεσματικότητα της μεθοδολογίας.

Παρά το γεγονός ότι δεν αναδείχθηκε μία οικογένεια η οποία να προσφέρει καθολική λύση, όπως άλλωστε αναμενόταν, η έρευνα έδειξε ότι σύνολα δεδομένων με συγκεκριμένα χαρακτηριστικά φαίνεται να ωφελούνται από συγκεκριμένες οικογένειες. Ασύμμετρες κατανομές σχετίζονται με τις οικογένειες Box-Cox και Dual, ενώ κατανομές με παχιές ουρές ωφελούνται περισσότερο από την οικογένεια μετασχηματισμών Modulus.

Σε δεύτερη φάση, η σχετική μεθοδολογία επεκτάθηκε σε προβλήματα πολλών επεξηγηματικών μεταβλητών ενσωματώνοντας και τη Μπεϋζιανή επιλογή μεταβλητών καθώς αποτελεί αναπόσπαστο κομμάτι της αντιμετώπισης και διαχείρισης κάθε πραγματικού προβλήματος με πλήθος επεξηγηματικών μεταβλητών. Η πρότερη κατανομή δύναμης, με τη συνακόλουθη χρήση φανταστικών δεδομένων, δε μπορεί εύκολα να επεκταθεί στο νέο σύνολο προβλημάτων υπό εξέταση. Καταφύγαμε, λοιπόν, στη χρήση εναλλακτικών μορφών του παράγοντα Bayes και συγκεκριμένα στη χρήση του ενδογενούς παράγοντα Bayes και του κλασματικού παράγοντα Bayes.

Η εφαρμογή της διευρυμένης μεθόδου γίνεται σε δύο γνωστά σύνολα δεδομένων (σύνολα Hald και Highway) που έχουν χρησιμοποιηθεί σε ποικίλα άρθρα με θεματολογία σχετική της Μπεϋζιανής επιλογής μοντέλου και μεταβλητών. Το πρώτο σύνολο δεδομένων έχει την ιδιαιτερότητα ότι αποτελείται από ζεύγη μεταβλητών που μοιράζονται πολύ ισχυρές συσχετίσεις, ενεργοποιώντας θέματα πολυσυγγραμικότητας, ενώ παράλληλα αποτελεί ένα πρόβλημα ελαχίστου μεγέθους δείγματος που σχεδόν αγγίζει τα όρια ενός 'small n - large p' προβλήματος. Το δεύτερο σύνολο δεδομένων είναι μέτριου μεγέθους με αρκετές επεξηγηματικές μεταβλητές, στο οποίο η ορθή επιλογή μεταβλητών είναι πρωτεύουσας σημασίας. Μέσα από τα σχετικά αποτελέσματα, γίνεται άμεση σύγκριση των διαφόρων προσεγγίσεων Μπεϋζιανής επιλογής μοντέλου που εφαρμόζονται και αναδεικνύονται διάφορες όχι ευρέως γνωστές δυσκολίες στην εφαρμογή του ενδογε-

νούς παράγοντα Bayes σε σχέση με τον κλασματικό παράγοντα Bayes.

Σημειώνεται ότι η εφαρμογή όλων των παραπάνω μεθόδων βασίστηκε στην κατασκευή κατάλληλων αλγορίθμων με χρήση του ανοιχτού (*open-source*) στατιστικού λογισμικού *R software for statistical computing*, version 3.1.2 (*http://www.R-project.org/*).

Ακολούθως, περιγράφεται συνοπτικά η δομή της παρούσας διατριβής. Το Κεφάλαιο 1 αποτελεί μια εισαγωγή στις οικογένειες μετασχηματισμών με βάση τη διεθνή βιβλιογραφία, δίνοντας ιδιαίτερη έμφαση στις οικογένειες που μας ενδιαφέρουν στη διατριβή αυτή (Box-Cox, Modulus, Yeo-Johnson και Dual) αποκαλύπτοντας ομοιότητες και διαφορές μεταξύ τους. Στο Κεφάλαιο 2 παρουσιάζεται η προτεινόμενη προσέγγιση Μπεϋζιανής επιλογής μετασχηματισμού και σχετικής συμπερασματολογίας. Εδώ εντάσσεται η επιλογή πρότερων κατανομών και περιλαμβάνει δυο εναλλακτικές προσεγγίσεις. Σχετικές εφαρμογές εκτίθενται στο Κεφάλαιο 3 με βάση μονομεταβλητά σύνολα προσομοιωμένων δεδομένων επιλεγμένα ειδικά με στόχο την ανάδειξη της μεθοδολογίας του Κεφαλαίου 2. Στο Κεφάλαιο 4 περιγράφεται η επέκταση της μεθοδολογίας από τα μονομεταβλητά προβλήματα σε μοντέλα που περιέχουν και πλήθος επεξηγηματικών μεταβλητών εισάγοντας και την έννοια της επιλογής μεταβλητών, αναγνωρίζοντας ταυτόχρονα σχετικά εμπόδια που προκύπτουν και προτείνοντας νέες λύσεις προς αντιμετώπιση αυτών. Ακολουθεί το Κεφάλαιο 5 όπου χρησιμοποιούνται γνωστά σύνολα δεδομένων από τη βιβλιογραφία, αλλά και προσομοιωμένα σύνολα δεδομένων, ώστε να ελεγχθεί η αποδοτικότητα και η εφαρμοσιμότητα της προτεινόμενης μεθοδολογίας παρουσία πολλών επεξηγηματικών μεταβλητών. Τέλος, το Κεφάλαιο 6 περιέχει την τελική συζήτηση πάνω στα αποτελέσματα και τη συνεισφορά της παρούσας διατριβής, κάνοντας ειδική αναφορά και σε πιθανές μελλοντικές επεκτάσεις της έρευνας πάνω στο θέμα.

Κεφάλαιο 1

Οικογένειες Μετασχηματισμών

1.1 Εισαγωγή

Το παρόν κεφάλαιο διερευνά τη βιβλιογραφία σχετικά με τους μετασχηματισμούς που μπορούν να εφαρμοστούν στις τιμές εξαρτημένων ή/και ανεξάρτητων μεταβλητών στα πλαίσια γραμμικών στατιστικών μοντέλων.

Η προσέγγιση στην οποία κυρίως εστιάζουμε σε αυτή την σύντομη βιβλιογραφική έρευνα που ακολουθεί είναι Μπεϋζιανή, αν και όχι κατά αποκλειστικότητα: στις αντίστοιχες περιπτώσεις που αναφέρονται παρακάτω, η διαμόρφωση των εκ των προτέρων κατανομών για τις άγνωστες παραμέτρους των μοντέλων είναι προαπαιτούμενο για την μετέπειτα εκτίμηση a posteriori κατανομών. Εκτός των Μπεϋζιανών προσεγγίσεων, αναφέρονται και επιλεγμένα άρθρα που βασίζονται μεν στην κλασική στατιστική, αλλά συγκεντρώνουν αυξημένο ενδιαφέρον στο θέμα του μετασχηματισμού δεδομένων. Τα συγκεκριμένα άρθρα μπορούν να αποτελέσουν ένα σημείο εκκίνησης για μελλοντική έρευνα, ακόμη και με επέκταση στην περιοχή της Μπεϋζιανής συλλογιστικής.

Γενικά, ένας μετασχηματισμός των τιμών μιας μεταβλητής μπορεί να έχει ως στόχο τη γραμμικοποίηση μη γραμμικών μοτίβων σε ένα μοντέλο (π.χ. ώστε να ισχύει η προϋπόθεση γραμμικότητας σε μοντέλα γραμμικής παλινδρόμησης) ή τη διόρθωση της μη κανονικότητας των σφαλμάτων ενός μοντέλου το οποίο συνεπάγεται διόρθωση της δομής των σφαλμάτων. Πριν τη δεκαετία του 1960, η έρευνα στο συγκεκριμένο πεδίο αφορούσε σχεδόν αποκλειστικά μετασχηματισμούς των τιμών της μεταβλητής απόκρισης.

Κατά τη διάρκεια των τελευταίων πέντε περίπου δεκαετιών, το ενδιαφέρον έχει διευρυνθεί προς μετασχηματισμούς που αφορούν και τις τιμές των επεξηγηματικών μεταβλητών του μοντέλου εκτός από τη μεταβλητή απόκρισης. Έτσι, για παράδειγμα, μπορεί κάποιος να εστιάσει στην εύρεση ενός κατάλληλου συνόλου μετασχηματισμών που να οδηγεί σε ένα απλούστερο και φειδωλό μοντέλο ή ενός συνόλου μετασχηματισμών που να βελτιώνει τη συμπερασματολογία ή/και την προβλεπτική ικανότητα ενός μοντέλου δίνοντας πιο εύρωστους (*robust*) εκτιμητές των βασικών παραμέτρων ενδιαφέροντος.

Σε αυτό το σημείο, είναι σημαντικό να κάνουμε μια σύντομη παρένθεση για τη διάκριση μεταξύ των εννοιών της συμπερασματολογίας και της πρόβλεψης. Η πρώτη εστιάζει στη διερεύνηση ενός φαινομένου μέσα από την ανάδειξη ενός συνόλου επεξηγηματικών μεταβλητών που μπορούν να εξηγήσουν λογικά το υπό μελέτη φαινόμενο, ενώ η δεύτερη σχετίζεται με την κατασκευή μιας εξίσωσης που να μπορεί να προβλέπει με ακρίβεια ένα συγκεκριμένο χαρακτηριστικό (ή σύνολο χαρακτηριστικών) ενός πληθυσμού (το λεγόμενο *out-of-sample performance* της προβλεπτικής εξίσωσης) χωρίς να ασχολείται με το αν οι μεταβλητές της εξίσωσης έχουν οποιαδήποτε εννοιολογική σημασία ή σχέση με το υπό μελέτη φαινόμενο.

Όταν η ερμηνεία συγκαταλέγεται ανάμεσα στους βασικούς στόχους μιας έρευνας, η μετατροπή των τελικών αποτελεσμάτων (π.χ. των εκτιμητών των συντελεστών) πίσω στην αρχική κλίμακα των δεδομένων, δηλαδή στην κλίμακα πριν από την εφαρμογή οποιωνδήποτε μετασχηματισμών, είναι επιβεβλημένη. Στην ουσία, πρόκειται για μετατροπή σε άλλο μετρικό σύστημα, κάτι που σίγουρα ενέχει συστηματικό σφάλμα. Με το συγκεκριμένο ζήτημα έχουν ασχοληθεί, μεταξύ άλλων, οι Taylor (1985) και Stow, Reckhow & Qian (2006). Μια πρόσφατη εναλλακτική, μέσω των λεγόμενων γενικευμένων συντελεστών παλινδρόμησης, προτείνεται στο άρθρο των Gottardo & Raftery (2007) του οποίου λεπτομέρειες παρουσιάζονται αργότερα στο παρόν κεφάλαιο.

Είναι εμφανές ότι το μέγεθος του συνόλου μετασχηματισμών αντιστοιχεί στο πλήθος των επιπρόσθετων παραμέτρων υπό εκτίμηση. Αν το μέγεθος αυτό είναι μεγαλύτερο της μονάδας, τότε οι τιμές περισσότερων της μίας μεταβλητών προορίζονται για μετασχηματισμό. Όπως ήδη αναφέρθηκε, είναι στην ευχέρεια του ερευνητή αν θα μετασχηματίσει μόνο τις τιμές της μεταβλητής απόκρισης, μόνο κάποιων ή όλων των επεξηγηματικών

μεταβλητών ή και τα δύο.

Ας σημειωθεί ότι ο χώρος $\mathcal{M}_\lambda$ της παραμέτρου μετασχηματισμού μπορεί να είναι είτε διακριτός είτε συνεχής. Η πρώτη περίπτωση έχει το πλεονέκτημα ότι ο χώρος είναι λιγότερο πολύπλοκος (οδηγώντας σε χαμηλότερο υπολογιστικό βάρος) και οι επιλογές της παραμέτρου είναι ευκολότερα ερμηνεύσιμες, ενώ στη δεύτερη περίπτωση έχουμε συνήθως πιο ακριβείς και ρεαλιστικές επιλογές τιμών της παραμέτρου μετασχηματισμού. Μια τελευταία παρατήρηση είναι ότι οι όροι *οικογένεια μετασχηματισμών* και *κλάση μετασχηματισμών* χρησιμοποιούνται εναλλάξ στο κείμενο της παρούσας διατριβής.

1.2 Δύο Εισαγωγικά Άρθρα

Στις αρχές της δεκαετίας του 60', δύο σημαίνοντα άρθρα δημοσιεύθηκαν: το πρώτο ήταν από τους Box & Cox (1964) και το δεύτερο από τους Box & Tidwell (1962). Και τα δύο αυτά άρθρα παρουσιάζουν εφαρμογές σχετιζόμενες κυρίως με πειραματικούς σχεδιασμούς, όπως επέτασσε η μόδα των ημερών εκείνων στον χώρο της στατιστικής. Το άρθρο των Box & Cox περιγράφει μια απλή παραμετρική κλάση μετασχηματισμών δεδομένων η οποία είναι αρκετά εύκολη στη χρήση και αρχικά προοριζόταν μόνο για τη μεταβλητή απόκρισης Y ενός γραμμικού μοντέλου. Θεωρώντας την παράμετρο μετασχηματισμού λ, η κλάση μετασχηματισμών που προτάθηκε για μια τυχαία παρατήρηση y είναι η εξής:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}, \; y \in \mathbb{R}^+.$$

Αυτή η κλάση αποτελεί επέκταση της πολύ απλούστερης μονότονης συνάρτησης του Tukey (1957) η οποία, όμως, είχε ασυνέχεια στο σημείο $\lambda = 0$, κάτι που αναγνωρίζεται ως σημαντικό ελάττωμα για μια συνάρτηση μετασχηματισμού. Αντίθετα, οι Box & Cox όρισαν ότι για την περίπτωση όπου $\lambda \neq 0$ ο μετασχηματισμός δεν συνίσταται σε μια απλή ύψωση της τ.μ. Y σε δύναμη, αλλά έχουμε μετακίνηση κατά -1 και διαίρεση με λ. Αυτή η προσαρμογή έγινε ακριβώς στη βάση αποφυγής της ασυνέχειας. Μια ενδιαφέρουσα απλούστευση αφορά την ανάλυση διακύμανσης (ANOVA), όπου μπορεί κανείς να υποθέσει ότι $y^{(\lambda)} = y^\lambda$ εφόσον τα αποτελέσματα είναι αμετάβλητα για γραμμικούς μετα-

σχηματισμούς. Η προϋπόθεση κάτω από την οποία λειτουργεί η μεθοδολογία του εν λόγω άρθρου είναι ότι για κάθε σταθερή τιμή της παραμέτρου λ, η μετασχηματισμένη τ.μ. $Y^{(\lambda)}$ είναι μια μονότονη συνάρτηση της τ.μ. Y (τουλάχιστον για τις τιμές της Y που έχουν μη μηδενική πιθανότητα). Ακόμη μια σημαντική προϋπόθεση είναι ότι για κάποια κατάλληλη αλλά άγνωστη τιμή της παραμέτρου μετασχηματισμού λ για το σύνολο μετασχηματισμένων παρατηρήσεων $\boldsymbol{y}^{\lambda} = \left(y_1^{\lambda}, \ldots, y_n^{\lambda}\right)$ μεγέθους n, τα κατάλοιπα του μοντέλου ικανοποιούν τις προϋποθέσεις της γραμμικής παλινδρόμησης εκτός αυτής της γραμμικότητας: ανεξαρτησία σφαλμάτων, κανονικότητα σφαλμάτων, ομοσκεδαστικότητα.

Στο αρχικό αυτό άρθρο των Box & Cox, για την εκτίμηση της παραμέτρου μετασχηματισμού λ εφαρμόστηκαν ξεχωριστά μια πρώτη προσέγγιση βασισμένη στη μέγιστη πιθανοφάνεια και μια δεύτερη, Μπεϋζιανή, προσέγγιση. Υποτίθεται ότι η τ.μ. Y παίρνει τιμές στον αυστηρά θετικό άξονα, δηλαδή εξαιρουμένου και του μηδενός. Παρόλα αυτά, έχουν αναφερθεί απλές τεχνικές για να μπορέσουν να ενσωματωθούν και τιμές του αρνητικού άξονα, με το συγκεκριμένο άρθρο να αναφέρει την μετατόπιση των τιμών της κατανομής προς τα δεξιά προσθέτοντας μια αρκούντως μεγάλη ποσότητα στο διάνυσμα παρατηρήσεων της μεταβλητής απόκρισης (*shifted Box-Cox*). Μια ακόμη τεχνική για το ίδιο ζήτημα σχετίζεται με τον μετασχηματισμό *signed power* των Bickel & Docksum (1981) που περιγράφεται στην επόμενη ενότητα. Αρκετά αντεπιχειρήματα έχουν αναπτυχθεί σχετικά με την τεχνική shifted Box-Cox (ή αλλιώς μετατοπισμένος Box-Cox μετασχηματισμός). Ένα από αυτά έγκειται στο ότι τα ασυμπτωτικά αποτελέσματα της θεωρίας μέγιστης πιθανοφάνειας είναι πλέον ακατάλληλα προς χρήση εφόσον το εύρος τιμών της μετασχηματισμένης μεταβλητής εξαρτάται άμεσα από την σταθερά μετατόπισης, η επιλογή της οποίας γίνεται με αρκετά αυθαίρετο τρόπο (Yeo & Johnson 2000). Επιπλέον, η επιλογή της σταθεράς μετατόπισης είναι πιθανό να επηρεάσει την επιλογή της παραμέτρου μετασχηματισμού λ. Αυτό το θέμα έχει αναφερθεί σε πολύ λίγα άρθρα και έχει διερευνηθεί από ακόμη λιγότερα (Stahel 2002, π.χ.). Στα πλαίσια μιας προκαταρκτικής διερεύνησης που διεξάγαμε εμείς, το μοτίβο αυτό φαίνεται να επιβεβαιώνεται και μάλιστα όσο περισσότερο αυξάνει η τιμή της σταθεράς μετατόπισης τόσο πιο μεγάλη είναι η τιμή εκτίμησης της παραμέτρου λ. Εν γένει, η πιο ασφαλής και απλή επιλογή για τη σταθερά μετατόπισης είναι τιμές κοντινές στην απόλυτη ελάχιστη παρατηρούμενη τιμή $|\min(\boldsymbol{y})|$.

Η κλάση μετασχηματισμών Box-Cox εφαρμοζόμενη στην τ.μ. Y αλλάζει την κατανομή των σφαλμάτων. Ο σκοπός είναι η αλλαγή αυτή να οδηγήσει σε μια κατεύθυνση διόρθωσης/ρύθμισης των υποθέσεων του μοντέλου. Το μετασχηματισμένο μοντέλο έχει τη γενική μορφή (1.1) όπου $\mathbf{Y}^{(\lambda)}$ είναι το τυχαίο δείγμα της μετασχηματισμένης μεταβλητής απόκρισης μήκους n, $\mathbf{X}$ είναι ο πίνακας σχεδιασμού , $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ είναι το διάνυσμα παραμέτρων διάστασης $(p+1)$ και τα σφάλματα $\varepsilon_i, i = 1, \ldots, n$ είναι ανεξάρτητα μεταξύ τους:

$$\mathbf{Y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right). \tag{1.1}$$

Η μέθοδος μέγιστης πιθανοφάνειας (*maximum likelihood estimation*, MLE) βασίζεται στην κατανομή X^2. Ο επονομαζόμενος κανονικοποιημένος (ή κλιμακωτός - *scaled*) μετασχηματισμός $\mathbf{z}^{(\lambda)} = \frac{\boldsymbol{y}^{(\lambda)}}{|J(\boldsymbol{y},\lambda)^{1/n}|}$, όπου η ποσότητα $|J(\boldsymbol{y}, \lambda)|^{1/n}$ ισούται με τον δειγματικό γεωμετρικό μέσο της Ιακωβιανής του μετασχηματισμού από το διάνυσμα των αρχικών παρατηρήσεων $\boldsymbol{y}$ στο μετασχηματισμένο διάνυσμα $\boldsymbol{y}^{(\lambda)}$, χρησιμοποιείται επίσης κατά τη διαδικασία μεγιστοποίησης, η οποία περιλαμβάνει την ελαχιστοποίηση του αθροίσματος τετραγώνων των υπολοίπων του μοντέλου για τη νέα αυτή μεταβλητή. Σε μη διανυσματική μορφή έχουμε

$$z^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda \cdot g(\boldsymbol{y})^{\lambda-1}}, & \text{if } \lambda \neq 0 \\ \log(y) \cdot g(\boldsymbol{y}), & \text{if } \lambda = 0 \end{cases}.$$

όπου με $g(\boldsymbol{y}) = \left(\prod_{i=1}^{n} y_i\right)^{1/n}$ συμβολίζουμε τον δειγματικό γεωμετρικό μέσο των αρχικών παρατηρήσεων. Έτσι επιδιώκεται η εξάλειψη, κατά κάποιον τρόπο, της εξάρτησης της κλίμακας από την παράμετρο λ όταν συγκρίνουμε ποσότητες που μας ενδιαφέρουν, με την αιτιολογία ότι διατηρείται το ίδιο μετρικό σύστημα για το $z^{(\lambda)}$ όπως και για το μη μετασχηματισμένο y. Ο Li (2005) αναφέρει ότι ο παραπάνω κανονικοποιημένος μετασχηματισμός $\mathbf{z}^{(\lambda)}$ ενδέχεται να μειώσει τη δεσμευμένη διακύμανση των συντελεστών παλινδρόμησης, δεδομένης της παραμέτρου μετασχηματισμού, καθώς και να διατηρήσει τις τιμές των στατιστικών t και F αμετάβλητες σε σχέση με τις μη μετασχηματισμένες τιμές. Σχετικά με τη Μπεϋζιανή προσέγγιση, κατάλληλες ομοιόμορφες κατανομές χρησιμοποιούνται για όλες τις a priori κατανομές που εμπλέκονται για τις παραμέτρους $\boldsymbol{\beta}$ και $\log(\sigma^2)$ δεδομένης της παραμέτρου λ. Από την κλασσική και από τη Μπεϋζιανή προσέγγιση προκύπτουν δύο εξισώσεις για την περιθώρια κατανομή του λ: μία για τη μεγιστο-

ποιημένη λογαριθμική πιθανοφάνεια

$$l_1(\lambda) = -\frac{1}{2} n \log\left(\frac{RSS_{\mathbf{z}^{(\lambda)}}}{n}\right)$$

και μια για τον λογάριθμο της συνεισφοράς στην a posteriori κατανομή του λ

$$l_2(\lambda) = -\frac{1}{2} \nu_r \log\left(\frac{RSS_{\mathbf{z}^{(\lambda)}}}{\nu_r}\right)$$

όπου $RSS_{\mathbf{z}^{(\lambda)}}$ είναι το άθροισμα τετραγώνων των υπολοίπων για το κανονικοποιημένο διάνυσμα $\mathbf{z}^{(\lambda)}$. Οι ανωτέρω σχέσεις διαφέρουν μόνο στον εκτιμητή της διακύμανσης (ο οποίος σχετίζεται με το n στην περίπτωση της μεθόδου μέγιστης πιθανοφάνειας και με τους βαθμούς ελευθερίας των υπολοίπων $\nu_r = n - \text{rank}(\mathbf{X}) \leq n - (p+1)$ στην περίπτωση της Μπεϋζιανής προσέγγισης), ενώ είναι και οι δύο μονότονες συναρτήσεις της εκτιμήτριας του μέσου τετραγωνικού σφάλματος της ANOVA. Επομένως, πρόκειται για σχεδόν ισοδύναμες εκφράσεις που ως επί το πλείστον παράγουν πολύ παρόμοια αποτελέσματα.

Ένα από τα βασικά πλεονεκτήματα που προσφέρει η κλάση μετασχηματισμών δύναμης Box-Cox είναι ότι ο όρος της Ιακωβιανής (1.2) μπορεί να υπολογιστεί πολύ εύκολα και το ίδιο και η συνάρτηση πιθανοφάνειας σε σχέση με τις αρχικές μη μετασχηματισμένες παρατηρήσεις:

$$|J(\boldsymbol{y}, \lambda)| = \prod_{i=1}^{n} \left| \frac{dy_i^{(\lambda)}}{dy_i} \right|. \tag{1.2}$$

Όπως αναφέρθηκε και προηγουμένως, η Ιακωβιανή μπορεί να εκφραστεί και σαν συνάρτηση του δειγματικού γεωμετρικού μέσου $g(\boldsymbol{y})$ ως εξής:

$$|J(\boldsymbol{y}, \lambda)| = \left(\prod_{i=1}^{n} y_i^{\lambda - 1} \right)^{\frac{1}{n}} = g(\boldsymbol{y})^{\lambda - 1}. \tag{1.3}$$

Το δεύτερο σημαίνον άρθρο που προαναφέραμε (Box & Tidwell 1962) δημοσιεύθηκε δύο χρόνια νωρίτερα και πρότεινε μετασχηματισμούς δύναμης για τις τιμές των επεξηγηματικών μεταβλητών σε καθαρά μη Μπεϋζιανή λογική. Η μεθοδολογία που προτάθηκε σε αυτό οδηγεί σε ένα μη γραμμικό μοντέλο, αν και η σχέση μεταξύ των συμμεταβλητών και της μεταβλητής απόκρισης ενδέχεται να γραμμικοποιείται μέσω των μετασχηματισμών. Η γενική μορφή του μοντέλου ακολουθεί παρακάτω, με τις παραμέτρους μετασχηματισμού $\lambda_1, \ldots, \lambda_p$ να αντιστοιχούν στις επεξηγηματικές μεταβλητές πλήθους p του μοντέλου:

$$y_i = \beta_0 + \beta_1 x_{i1}^{\lambda_1} + \ldots + \beta_p x_{ip}^{\lambda_p} + \varepsilon_i \,, \qquad \varepsilon_i \overset{i.i.d}{\sim} N\left(0, \sigma^2\right).$$

Μέσω μιας επαναληπτικής διαδικασίας, η περιγραφή της οποίας παραλείπεται εδώ, παράγονται εκτιμητές μέγιστης πιθανοφάνειας για τις παραμέτρους $\lambda_1, \ldots, \lambda_p$. Η έλλειψη σύγκλισης μπορεί να αποτελέσει συχνά ένα πρόβλημα στη διαδικασία αυτή, διαφορετικά λίγες μόνο επαναλήψεις (συνήθως τρεις) αρκούν, σύμφωνα με τους συγγραφείς.

Στο άρθρο αναφέρεται επίσης ότι το σύνολο μετασχηματισμών θα μπορούσε να είναι πιο περίπλοκο αν κάποιος το επιθυμούσε. Συγκεκριμένα, κάθε μετασχηματισμένη μεταβλητή θα μπορούσε να είναι μια συνάρτηση όλων των αρχικών επεξηγηματικών μεταβλητών $X_1, X_2, \ldots, X_p$ αντί να είναι συνάρτηση μόνο μιας αρχικής μεταβλητής. Συνεπώς, μια πιο γενική μορφή ενός μετασχηματισμού T_i θα μπορούσε να είναι η εξής:

$$T_i = T_i\left(X_1, X_2, \ldots, X_p\right), \qquad i = 1, 2, \ldots, p.$$

Κατά αυτόν τον τρόπο, η έννοια των αλληλεπιδράσεων εισάγεται στο πεδίο των μετασχηματισμών δεδομένων.

1.3 Μεταγενέστερο Έργο

1.3.1 Τροποποιήσεις των συναρτήσεων μετασχηματισμών

Όπως έχει ήδη αναφερθεί στην Ενότητα 1.2, ένα βασικό περιοριστικό σημείο στην τυπική μέθοδο μετασχηματισμού Box-Cox είναι το γεγονός ότι η τ.μ. Y πρέπει να είναι αυστηρά θετική. Αν τα δεδομένα μας εκτείνονται στον πραγματικό άξονα τιμών, τότε μπορούμε να υπερβούμε το εμπόδιο προσθέτοντας μια αρκούντως μεγάλη σταθερά στις παρατηρούμενες τιμές. Οι Bickel & Docksum (1981) ήταν οι πρώτοι που πρότειναν μια εναλλακτική σε αυτή την τεχνική, υπό το όνομα του μετασχηματισμού *signed power*. Για $\lambda \in \mathbb{R}^+$, ο μετασχηματισμός προσαρμόζεται έτσι ώστε να δέχεται και αρνητικές τιμές y, χρησιμοποιώντας τη συνάρτηση $\text{sign}(y)$ με τέτοιο τρόπο ώστε και μη φραγμένες κατανομές να μπορούν να καλυφθούν (π.χ. η κανονική κατανομή). Ο τυπικός μετασχηματισμός Box-Cox αποτελεί ειδική περίπτωση της παρακάτω έκφρασης για θετικό y και θετικό λ:

$$y^{(\lambda)} = \frac{|y|^{\lambda}\,\text{sign}(y) - 1}{\lambda}, \quad \lambda > 0,\ y \in \mathbb{R} \tag{1.4}$$

όπου $\text{sign}(y) = -1$ για αρνητικό y και $\text{sign}(y) = +1$ για θετικό y.

Ο μετασχηματισμός (1.4) μπορεί να θεωρηθεί σαν μια επέκταση του *Modulus transformation* που εισήγαγαν οι John & Draper (1980). Ο μετασχηματισμός Modulus φαίνεται να λειτουργεί καλά όταν προϋπάρχει ήδη κάποια συμμετρία στα δεδομένα:

$$y^{(\lambda)} = \begin{cases} \frac{\text{sign}(y)\left[(|y|+1)^{\lambda}-1\right]}{\lambda} & \lambda \neq 0 \\ \text{sign}(y)\log\left(|y|+1\right) & \lambda = 0 \end{cases}, \; y \in \mathbb{R}.$$

Μια ακόμη διαφοροποίηση του τυπικού μετασχηματισμού Box-Cox προτάθηκε στο άρθρο των Yeo & Johnson (2000) με ιδιαίτερη έμφαση στην άμβλυνση του βαθμού ασυμμετρίας μιας κατανομής. Ακολουθεί η μορφή του μετασχηματισμού των Yeo-Johnson, η οποία ουσιαστικά αποτελεί μια εξομαλυμένη (*smoothed*) εναλλακτική του μετασχηματισμού Modulus (John & Draper 1980):

$$y^{(\lambda)} = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda} & y \geq 0, \lambda \neq 0 \\ \log(y+1) & y \geq 0, \lambda = 0 \\ -\frac{(-y+1)^{2-\lambda}-1}{2-\lambda} & y < 0, \lambda \neq 2 \\ -\log(-y+1) & y < 0, \lambda = 2 \end{cases}. \tag{1.5}$$

Η εν λόγω παραλλαγή επιχειρεί επίσης να άρει το πρόβλημα των κάτω φραγμένων τιμών της τ.μ. Y διασφαλίζοντας παράλληλα τη συνέχεια της συνάρτησης μετασχηματισμού. Στη βάση ενός παραδείγματος για μια μίξη μιας κανονικής και μιας γάμμα κατανομής, υποστηρίζεται ότι ο μετασχηματισμός *signed power* προορίζεται για τη διόρθωση της κύρτωσης και άρα είναι μάλλον ακατάλληλος για την άμβλυνση της ασυμμετρίας μιας κατανομής. Ο μετασχηματισμός των Yeo-Johnson αποτελεί βελτίωση του μετασχηματισμού του Manly (1976) που επίσης εστίαζε στην άμβλυνση της ασυμμετρίας, αλλά δε δεχόταν παρατηρήσεις στο μήκος του πραγματικού άξονα τιμών:

$$y^{(\lambda)} = \begin{cases} \frac{\exp(\lambda y)-1}{\lambda} & \lambda \neq 0 \\ y & \lambda = 0 \end{cases}, \; y \in \mathbb{R}^{+}.$$

Η αποτελεσματικότητα του εν λόγω μετασχηματισμού αυτού σε δικόρυφες κατανομές σχήματος U (*U-shaped*) αξιολογείται ως μάλλον χαμηλή.

Ο προτεινόμενος μετασχηματισμός (1.5) των Yeo-Johnson βασίζεται στην έννοια της σχετικής ασυμμετρίας (*relative skewness*) (van Zwet 1964) και στο γεγονός ότι ο τυπικός μετασχηματισμός Box-Cox παρουσιάζει σημείο καμπής για $\lambda = 1$. Συγκεκριμένα,

ο μετασχηματισμός Box-Cox είναι κοίλος στο y κάτω από το σημείο καμπής $\lambda = 1$ και κυρτός στο y πάνω από το σημείο αυτό. Θέματα προκύπτουν στην περίπτωση όπου η τ.μ. Y παίρνει και θετικές και αρνητικές τιμές, αλλιώς ο μετασχηματισμός *sign power* θα ήταν αρκετός για προβλήματα ασυμμετρίας. Η πρόσθεση μιας μονάδας στο y επιτρέπει τη διατήρηση του πρόσημου της μη μετασχηματισμένης παρατήρησης y. Επίσης, επιτρέπει την ενσωμάτωση του Ταυτοτικού μετασχηματισμού στην περίπτωση όπου $\lambda = 1$. Για θετικές τιμές του y, ο μετασχηματισμός είναι ισοδύναμος με τον μετασχηματισμό Modulus. Η μόνη διαφορά είναι ότι για $y < 0$ έχουμε $(2 - \lambda)$ αντί για λ. Αυτό συμβαίνει αν θεωρήσουμε διαφορετικές τιμές λ_+, λ_- της παραμέτρου μετασχηματισμού για τον θετικό και τον αρνητικό άξονα και επιθυμούμε και την ύπαρξη συνέχειας στο $y = 0$, οδηγώντας στη συνθήκη $(\lambda_+) + (\lambda_-) = 2$. Το κριτήριο που χρησιμοποιείται εδώ για την επιλογή μιας κατάλληλης τιμής για το λ είναι η ελαχιστοποίηση της πληροφορίας Kullback-Leibler (KL) (μέσω εκτιμητριών μέγιστης πιθανοφάνειας για το λ) και συγκεκριμένα η ελαχιστοποίηση της απόστασης KL μεταξύ της κανονικής κατανομής και της μετασχηματισμένης κατανομής που περιέχει την παράμετρο λ.

Μια ακόμη πολύ ενδιαφέρουσα ιδέα περιγράφεται εύγλωττα στο άρθρο του Yang (2006) όπου, για μια ακόμη φορά, μόνο αυστηρά θετικές παρατηρήσεις επιτρέπονται. Ο μετασχηματισμός Dual σχεδιάστηκε για την επίλυση του προβλήματος που σχετίζεται με την περικομμένη κατανομή των μετασχηματισμένων δεδομένων επεκτείνοντας το αντίστοιχο προβληματικό φράγμα τους. Δεν υπάρχει πλέον ουδέτερη τιμή της παραμέτρου λ_T, δηλαδή δεν υπάρχει τιμή της παραμέτρου μετασχηματισμού που να αντιστοιχεί σε απουσία μετασχηματισμού, θεωρώντας ότι όλα τα διανύσματα παρατηρήσεων είναι a priori μη κανονικά. Για τιμές του λ_T κοντά στο μηδέν, ο μετασχηματισμός Dual προσεγγίζει τον μετασχηματισμό Box-Cox. Εξαιτίας της συμμετρίας της συνάρτησης μετασχηματισμού γύρω από την τιμή $\lambda_T = 0$, θεωρούνται μόνο θετικές τιμές της παραμέτρου μετασχηματισμού, ως ισοδύναμες με το $-\lambda_T$:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda} - y^{-\lambda}}{2\lambda}, \lambda > 0 \\ \log(y), \lambda = 0 \end{cases} \quad y > 0.$$

Μια παρατήρηση είναι ότι σε κάποιους από τους προαναφερθέντες μετασχηματισμούς (όπως ο τυπικός και ο μετατοπισμένος Box-Cox, καθώς και ο Modulus μετασχηματισμός),

το εύρος των μετασχηματισμένων τιμών επηρεάζεται από το πρόσημο της παραμέτρου λ. Το Διάγραμμα 2.1 της Ενότητας 2.3 απεικονίζει τη σύνδεση μεταξύ των βασικών οικογενειών μετασχηματισμών που αναφέρθηκαν μέχρι τώρα και που θα χρησιμοποιηθούν στη συνέχεια για την ανάπτυξη της μεθοδολογίας μας.

1.3.2 Πρότερες κατανομές

Όπως είδαμε και προηγουμένως όσον αφορά το αρχικό άρθρο των Box & Cox, η επιλογή της πρότερης κατανομής αφορά τον καθορισμό της μορφής $\pi(\lambda, \boldsymbol{\beta}, \sigma)$, όπου $\boldsymbol{\beta}$ είναι το διάνυσμα των παραμέτρων παλινδρόμησης του μετασχηματισμένου μοντέλου. Επιπλέον, στην περίπτωση όπου η πρότερη κατανομή του λ θεωρηθεί ομοιόμορφη στον χώρο τιμών $\mathcal{M}_\lambda$ του λ, τότε το πρόβλημα απλουστεύεται περαιτέρω και αρκεί να επιλεγεί μια δεσμευμένη πρότερη κατανομή της μορφής $\pi(\boldsymbol{\beta}, \sigma|\lambda)$.

Η συνεισφορά του Pericchi (1981) συνίσταται στην άρση της εξάρτησης της πρότερης κατανομής από τα δεδομένα, σε αντίθεση με την πρότερη κατανομή που χρησιμοποιήθηκε από τους Box & Cox (1964) η οποία παρόλα αυτά είναι μη πληροφοριακή και παράγει τα ίδια αποτελέσματα με την προσέγγιση που βασίζεται στη μέθοδο μέγιστης πιθανοφάνειας. Επίσης, παρουσιάζονται κριτήρια για τη διόρθωση των προϋποθέσεων της κανονικότητας, της προσθετικότητας (*additivity*) και της ομοσκεδαστικότητας, τα οποία ουσιαστικά είναι μονότονες παραλλαγές γνωστών στατιστικών ελεγχοσυναρτήσεων στη βάση της προσέγγισης πιθανοφάνειας (πηλίκο πιθανοφανειών Neyman-Pearson, πηλίκο F και στατιστικό L_1 των Neyman-Pearson αντίστοιχα). Με άλλα λόγια, η παράμετρος λ μπορεί να εκτιμηθεί ούτως ώστε να ικανοποιεί ένα συγκεκριμένο κριτήριο (ή έναν συνδυασμό κριτηρίων). Για παράδειγμα, θα μπορούσε να επιλεγεί μια τιμή του λ ώστε να ελαχιστοποιείται η F τιμή που αντιστοιχεί στους βαθμούς ελευθερίας για τη μη προσθετικότητα (*non additivity*).

Ακολούθως, δύο μορφές πρότερης κατανομής παρουσιάζονται στις οποίες ο δείκτης $\{BC\}$ συμβολίζει την πρότερη κατανομή κατά Box-Cox και ο δείκτης $\{P\}$ την πρότερη κατανομή κατά Pericchi.

Για $\lambda = 1$:

$$\pi_{BC}(\boldsymbol{\beta}, \sigma, \lambda = 1) = \pi(\boldsymbol{\beta}, \sigma|\lambda = 1)\ \pi(\lambda = 1) \propto \pi(\lambda = 1) \cdot \frac{1}{\sigma}$$

$$\pi_P(\boldsymbol{\beta}, \sigma, \lambda = 1) \propto \mathbf{I}_n(\boldsymbol{\beta}, \sigma)^{1/2}\, \pi(\lambda = 1) \propto \pi(\lambda = 1) \cdot \frac{1}{\sigma^{r\mathbf{x}+1}}$$

όπου η ποσότητα $\mathbf{I}_n$ αντιπροσωπεύει τον πίνακα της πληροφορίας κατά Fisher και $r_{\mathbf{X}}$ είναι ο βαθμός (*rank*) του πίνακα σχεδιασμού $\mathbf{X}$. Η πρότερη κατανομή κατά Pericchi βασίζεται στον κανόνα του Jeffreys και δεν υποθέτει πρότερη εξάρτηση μεταξύ των παραμέτρων. Η πρότερη κατανομή κατά BC η οποία υποθέτει ομοιόμορφες πρότερες κατανομές για τα $\boldsymbol{\beta}$ και $\log \sigma$ δεδομένου του λ.

Για κάθε λ:

$$\pi_{BC}(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) / \left(\sigma \cdot |J(\boldsymbol{y}, \lambda)|^{r\mathbf{x}/n}\right)$$

$$\pi_P(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) / (\sigma^{r\mathbf{x}+1})$$

όπου $|J(\boldsymbol{y}, \lambda)|^{1/n} = \left(\prod_{i=1}^{n} y_i^{\lambda-1}\right)^{1/n}$ είναι ο γεωμετρικός μέσος της Ιακωβιανής (ποσότητα αντιπροσωπευτική της κλίσης (*gradient*) του $y^{(\lambda)}$ επί του y). Ο όρος αυτός δείχνει την εξάρτηση της πρότερης κατανομής κατά BC από τα δεδομένα. Είναι εμφανές ότι η πρότερη κατανομή κατά Pericchi δεν εμπεριέχει παρόμοιο όρο και συνεπώς είναι ανεξάρτητη της έκβασης (*outcome-independent*). Η πρότερη κατανομή του λ συνήθως θεωρείται ομοιόμορφη επί του εύρους των δυνατών τιμών λ.

Η εκ των υστέρων κατανομή που προκύπτει με βάση την πρότερη κατανομή του Pericchi ισούται με την εκ των υστέρων κατανομή που προκύπτει από μια συνήθη κανονική-αντίστροφη-γάμμα κατανομή (*normal-inverse-gamma*, NIG) με τιμές για τις αντίστοιχες υπερπαραμέτρους πολύ κοντά στο μηδέν, άρα μη πληροφοριακή. Επιπλέον, ο Pericchi ισχυρίζεται ότι η συγκεκριμένη πρότερη κατανομή σχετίζεται με σταθερότητα κλίμακας (*scale invariance*).

Όλα τα παραπάνω ισχύουν υπό την προϋπόθεση ότι υπάρχει μια τιμή του λ για την οποία πληρούνται ταυτόχρονα τα κριτήρια της κανονικότητας, της προσθετικότητας και της ομοσκεδαστικότητας. Το διάστημα της μέγιστης εκ των υστέρων πιθανότητας (*maximum a posteriori interval*) για την παράμετρο λ υπό την πρότερη κατανομή του Pericchi είναι πιο στενό σε σχέση με το διάστημα που αντιστοιχεί στην πρότερη κατανομή κατά BC.

Τρία χρόνια μετά τη δημοσίευση της δουλειάς του Pericchi, δημοσιεύθηκε ένα άρθρο του Sweeting (1984) σχετικά με την επιλογή πρότερης κατανομής έχοντας σαν σημείο έναρξης ένα κανονικό ομοσκεδαστικό γραμμικό μοντέλο. Το άρθρο επιχειρεί να συμβιβάσει τις προσεγγίσεις του Pericchi και των Box-Cox όσον αφορά το θέμα της εξάρτησης

της μη πληροφοριακής πρότερης κατανομής από τα δεδομένα, τονίζοντας ότι και η πρώτη προσέγγιση υποφέρει από διάφορα μειονεκτήματα. Συγκεκριμένα, ισχυρίζεται ότι αν κανείς υπολογίσει το όριο μιας γνήσιας συζυγούς πρότερης κατανομής με σκοπό να κατασκευάσει μια μη γνήσια πρότερη κατανομή μεγάλης διακύμανσης, η πρότερη κατανομή που προκύπτει δεν είναι συνεπής από διάφορες απόψεις, π.χ. εξαιτίας της μη ρεαλιστικής επίδρασης της συζυγούς πρότερης κατανομής σε αυτήν.

Ας θεωρήσουμε στο εξής ότι ο πίνακας σχεδιασμού είναι πλήρους βαθμού. Η νέα πρότερη κατανομή που προτείνεται έχει την ακόλουθη μορφή:

$$\pi(\lambda, \beta_0, \boldsymbol{\beta}_{\setminus 0}, \sigma) \propto \frac{\pi(\lambda)}{\sigma\,(1+\lambda\beta_0)^{p+1-\beta_0/\lambda}}$$

όπου β_0 είναι η παράμετρος που αντιστοιχεί στον σταθερό όρο και $\boldsymbol{\beta}_{\setminus 0}$ είναι συνήθεις παράμετροι των επεξηγηματικών μεταβλητών του μετασχηματισμένου μοντέλου έτσι ώστε η ποσότητα $\mu = \mathrm{E}\big[Y^\lambda | X_1 = 0\big] = 1 + \lambda\beta_0$ να ακολουθεί κανονική κατανομή. Δυστυχώς, όλες οι προσεγγίσεις που παρουσιάστηκαν ως τώρα για την κατασκευή της πρότερης κατανομής αγνοούν το γεγονός ότι η τ.μ. Y κινείται αναγκαστικά στον αυστηρά θετικό άξονα για τον μετασχηματισμό Box-Cox.

Ο Sweeting (1985) δίνει έμφαση στο πρόβλημα της ασυνέπειας που προκύπτει λόγω της επιλογής της πρότερης κατανομής, εννοώντας το πρόβλημα της μη αναγνωρισιμότητας (*non identifiability problem*) σε μια γειτονιά του λ. Η διαδικασία εκμαίευσης της πρότερης πληροφορίας θα πρέπει να λαμβάνει υπόψη μεταξύ άλλων τις αλλαγές στην κλίμακα των μεταβλητών. Στο άρθρο δίνονται παραδείγματα πρότερων κατανομών για τις περιπτώσεις του τυπικού και μετατοπισμένου Box-Cox μετασχηματισμού, όπως και για την περίπτωση του αναδιπλούμενου μετασχηματισμού δύναμης (*folded power transformation*) που είναι έγκυρος μόνο για $y \in (0,1)$ και παρουσιάζεται εδώ:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - (1-y)^\lambda}{\lambda} & \lambda \neq 0 \\ \log(y/(1-y)) & \lambda = 0 \end{cases}, \quad y \in (0,1).$$

Η πρότερη κατανομή που προτείνεται για τον μετασχηματισμό Box-Cox είναι η εξής:

$$\pi(\lambda, \boldsymbol{\beta}, \sigma) \propto q(\boldsymbol{\beta}_{\setminus 0})\sigma^{r-1}(1+\lambda\beta_0)^{-(r+1)}$$

όπου $q(\boldsymbol{\beta}_{\setminus 0})$ είναι η δεσμευμένη πρότερη κατανομή του $\boldsymbol{\beta}_{\setminus 0}$ δεδομένου του $\lambda = \lambda_1$, ενώ η επιλογή της παραμέτρου r αφήνεται ολότελα στον εκάστοτε ερευνητή.

1.3.2.1 Εξηγώντας τη λογική πίσω από τις πρότερες κατανομές κατά Box-Cox και κατά Pericchi

Ας εστιάσουμε στις δύο βασικές εναλλακτικές μορφές για την από κοινού πρότερη κατανομή $\pi(\boldsymbol{\beta}, \sigma, \lambda)$ οι οποίες, στην πραγματικότητα, είναι παραλλαγές της γενικής μη γνήσιας πρότερης κατανομής του Jeffreys. Η πρώτη πρότερη προσέγγιση από τους Box & Cox (1964) έχει τη μορφή:

$$\pi_{BC}(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) \cdot \frac{1}{\sigma} \cdot |J(\boldsymbol{y}, \lambda)|^{-\frac{p+1}{n}} \tag{1.6}$$

όπου, μόνο για την οικογένεια BC, ο όρος που σχετίζεται με την Ιακωβιανή μπορεί να γραφτεί σαν συνάρτηση του γεωμετρικού μέσου των παρατηρήσεων $g(\boldsymbol{y})$, βλέπε (1.3):

$$\pi_{BC}(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) \cdot \frac{1}{\sigma} \cdot g(\boldsymbol{y})^{-(\lambda-1)(p+1)}.$$

Ας αναλύσουμε εν συντομία την κατασκευή αυτής της πρότερης κατανομής. Στην περίπτωση όπου ένας μετασχηματισμός των δεδομένων κριθεί μη αναγκαίος, τότε μια μη πληροφοριακή πρότερη κατανομή για την παράμετρο θέσης $\boldsymbol{\beta}$ είναι ανάλογη της μονάδας και για την παράμετρο κλίμακας σ είναι σ^{-1}. Συνολικά έχουμε:

$$\pi\left(\boldsymbol{\beta}, \sigma, \lambda = 1\right) = \pi\left(\boldsymbol{\beta}, \sigma | \lambda = 1\right) \cdot \pi(\lambda = 1) \propto \pi(\lambda = 1) \cdot \sigma^{-1}.$$

Στο σημείο αυτό γίνεται η υπόθεση ότι ο μετασχηματισμός των παρατηρήσεων $\boldsymbol{y}$ είναι γραμμικός επί του εύρους των παρατηρήσεων αυτών. Συνεπώς,

$$\mathrm{E}\left(y_i^{(\lambda)}\right) \simeq \alpha_\lambda + b_\lambda \mathrm{E}\left(y_i\right)$$

όπου b_λ είναι μια ποσότητα που αντιπροσωπεύει την κλίση $\frac{\mathrm{d}Y^{(\lambda)}}{\mathrm{d}Y}$. Δεδομένου ότι όλοι οι όροι της παραμέτρου θέσης $\boldsymbol{\beta}$ θα πρέπει να πολλαπλασιαστούν με b_λ στην περίπτωση ανάγκης για μετασχηματισμό, η πρότερη κατανομή της παραμέτρου αυτής γίνεται $\pi\left(\boldsymbol{\beta}\right) = |b_\lambda|^{-(p+1)}$. Επιλέγοντας για την κλίση έναν όρο σχετικό με την Ιακωβιανή, όπως $b_\lambda = |J\left(\boldsymbol{y}, \lambda\right)|^{\frac{1}{n}}$, μας οδηγεί στη μορφή (1.6). Ένα μειονέκτημα της προσέγγισης αυτής, όπως αναγνωρίζεται και από τους συγγραφείς της, είναι ότι η τελευταία αυτή επιλογή είναι αυθαίρετη, αν και πολύ βολική. Η δεύτερη πρότερη κατανομή αποδίδεται στον Pericchi (1981):

$$\pi_P(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) \cdot \frac{1}{\sigma^{p+2}}.$$

και βασίζεται στην ίδια ακριβώς λογική με τους Box & Cox με τη μόνη διαφορά ότι ο Pericchi χρησιμοποιεί την πρότερη κατανομή του Jeffreys για το διάνυσμα παραμέτρων $(\boldsymbol{\beta}, \sigma)$. Το πλεονέκτημα εδώ είναι ότι η πρότερη κατανομή δεν είναι εξαρτημένη από τις παρατηρήσεις σε αντίθεση με την $\pi_{BC}(\boldsymbol{\beta}, \sigma, \lambda)$. Υπενθυμίζεται ότι η πρότερη κατανομή του Pericchi οδηγεί σε μια εκ των υστέρων μορφή που είναι ισοδύναμη με την αντίστοιχη *profile* λογαριθμημένη πιθανοφάνεια που θα προέκυπτε από τη μέθοδο μέγιστης πιθανοφάνειας.

1.3.3 Διάφορα θέματα

1.3.3.1 Συμπερασματολογία και πρόβλεψη

Το άρθρο των Stow, Reckhow & Qian (2006) χειρίζεται το θέμα της στατιστικής ανάλυσης μετά τον μετασχηματισμό των δεδομένων και συγκεκριμένα ασχολείται με το συστηματικό σφάλμα που πηγάζει από τον αναμετασχηματισμό των αποτελεσμάτων πίσω στην αρχική κλίμακα χάριν ερμηνείας στο Μπεϋζιανό μοντέλο - γραμμικό ή μη γραμμικό. Αυτό το είδος σφάλματος οφείλεται στη μη γραμμικότητα των μετασχηματισμών η οποία με τη σειρά της συχνά δημιουργεί μεταβλητότητα στον μέσο όρο της κατανομής της μεταβλητής απόκρισης. Σημειώνεται ότι μόνο ο μετασχηματισμός Log λαμβάνεται υπόψη στο συγκεκριμένο άρθρο, αν και αυτό δεν αποτελεί περιορισμό για τη λογική που αναπτύσσεται εκεί. Θεωρώντας το απλό γραμμικό μοντέλο, η συνήθης εξίσωση μετασχηματισμού είναι η εξής:

$$\log(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log(X) \Leftrightarrow \hat{Y} = \exp(\hat{\beta}_0) \cdot X^{\hat{\beta}_1}.$$

Μια μερική λύση για τη μείωση του συστηματικού σφάλματος που αναφέρθηκε είναι να πολλαπλασιαστεί η μετασχηματισμένη εξίσωση με $\exp(0.5 \cdot S^2)$, όπου εμπλέκεται η δειγματική διακύμανση των μετασχηματισμένων δεδομένων, αν και η συγκεκριμένη τεχνική δεν αρκεί για προβλήματα εκτός του γραμμικού μοντέλου με την κλασική προϋπόθεση ότι η δομή των σφαλμάτων του μοντέλου ακολουθεί τυποποιημένη κανονική κατανομή. Υποστηρίζεται ότι ένας πιο ρεαλιστικός αναμετασχηματισμός πίσω στην αρχική κλίμακα θα μπορούσε να είναι ο εξής:

$$Y = \exp(\hat{\beta}_0) \cdot X^{\hat{\beta}_1} \cdot \exp(\varepsilon)$$

λαμβάνοντας υπόψη το σφάλμα του λογαριθμημένου μοντέλου και η ολική εκ των υστέρων κατανομή θα μπορούσε απλά να αποτελέσει το όρισμα της εκθετικής συνάρτησης.

Ας σημειωθεί εδώ ότι οι Hinkley & Runger (1984) ισχυρίζονται ότι η σύγκριση της τιμής μιας παραμέτρου του μοντέλου για διαφορετικές τιμές του λ και η αντίστοιχη διακύμανση που προκύπτει δεν έχουν φυσική ερμηνεία και επομένως η διαδικασία που ακολουθείται στο άρθρο των Bickel & Docksum (1981) είναι ακατάλληλη. Στο άρθρο τους οι Bickel & Docksum (1981) είχαν υποστηρίξει ότι η ασυμπτωτική μη δεσμευμένη διακύμανση των εκτιμητών των παραμέτρων του μετασχηματισμένου μοντέλου είναι αυξημένη κατά ένα πολύ μεγάλο παράγοντα σε σχέση με την αντίστοιχη διακύμανση δεσμευμένη ως προς την παράμετρο μετασχηματισμού (δηλαδή για συγκεκριμένη, γνωστή τιμή του λ). Επομένως, θεώρησαν ότι η συμπερασματολογία για το διάνυσμα των παραμέτρων παλινδρόμησης θα έπρεπε να λάβει υπόψη της την αβεβαιότητα σχετικά με την αληθινή τιμή της παραμέτρου λ. Το άρθρο αυτό αποτέλεσε λίγο αργότερα αντικείμενο σφοδρής κριτικής όχι μόνο από τους Hinkley & Runger (1984), αλλά και από τους Box & Cox (1982). Το βασικό αντεπιχείρημα ήταν ότι, στην πράξη, η τιμή του λ θεωρείται γνωστή κατά τη συμπερασματολογία σχετικά με τις παραμέτρους παλινδρόμησης, καθώς η διαδικασία αυτή έχει νόημα μόνο εφόσον η κλίμακα των δεδομένων είναι γνωστή. Οι Taylor, Cumberland & Meng (1996) εστιάζουν την έρευνά τους σε ποσότητες που εμπλέκονται σε μοντέλα ANOVA οι οποίες ερμηνεύονται ανεξάρτητα της επιλεγόμενης τιμής του λ, όπως ο συντελεστής ενδοσυσχέτισης (*intraclass correlation coefficient*) ρ ή ένα προβλεπόμενο ποσοστημόριο q. Καταλήγουν ότι η διακύμανση τέτοιων ποσοτήτων αυξάνει μόνο κατά ελάχιστα όταν η παράμετρος λ θεωρείται γνωστή σε σχέση με το αν η παράμετρος θεωρείται άγνωστη και άρα προς εκτίμηση. Αναφέρουν ακόμη ότι οι παράμετροι λ και ρ χαρακτηρίζονται από ασυμπτωτική ορθογωνιότητα (δηλ. από πρακτικά μηδενική συσχέτιση). Παρόλα αυτά, χρειάζεται προσοχή όσον αφορά παραμέτρους που σχετίζονται με τις ουρές της κατανομής. Προηγουμένως, οι Carroll & Ruppert (1981) είχαν επίσης συμπεράνει ότι το κόστος για τη συμπερασματολογία μετά την εκτίμηση του λ εμφανίζεται να είναι εν γένει χαμηλό. Εισήγαγαν, κατά τρόπο ανάλογο με τους Bickel & Docksum (1981), το πηλίκο των ασυμπτωτικών διακυμάνσεων όπως φαίνεται στην (1.7) σαν ένα

εργαλείο ποσοτικοποίησης του μέσου κόστους εκτίμησης της παραμέτρου λ:

$$\frac{\text{Var}\left[Q\left(\hat{\boldsymbol{\beta}}, \hat{\lambda}\right)\right]}{\text{Var}\left[Q\left(\hat{\boldsymbol{\beta}}, \lambda = \lambda_0\right)\right]} \tag{1.7}$$

όπου Q είναι η εκτιμώμενη διάμεσος της μεταβλητής απόκρισης δεδομένων των τιμών των επεξηγηματικών μεταβλητών όπως προκύπτει από τη συνάρτηση αντίστροφου μετασχηματισμού πίσω στην αρχική κλίμακα των δεδομένων. Έδειξαν ότι το πηλίκο αυτό είναι μεγαλύτερο της μονάδας αλλά όχι με μεγάλη απόκλιση από τη μονάδα, άρα η αντίστοιχη διακύμανση της εκτίμησης της διαμέσου είναι μόνο λίγο μεγαλύτερη όταν η παράμετρος μετασχηματισμού θεωρείται άγνωστη σε σχέση με όταν θεωρείται γνωστή με τιμή $\lambda = \lambda_0$.

Η ομοσκεδαστικότητα αποτελεί ένα ακόμη μεγάλο κεφάλαιο όσον αφορά τις συνήθεις προϋποθέσεις πολλών στατιστικών μεθόδων. Όταν η συγκεκριμένη ιδιότητα απουσιάζει από ένα δείγμα παρατηρήσεων, ο συντελεστής συσχέτισης του Pearson ενδέχεται να είναι πλασματικά αυξημένος (το ίδιο και ο αντίστοιχος εκτιμητής καλής προσαρμογής του μοντέλου). Άλλα παραδείγματα όπου η ιδιότητα της ομοσκεδαστικότητας είναι κομβική περιλαμβάνουν κάποιους αλγόριθμους αναγνώρισης προτύπων (*pattern recognition*), όπως η διακριτική ανάλυση του Fisher, και διαδικασίες που σχετίζονται με την ANOVA (δηλ. για την επιλογή κατάλληλων τύπων t-test και post-hoc ελεγχοσυναρτήσεων ή απλά για την διασφάλιση της ευρωστίας ενός μοντέλου).

Ο Sakia (1992) εκτιμά ότι, εν γένει, η τυπική Box-Cox οικογένεια μετασχηματισμών δεν οδηγεί σε ένα μοντέλο όπου να πληρούνται συγχρόνως και οι τρεις βασικές προϋποθέσεις (N-H-A, από τα αρχικά των αγγλικών όρων για την κανονικότητα, την ομοσκεδαστικότητα και την προσθετικότητα). Συνεπώς, υπάρχει άπλετος χώρος για καινοτόμο έρευνα στο πεδίο αυτό.

Για παράδειγμα, οι εκτιμήτριες μέγιστης πιθανοφάνειας θεωρούνται εύρωστες σε αποκλίσεις από την κανονικότητα δεδομένης της ύπαρξης σημαντικού βαθμού συμμετρίας στην κατανομή των σφαλμάτων, αλλά δε θεωρούνται ιδιαίτερα εύρωστες σε περιπτώσεις ετεροσκεδαστικότητας. Μια εν δυνάμει λύση στο πρόβλημα αυτό ίσως να σχετίζεται με την διερεύνηση της σχέσης μεταξύ της διακύμανσης και της μέσης τιμής. Οι Yeo & Johnson (2000) υποθέτουν ότι η διακύμανση σ^2 μιας μετασχηματισμένης μεταβλητής είναι αύξουσα συνάρτηση του αντίστοιχου μέσου $\mu = \beta_0$, ενώ για την αντίστοιχη διακύ-

μανση της τ.μ. $Y^{(\lambda)}$ θεωρούν ότι Var $\left[Y^{(\lambda)}\right] \approx \sigma^2(\mu) \cdot f(\lambda, \mu)$, όπου

$$f(\lambda, \mu) = \begin{cases} (\mu + 1)^{2(\lambda - 1)} & \mu \geq 0 \\ (-\mu + 1)^{2(1-\lambda)} & \mu < 0 \end{cases}.$$

Αυτό υποτίθεται ότι μπορεί να συμβάλλει στην επιλογή μιας τιμής για το λ που θα οδηγήσει σε μια σταθεροποιημένη διακύμανση της τ.μ. $Y^{(\lambda)}$ λαμβάνοντας υπόψη ότι η $f(\lambda, \mu)$ είναι αύξουσα συνάρτηση για $\lambda < 1$.

Η ανωτέρω εξίσωση για τη διακύμανση θυμίζει πολύ τη σχετική εξίσωση που παρουσίασε ο Sarkar (1985) βε βάση τον μετασχηματισμό Box-Cox. Ας υποθέσουμε ότι έχουμε ένα σύνολο παρατηρήσεων και ότι το αντίστοιχο μοντέλο διακρίνεται από ετεροσκεδαστικότητα. Έστω ακόμη ότι

$$\text{Var}[Y] = \sigma^2(\text{E}[Y])^{\epsilon} \tag{1.8}$$

για κάποια μη μηδενική ποσότητα ϵ, με θετική αναμενόμενη τιμή $\text{E}[Y]$ και με τιμή σ^2 που θεωρητικά αντικατοπτρίζει την κοινή διακύμανση (δηλ. τη σταθερά ομοσκεδαστικότητας). Η αιτία της ετεροσκεδαστικότητας είναι προφανώς ο όρος της αναμενόμενης τιμής.

Ακολουθεί η διαδικασία σταθεροποίησης της διακύμανσης κατά Bartlett (1947). Βασισμένη στην ανάπτυξη κατά Taylor, συσχετίζει τη διακύμανση της μετασχηματισμένης μεταβλητής με τη μεταβλητότητα της αρχικής μεταβλητής μέσω της αναμενόμενης τιμής της αρχικής μεταβλητής (χρησιμοποιώντας δειγματικούς εκτιμητές):

$$\text{Var}\left[Y^{(\lambda)}\right] = \text{Var}[Y](\text{E}[Y])^{2(\lambda-1)}. \tag{1.9}$$

Συνδυάζοντας τις δύο τελευταίες εξισώσεις (1.8) και (1.9), καταλήγουμε στην εξής σχέση:

$$\text{Var}\left[Y^{(\lambda)}\right] = \sigma^2(\text{E}[Y])^{2(\lambda-1)+\epsilon}.$$

Η παράμετρος μετασχηματισμού λ μπορεί, επομένως, να επιλεγεί έτσι ώστε

$$2(\lambda - 1) + \epsilon = 0$$

με σκοπό να εξαλειφθεί ο δεύτερος όρος που προκαλεί την ανεπιθύμητη μεταβλητότητα και να επέλθει ομοσκεδαστικότητα Var $\left[Y^{(\lambda)}\right] = \sigma^2$. Άλλες παλαιότερες προτάσεις επί της σχέσης μεταξύ διακύμανσης και μέσης τιμής απαντώνται σε άρθρα οικονομετρίας, μερικά εκ των οποίων αναφέρονται στο άρθρο του Sakia (1992).

1.4 Πρόσφατο Μπεϋζιανό Έργο

Τρία σχετικά πρόσφατα άρθρα ερευνούν το πρόβλημα του μετασχηματισμού δεδομένων στο γραμμικό μοντέλο υπό το πρίσμα της Μπεϋζιανής συλλογιστικής. Και τα τρία αυτά άρθρα ενσωματώνουν κάποια τεχνική επιλογής μεταβλητών.

Στο άρθρο των Hoeting & Ibrahim (1998), προτείνεται μια αρκετά ξεκάθαρη μεθοδολογία για την ταυτόχρονη επιλογή μεταβλητών και μετασχηματισμών. Οι μετασχηματισμοί αφορούν τις τιμές των επεξηγηματικών μεταβλητών μόνο. Ένα σημαντικό μειονέκτημα της προταθείσας διαδικασίας είναι ότι είναι αποδοτική μόνο για μικρό p, δηλαδή μόνο για περιορισμένο αριθμό πιθανών συμμεταβλητών (ενδεικτικά, λιγότερες από 15) καθώς σε αντίθετη περίπτωση είναι υπολογιστικά ασύμφορη. Το ζήτημα της ταυτόχρονης διεξαγωγής των δύο διαδικασιών είναι κομβικό. Μπορεί εύκολα να διαπιστωθεί ότι οι εκτιμήτριες των διαφόρων παραμέτρων (δηλ. των παραμέτρων μετασχηματισμού και των κλασικών συντελεστών παλινδρόμησης) καθώς και η στατιστική σημαντικότητα αυτών επηρεάζονται άμεσα από τη σειρά με την οποία κανείς εκτελεί την επιλογή μεταβλητών και την επιλογή των μετασχηματισμών. Για τους μετασχηματισμούς επιλέγεται η οικογένεια Box-Cox πάνω στις τιμές των συμμεταβλητών, αν και αυτό δεν είναι περιοριστικό και η μεθοδολογία μπορεί να εφαρμοστεί και για άλλες οικογένειες μετασχηματισμών.

Το προβλεπτικό κριτήριο που χρησιμοποιείται στην επιλογή μεταβλητών κατασκευάστηκε από τους Laud & Ibrahim (1995). Εκφράζει την απόσταση των προβλέψεων ενός επαναληπτικού πειράματος (ιδίου πίνακα σχεδιασμού $\mathbf{X}$) από τις τρέχουσες τιμές $\boldsymbol{y}$, συμπεριλαμβάνοντας και τη μεταβλητότητα αυτών. Εμφανώς, όσο μικρότερη είναι η απόσταση αυτή, τόσο πιο βελτιωμένη θεωρείται η προβλεπτική ικανότητα του μοντέλου. Άρα, στη βάση της προβλεπτικής κατανομής του νέου διανύσματος αποκρίσεων $\tilde{\boldsymbol{y}}$ ενός επαναληπτικού πειράματος, το λεγόμενο L_M κριτήριο δίνεται από την παρακάτω εξίσωση:

$$\begin{aligned} L_M &= \left(\mathrm{E}\left[(\tilde{\boldsymbol{y}} - y)'(\tilde{\boldsymbol{y}} - y) \right] \right)^{1/2} \\ &= \Big(\sum_{i=1}^{n} \left(\mathrm{E}\left[\tilde{y}_i \right] - y_i \right)^2 + \mathrm{Var}\left[\tilde{y}_i \right] \Big)^{1/2}. \end{aligned}$$

Εξετάζονται μη πληροφοριακές (Jeffreys) αλλά και πληροφοριακές πρότερες κατανομές με βάση τα δεδομένα (παρόμοιες με αυτές του Zellner). Το μόνο θολό σημείο στην όλη διαδικασία είναι ότι για να κατασκευάσει κανείς την πληροφοριακή πρότερη κατα-

νομή έτσι όπως αυτή προτείνεται, θα πρέπει να μαντέψει αυθαίρετα ένα αρχικό διάνυσμα αποκρίσεων y_0, ενώ υπάρχει και μια ποινή που εμπλέκεται στους υπολογισμούς σχετική με την εμπιστοσύνη που αποδίδει ο ερευνητής στην αυθαίρετη αυτή εικασία που έχει κάνει. Τα αποτελέσματα ξεκάθαρα δείχνουν ευαισθησία του επιλεγόμενου μοντέλου στη βάση της πρότερης αυτής εικασίας του ερευνητή. Ένα σκορ βαθμονόμησης (*calibration score*) για το κριτήριο L_M συνυπολογίζει τη μεταβλητότητα στις τιμές του κριτηρίου για τα διάφορα μοντέλα και συγκεκριμένα την τυπική απόκλιση αυτών.

Ο προτεινόμενος αλγόριθμος αποτελείται από τα ακόλουθα τέσσερα βήματα.

1. Κατασκεύασε τα $|\mathcal{M}_p| = 2^p$ πιθανά μοντέλα δεδομένων των p επεξηγηματικών μεταβλητών. Εδώ ο μοντελοχώρος είναι διακριτός.

2. Για κάθε μοντέλο $M \in \mathcal{M}_p$, βρες το σύνολο (διάνυσμα) μετασχηματισμών της οικογένειας Box-Cox που ελαχιστοποιεί το κριτήριο L_M. Εδώ ο χώρος τιμών της παραμέτρου μετασχηματισμού θεωρείται συνεχής χωρίς αυτό να είναι δεσμευτικό.

3. Για κάθε μοντέλο $M \in \mathcal{M}_p$, υπολόγισε ένα σκορ σύγκρισης βάσει του μοντέλου M^* που ελαχιστοποιεί το κριτήριο L_M μέσω της εξίσωσης:

$$\psi_M = \frac{L_M - L_{M^*}}{S_{L_{M^*}}} = \frac{L_M - L_{M^*}}{\left(\text{Var}\left[L_{M^*}(Y)\right]\right)^{1/2}}.$$

   Αν ο χώρος των μετασχηματισμών είναι συνεχής, τότε η ελαχιστοποίηση επιτυγχάνεται μέσω αριθμητικών μεθόδων.

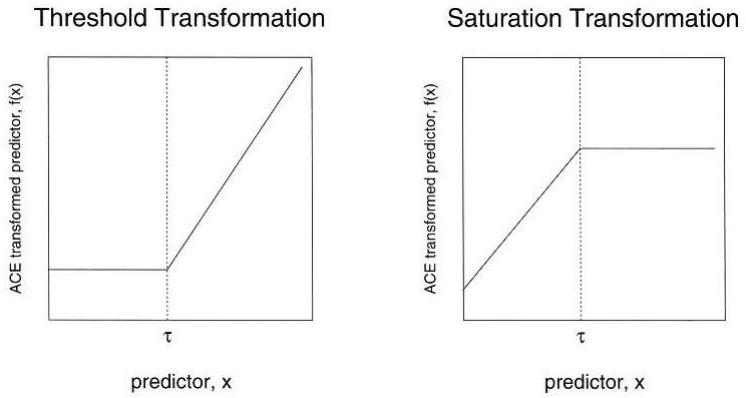
4. Επίλεξε ένα μοντέλο συνυπολογίζοντας την τιμή του κριτηρίου L_M και την τιμή του σκορ σύγκρισης ψ_M (γενικός κανόνας για το σκορ αυτό: να μην υπερβαίνει τις δύο μονάδες).

Επομένως, μπορεί κανείς να επιλέξει ένα μοντέλο που έχει ικανοποιητικά καλή προβλεπτική συμπεριφορά (χαμηλή τιμή L_M) ή να επιλέξει ένα αρκούντως φειδωλό και όχι υπερβολικά πολύπλοκο μοντέλο με χαμηλή τιμή ψ_M. Αν η μεθοδολογία Μπεϋζιανής στάθμισης μοντέλων (*Bayesian model averaging*, BMA) είχε ενσωματωθεί στην προταθείσα μεθοδολογία, τότε η προβλεπτική απόδοση του επιλεγόμενου μοντέλου αναμένεται να ήταν ακόμη πιο βελτιωμένη. Ακόμη, ο βαθμός ευρωστίας του κριτηρίου L δεν έχει ακόμη ελεγχθεί στην περίπτωση που οι προϋποθέσεις του γραμμικού μοντέλου παραβιάζονται. Το-

νίζεται, επίσης, από τους συγγραφείς ότι μια ταυτόχρονη διαδικασία μένει να αναπτυχθεί για την αντιμετώπιση προβλημάτων επιλογής μεταβλητών και μετασχηματισμών ταυτόχρονα με τον έλεγχο των προϋποθέσεων του μοντέλου.

Σε μια μετέπειτα ερευνητική δουλειά, οι Hoeting, Raftery & Madigan (2002) θεωρούν μετασχηματισμούς για τις τιμές των επεξηγηματικών μεταβλητών, καθώς και της μεταβλητής απόκρισης του μοντέλου. Από την άλλη μεριά και σε αντίθεση με το προηγούμενο άρθρο, η έννοια των ταυτόχρονων διαδικασιών δεν περιλαμβάνεται στη δουλειά αυτή. Η κλάση μετασχηματισμών Box-Cox χρησιμοποιείται για τη μεταβλητή απόκρισης, ενώ για τις συμμεταβλητές εξετάζονται οι μετασχηματισμοί σημείου αλλαγής (*change-point transformations*).

Το κύριο πλεονέκτημα των μετασχηματισμών σημείου αλλαγής, εκτός του ότι είναι αρκετά απλοί στη σύλληψη, είναι ότι δεν επιφέρουν αλλαγές στην κλίμακα των συντελεστών παλινδρόμησης $\boldsymbol{\beta}$. Έτσι, η ερμηνεία των αποτελεσμάτων απλοποιείται. Στους μετασχηματισμούς σημείου αλλαγής η αναμενόμενη τιμή της μεταβλητής απόκρισης παραμένει σταθερή κατά προσέγγιση για τιμές των συμμεταβλητών πάνω (ή κάτω) από ένα ορισμένο επίπεδο. Άρα, υπάρχουν δύο εναλλακτικές σε έναν μετασχηματισμό σημείου αλλαγής: ο μετασχηματισμός κατωφλιού (*threshold transformation*) και ο μετασχηματισμός κορεσμού (*saturation transformation*) (Διάγραμμα 1.1). Εντούτοις, μειονέκτημα των μετασχηματισμών αυτών αποτελεί η αποκλειστική χρήση τους αναφορικά με συμμεταβλητές που επιδέχονται μονότονους μετασχηματισμούς.



Διάγραμμα 1.1: a. Μετασχηματισμός κατωφλιού, b. Μετασχηματισμός κορεσμού.

Οι μετασχηματισμοί σημείου αλλαγής ορίζονται μέσα από τις παρακάτω εξισώσεις, όπου η αριστερή εξίσωση αφορά το Διάγραμμα 1.1(a) ενώ η εξίσωση στα δεξιά αφορά το

Διάγραμμα 1.1(b):

$$x_{(\tau)} = \begin{cases} 0, & \text{if } x \leq \tau \\ (x-\tau), & \text{if } x > \tau \end{cases}, \qquad x_{(\tau)} = \begin{cases} (x-\tau), & \text{if } x \leq \tau \\ 0, & \text{if } x > \tau \end{cases}.$$

Μέσα από κάποιες αρχικές γραφικές παραστάσεις μονότονων μετασχηματισμών συναρτήσει των αμετασχημάτιστων δεδομένων, ξεχωρίζουμε τις μεταβλητές των οποίων οι παρατηρήσεις φαίνεται να επιδέχονται μετασχηματισμούς σημείου αλλαγής. Η γενική μορφή ενός τέτοιου μετασχηματισμού παρουσιάζεται παρακάτω, με την παράμετρο τ να δηλώνει τη θέση του σημείου αλλαγής:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x_{(\tau)} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

Παρατηρούμε ότι, σε έναν τέτοιο μετασχηματισμό, η αρχική x καθώς και η μετασχηματισμένη $x_{(\tau)}$ συνυπάρχουν στο τελικό μοντέλο.

Η εκτίμηση του σημείου θέσης τ μπορεί να γίνει με τη βοήθεια ενός προσεγγιστικού παράγοντα Bayes (B_{10}) μεταξύ του μοντέλου με τον μετασχηματισμό σημείου αλλαγής προς το μοντέλο χωρίς τον μετασχηματισμό αυτόν, όπου $\tau_k,\ k = 1, \ldots, \tilde{T}$ είναι μια σειρά από $\tilde{T}$ πιθανές θέσεις για το τ, $\pi(\tau_k)$ είναι η αντίστοιχη πρότερη κατανομή, d είναι η διαφορά των βαθμών ελευθερίας που εμπλέκονται στη σύγκριση των δύο μοντέλων και R^2 είναι ο συντελεστής προσδιορισμού του μοντέλου που αντιστοιχεί στην τιμή τ_k:

$$B_{10} \approx \tilde{T}^{-\frac{d}{2}} \sum_{k=1}^{\tilde{T}} \left[1 - R^2(\tau_k)\right]^{-\frac{\tilde{T}}{2}} \pi(\tau_k).$$

Συνήθως, τα πιθανά σημεία θέσης τίθενται ίσα με τις δειγματικές τιμές της αντίστοιχης επεξηγηματικής μεταβλητής, εξαιρώντας την ελάχιστη και τη μέγιστη τιμή.

Εδώ εφαρμόζεται η τεχνική BMA, με βάση τις εκ των υστέρων πιθανότητες των μοντέλων, η οποία αποτελεί ένα εργαλείο περαιτέρω βελτίωσης της προβλεπτικής ικανότητας του μοντέλου, όπως έχει δειχθεί μέσα από διαδικασίες διασταυρούμενης επικύρωσης (*cross validation techniques*). Η χρήση του αλγόριθμου MC^3 απλοποιεί σημαντικά την εξερεύνηση του μοντελοχώρου. Ο αλγόριθμος αυτός αρχικά εισήχθη από τους Madigan & York (1995). Για περισσότερες πληροφορίες, βλέπε Ntzoufras (2009). Παρόλα αυτά, θέματα σύγκλισης παραμένουν στον αλγόριθμο MC^3, καθώς και ζητήματα επαρκούς εξερεύνησης του μοντελοχώρου.

Όσον αφορά τα μειονεκτήματα της μεθοδολογίας του άρθρου των Hoeting, Raftery & Madigan (2002), θα αναφέρουμε ότι η αλλαγή στην κλίμακα της μεταβλητής απόκρισης λόγω μετασχηματισμού δεν αντιμετωπίζεται με κάποιον τρόπο. Όπως σημειώνουν οι Gottardo & Raftery (2007), η εκτίμηση κάθε σημείου θέσης (τ) εκτελείται μέσα από μια προσεγγιστική διαδικασία η οποία δε λαμβάνει υπόψη την αλλαγή στην κλίμακα λόγω μετασχηματισμών της απόκρισης και έτσι είναι αδύνατο να σταθμιστούν οι εκτιμώμενες τιμές με βάση όλους τους πιθανούς μετασχηματισμούς δύναμης.

Στη συνέχεια, το άρθρο των Gottardo & Raftery (2007) δίνει νέες κατευθύνσεις στο πρόβλημα της επιλογής μοντέλου, καθώς εκτός από την επιλογή μεταβλητής και μετασχηματισμού, αντιμετωπίζει και το θέμα του εντοπισμού ακραίων τιμών μέσω μιας κατανομής Student αγνώστων βαθμών ελευθερίας. Παρόλα αυτά, η έννοια της ταυτόχρονης διαχείρισης αυτών των όψεων του προβλήματος δεν απαντάται εδώ.

Η οικογένεια Box-Cox χρησιμοποιείται για τα δεδομένα της μεταβλητής απόκρισης καθώς και για τις επεξηγηματικές μεταβλητές. Επομένως, μόνο θετικές τ.μ. λαμβάνονται υπόψη, αν και η διαδικασία θα μπορούσε εύκολα να συμπεριλάβει τ.μ. που εκτείνονται στον πραγματικό άξονα τιμών μέσω του μετατοπισμένου μετασχηματισμού ή άλλων εναλλακτικών προσεγγίσεων.

Η βασική συνεισφορά του εν λόγω άρθρου είναι η εισαγωγή των λεγόμενων γενικευμένων συντελεστών παλινδρόμησης (*generalized regression coefficients*), συμβολιζόμενων ως β_j^G, $j = 1, \ldots, p$. Αυτοί οι γενικευμένοι συντελεστές έχουν παρόμοια ερμηνεία με τους συνήθεις συντελεστές παλινδρόμησης, αλλά δεν εξαρτώνται από τους μετασχηματισμούς. Αυτό είναι ιδιαίτερα βοηθητικό όσον αφορά την ερμηνευσιμότητα των μετασχηματισμένων μεταβλητών οι οποίες δεν επηρεάζονται από την επιβεβλημένη αλλαγή κλίμακας. Ας θεωρήσουμε το τυπικό μετασχηματισμένο γραμμικό μοντέλο, όπου λ και $\lambda_j, j = 1, \ldots, p$, είναι παράμετροι μετασχηματισμού προς εκτίμηση που αντιστοιχούν στη μεταβλητή απόκρισης και στις επεξηγηματικές μεταβλητές αντίστοιχα και $X_{ij}^{(\lambda_j)}$ είναι ο μετασχηματισμός του τυχαίου δείγματος της X_j:

$$y_i^{(\lambda)} = \beta_0 + \sum_{j=1}^{p} X_{ij}^{(\lambda_j)} \beta_j + \varepsilon_i, \quad \varepsilon_i | \psi \sim N(0, \psi^{-1}).$$

Ένας επιπλέον παράγοντας ω_i εισάγεται στο μοντέλο ώστε να βοηθήσει στον εντοπισμό

πιθανών ακραίων τιμών:

$$y_i^{(\lambda)} = \beta_0 + \sum_{j=1}^{p} X_{ij}^{(\lambda_j)}\beta_j + \frac{\varepsilon_i}{\sqrt{\omega_i}}, \quad \varepsilon_i|\psi \sim N(0, \psi^{-1}),\ \omega_i \sim \Gamma(\nu/2, \nu/2).$$

Υποθέτουμε ότι τα σφάλματα ε_i και οι όροι ω_i είναι ανεξάρτητα και επομένως το ακόλουθο πηλίκο ακολουθεί Student κατανομή:

$$\frac{\varepsilon_i}{\sqrt{\omega_i}} \sim t_{(\nu,0,\psi^{-1})}.$$

Το πλεονέκτημα αυτού του τεχνάσματος είναι ότι τα δειγματικά σφάλματα δεδομένων των τιμών των ω_i είναι και πάλι κανονικά κατανεμημένα αλλά με διαφορετική διακύμανση. Η πρότερη κατανομή της άγνωστης παραμέτρου ν (βαθμοί ελευθερίας) θεωρείται ομοιόμορφη στο διάστημα $\{1, \ldots, 100\}$. Αναφορικά με τους συνήθεις συντελεστές παλινδρόμησης αφού λάβει χώρα ο μετασχηματισμός, η ερμηνευσιμότητά τους χάνεται εξαιτίας των αλλαγών που επέρχονται στην κλίμακα των μεταβλητών. Άρα, το αποτέλεσμα μιας MCMC διαδικασίας ενδέχεται να μην έχει νόημα από άποψη συμπερασματολογίας. Με βάση τη διάμεσο της μεταβλητής απόκρισης, οι γενικευμένοι συντελεστές παλινδρόμησης κατασκευάζονται όπως φαίνεται παρακάτω. Προτού λάβει χώρα ο μετασχηματισμός, η διάμεσος της μεταβλητής απόκρισης (για κάθε παρατήρηση, παραλείποντας τον δείκτη i) ισούται με:

$$\text{MED}(Y|\mathbf{X}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \Leftrightarrow$$

$$\frac{\text{d MED}(Y|\mathbf{X})}{\text{d}X_j} = \beta_j,\ j = 1, \ldots, p.$$

Μετά τον μετασχηματισμό, για κάθε παρατήρηση i παίρνουμε:

$$\text{MED}\left(\frac{Y^{\lambda}-1}{\lambda}\middle|\mathbf{X}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j\left(\frac{X_j^{\lambda_j}-1}{\lambda_j}\right) \Leftrightarrow$$

$$\text{MED}(Y|\mathbf{X}) = \left[1 + \lambda\left\{\beta_0 + \sum_{j=1}^{p}\beta_j\left(\frac{X_j^{\lambda_j}-1}{\lambda_j}\right)\right\}\right]^{\frac{1}{\lambda}} \Leftrightarrow$$

$$\frac{\text{d MED}(Y|\mathbf{X})}{\text{d}X_j} = \beta_j X_j^{\lambda_j - 1}\left[1 + \lambda\left\{\beta_0 + \sum_{j=1}^{p}\beta_j\left(\frac{X_j^{\lambda_j}-1}{\lambda_j}\right)\right\}\right]^{\frac{1}{\lambda}-1}.$$

Η ποσότητα $\frac{\text{d MED}(Y|\mathbf{X})}{\text{d}X_j}$ δεν εξαρτάται από τους μετασχηματισμούς εφόσον αναφέρεται στα αρχικά δεδομένα και επομένως μπορεί να ερμηνευθεί με κλασικούς όρους, δηλαδή

όπως οι συνήθεις συντελεστές του γραμμικού μοντέλου παλινδρόμησης, στην αρχική κλίμακα των δεδομένων. Άρα, αν θέλαμε να ορίσουμε ένα μέτρο της περιθώριας αλλαγής της τ.μ. Y που προκαλείται από μια αλλαγή στην X_j (ceteris paribus), θα μπορούσαμε να χρησιμοποιήσουμε τη δειγματική μέση τιμή της ανωτέρω ποσότητας η οποία ονομάζεται γενικευμένος συντελεστής παλινδρόμησης β_j^G:

$$\beta_j^G = \frac{1}{n}\sum_{i=1}^{n} \beta_j X_{ij}^{\lambda_j - 1} \left[1 + \lambda \left\{\beta_0 + \sum_{j=1}^{p} \beta_j \left(\frac{X_{ij}^{\lambda_j} - 1}{\lambda_j}\right)\right\}\right]^{\frac{1}{\lambda}-1}.$$

Ο προτεινόμενος αλγόριθμος αποτελείται από πολλαπλά βήματα Metropolis-Hastings (MH) και Gibbs διαδικασιών. Περαιτέρω έρευνα είναι αναγκαία προκειμένου ο αλγόριθμος αυτός να λειτουργήσει πιο αποδοτικά σε περιπτώσεις μεγάλου p καθώς και σε περιπτώσεις πολυσυγγραμικότητας. Ως συνήθως, οι μετασχηματισμοί είναι κατάλληλοι μόνο για αυστηρά θετικές τιμές των μεταβλητών.

1.4.1 Πρότερες κατανομές για την παράμετρο μετασχηματισμού με βάση τη βιβλιογραφία

Ακολουθεί ένας πολύ σύντομος αλλά αντιπροσωπευτικός κατάλογος πρότερων κατανομών που έχουν χρησιμοποιηθεί κατά καιρούς για την παράμετρο μετασχηματισμού λ, αλλά και για τις σχετικές με αυτήν παραμέτρους μ (ή $\boldsymbol{\beta}$ για μοντέλα με συμμεταβλητές) και σ^2, στα πλαίσια της Μπεϋζιανής επιλογής μοντέλου. Σε αυτήν την ενότητα, η παράμετρος λ αναφέρεται μόνο στην οικογένεια Box-Cox, εφόσον η οικογένεια αυτή χρησιμοποιείται σχεδόν αποκλειστικά σε προβλήματα που συμπεριλαμβάνουν μετασχηματισμούς δεδομένων. Επιπλέον, δεν έχει εντοπιστεί δημοσιευμένο υλικό που να παρουσιάζει κάποιου είδους σύγκριση μεταξύ οικογενειών μετασχηματισμών. Για λόγους πληρότητας, ο παρακάτω κατάλογος περιλαμβάνει και κάποια πληροφορία που ενδεχομένως έχει αναφερθεί προηγουμένως στο παρόν κεφάλαιο. Τα επιστημονικά άρθρα παρουσιάζονται σε χρονολογική σειρά και όχι σε σειρά σημαντικότητας. Αυτό επιτρέπει να δωθεί έμφαση στο ότι δεν ανιχνεύεται κάποιο χρονικό μοτίβο σε σχέση π.χ. με την πολυπλοκότητα των πρότερων κατανομών που χρησιμοποιούνται.

Στην πλειοψηφία των περιπτώσεων που ακολουθούν, οι ερευνητές χρησιμοποιούν είτε μια επίπεδη μη γνήσια πρότερη κατανομή για την παράμετρο λ, συνήθως της μορφής

$\pi(\lambda) \propto 1$, είτε μια ομοιόμορφη πρότερη κατανομή $U(-h, h)$ όπου h είναι μια θετική ποσότητα λογικού μεγέθους, είτε και απλά μια διακριτή κατανομή πάνω σε ένα σύνολο διακριτών τιμών για το λ. Τιμές του h που απαντώνται συνηθέστερα είναι $h = 1, h = 4$ και ίσως $h = 20$ ως μια πιο ακραία επιλογή. Η πρότερη κατανομή $\pi(\boldsymbol{\beta}, \sigma^2)$ των παραμέτρων του μετασχηματισμένου γραμμικού μοντέλου έχει ως επί το πλείστον μια μορφή σχετική με την πρότερη κατανομή του Jeffreys.

◇ **Box & Cox (1964)** Στα πλαίσια του γραμμικού μοντέλου με p συμμεταβλητές (δηλαδή με $p + 1$ συντελεστές στο μοντέλο συμπεριλαμβανομένου του σταθερού όρου), η πρότερη κατανομή έχει τη γενική μορφή:

$$\pi(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) \cdot \sigma^{-1} \cdot g(\boldsymbol{y})^{-(\lambda-1)(p+1)},$$

όπου $g(\boldsymbol{y})$ είναι ο γεωμετρικός μέσος των παρατηρήσεων και ο όρος $\pi(\lambda)$ λαμβάνεται ανάλογος της μονάδας. Ισοδύναμα,

$$\pi(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) \cdot \sigma^{-1} \cdot |J(\boldsymbol{y}, \lambda)|^{-(p+1)/n},$$

όπου $|J(\boldsymbol{y}, \lambda)| = \prod_{i=1}^{n} y_i^{\lambda-1}$ είναι η συνήθης Ιακωβιανή του μετασχηματισμού Box-Cox.

◇ **Pericchi (1981)** Σε μια παρόμοια λογική με τους Box και Cox, ο Pericchi κατασκεύασε την πρότερη κατανομή

$$\pi(\boldsymbol{\beta}, \sigma, \lambda) \propto \pi(\lambda) \cdot \sigma^{-(p+2)}$$

η οποία δεν έχει εξάρτηση από τα δεδομένα. Επίσης, θεωρεί ότι η κατανομή $\pi(\lambda)$ είναι ομοιόμορφη στον χώρο τιμών $\mathcal{M}_\lambda$ της παραμέτρου μετασχηματισμού.

◇ **Sweeting (1984)** Η χρησιμοποιούμενη πρότερη κατανομή έχει τη μορφή:

$$\pi(\boldsymbol{\beta}, \sigma, \lambda) \propto \left\{(1 + \lambda\beta_0)^{(p+1-\beta_0)/\lambda}\, \sigma\right\}^{-1} \pi(\lambda)$$

όπου $\pi(\lambda) \propto 1$, β_0 είναι ο σταθερός όρος του μοντέλου και $\boldsymbol{\beta}$ είναι ένα $p \times 1$ διάνυσμα συντελεστών των επεξηγηματικών μεταβλητών. Ακόμη, μια ομοιόμορφη κατανομή για την παράμετρο λ αναφέρεται, παρότι δε χρησιμοποιείται στις εφαρμογές που παρουσιάζονται.

◇ **De Oliveira, Kedem & Short (1997)** Οι ερευνητές, αν και αναφέρονται εκτενώς στις πρότερες κατανομές των Box-Cox και του Pericchi και συζητούν θέματα για τη συνέπεια μιας πρότερης κατανομής, τελικά χρησιμοποιούν την κατανομή:

$$\pi(\lambda) = U(-3, 3).$$

◇ **Hoeting & Ibrahim (1998)** Αναφέρεται η χρήση ενός διακριτού συνόλου μετασχηματισμών δύναμης $\alpha = (-1, 0, 0.5, 1, 2)$ κάτω από μη πληροφοριακές πρότερες κατανομές για το διάνυσμα παραμέτρων $(\boldsymbol{\beta}, \sigma^{-1})$.

◇ **Hoeting, Raftery & Madigan (2002)** Εδώ, όσον αφορά τη μεταβλητή απόκρισης στο γραμμικό μοντέλο με διάνυσμα παραμέτρων $\boldsymbol{\beta}$ διάστασης $(p+1)$ και άγνωστη διακύμανση σφαλμάτων σ^2 επιβάλλονται μετασχηματισμοί Box-Cox, ενώ μετασχηματισμοί σημείου αλλαγής εφαρμόζονται για τις τιμές των επεξηγηματικών μεταβλητών. Λόγω περιορισμένης πρότερης πληροφορίας, επιλέγεται η συνήθης κλάση των κανονικών-γάμμα συζυγών κατανομών (βλέπε σελίδες 490-491 του σχετικού άρθρου):

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{V}), \quad \frac{\nu\lambda}{\sigma^2} \sim \chi^2_\nu$$

όπου $\nu, \lambda, \mathbf{V}_{(p+1)\times(p+1)}, \mu_{(p+1)\times 1}$ είναι υπερπαράμετροι που χρειάζεται να προσδιοριστούν. Οι συντελεστές $\beta_i, i = 1, \ldots, p$ των επεξηγηματικών μεταβλητών θεωρούνται εκ των προτέρων ανεξάρτητοι και

$$\mu = (\hat{\beta}_0, 0, \ldots, 0)$$

με $\hat{\beta}_0$ τον εκτιμητή ελαχίστων τετραγώνων του σταθερού όρου. Ακόμη,

$$\mathbf{V} = \sigma^2 \cdot \mathbf{I}_D(s_Y^2, \phi^2 s_1^{-2}, \phi^2 s_2^{-2,\ldots,\phi^2 s_p^{-2}})$$

όπου $\mathbf{I}_D()$ συμβολίζει έναν διαγώνιο πίνακα με το πρώτο στοιχείο να είναι η δειγματική διακύμανση της μεταβλητής απόκρισης, s_i^2 να είναι η δειγματική διακύμανση της τ.μ. X_i και ϕ να είναι μια ακόμη υπερπαράμετρος.

Μια σημαντική παρατήρηση είναι ότι οι ανωτέρω πρότερες κατανομές εξαρτώνται από τα δεδομένα. Για παράδειγμα, οι τιμές

$$\nu = 2.58, \lambda = 0.28, \phi = 2.85$$

στο άρθρο επιλέγονται με βάση ένα κριτήριο μεγιστοποίησης του $\sigma^2 \leq 1$, ενώ παράλληλα απαιτείται η πιθανότητα του διανύσματος συντελεστών $\boldsymbol{\beta}$ να είναι επίπεδη πάνω στον μοναδιαίο υπερκύβο $[-1, 1]^p$ και η πιθανότητα του σ^2 να είναι επίπεδη πάνω στο διάστημα $(\epsilon, 1)$ για κάποια μικρή τιμή του ϵ. Η παράμετρος μετασχηματισμού λ επιλέγεται από το διακριτό σύνολο

$$(-1, 0, 0.5, 1).$$

Σχετικά με τον μοντελοχώρο, όλα τα μοντέλα θεωρούνται ως εκ των προτέρων ισοπίθανα.

◇ **Lee et al. (2005)** Ξανά επιλέγεται μια ομοιόμορφη κατανομή

$$\pi(\lambda) = U(-4, 4)$$

για την παράμετρο μετασχηματισμού λ.

◇ **Fan, Wang & Balakrishnan (2008)** Και πάλι χρησιμοποιείται μια κατανομή

$$\pi(\lambda) = U(-h, h)$$

όπου h μια φραγμένη θετική πραγματική τιμή.

◇ **Wang (2008)** Σε αυτή την ερευνητική εργασία,

$$\pi(\lambda) = U(-0.5, 1.2)$$

και

$$\pi(\lambda) = U(0, 1.2)$$

για έναν έλεγχο ευαισθησίας. Γενικά, αν $\boldsymbol{\beta}$ είναι το διάνυσμα συντελεστών του μοντέλου, τότε $\pi(\boldsymbol{\beta}) \propto c$ όπου c είναι μια σταθερά.

◇ **Klein Entink, van der Linden & Fox (2009)** Οι ερευνητές χρησιμοποιούν την απλή πρότερη κατανομή

$$\pi(\lambda) = U(-4, 4).$$

◇ **Gottardo & Raftery (2009)** Χρησιμοποιείται

$$\pi(\lambda) = U[-1, 1],$$

καθώς και

$$\pi(\lambda_j) = U[-1, 1], j = 1, \ldots, p$$

επιτρέποντας μετασχηματισμούς των δεδομένων των επεξηγηματικών μεταβλητών καθώς και της μεταβλητής απόκρισης. Προτείνεται ακόμη ότι, για τον έλεγχο της αναγκαιότητας ενός μετασχηματισμού, θα μπορούσε κανείς να χρησιμοποιήσει μια κατανομή μοναδιαίας μάζας στην τιμή $\lambda = 1$ και μια συνεχή κατανομή $\pi(\lambda)$. Το κύριο μοντέλο που χρησιμοποιούν οι ερευνητές είναι:

$$y_i^{(\lambda)} = \beta_0 + \sum_{j=1}^{p} X_{ij}^{(\lambda_j)} \beta_j + \varepsilon_i, \quad \varepsilon_i | \psi \sim N(0, \psi^{-1}).$$

Μια μίξη μιας κανονικής κατανομής και μιας κατανομής μοναδιαίας μάζας στο μηδέν χρησιμοποιείται για περαιτέρω μοντελοποίηση των συντελεστών:

$$\beta_j | \lambda, \lambda_j, \sigma_\beta \sim (1 - \omega)\delta_0 + \omega N\left(0, \frac{S^2_{g_\lambda(\mathbf{Y})}}{S^2_{g_{\lambda_j}(\mathbf{X}_j)}} \sigma^2_\beta\right), \; j = 1, \ldots, p$$

όπου w είναι το πρότερο βάρος της μεταβλητής που θα ενσωματωθεί στο μοντέλο και θεωρητικά ακολουθεί μια βήτα κατανομή, $S^2_{\mathbf{z}}$ είναι η δειγματική διακύμανση του διανύσματος $\mathbf{z}$ και σ^2_β είναι μια παράμετρος κοινής διακύμανσης ομοιόμορφη στο διάστημα $[0, 1]$. Επίσης, χρησιμοποιείται η πρότερη κατανομή

$$\beta_0 | \lambda \propto [S_{g_\lambda}(\mathbf{Y})]^{-1}$$

ώστε να διορθωθεί η αλλαγή που επέρχεται στην κλίμακα λόγω της παραμέτρου λ. Τέλος, $\pi(\psi) \propto \psi^{-1}$. Προτείνεται και μια πιο εύρωστη εναλλακτική του ανωτέρω πλαισίου, αν και η κατασκευή των πρότερων κατανομών δεν αλλάζει στον πυρήνα της.

◇ **Freni & Mannina (2010)** Στο συγκεκριμένο άρθρο, οι ερευνητές δεν ενδιαφέρονται ιδιαίτερα για την πρότερη κατανομή της παραμέτρου μετασχηματισμού, αλλά μόνο για την πρότερη κατανομή που σχετίζεται με μεταβλητές κύριου ενδιαφέροντος (όπου θεωρούν εμπειρικές, ομοιόμορφες και κανονικές/λογαριθμοκανονικές κατανομές). Αναφέρεται λακωνικά ότι ακολουθείται η προσέγγιση των Box και Cox.

◇ **Yang, Christensen & Sorensen (2011)** Η πρότερη κατανομή της παραμέτρου λ είναι

$$\pi(\lambda) = U(-3, 3).$$

1.5 Συμπεράσματα

Θα πρέπει να αναγνωριστεί ότι η αλματώδης ανάπτυξη των υπολογιστικών συστημάτων έχει συνεισφέρει ποικιλοτρόπως στην αντιμετώπιση ιδιαίτερα πολύπλοκων υπολογιστικών προβλημάτων, όπως αυτό της ταυτόχρονης επιλογής μεταβλητών και μετασχηματισμών δεδομένων, για να μην επεκταθούμε στην επιλογή και άλλων δομικών χαρακτηριστικών των στατιστικών μοντέλων (π.χ. συνδετικών συναρτήσεων).

Είναι ευρέως αποδεκτό ότι ο εντοπισμός παρατηρήσεων υψηλής επιρροής (*influential cases*) ή/και ακραίων τιμών είναι υψίστης σημασίας, καθώς η ύπαρξή τους ενδέχεται να οδηγήσει σε παραπλανητικά αποτελέσματα (π.χ. σε ακατάλληλες τιμές για τις εμπλεκόμενες παραμέτρους μετασχηματισμού), ενώ μετά την εφαρμογή ενός μετασχηματισμού μπορεί να μην ανιχνεύονται πλέον. Εκτός από το άρθρο των Gottardo & Raftery (2007) που περιγράφτηκε νωρίτερα, ελάχιστα άρθρα έχουν δημοσιευθεί τα οποία να συνδυάζουν τις έννοιες της ανίχνευσης ακραίων τιμών και του μετασχηματισμού, βλέπε για παράδειγμα Attkinson (1982*b*), Attkinson (1982*a*), Attkinson (1985), Attkinson (1986) και Cook & Wang (1983).

Στο Κεφάλαιο 2 που ακολουθεί, συγκεκριμένες οικογένειες μετασχηματισμών από αυτές που ανιχνεύθηκαν στη βιβλιογραφία επιλέγονται με σκοπό την ανάπτυξη και παρουσίαση μιας καινοτόμου μεθοδολογίας σχετικά με το πρόβλημα της επιλογής μετασχηματισμού.

Κεφάλαιο 2

Μπεϋζιανή Επιλογή Οικογένειας Μετασχηματισμών για Μονομεταβλητά Προβλήματα

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο, αναπτύσσεται με λεπτομέρειες η Μπεϋζιανή προσέγγιση του προβλήματος της επιλογής μετασχηματισμού. Εστιάζουμε στη Μπεϋζιανή συμπερασματολογία για συγκεκριμένες οικογένειες μετασχηματισμών και συγκεκριμένους μετασχηματισμούς μέσα σε αυτές, καθώς επίσης και στον προσδιορισμό των πρότερων κατανομών και τον υπολογισμό της εκ των υστέρων κατανομής.

Όσον αφορά στην επιλογή μοντέλου, ακολουθείται μια διαδικασία δύο βημάτων. Αρχικά, επιλέγουμε τη βέλτιστη οικογένεια μετασχηματισμών T για ένα δεδομένο σύνολο παρατηρήσεων και, σε ένα δεύτερο στάδιο, επιλέγουμε τη βέλτιστη τιμή της παραμέτρου μετασχηματισμού λ_T δεδομένης της οικογένειας T.

Το υπόλοιπο κεφάλαιο είναι οργανωμένο ως ακολούθως. Στην Ενότητα 2.2 περιγράφεται το ερευνητικό κίνητρο της παρούσας ερευνητικής δουλειάς με ειδική μνεία σε σημαίνοντα άρθρα της βιβλιογραφίας, ενώ δίνεται έμφαση στα καινοτόμα στοιχεία της μεθοδολογίας που αναπτύσσουμε. Η Ενότητα 2.3 ξεδιαλέγει τις συγκεκριμένες οικογένειες μετασχηματισμών με τις οποίες θα ασχοληθούμε στο κεφάλαιο αυτό και αποκαλύπτει τις

μεταξύ τους διαφορές και ομοιότητες. Η Ενότητα 2.4 καταπιάνεται με την προσέγγιση της Μπεϋζιανής συμπερασματολογίας και επιλογής μετασχηματισμού. Ο προσδιορισμός της πρότερης κατανομής παρουσιάζεται λεπτομερώς σε αυτή την ενότητα στη βάση δύο διαφορετικών προσεγγίσεων. Η Ενότητα 2.5 περιέχει κάποιες συμπερασματικές παρατηρήσεις και πιθανές ερευνητικές επεκτάσεις υπό εξέταση.

2.2 Ερευνητικό Κίνητρο

Η κανονικότητα αποτελεί θεμελιώδη προϋπόθεση για μια πληθώρα στατιστικών μοντέλων, από τους απλούς στατιστικούς ελέγχους υποθέσεων μέχρι την ανάλυση γραμμικής παλινδρόμησης, οικονομετρικά μοντέλα και προβλήματα ελέγχου ποιότητας. Παρότι ο περιορισμός της κανονικότητας μπορεί πλέον να αρθεί στη σύγχρονη Μπεϋζιανή ανάλυση, καθώς πιο εξελιγμένες μέθοδοι ή μη-παραμετρικές προσεγγίσεις έχουν αναπτυχθεί, η προϋπόθεση της κανονικότητας μπορεί να απλουστεύσει σημαντικά και να επιταχύνει την υπολογιστική διαδικασία εξαιτίας της σύνδεσής της με συζυγείς μορφές πρότερων κατανομών. Κάτι τέτοιο είναι ιδιαίτερα βοηθητικό σε μεγάλα σύνολα δεδομένων, τα οποία αποκτούν ολοένα και πιο κεντρικό ενδιαφέρον στη σύγχρονη επιστήμη. Παρομοίως, όταν τα δεδομένα του προβλήματος καταφθάνουν ανά τακτά χρονικά διαστήματα, η διαδοχική τους ανάλυση είναι αρκετά άμεση σε συζυγείς περιπτώσεις, βλέπε για παράδειγμα το άρθρο των Zamba, Tsiamyrtzis & Hawkins (2008) για μια Μπεϋζιανή εφαρμογή στα πλαίσια του ελέγχου ποιότητας. Επιπλέον, παρόλο που η βιβλιογραφία αφθονεί σε κλασικούς ελέγχους κανονικότητας (Razali & Wah 2011, Thode 2002, για παράδειγμα), φαίνεται να υπάρχει έλλειψη ομόλογων Μπεϋζιανών διαδικασιών. Ένας απλός τρόπος για να διενεργήσουμε έναν τέτοιο έλεγχο είναι να εξετάσουμε την ανάγκη (ή μη) για έναν μετασχηματισμό των δεδομένων υπό τη Μπεϋζιανή συλλογιστική. Με αυτήν τη λογική, η προτεινόμενη μεθοδολογία αυτού του κεφαλαίου μπορεί κάλλιστα να χρησιμοποιηθεί σαν ένας Μπεϋζιανός έλεγχος κανονικότητας.

Υπό τη Μπεϋζιανή ομπρέλα, θα εντρυφήσουμε στο πρόβλημα της επιλογής οικογένειας μετασχηματισμού. Τέσσερις μονοπαραμετρικές οικογένειες μετασχηματισμών λαμβάνονται υπόψη με σκοπό την κανονικοποίηση μιας μεταβλητής απόκρισης Y. Συγκεκρι-

μένα, η μεθοδολογία του παρόντος κεφαλαίου περιλαμβάνει τις οικογένειες *Box-Cox* (Box & Cox 1964), *Modulus* (John & Draper 1980), *Yeo & Johnson* (Yeo & Johnson 2000) και *Dual* (Yang 2006).

Αρκετοί ερευνητές μέχρι στιγμής έχουν συνεισφέρει στην περιοχή των Μπεϋζιανών μετασχηματισμένων μοντέλων, όπως πήραμε ήδη μια γεύση στο Κεφάλαιο 1. Σε αυτό το σημείο, θα δώσουμε συνοπτική πληροφορία για κάποιες περασμένες ερευνητικές δουλειές με ιδιαίτερο ενδιαφέρον για την μεθοδολογία που αναπτύξαμε και που ακολουθεί αργότερα. Στο άρθρο του, ο Pericchi (1981) θεώρησε το γραμμικό μοντέλο παλινδρόμησης μετασχηματισμένο κατά Box & Cox (1964) και πρότεινε μια μη πληροφοριακή πρότερη κατανομή η οποία δεν εξαρτάται από τα εκάστοτε δεδομένα. Κατά αυτόν τον τρόπο, κατάφερε όχι μόνο να αντλήσει τη βέλτιστη τιμή της παραμέτρου μετασχηματισμού που αντιστοιχεί στην κανονικότητα, αλλά και να διορθώσει θέματα σχετικά με τις προϋποθέσεις της ομοσκεδαστικότητας και της προσθετικότητας. Ο Sweeting (1984, 1985) επίσης διερεύνησε μια μη-εμπειρική πρότερη κατανομή της βασικής παραμέτρου μετασχηματισμού λ υπό την οικογένεια Box-Cox σε περίπτωση που υπάρχει ασαφής (επίπεδη) πρότερη πληροφορία για τις υπόλοιπες παραμέτρους του μοντέλου, αλλά επιπλέον ισχυρίστηκε ότι η μεθοδολογία του απέφυγε κάποιες ανεπιθύμητες ιδιότητες της μεθόδου του Pericchi. Κατά βάση, ενδιαφέρθηκε για το ζήτημα της μη-ταυτοποίησης (non-identifiability) σε μια γειτονιά του λ λαμβάνοντας υπόψη ότι οι παράμετροι του μοντέλου πρέπει να είναι a priori ανεξάρτητες του λ για κάθε τιμή λ_0 σε μια γειτονιά του λ.

Στο άρθρο τους, οι Hoeting, Raftery & Madigan (2002) μελέτησαν πολυμεταβλητά προβλήματα, πάλι στα πλαίσια του Μπεϋζιανού κανονικού γραμμικού μοντέλου. Η μεθοδολογία τους ενσωματώνει την ταυτόχρονη επιλογή μεταβλητών και μετασχηματισμών. Οι τιμές των επεξηγηματικών μεταβλητών μετασχηματίστηκαν μέσω ενός μετασχηματισμού σημείου αλλαγής, ενώ για τη μεταβλητή απόκρισης χρησιμοποιήθηκε η οικογένεια Box-Cox. Δεδομένου ότι ο κύριος στόχος τους ήταν η βελτιστοποίηση της προβλεπτικής ικανότητας του τελικού μοντέλου, εφαρμόστηκε η τεχνική Μπεϋζιανής στάθμισης μοντέλων BMA μέσω του αλγόριθμου MC^3 (Madigan & York 1995) ώστε να μειωθεί η αβεβαιότητα που συνοδεύει το εκάστοτε στατιστικό μοντέλο. Μια επίσης ενδιαφέρουσα προσέγγιση προτάθηκε από τους Gottardo & Raftery (2009) και συνδυάζει την επιλογή μοντέ-

λου, την επιλογή μετασχηματισμών με βάση την οικογένεια Box-Cox και την ανίχνευση ακραίων τιμών ταυτόχρονα. Όπως ήδη έχει αναφερθεί στο Κεφάλαιο 1, η έννοια των γενικευμένων συντελεστών παλινδρόμησης εισάγεται ώστε να λυθούν προβλήματα ερμηνείας και συμπερασματολογίας που ανακύπτουν λόγω της ύπαρξης μετασχηματισμών στο τελικό μοντέλο. Ακριβώς επειδή οι νέοι αυτοί συντελεστές αποτελούν παραμέτρους ανεξάρτητες από μετασχηματισμούς, διευκολύνουν τη στατιστική συμπερασματολογία καθώς επιδέχονται της ίδιας ερμηνείας με τους κλασικούς συντελεστές παλινδρόμησης στην κλίμακα μέτρησης των αρχικών μη μετασχηματισμένων δεδομένων.

Στη διεθνή βιβλιογραφία, ο όρος *επιλογή μετασχηματισμού* αφορά την επιλογή της βέλτιστης τιμής της παραμέτρου μετασχηματισμού στα πλαίσια μιας συγκεκριμένης οικογένειας (ως επί το πλείστον της οικογένειας Box-Cox). Μέχρι στιγμής, δε γνωρίζουμε να υπάρχει κάποια δημοσιευμένη επιστημονική εργασία, Μπεϋζιανή ή μη, στην οποία να αντιπαρατίθενται ή/και να συγκρίνονται διαφορετικές οικογένειες μετασχηματισμών. Η συνεισφορά της παρούσας ερευνητικής δουλειάς αφορά κυρίως την επέκταση της έννοιας της επιλογής μετασχηματισμού ώστε να συμπεριλάβει και τη διαδικασία σύγκρισης οικογενειών μετασχηματισμών. Πιο συγκεκριμένα, εισάγουμε μια διαδικασία δύο βημάτων κατά την οποία σε ένα πρώτο επίπεδο επιλέγεται η οικογένεια μετασχηματισμών και σε ένα δεύτερο επίπεδο προσδιορίζεται η τιμή της παραμέτρου μετασχηματισμού δεδομένης της οικογένειας του πρώτου βήματος. Λόγω του ότι η προτεινόμενη μεθοδολογία εντάσσεται στα πλαίσια της Μπεϋζιανής συλλογιστικής, είναι πολύ σημαντική η επιλογή κατάλληλων πρότερων κατανομών. Στην περίπτωσή μας, η διαδικασία αυτή περιπλέκεται περισσότερο καθώς οι πρότερες κατανομές της παραμέτρου μετασχηματισμού λ_T για κάθε οικογένεια μετασχηματισμού T οφείλουν να είναι συμβατές μεταξύ τους ώστε να είναι σύμφωνες με τη διαφορετική ερμηνεία του λ_T δεδομένου του T. Επομένως, η συμβατότητα των πρότερων κατανομών είναι ένα θεμελιώδες ζήτημα στο πρόβλημα της οικογένειας μετασχηματισμού έτσι όπως τίθεται στα πλαίσια της παρούσας διατριβής.

2.3 Οικογένειες Μετασχηματισμών

Οι εξής τέσσερις οικογένειες μετασχηματισμών έχουν επιλεγεί προς σύγκριση: Box-Cox, Modulus, Yeo & Johnson και Dual οι οποίες είναι όλες μονοπαραμετρικές όπως περιγράψαμε αναλυτικά στο Κεφάλαιο 1. Εδώ, θεωρούμε τον αντίστοιχο παραμετρικό χώρο ως συνεχή ώστε να επιτύχουμε μεγαλύτερη ακρίβεια για τους παραγόμενους εκτιμητές των μοντέλων, αν και έτσι αυξάνει η υπολογιστική πολυπλοκότητα.

Ας παρουσιάσουμε τον βασικό συμβολισμό του κεφαλαίου. Κάθε οικογένεια συνοδεύεται από έναν δείκτη μετασχηματισμού T και περιέχει μια παράμετρο μετασχηματισμού λ_T. Συμβολίζουμε με $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ τα αρχικά (παρατηρούμενα) δεδομένα και με $\boldsymbol{y}^{(\lambda_T)} = \left(y_1^{(\lambda_T)}, \ldots, y_n^{(\lambda_T)}\right)^\top$ τα μετασχηματισμένα δεδομένα για μια συγκεκριμένη τιμή του λ_T δεδομένης μιας συγκεκριμένης οικογένειας μετασχηματισμού T. Σκοπός είναι να αναγνωρίσουμε ποιο $\boldsymbol{y}^{(\lambda_T)}$ μπορεί να υποτεθεί ότι αποτελεί δείγμα από μια κανονική κατανομή με παραμέτρους (μ_T, σ_T^2) για κάποια κατάλληλη τιμή της παραμέτρου μετασχηματισμού λ_T και για κατάλληλη οικογένεια μετασχηματισμών T.

Για οποιαδήποτε οικογένεια T, η πιθανοφάνεια των αρχικών δεδομένων $\boldsymbol{y}$ καθορίζεται πλήρως μέσω του αντίστροφου μετασχηματισμού $\boldsymbol{y}^{(\lambda_T)} \to \boldsymbol{y}$. Αποτελείται, δηλαδή, από την πιθανοφάνεια των μετασχηματισμένων δεδομένων πολλαπλασιασμένη επί την απόλυτη τιμή της αντίστοιχης Ιακωβιανής $\left|J(\boldsymbol{y}, \lambda_T|T)\right| = \prod_{i=1}^n \left|\frac{\partial y_i^{(\lambda_T)}}{\partial y_i}\right|$. Επομένως, η εν λόγω πιθανοφάνεια δίνεται από τη σχέση:

$$
\begin{aligned}
f\left(\boldsymbol{y}|\mu_T, \sigma_T^2, \lambda_T, T\right) &= \\
&= \left(2\pi\sigma_T^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_T^2}\sum_{i=1}^n \left(y_i^{(\lambda_T)} - \mu_T\right)^2\right) \times \prod_{i=1}^n \left|\frac{\partial y_i^{(\lambda_T)}}{\partial y_i}\right|. \quad (2.1)
\end{aligned}
$$

Οι μαθηματικοί τύποι των υπό σύγκριση οικογενειών μετασχηματισμών με τις αντίστοιχες Ιακωβιανές ορίζουσες $|J(\boldsymbol{y}, \lambda_T|T)|$ σε απόλυτη τιμή συνοψίζονται στον Πίνακα 2.1. Ο *Ταυτοτικός* (Id) και ο *Λογαριθμικός* (Log) μετασχηματισμός έχουν προστεθεί στις οικογένειες προς σύγκριση αυξάνοντας το πλήθος τους σε έξι. Ο μετασχηματισμός Id υποδηλώνει ότι δεν υπάρχει αναγκαιότητα μετασχηματισμού των δεδομένων εφόσον τα αρχικά δεδομένα (και ο όποιος γραμμικός μετασχηματισμός αυτών) περιγράφονται επαρκώς από την κανονική κατανομή.

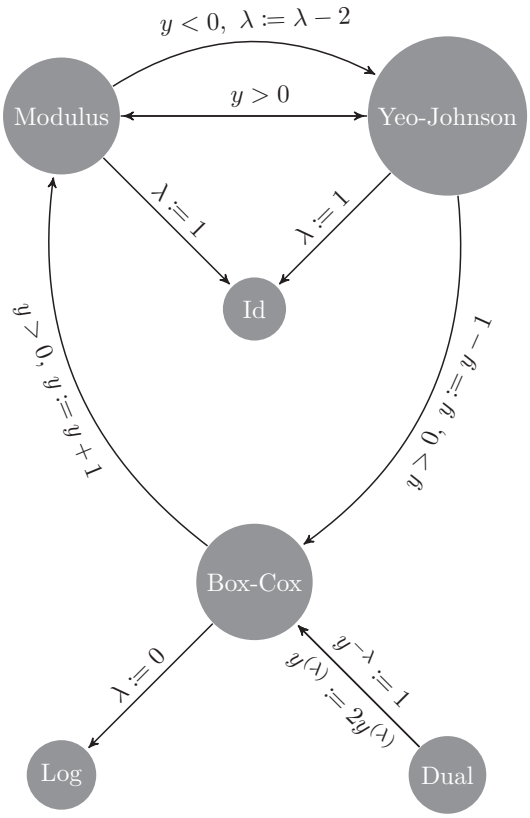
Όσον αφορά τους μετασχηματισμούς BC, Dual και Log, για να διαχειριστούμε δεδο-

Πίνακας 2.1: Οι έξι οικογένειες μετασχηματισμών και οι αντίστοιχες Ιακωβιανές $\left|J(\boldsymbol{y}, \lambda_T|T)\right|$ σε απόλυτη τιμή. Όπου δεν ορίζεται διαφορετικά, $y_i \in \mathbb{R}$.

Family T	$y_i^{(\lambda_T)}$	$\left\|J(\boldsymbol{y}, \lambda_T\|T)\right\|$
Id	$= y_i$	$= 1$
Log	$= \log(y_i)\ , \quad y_i > 0$	$= \prod_{i=1}^{n} \left(y_i^{-1}\right)$
Box-Cox	$= \begin{cases} \frac{y_i^{\lambda_T}-1}{\lambda_T}\,, & \lambda_T \neq 0 \\ \log(y_i)\,, & \lambda_T = 0 \end{cases} \quad y_i > 0$	$= \prod_{i=1}^{n} \left(y_i^{\lambda_T - 1}\right)$
Modulus	$= \begin{cases} \frac{\text{sign}(y_i)\left[(\lvert y_i\rvert+1)^{\lambda_T}-1\right]}{\lambda_T}, & \lambda_T \neq 0 \\ \text{sign}(y_i)\log\left(\lvert y_i\rvert+1\right), & \lambda_T = 0 \end{cases}$	$= \prod_{i=1}^{n} \left(\lvert y_i\rvert + 1\right)^{\lambda_T - 1}$
Yeo & Johnson	$= \begin{cases} \frac{(y_i+1)^{\lambda_T}-1}{\lambda_T}\,, & y_i \geq 0, \lambda_T \neq 0 \\ \log(y_i+1)\,, & y_i \geq 0, \lambda_T = 0 \\ -\frac{(-y_i+1)^{2-\lambda_T}-1}{2-\lambda_T}\,, & y_i < 0, \lambda_T \neq 2 \\ -\log(-y_i+1)\,, & y_i < 0, \lambda_T = 2 \end{cases}$	$= \begin{cases} \prod_{i=1}^{n} (y_i+1)^{\lambda_T - 1}\ , & y_i \geq 0 \\ \prod_{i=1}^{n} (-y_i+1)^{1-\lambda_T}\ , & y_i < 0 \end{cases}$
Dual	$= \begin{cases} \frac{y_i^{\lambda_T}-y_i^{-\lambda_T}}{2\lambda_T}\,, & \lambda_T > 0 \\ \log(y_i)\,, & \lambda_T = 0 \end{cases} \quad y_i > 0$	$= \prod_{i=1}^{n} \frac{y_i^{\lambda_T - 1}+y_i^{-\lambda_T - 1}}{2}$

Σημείωση: $\text{sign}(y_i) = -1$ για αρνητικό y_i και $\text{sign}(y_i) = +1$ για θετικό y_i.

μένα που ανήκουν στο σύνολο $\mathbb{R}$ των πραγματικών αριθμών μπορούμε να μετατοπίσουμε τα δεδομένα προς τα δεξιά προσθέτοντας μια αρκούντως μεγάλη σταθερά $\xi > |\min(\boldsymbol{y})|$. Μια απλή προσέγγιση θα ήταν να θέσουμε τη σταθερά αυτή ίση με $\xi = |\min(\boldsymbol{y})| + \epsilon$, όπου ϵ συμβολίζει μια μικρή θετική ποσότητα. Απουσία μετατόπισης αντιστοιχεί στην τιμή $\xi = 0$. Υπενθυμίζεται ότι ο μετασχηματισμός Modulus των John & Draper (1980) αίρει τον περιορισμό για θετικές παρατηρήσεις και προτείνεται ως κατάλληλος όταν προϋπάρχει κάποια συμμετρία. Αντικαθιστώντας το y_i με $y_i + 1$ περνάμε από την οικογένεια Box-Cox στην οικογένεια Modulus για θετικό y_i. Η οικογένεια Yeo-Johnson επίσης λειτουργεί με πραγματικές τιμές του διανύσματος απόκρισης και προτείνεται για ασύμμετρα



Διάγραμμα 2.1: Σύνδεση των υπό μελέτη οικογενειών μετασχηματισμών. Το σύμβολο $:=$ υποδηλώνει αντικατάσταση του αριστερού μέρους από το δεξί μέρος.

δείγματα παρατηρήσεων. Για θετικό y_i, ο μετασχηματισμός YJ είναι ισοδύναμος με τον μετασχηματισμό Modulus και επομένως και με τον Box-Cox αν κάθε y_i αντικατασταθεί με $y_i - 1$. Τέλος, η οικογένεια Dual δέχεται επίσης μόνο θετικές παρατηρήσεις και δεν υπάρχει ουδέτερη τιμή για την παράμετρο λ_T που να αντιστοιχεί στον ταυτοτικό μετασχηματισμό σε αντίθεση με την τιμή της μονάδας που αντιστοιχεί στις υπόλοιπες υπό μελέτη οικογένειες. Εμπειρικά στοιχεία, με βάση κανονικά δείγματα ποικίλου μεγέθους, δείχνουν ότι η παράμετρος λ_T τείνει να κινείται στο διάστημα $1.10 - 1.30$. Για τιμές του λ_T κοντά στο μηδέν, ο μετασχηματισμός Dual προσεγγίζει τον μετασχηματισμό Box-Cox.

Κατασκευάσαμε το Διάγραμμα 2.1 ώστε να απεικονίσουμε με συνεκτικό και γλαφυρό τρόπο τις σχέσεις μεταξύ των διαφόρων οικογενειών μετασχηματισμών. Ας σημειωθεί ότι, πέρα από την οικογένεια Box-Cox, η διαδικασία μετατόπισης των παρατηρήσεων στον θετικό άξονα μπορεί να εφαρμοστεί και στην περίπτωση των μετασχηματισμών Dual και

Log αν χρειαστεί.

2.4 Μπεϋζιανή Μοντελοποίηση

Σε αυτήν την ενότητα, μπαίνουμε στον πυρήνα του τρέχοντος κεφαλαίου και περιγράφουμε το πρόβλημα επιλογής μετασχηματισμού με αυστηρά Μπεϋζιανούς όρους. Έμφαση δίνεται στη Μπεϋζιανή συμπερασματολογία και στον μηχανισμό επιλογής της βέλτιστης οικογένειας μετασχηματισμών αφού γίνει ο προσδιορισμός κατάλληλων πρότερων κατανομών και η εξαγωγή της εκ των υστέρων κατανομής. Σχετικά με την επιλογή μοντέλου, ακολουθείται μια διαδικασία δύο βημάτων. Σε πρώτη φάση επιλέγεται η βέλτιστη οικογένεια μετασχηματισμών για δεδομένο σύνολο παρατηρήσεων και σε δεύτερη φάση επιλέγεται η βέλτιστη τιμή της παραμέτρου μετασχηματισμού λ_T δεδομένης της οικογένειας $T \in \mathcal{T}$, όπου

$$\mathcal{T} = \{\text{Id}, \text{Log}, \text{BC}, \text{Mod}, \text{YJ}, \text{Dual}\} \tag{2.2}$$

είναι ο μοντελοχώρος των οικογενειών.

2.4.1 Διαμόρφωση των πρότερων κατανομών

Στα πλαίσια της Μπεϋζιανής προσέγγισης, η κατασκευή μιας πρότερης κατανομής για τον δείκτη του μοντελοχώρου $T \in \mathcal{T}$ όπως ορίστηκε στην (2.2) καθώς και για το διάνυσμα παραμέτρων $\boldsymbol{\theta}_T = (\mu_T, \sigma_T^2, \lambda_T)^\top$, υπό την εκάστοτε οικογένεια T, παίζει θεμελιώδη ρόλο στην όλη διαδικασία μοντελοποίησης. Σημειώνεται, σε αυτό το σημείο, ότι οι παράμετροι μ_T, σ_T^2 θα έπρεπε να εμπεριέχουν και έναν δείκτη λ_T πέραν του δείκτη T, αλλά αυτό αποφεύχθηκε για λόγους απλοποίησης του συμβολισμού.

Όσον αφορά την πρότερη πιθανότητα εμφάνισης κάθε μιας εκ των έξι οικογενειών μετασχηματισμών, χρησιμοποιείται μια διακριτή ομοιόμορφη κατανομή στον χώρο $\mathcal{T}$ ώστε να δηλώσουμε την a priori άγνοιά μας:

$$\pi(T) = \frac{1}{|\mathcal{T}|} = \frac{1}{6}\,. \tag{2.3}$$

Η ομοιόμορφη κατανομή στον μοντελοχώρο υιοθετείται συχνά σαν σημείο αναφοράς. Η βιβλιογραφία που σχετίζεται με άλλες μορφές πρότερων κατανομών για τον μο-

ντελοχώρο είναι περιορισμένη σε προβλήματα επιλογής μεταβλητών (βλέπε για παράδειγμα Cui & George (2008) και Dellaportas, Forster & Ntzoufras (2012)), σε προβλήματα πολλαπλών συγκρίσεων (βλέπε για παράδειγμα Scott & Berger (2006) και Scott & Berger (2010)) και πρόσφατα σε προβλήματα ανάλυσης κατά συστάδες (*cluster analysis*) (Casella, Moreno & Girón (2014)). Θεωρούμε ότι η επιλογή μας είναι συνετή για δύο λόγους:

(i) στην περίπτωση όπου τα δεδομένα υποστηρίζουν την κανονική κατανομή (η οποία αποτελεί ειδική περίπτωση στις περισσότερες οικογένειες μετασχηματισμών υπό σύγκριση), τότε ο μετασχηματισμός Id θα υποστηρίζεται σθεναρά εκ των υστέρων εφόσον ο παράγοντας Bayes επιβάλλει ποινή για κάθε επιπλέον παράμετρο που εισάγεται στο μοντέλο αλλά και για αποκλίσεις από την πρότερη κατανομή η οποία είναι κεντραρισμένη στο κανονικό μοντέλο σύμφωνα με τη μέθοδό μας,

(ii) στην περίπτωση όπου δύο οικογένειες μετασχηματισμών ίδιας διάστασης παραμέτρων είναι ισοδύναμες, τότε αναμένουμε να λάβουν εκ των υστέρων ίσο βάρος, το οποίο δε θα μπορούσε να συμβεί αν τα πρότερα βάρη δεν ήταν ίσα κατανεμημένα.

Για την πρότερη κατανομή των παραμέτρων $\boldsymbol{\theta}_T$ του μετασχηματισμένου μοντέλου χρησιμοποιούμε την ακόλουθη ιεραρχική δομή:

$$\pi(\boldsymbol{\theta}_T|T) = \pi\left(\mu_T, \sigma_T^2 \middle| \lambda_T, T\right) \pi(\lambda_T|T).$$

Σχετικά με τις παραμέτρους θέσης και κλίμακας (μ_T, σ_T^2), για λόγους συζυγίας, χρησιμοποιούμε την κανονική-αντίστροφη-γάμμα (NIG) πρότερη κατανομή με την ακόλουθη ιεραρχική μορφή:

$$\begin{aligned} \pi\left(\mu_T, \sigma_T^2|\lambda_T, T\right) &= NIG\left(\mu_T, \sigma_T^2; \mu_0, k_0^{-1}, \alpha_0, \beta_0\right) \\ &= N\left(\mu_T; \mu_0, \frac{\sigma_T^2}{k_0}\right) IG\left(\sigma_T^2; \alpha_0, \beta_0\right). \end{aligned} \tag{2.4}$$

Οι τυπικές τιμές των εμπλεκόμενων υπερπαραμέτρων, υπό την απουσία πρότερης πληροφορίας, είναι $\mu_0 = 0$, $k_0 = 0.001$, $\alpha_0 = 0.001$ και $\beta_0 = 0.001$.

Για τη βασική παράμετρο λ_T προτείνουμε τη χρήση δύο ξεχωριστών πρότερων κατανομών, τις οποίες περιγράφουμε αναλυτικά στις δύο επόμενες υποενότητες. Υπάρχουν

δύο κομβικά ζητήματα που ελήφθησαν υπόψη για τον σχηματισμό των κατανομών αυτών. Από τη μία μεριά, υπάρχει το ζήτημα της συμβατότητας των πρότερων κατανομών του λ_T το οποίο είναι δύσκολο εξαιτίας της διαφορετικής ερμηνείας της εν λόγω παραμέτρου ανάλογα με την εκάστοτε οικογένεια. Επί παραδείγματι, η τιμή $\lambda_T = 0$ λαμβάνει διαφορετικό νόημα ανάλογα με την οικογένεια στην οποία αναφέρεται και μπορεί να αντιστοιχεί στον λογάριθμο των αρχικών δεδομένων ή στον αρνητικό λογάριθμο των μετατοπισμένων δεδομένων (βλέπε Πίνακα 2.1). Παρομοίως, η τιμή $\lambda_T = 1$ ενδέχεται να οδηγεί στον Ταυτοτικό μετασχηματισμό για τις οικογένειες Modulus και YJ ή σε μια απλή μετατόπιση των δεδομένων κατά μια ποσότητα ή ακόμη και σε μη τετριμμένο μετασχηματισμό (δηλαδή διάφορο του Ταυτοτικού) στην περίπτωση της οικογένειας Dual. Επομένως, οι πρότερες κατανομές που σχετίζονται με το λ_T θα πρέπει να έχουν ένα κοινό υπόβαθρο για όλους τους δείκτες $T \in \mathcal{T}$. Το ευρέως γνωστό παράδοξο των Lindley-Bartlett (Lindley 1957, Bartlett 1957) είναι ένα δεύτερο ζήτημα που αξίζει προσοχής και συνδέεται με την επιλογή των πρότερων κατανομών. Σύμφωνα με αυτό, η σύγκριση μοντέλων είναι ευαίσθητη στην επιλογή της a priori διακύμανσης εφόσον πολύ μεγάλη διασπορά ενδέχεται να επιφέρει παραπλανητικά αποτελέσματα υποστηρίζοντας το αξίωμα του απλούστερου μοντέλου και να αναδείξει πλασματικά τον Ταυτοτικό ή τον Λογαριθμικό μετασχηματισμό ανεξάρτητα από το πού οδηγούν τα δεδομένα του εκάστοτε προβλήματος.

Για τους ανωτέρω λόγους, υιοθετήθηκε η προσέγγιση της πρότερης κατανομής δύναμης στη βάση ενός συνόλου φανταστικών δεδομένων $\boldsymbol{y}^*$ (Ibrahim & Chen 2000). Έτσι, η συμβατότητα μεταξύ των διαφορετικών οικογενειών εισάγεται αυτόματα χρησιμοποιώντας ένα κοινό διάνυσμα φανταστικών δεδομένων αφού οι πρότερες κατανομές δεν είναι τίποτε άλλο παρά ενημερωμένες (*rescaled*) ύστερες κατανομές βασισμένες στα φανταστικά δεδομένα $\boldsymbol{y}^*$. Παρόμοιες προσεγγίσεις έχουν εφαρμοστεί στα πλαίσια γνωστών προβλημάτων επιλογής μοντέλου, όπως η ιδέα της g-πρότερης κατανομής του Zellner (1986), και στα γραφικά μοντέλα (Ntzoufras & Tarantola 2013, για παράδειγμα). Ορισμένες ενδιαφέρουσες ιδιότητες της κλάσης των πρότερων κατανομών δύναμης περιγράφονται στο άρθρο των Ibrahim, Chen & Sinha (2003). Παράλληλα με την πρώτη αυτή προσέγγιση, χρησιμοποιείται εναλλακτικά μια κανονική πρότερη κατανομή μοναδιαίας πληροφορίας (ή μια λογαριθμοκανονική πρότερη κατανομή στην περίπτωση του μετα-

σχηματισμού Dual) όπου πάλι επικαλούμαστε το σύνολο φανταστικών δεδομένων $\boldsymbol{y}^*$. Η εναλλακτική αυτή προσέγγιση απλοποιεί σημαντικά το υπολογιστικό κόστος σε σχέση με την πρώτη προσέγγιση, αφού προσεγγίζεται μόνο ένα ολοκλήρωμα αντί για δύο για την εκτίμηση της περιθώριας κατανομής κάτω από την εκάστοτε οικογένεια μετασχηματισμών. Περισσότερες λεπτομέρειες ακολουθούν στη συνέχεια.

2.4.1.1 Πρότερη κατανομή δύναμης για το λ_T

Στην ενότητα αυτή, χρησιμοποιούμε την έννοια της πρότερης κατανομής δύναμης των Ibrahim & Chen (2000) ώστε να προσδιορίσουμε την πρότερη κατανομή για την παράμετρο μετασχηματισμού λ_T. Η πρότερη κατανομή δύναμης που προκύπτει συμβολίζεται Prior A.

Κάτω από οποιαδήποτε οικογένεια T με συγκεκριμένη παράμετρο λ_T, η πιθανοφάνεια ενός διανύσματος παρατηρήσεων $\boldsymbol{y}$ μεγέθους n, περιθωριοποιημένη ως προς τις υπόλοιπες παραμέτρους εκτός του λ_T, με βάση τις (2.1) και (2.4) δίνεται από:

$$\begin{aligned} f(\boldsymbol{y}|\lambda_T, T) &= \int f\left(\boldsymbol{y}\middle|\mu_T, \sigma_T^2, \lambda_T, T\right) \pi(\mu_T, \sigma_T^2|T) d\mu_T d\sigma_T^2 \\ &= f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\lambda_T, T\right) \times \prod_{i=1}^{n} \left|\frac{\partial y_i^{(\lambda_T)}}{\partial y_i}\right| \end{aligned} \tag{2.5}$$

με

$$f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\lambda_T, T\right) \propto \left(\beta_0 + \frac{(n-1)S_T^2}{2} + \frac{nk_0\left(\overline{\boldsymbol{y}^{(\lambda_T)}} - \mu_0\right)^2}{2(k_0+n)}\right)^{-\left(\alpha_0+\frac{n}{2}\right)} \tag{2.6}$$

όπου S_T^2 είναι η δειγματική διακύμανση των μετασχηματισμένων δεδομένων. Περισσότερες λεπτομέρειες για τον ακριβή υπολογισμό της (2.5) ακολουθούν αμέσως.

Η περιθώρια πιθανοφάνεια των αρχικών παρατηρήσεων δεδομένου του (λ_T, T) είναι:

$$\begin{aligned} &f\left(\boldsymbol{y}|\lambda_T, T\right) = f\left(\boldsymbol{y}^{(\lambda_T)}|\lambda_T, T\right) \cdot \left|J(\boldsymbol{y}, \lambda_T|T)\right| \\ &= \left|J(\boldsymbol{y}, \lambda_T|T)\right| \iint f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\mu_T, \sigma_T^2, \lambda_T, T\right) \cdot \pi\left(\mu_T, \sigma_T^2|\lambda_T, T\right) \, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\ &= \left|J(\boldsymbol{y}, \lambda_T|T)\right| \iint f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\mu_T, \sigma_T^2, \lambda_T, T\right) \cdot N\left(\mu_T|\mu_0, \frac{\sigma_T^2}{k_0}\right) IG\left(\sigma_T^2|\alpha_0, \beta_0\right) \, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\ &= \left|J(\boldsymbol{y}, \lambda_T|T)\right| \iint \left(2\pi\sigma_T^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_i \left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\sigma_T^2}\right) \cdot \left(2\pi\sigma_T^2\right)^{-\frac{1}{2}} k_0^{\frac{1}{2}} \end{aligned}$$

$$\cdot \exp\left(-\frac{k_0(\mu_T-\mu_0)^2}{2\sigma_T^2}\right)\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\sigma_T^2\right)^{-(\alpha_0+1)}\exp\left(-\frac{\beta_0}{\sigma_T^2}\right)\mathrm{d}\mu_T\mathrm{d}\sigma_T^2$$

$$= \left|J(\boldsymbol{y},\lambda_T|T)\right|\cdot\overbrace{(2\pi)^{-\frac{n+1}{2}}k_0^{\frac{1}{2}}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}}^{C}\iint\left(\sigma_T^2\right)^{-\frac{n+1}{2}-\alpha_0-1}\exp\left(-\frac{\beta_0}{\sigma_T^2}\right)\exp\left(-\frac{k_0(\mu_T-\mu_0)^2}{2\sigma_T^2}\right)$$

$$\exp\left(-\frac{\overbrace{\sum_{i=1}^{n}\left(y_i^{(\lambda_T)}-\overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2}^{(n-1)S_T^2}}{2\sigma_T^2}\right)\cdot\exp\left(-\frac{n\left(\overline{\boldsymbol{y}^{(\lambda_T)}}-\mu_T\right)^2}{2\sigma_T^2}\right)\mathrm{d}\mu_T\mathrm{d}\sigma_T^2$$

$$= \left|J(\boldsymbol{y},\lambda_T|T)\right|\cdot C\int\left(\sigma_T^2\right)^{-\frac{n+1}{2}-\alpha_0-1}\exp\left(-\frac{2\beta_0+(n-1)S_T^2}{2\sigma_T^2}\right)$$

$$\cdot\int\exp\left(-\frac{n\left(\overline{\boldsymbol{y}^{(\lambda_T)}}-\mu_T\right)^2+k_0(\mu_T-\mu_0)^2}{2\sigma_T^2}\right)\mathrm{d}\mu_T\mathrm{d}\sigma_T^2$$

$$= \left|J(\boldsymbol{y},\lambda_T|T)\right|\cdot C\int\left(\sigma_T^2\right)^{-\frac{n+1}{2}-\alpha_0-1}\exp\left(-\frac{2\beta_0+(n-1)S_T^2}{2\sigma_T^2}\right)$$

$$\cdot\int\exp\left(-\frac{(k_0+n)\mu_T^2+\left(-2k_0\mu_0-2n\overline{\boldsymbol{y}^{(\lambda_T)}}\right)\mu_T+k_0\mu_0^2+n\overline{\boldsymbol{y}^{(\lambda_T)}}^2}{2\sigma_T^2}\right)\mathrm{d}\mu_T\mathrm{d}\sigma_T^2$$

$$= \left|J(\boldsymbol{y},\lambda_T|T)\right|\cdot C\int\left(\sigma_T^2\right)^{-\frac{n+1}{2}-\alpha_0-1}\exp\left(-\frac{2\beta_0+(n-1)S_T^2}{2\sigma_T^2}\right)$$

$$\cdot\int\exp\left(-\frac{(k_0+n)\left[\mu_T^2-2\left(\frac{k_0\mu_0+n\overline{\boldsymbol{y}^{(\lambda_T)}}}{k_0+n}\right)\mu_T\pm\left(\frac{k_0\mu_0+n\overline{\boldsymbol{y}^{(\lambda_T)}}}{k_0+n}\right)^2\right]+k_0\mu_0^2+n\overline{\boldsymbol{y}^{(\lambda_T)}}^2}{2\sigma_T^2}\right)$$

$$\mathrm{d}\mu_T\mathrm{d}\sigma_T^2$$

$$= \left|J(\boldsymbol{y},\lambda_T|T)\right|\cdot C\int\left(\sigma_T^2\right)^{-\frac{n+1}{2}-\alpha_0-1}\exp\left(-\frac{2\beta_0+(n-1)S_T^2}{2\sigma_T^2}\right)$$

$$\cdot\int\exp\left(-\frac{(k_0+n)\left(\mu_T-\frac{k_0\mu_0+n\overline{\boldsymbol{y}^{(\lambda_T)}}}{k_0+n}\right)^2+\frac{nk_0\left(\mu_0-\overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2}{k_0+n}}{2\sigma_T^2}\right)\mathrm{d}\mu_T\mathrm{d}\sigma_T^2$$

$$= \left|J(\boldsymbol{y},\lambda_T|T)\right|\cdot C\int\left(\sigma_T^2\right)^{-\frac{n+1}{2}-\alpha_0-1}\exp\left(-\frac{2\beta_0+(n-1)S_T^2+\frac{nk_0\left(\mu_0-\overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2}{k_0+n}}{2\sigma_T^2}\right)$$

$$\cdot \left(\int \exp\left(-\frac{(k_0+n)\left(\mu_T - \frac{k_0\mu_0 + n\overline{\boldsymbol{y}^{(\lambda_T)}}}{k_0+n}\right)^2}{2\sigma_T^2} \right) \mathrm{d}\mu_T \right) \mathrm{d}\sigma_T^2$$

$$= \frac{\left|J(\boldsymbol{y}, \lambda_T|T)\right| \cdot C}{(k_0+n)^{-\frac{1}{2}}} \int \left(\sigma_T^2\right)^{-\frac{n+2\alpha_0}{2}-1} \exp\left(-\frac{2\beta_0 + (n-1)S_T^2 + \frac{nk_0\left(\mu_0 - \overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2}{k_0+n}}{2\sigma_T^2} \right) \mathrm{d}\sigma_T^2$$

$$= \frac{\left|J(\boldsymbol{y}, \lambda_T|T)\right| \cdot C}{(k_0+n)^{-\frac{1}{2}}} \Gamma\left(\frac{n+2\alpha_0}{2}\right) \left(\beta_0 + \frac{(n-1)S_T^2}{2} + \frac{nk_0\left(\mu_0 - \overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2}{2(k_0+n)} \right)^{-\frac{n+2\alpha_0}{2}}$$

$$= \left|J(\boldsymbol{y}, \lambda_T|T)\right| \cdot C' \left(\beta_0 + \frac{(n-1)S_T^2}{2} + \frac{nk_0\left(\mu_0 - \overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2}{2(k_0+n)} \right)^{-\frac{n+2\alpha_0}{2}}$$

με C' να είναι η κοινή σταθερά που διατηρείται αναλλοίωτη μεταξύ των οικογενειών μετασχηματισμών.

Τώρα, ας συμβολίσουμε με $\boldsymbol{y}^*$ ένα σύνολο φανταστικών δεδομένων μεγέθους n^*. Η πρότερη κατανομή δύναμης του λ_T, υπό την οικογένεια T, ορίζεται ως εξής:

$$\pi_A\left(\lambda_T|\boldsymbol{y}^*, T\right) \propto f\left(\boldsymbol{y}^*|\lambda_T, T\right)^\delta \pi^N(\lambda_T|T). \tag{2.7}$$

Υψώνουμε την ποσότητα $f\left(\boldsymbol{y}^*\middle|\lambda_T, T\right)$, που είναι η περιθωριοποιημένη συνάρτηση πιθανοφάνειας δεδομένου του λ_T για τα φανταστικά δεδομένα $\boldsymbol{y}^*$, σε μια δύναμη $0 < \delta \leq 1$. Η παράμετρος δ ονομάζεται παράμετρος δύναμης (*power parameter*) και η υψωμένη συνάρτηση στη δ ονομάζεται πιθανοφάνεια δύναμης (*power likelihood*). Η παράμετρος δ ουσιαστικά μειώνει την επίδραση των φανταστικών δεδομένων. Θεωρούμε ότι η παράμετρος δ είναι ίση με τον αντίστροφο του μεγέθους δείγματος n^* των φανταστικών δεδομένων ώστε να εξασφαλίσουμε ότι η επίδραση του $\boldsymbol{y}^*$ στην πρότερη κατανομή δύναμης θα έχει μοναδιαία βαρύτητα. Επιπλέον, για να ορίσουμε πλήρως τη πρότερη κατανομή δύναμης, ξεκινάμε από μια πρότερη κατανομή βάσης ή αναφοράς $\pi^N(\lambda_T|T)$ για την παράμετρο λ_T. Αυτή η κατανομή λαμβάνεται συνήθως διάχυτη ώστε να αντανακλά την απουσία γνώσης για το λ_T προτού παρατηρηθούν τα φανταστικά δεδομένα. Εδώ επιλέγουμε $\pi^N(\lambda_T|T) = N\left(\lambda_T|m_0, s_0^2\right)$, με $m_0 = 1$ και $s_0 = 100$. Χρησιμοποιώντας τη (2.6), η τελική μορφή της πρότερης κατανομής δύναμης για το λ_T, υπό την οικογένεια T, είναι

η ακόλουθη:

$$\pi_A\left(\lambda_T|\boldsymbol{y}^*,T\right) \propto \left(\beta_0+\frac{(n^*-1)S_T^{*^2}}{2}+\frac{n^*k_0\left(\overline{\mathbf{y}^{*(\lambda_T)}}-\mu_0\right)^2}{2(k_0+n^*)}\right)^{-\left(\frac{\alpha_0}{n^*}+\frac{1}{2}\right)} \times\left(\prod_{i=1}^{n^*}\left|\frac{\partial y_i^{*(\lambda_T)}}{\partial y_i^*}\right|\right)^{\frac{1}{n^*}} \times N\left(\lambda_T|m_0,s_0^2\right) \tag{2.8}$$

όπου $S_T^{*^2}$ είναι η δειγματική διακύμανση των μετασχηματισμένων φανταστικών δεδομένων.

Σχετικά με το διάνυσμα $\boldsymbol{y}^*$, ιδανικά αντιπροσωπεύει ιστορικά δεδομένα από παλαιότερες σχετικές έρευνες ή πληροφορία από εξειδικευμένους ερευνητές της εκάστοτε ερευνητικής περιοχής. Σε περίπτωση που κάτι τέτοιο δεν είναι διαθέσιμο, μπορεί κανείς να θεωρήσει φανταστικά δεδομένα που να υποστηρίζουν τη μηδενική υπόθεση της έρευνας ή κάποιο μοντέλο αναφοράς, συνάδοντας με τη λεγόμενη δύσπιστη πρότερη προσέγγιση (*skeptical prior*) όπως την περιέγραψαν οι Spiegelhalter, Abrams & Myles (2004). Εδώ προτείνουμε τα δεδομένα $\boldsymbol{y}^*$ να προέρχονται από μια κανονική κατανομή $N\left(\mu_s,\sigma_s^2\right)$ η οποία αντικατοπτρίζει τον Ταυτοτικό μετασχηματισμό που θα θέλαμε ιδανικά να επιλέξουμε. Μια εναλλακτική ιδέα θα ήταν να χρησιμοποιήσουμε τα πραγματικά δεδομένα $\boldsymbol{y}$ ως φανταστικά με αποτέλεσμα τη δημιουργία μιας εμπειρικής πρότερης κατανομής ελάχιστης επιρροής (βλέπε Ntzoufras 2009). Σε κάθε περίπτωση, τυποποιούμε τα αρχικά δεδομένα $\boldsymbol{y}$ πριν αυτά υποβληθούν σε μετασχηματισμό. Συνεπώς, είναι εύλογο να επιλέξουμε $\mu_s=0,\sigma_s^2=1$ για την κατανομή των φανταστικών δεδομένων. Αν πρότερη πληροφορία για την παράμετρο λ_T είναι διαθέσιμη, τότε η πληροφορία αυτή μπορεί να ενσωματωθεί στην a priori κατανομή χρησιμοποιώντας τα φανταστικά δεδομένα μέσω μιας τεχνικής αντίστροφου μετασχηματισμού με βάση τη σειρά Taylor (βλέπε Ενότητα 3.3.1).

Όσον αφορά το σχήμα της, η πρότερη κατανομή δύναμης δεν παρουσιάζει συμμετρία. Παρόλα αυτά, η αντίστοιχη κορυφή είναι αρκετά σταθερή και ακριβής. Συχνά, η κατανομή $\pi_A\left(\lambda_T|\boldsymbol{y}^*,T\right)$ δεν έχει κλειστή μορφή και επομένως η σταθερά κανονικοποίησης είναι προς εκτίμηση μέσω εναλλακτικών υπολογιστικών εργαλείων.

2.4.1.2 Κανονική πρότερη κατανομή για το λ_T με ερμηνεία μοναδιαίας πληροφορίας

Θέλοντας να φτάσουμε σε μια κλειστή μορφή της a priori κατανομής της παραμέτρου λ_T, σε αντίθεση με την (2.8), εισάγουμε ένα εναλλακτικό πλαίσιο πρότερης κατανομής (Prior B). Προς απλούστευση της διαδικασίας διαμόρφωσης του μοντέλου, προτιμήθηκε μια κανονική πρότερη κατανομή με ερμηνεία μοναδιαίας πληροφορίας κατά προσέγγιση, ως μια κατανομή χαμηλής πληροφορίας. Κατά αυτόν τον τρόπο, ο υπολογισμός της περιθώριας πιθανοφάνειας γίνεται απλούστερος σε σχέση με την προσέγγιση της Ενότητας 2.4.1.1, εφόσον μόνο ένα ολοκλήρωμα μένει να εκτιμηθεί αντί για δύο (βλέπε Ενότητα 2.4.3 για λεπτομέρειες). Κάτω από τις οικογένειες Box-Cox, Modulus και Yeo & Johnson, εισάγουμε μια κανονική κατανομή $\pi_B\left(\lambda_T|\boldsymbol{y}^*, T\right) = N\left(\lambda_T|\mu_{\lambda_T}, \sigma^2_{\lambda_T}, T\right)$ με μέσο μ_{λ_T} και διακύμανση $\sigma^2_{\lambda_T}$. Υπό την οικογένεια Dual, η κανονική a priori κατανομή αφορά την παράμετρο $\log \lambda_T$ αντί της λ_T ώστε ο χώρος τιμών της παραμέτρου μετασχηματισμού να επεκταθεί σε όλο τον πραγματικό άξονα. Συνεπώς, αποκλειστικά για τη συγκεκριμένη οικογένεια, η παράμετρος λ_T ακολουθεί a priori μια λογαριθμοκανονική κατανομή $LN\left(\lambda_T|\mu_{\log\lambda_T}, \sigma^2_{\log\lambda_T}, T\right)$. Ο πρότερος μέσος ισούται με μονάδα και αντιστοιχεί στη μηδενική υπόθεση της κανονικότητας του $\boldsymbol{y}$ τουλάχιστον για τις τρεις πρώτες οικογένειες μετασχηματισμών (Box-Cox, Modulus, Yeo & Johnson). Σύμφωνα με εμπειρικά στοιχεία που παραγάγαμε με βάση προσομοιωμένα δεδομένα από την τυπική κατανομή με διάφορα μεγέθη δείγματος, η τιμή του λ_T που αντιστοιχεί στην κανονικότητα για την οικογένεια Dual κυμαίνεται κοντά στο $\widehat{\lambda}_D = 1.2$. Στα πλαίσια ενοποίησης των εξισώσεων, εισάγουμε μια νέα παράμετρο για το υπόλοιπο της παρούσας ενότητας:

$$\widetilde{\lambda}_T = \begin{cases} \lambda_T, & \text{για } T = \text{BC, Mod, YJ} \\ \log \lambda_T, & \text{για } T = \text{Dual} \end{cases} \tag{2.9}$$

και επομένως η προτεινόμενη πρότερη κατανομή για το $\widetilde{\lambda}_T$, υπό την εκάστοτε οικογένεια T, παίρνει την ακόλουθη μορφή:

$$\pi_B\left(\widetilde{\lambda}_T|\boldsymbol{y}^*, T\right) = N\left(\widetilde{\lambda}_T|\mu_{\widetilde{\lambda}_T}, \sigma^2_{\widetilde{\lambda}_T}, T\right) \tag{2.10}$$

με

$$\mu_{\widetilde{\lambda}_T} = \begin{cases} 1, & \text{για } T = \text{BC, Mod, YJ} \\ \log\left(\widehat{\lambda}_D\right), & \text{για } T = \text{Dual} \end{cases}. \tag{2.11}$$

Όσον αφορά την τυπική απόκλιση δεδομένης της οικογένειας T $\left(\sigma_{\widetilde{\lambda}_T}\right)$, απορρέει με βάση την παρατηρούμενη πληροφορία κατά Fisher της παραμέτρου ενδιαφέροντος για ένα σύνολο φανταστικών δεδομένων $\boldsymbol{y}^*$ η οποία υπολογίζεται στη μέση τιμή της παραμέτρου μετασχηματισμού, δηλαδή:

$$\sigma_{\widetilde{\lambda}_T} = \left[-\frac{\partial^2}{\partial \widetilde{\lambda}_T^2} \log f\left(\boldsymbol{y}^* \middle| \widetilde{\lambda}_T, T\right)^{1/n^*} \Bigg|_{\widetilde{\lambda}_T = \mu_{\widetilde{\lambda}_T}} \right]^{-\frac{1}{2}}. \tag{2.12}$$

Στη (2.12), η πιθανοφάνεια δεδομένου του $\widetilde{\lambda}_T$ για τα φανταστικά δεδομένα $\boldsymbol{y}^*$ υψώνεται στη δύναμη $\delta = (n^*)^{-1}$ ώστε να σχηματίσει μια πρότερη κατανομή μοναδιαίας πληροφορίας. Η πρότερη κατανομή που προκύπτει με αυτή την προσέγγιση μπορεί να φανεί χρήσιμη σαν προκαθορισμένη επιλογή κατά την απουσία πρότερης πληροφορίας, χρησιμοποιώντας μια παράμετρο δύναμης με τιμή $1/n^*$, όπως στην (2.12), και φανταστικά δεδομένα προερχόμενα από μια κανονική κατανομή, σύμφωνα πάντα με τη *skeptical prior* προσέγγιση των Spiegelhalter et al. (2004). Από την άλλη μεριά, όταν διατίθεται πρότερη πληροφορία, μπορεί να αξιοποιηθεί και να ενσωματωθεί στην πρότερη κατανομή μέσω των φανταστικών δεδομένων με κατάλληλα επιλεγμένη τιμή της παραμέτρου δύναμης.

Η τυπική απόκλιση για κάθε οικογένεια T δίνεται από τη σχέση

$$\sigma_{\widetilde{\lambda}_T} = \left(-\frac{q_T}{n^*} + \left[\frac{S^2_{\boldsymbol{w}_T - \boldsymbol{d}_T} + S_{\boldsymbol{z}_T \boldsymbol{r}_T}}{S^2_{\boldsymbol{z}_T}} - 2\left(\frac{S_{\boldsymbol{z}_T \boldsymbol{w}_T} - S_{\boldsymbol{z}_T \boldsymbol{d}_T}}{S^2_{\boldsymbol{z}_T}} \right)^2 \right] \right)^{-\frac{1}{2}} \tag{2.13}$$

όπου η δειγματική (αμερόληπτη) διακύμανση του $\mathbf{x}$ συμβολίζεται με $S^2_{\mathbf{x}}$ και η δειγματική συνδιακύμανση μεταξύ $\mathbf{x}$ και $\boldsymbol{y}$ συμβολίζεται με $S_{\mathbf{x}\boldsymbol{y}}$. Βλέπε Ενότητες 2.4.1.3 και 2.4.1.4 για τον αναλυτικό προσδιορισμό της (2.13) υπό τις οικογένειες Box-Cox και Dual (για τις υπόλοιπες οικογένειες οι υπολογισμοί είναι παρόμοιοι με αυτούς που αφορούν την οικογένεια Box-Cox και παραλείπονται). Το μετασχηματισμένο διάνυσμα $\boldsymbol{z}_T$ δίνεται από

$$\boldsymbol{z}_T = \begin{cases} (\boldsymbol{y}^* + \xi \mathbb{1}_{n^*})^{(\lambda_T)}, & \text{όπου } \lambda_T = 1, \text{ για } T = \text{BC} \\ (\boldsymbol{y}^* + \xi \mathbb{1}_{n^*})^{(\lambda_T)}, & \text{όπου } \lambda_T = \widehat{\lambda}_D, \text{ για } T = \text{Dual} \\ \boldsymbol{y}^{*(\lambda_T)}, & \text{όπου } \lambda_T = 1, \text{ αλλιώς} \end{cases}$$

όπου $\mathbb{1}_n$ είναι ένα διάνυσμα μήκους n με όλα τα στοιχεία ίσα με μονάδα και ξ είναι η παράμετρος μετατόπισης. Επιπλέον, ορίζουμε

$$\boldsymbol{d}_T = \begin{cases} |\boldsymbol{z}_T|, & \text{για } T = \text{YJ} \\ \boldsymbol{z}_T, & \text{αλλιώς} \end{cases},$$

$$q_T = \widehat{\lambda}_D \sum_{i=1}^{n^*} \frac{(y_i^* + \xi)^{2\widehat{\lambda}_D - 2} - (y_i^* + \xi)^{-2\widehat{\lambda}_D - 2} + 4\widehat{\lambda}_D (y_i^* + \xi)^{-2} \log(y_i^* + \xi)}{[\log(y_i^* + \xi \mathbb{1}_{n^*})]^{-1} \left[(y_i^* + \xi)^{\widehat{\lambda}_D - 1} + (y_i^* + \xi)^{-\widehat{\lambda}_D - 1} \right]^2}$$

για τον μετασχηματισμό Dual ή $q_T = 0$ για τους υπόλοιπους μετασχηματισμούς και

$$\boldsymbol{w}_T = \begin{cases} (\boldsymbol{y}^* + \xi \mathbb{1}_{n^*}) \circ \log(\boldsymbol{y}^* + \xi \mathbb{1}_{n^*}), & \text{για } T = \text{BC} \\ \text{sign}(\boldsymbol{y}^*) \circ (|\boldsymbol{y}^*| + \mathbb{1}_{n^*}) \circ \log(|\boldsymbol{y}^*| + \mathbb{1}_{n^*}), & \text{για } T = \text{Mod} \\ (|\boldsymbol{y}^*| + \mathbb{1}_{n^*}) \circ \log(|\boldsymbol{y}^*| + \mathbb{1}_{n^*}), & \text{για } T = \text{YJ} \\ \frac{1}{2}\left[(\boldsymbol{y}^* + \xi \mathbb{1}_{n^*})^{\widehat{\lambda}_D} + (\boldsymbol{y}^* + \xi \mathbb{1}_{n^*})^{-\widehat{\lambda}_D} \right] & \\ \qquad \circ \log(\boldsymbol{y}^* + \xi \mathbb{1}_{n^*}), & \text{για } T = \text{Dual} \end{cases}$$

όπου $\text{sign}(\boldsymbol{y}^*)$ είναι ένα διάνυσμα στοιχείων $\{+1, -1\}$ που εξαρτάται από το αν το i-οστό στοιχείο του $\boldsymbol{y}^*$ είναι θετικό ή αρνητικό και το σύμβολο $\circ$ συμβολίζει το γινόμενο Hadamard για τον πολλαπλασιασμό δύο διανυσμάτων στοιχείο προς στοιχείο. Τελικά, το r_T δίνεται από

$$\boldsymbol{r}_T = \begin{cases} \boldsymbol{w}_T \circ \log(\boldsymbol{y}^* + \xi \mathbb{1}_{n^*}) - 2(\boldsymbol{w}_T - \boldsymbol{z}_T), & \text{για } T = \text{BC} \\ \boldsymbol{w}_T \circ \log(|\boldsymbol{y}^*| + \mathbb{1}_{n^*}) - 2(\boldsymbol{w}_T - \boldsymbol{z}_T), & \text{για } T = \text{Mod} \\ \text{sign}(\boldsymbol{y}^*) \circ \boldsymbol{w}_T \circ \log(|\boldsymbol{y}^*| + \mathbb{1}_{n^*}) & \\ \quad -2\Big(\text{sign}(\boldsymbol{y}^*) \circ \boldsymbol{w}_T - \boldsymbol{z}_T \Big), & \text{για } T = \text{YJ} \\ \boldsymbol{z}_T \circ (\widehat{\lambda}_D)^2 \circ \log^2(\boldsymbol{y}^* + \xi \mathbb{1}_{n^*}) - (\boldsymbol{w}_T - \boldsymbol{z}_T), & \text{για } T = \text{Dual} \end{cases}.$$

2.4.1.3 Υπολογισμός της παραμέτρου κλίμακας υπό την Prior B για την οικογένεια Box-Cox

Οι υπολογισμοί που ακολουθούν αφορούν την οικογένεια Box-Cox, αλλά είναι παρεμφερείς και όσον αφορά τις οικογένειες Modulus και YJ. Ο δείκτης T διατηρείται για λόγους συνοχής και συνέπειας.

Θεωρώντας ένα σύνολο φανταστικών δεδομένων $\boldsymbol{y}^*$ μεγέθους n^*, η τυπική απόκλιση σ_{λ_T} υπό την οικογένεια T βασίζεται στην παρατηρούμενη πληροφορία κατά Fisher της παραμέτρου λ_T:

$$\sigma_{\lambda_T} = \left(-\frac{\partial^2}{\partial \lambda_T^2} \log f(\boldsymbol{y}^*|\lambda_T, T)^{1/n^*} \Big|_{\lambda_T=1} \right)^{-\frac{1}{2}}.$$

Σημειώνεται ότι η παρατηρούμενη πληροφορία κατά Fisher υπολογίζεται στο $\lambda_T = 1$ για την οικογένεια Box-Cox. Χωρίς βλάβη της γενικότητας (χ.β.γ.) θεωρούμε ότι το $\boldsymbol{y}^*$ συμβολίζει τα φανταστικά δεδομένα αφότου μετατοπιστούν στον θετικό άξονα. Με άλλα λόγια, αντί του $(\boldsymbol{y}^* + \xi \mathbb{1}_{n^*})$ χρησιμοποιούμε τον συμβολισμό $\boldsymbol{y}^*$ για λόγους διευκόλυνσης. Αρκεί να δείξουμε τους υπολογισμούς της δεύτερης παραγώγου της ποσότητας $\log f(\boldsymbol{y}^*|\lambda_T, T)^{1/n^*}$ ως προς λ_T. Η πιθανοφάνεια του $\boldsymbol{y}^*$ δεδομένου του λ_T παίρνει την ακόλουθη μορφή:

$$f\left(\boldsymbol{y}^*|\lambda_T, T\right) \propto \left|J(\boldsymbol{y}^*, \lambda_T|T)\right| \cdot f\left(\boldsymbol{y}^{*\,(\lambda_T)}|T\right).$$

Χρησιμοποιώντας μια διάχυτη NIG πρότερη κατανομή για τις παραμέτρους (μ_T, σ_T^2), η περιθώρια πιθανοφάνεια των μετασχηματισμένων δεδομένων προσεγγιστικά είναι:

$$\begin{aligned} f\left(\boldsymbol{y}^{*(\lambda_T)}|T\right) &\propto \left(\frac{(n^*-1)S_{\mathbf{z}}^{*2}}{2}\right)^{-\frac{n^*}{2}} \iff \\ \log f\left(\boldsymbol{y}^{*(\lambda_T)}|T\right) &\simeq -\frac{n^*}{2} \cdot \log\left(\frac{(n^*-1)S_{\mathbf{z}}^{*2}}{2}\right) + c, \end{aligned} \tag{2.14}$$

με $S_{\mathbf{z}}^{*2}$ να είναι η δειγματική διακύμανση των μετασχηματισμένων δεδομένων. Υπό την οικογένεια Box-Cox, η Ιακωβιανή του μετασχηματισμού είναι $\prod_{i=1}^{n^*} \left(y_i^*\right)^{\lambda_T - 1}$. Επομένως:

$$\begin{aligned} \frac{\partial \log f\left(\boldsymbol{y}^*|\lambda_T, T\right)^{\frac{1}{n^*}}}{\partial \lambda_T} &= \frac{1}{n^*}\sum_{i=1}^{n^*} \log y_i^* - \frac{n^*}{2n^*} \frac{\frac{\partial\left((n^*-1)S_{\mathbf{z}}^{*2}\right)}{\partial \lambda_T}}{(n^*-1)S_{\mathbf{z}}^{*2}} \\ &= \frac{1}{n^*}\sum_{i=1}^{n^*} \log y_i^* - \frac{n^*}{2n^*} \frac{\frac{\partial\left(\sum_{i=1}^{n^*}\left(\left(y_i^*\right)^{(\lambda_T)} - \overline{(\boldsymbol{y}^*)^{(\lambda_T)}}\right)^2\right)}{\partial \lambda_T}}{\sum_{i=1}^{n^*}\left(\left(y_i^*\right)^{(\lambda_T)} - \overline{(\boldsymbol{y}^*)^{(\lambda_T)}}\right)^2}. \end{aligned}$$

Εάν $z_i = \left(y_i^*\right)^{(\lambda_T)} = \frac{\left(y_i^*\right)^{\lambda_T} - 1}{\lambda_T}$ και $\bar{\mathbf{z}} = \overline{(\boldsymbol{y}^*)^{(\lambda_T)}} = \frac{1}{n^*}\sum_{i=1}^{n^*} z_i$, τότε:

$$\frac{\partial\left(\sum_{i=1}^{n^*}\left(\left(y_i^*\right)^{(\lambda_T)} - \overline{(\boldsymbol{y}^*)^{(\lambda_T)}}\right)^2\right)}{\partial \lambda_T} = \sum_{i=1}^{n^*} \frac{\partial\left[\left(\left(y_i^*\right)^{(\lambda_T)} - \overline{(\boldsymbol{y}^*)^{(\lambda_T)}}\right)^2\right]}{\partial \lambda_T}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n^*} 2\left(z_i - \overline{\mathbf{z}}\right) \frac{\partial\left(z_i - \overline{\mathbf{z}}\right)}{\partial \lambda_T} \\
&= \sum_{i=1}^{n^*} 2\left(z_i - \overline{\mathbf{z}}\right) \left(\frac{\partial z_i}{\partial \lambda_T} - \frac{\partial \overline{\mathbf{z}}}{\partial \lambda_T}\right).
\end{aligned}
$$

Εν γένει, $\mathbf{x} = (x_1, \ldots, x_{n^*})^T$ και η μέση τιμή του διανύσματος $\mathbf{x}$ συμβολίζεται με $\overline{\mathbf{x}} = \frac{1}{n^*}\sum_{i=1}^{n^*} x_i$. Οι παράγωγοι των z_i και $\overline{\mathbf{z}}$ ως προς λ_T είναι:

$$
\begin{aligned}
\frac{\partial z_i}{\partial \lambda_T} &= \frac{\partial\left(\frac{\left(y_i^*\right)^{\lambda_T}-1}{\lambda_T}\right)}{\partial \lambda_T} = \frac{\frac{\partial\left[\exp(\lambda_T \log y_i^*)-1\right]}{\partial \lambda_T}\lambda_T - \left[\left(y_i^*\right)^{\lambda_T}-1\right]}{{\lambda_T}^2} \\
&= \frac{\left(y_i^*\right)^{\lambda_T}\frac{\partial\left(\lambda_T \log y_i^*\right)}{\partial \lambda_T}\lambda_T - \left(y_i^*\right)^{\lambda_T}+1}{{\lambda_T}^2} \\
&= \frac{\left(y_i^*\right)^{\lambda_T}\lambda_T \log y_i^* - \left(y_i^*\right)^{\lambda_T}+1}{{\lambda_T}^2} \\
&= \frac{\left(y_i^*\right)^{\lambda_T}\log y_i^*}{\lambda_T} - \frac{\left(y_i^*\right)^{\lambda_T}-1}{{\lambda_T}^2} \\
&= \frac{w_i - z_i}{\lambda_T} \Rightarrow \\
\frac{\partial \overline{\mathbf{z}}}{\partial \lambda_T} &= \frac{\overline{\mathbf{w}} - \overline{\mathbf{z}}}{\lambda_T}
\end{aligned}
$$

όπου έχουμε θέσει $w_i = \left(y_i^*\right)^{\lambda_T}\log y_i^*$. Συνεπώς, η πρώτη παράγωγος ως προς λ_T διαμορφώνεται ως εξής:

$$
\frac{\partial \log f\left(\boldsymbol{y}^*|\lambda_T, T\right)^{\frac{1}{n^*}}}{\partial \lambda_T} = \frac{1}{n^*}\sum_{i=1}^{n^*}\log y_i^* \frac{\sum_{i=1}^{n^*}\left(z_i - \overline{\mathbf{z}}\right)\left(\frac{w_i - z_i}{\lambda_T} - \frac{\overline{\mathbf{w}} - \overline{\mathbf{z}}}{\lambda_T}\right)}{\sum_{i=1}^{n^*}\left(z_i - \overline{\mathbf{z}}\right)^2}.
$$

Προχωρούμε στον υπολογισμό της δεύτερης παραγώγου της ποσότητας που μας ενδιαφέρει ως προς λ_T:

$$
\frac{\partial^2}{\partial \lambda_T}\log f\left(\boldsymbol{y}^*|\lambda_T, T\right)^{\frac{1}{n^*}} = 0 - \frac{\partial}{\partial \lambda_T}\left(\frac{\sum_{i=1}^{n^*}\left(z_i - \overline{\mathbf{z}}\right)\left(\frac{w_i - z_i}{\lambda_T} - \frac{\overline{\mathbf{w}} - \overline{\mathbf{z}}}{\lambda_T}\right)}{\sum_{i=1}^{n^*}\left(z_i - \overline{\mathbf{z}}\right)^2}\right). \tag{2.15}
$$

Επιπλέον, έχουμε:

$$
\begin{aligned}
\sum_{i=1}^{n^*}\frac{\partial\left[\left(z_i - \overline{\mathbf{z}}\right)^2\right]}{\partial \lambda_T} &= 2\sum_{i=1}^{n^*}\left(z_i - \overline{\mathbf{z}}\right)\frac{\partial\left(z_i - \overline{\mathbf{z}}\right)}{\partial \lambda_T} = 2\sum_{i=1}^{n^*}\left(z_i - \overline{\mathbf{z}}\right)\left(\frac{\partial z_i}{\partial \lambda_T} - \frac{\partial \overline{\mathbf{z}}}{\partial \lambda_T}\right) \\
&= 2\sum_{i=1}^{n^*}\left(z_i - \overline{\mathbf{z}}\right)\left(\frac{w_i - z_i}{\lambda_T} - \frac{\overline{\mathbf{w}} - \overline{\mathbf{z}}}{\lambda_T}\right). \qquad (2.16)
\end{aligned}
$$

Ακόμη:

$$\begin{aligned}
\frac{\partial^2 z_i}{\partial {\lambda_T}^2} &= \frac{\partial\left(\frac{w_i - z_i}{\lambda_T}\right)}{\partial \lambda_T} = \frac{\frac{\partial (w_i - z_i)}{\partial \lambda_T}\lambda_T - (w_i - z_i)}{{\lambda_T}^2} \\
&= \frac{\left(\frac{\partial w_i}{\partial \lambda_T} - \frac{\partial z_i}{\partial \lambda_T}\right)\lambda_T - (w_i - z_i)}{{\lambda_T}^2} = \frac{\left(\frac{\partial\left({y_i^*}^{\lambda_T}\log y_i^*\right)}{\partial \lambda_T} - \frac{\partial z_i}{\partial \lambda_T}\right)\lambda_T - (w_i - z_i)}{{\lambda_T}^2} \\
&= \frac{\left(\log y_i^* \frac{\partial\left(\exp(\lambda_T \log y_i^*)\right)}{\partial \lambda_T} - \frac{\partial z_i}{\partial \lambda_T}\right)\lambda_T - (w_i - z_i)}{{\lambda_T}^2} \\
&= \frac{\left({y_i^*}^{\lambda_T}\log^2(y_i^*) - \frac{\partial z_i}{\partial \lambda_T}\right)\lambda_T - (w_i - z_i)}{{\lambda_T}^2} \\
&= \frac{\left(w_i \log y_i^* - \frac{w_i - z_i}{\lambda_T}\right)\lambda_T - (w_i - z_i)}{{\lambda_T}^2} \\
&= \frac{(\lambda_T \phi_i - w_i + z_i) - w_i + z_i}{{\lambda_T}^2} \\
&= \frac{\lambda_T \phi_i - 2w_i + 2z_i}{{\lambda_T}^2}
\end{aligned}$$

όπου $\phi_i = \frac{\partial w_i}{\partial \lambda_T} = w_i \log y_i^*$. Δεδομένου του ανωτέρω αποτελέσματος, έχουμε:

$$\begin{aligned}
&\sum_{i=1}^{n^*}\left[\left(\frac{w_i - z_i}{\lambda_T} - \frac{\overline{\mathbf{w}} - \overline{\mathbf{z}}}{\lambda_T}\right)^2 + (z_i - \overline{\mathbf{z}})\left(\frac{\phi_i}{\lambda_T} - \frac{2(w_i - z_i)}{{\lambda_T}^2} - \left[\frac{\overline{\boldsymbol{\phi}}}{\lambda_T} - \frac{2(\overline{\mathbf{w}} - \overline{\mathbf{z}})}{{\lambda_T}^2}\right]\right)\right] \\
= &\sum_{i=1}^{n^*}\left[\left(\frac{w_i - z_i}{\lambda_T} - \frac{\overline{\mathbf{w}} - \overline{\mathbf{z}}}{\lambda_T}\right)^2 + (z_i - \overline{\mathbf{z}})(r_i - \overline{\mathbf{r}})\right] \qquad (2.17)
\end{aligned}$$

όπου $r_i = \frac{\phi_i}{\lambda_T} - \frac{2(w_i - z_i)}{{\lambda_T}^2}$. Λαμβάνοντας υπόψη το αποτέλεσμα των (2.16) και (2.17), η (2.15) γίνεται:

$$\frac{\partial^2 \log f\left(\boldsymbol{y}^* | \lambda_T, T\right)^{\frac{1}{n^*}}}{\partial {\lambda_T}^2} =$$

$$\begin{aligned}
&= -\frac{\sum_{i=1}^{n^*}\left[\left(w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}})\right)^2 + {\lambda_T}^2 (z_i - \overline{\mathbf{z}})(r_i - \overline{\mathbf{r}})\right]\sum_{i=1}^{n^*}(z_i - \overline{\mathbf{z}})^2 - 2\left[\sum_{i=1}^{n^*}(z_i - \overline{\mathbf{z}})\left(w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}})\right)\right]^2}{{\lambda_T}^2\left[\sum_{i=1}^{n^*}(z_i - \overline{\mathbf{z}})^2\right]^2} \\
&= -\left[\frac{S^2_{\mathbf{w}-\mathbf{z}} + {\lambda_T}^2 S_{\mathbf{zr}}}{{\lambda_T}^2 S^2_{\mathbf{z}}} - 2\left(\frac{S_{\mathbf{zw}} - S^2_{\mathbf{z}}}{\lambda_T S^2_{\mathbf{z}}}\right)^2\right].
\end{aligned}$$

Σε αυτή την τελική έκφραση, χρησιμοποιήσαμε τα εξής:

$$\sum_{i=1}^{n^*}(z_i - \overline{\mathbf{z}})(w_i - z_i - \overline{\mathbf{w}} + \overline{\mathbf{z}}) = (n^* - 1)\left(S_{\mathbf{wz}} - S^2_{\mathbf{z}}\right),$$

$$\sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}})\,(r_i - \overline{\mathbf{r}}) = (n^* - 1) S_{\mathbf{zr}}$$

και

$$\sum_{i=1}^{n^*} (w_i - z_i - \overline{\mathbf{w}} + \overline{\mathbf{z}})^2 = (n^* - 1) S^2_{\mathbf{w-z}}$$

όπου η δειγματική (αμερόληπτη) διασπορά του $\boldsymbol{\alpha}$ συμβολίζεται με $S^2_{\boldsymbol{\alpha}}$ και η δειγματική συνδιασπορά μεταξύ των $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$ συμβολίζεται με $S_{\boldsymbol{\alpha\beta}}$.

Αντικαθιστώντας $\lambda_T = 1$ στην τελική έκφραση της δεύτερης παραγώγου της ποσότητας $\log f(\boldsymbol{y}^*|\lambda_T, T)^{1/n^*}$, παίρνοντας την αρνητική ποσότητα αυτής και υψώνοντάς τη στη δύναμη $-\frac{1}{2}$, παίρνουμε την τιμή της παραμέτρου κλίμακας σ_{λ_T} για την οικογένεια Box-Cox.

2.4.1.4 Υπολογισμός της παραμέτρου κλίμακας υπό την Prior B για την οικογένεια Dual

Οι υπολογισμοί αυτής της ενότητας αφορούν αποκλειστικά την οικογένεια Dual. Ο δείκτης T διατηρείται για λόγους συνέπειας όπως και προηγουμένως. Με βάση ένα σύνολο φανταστικών δεδομένων $\boldsymbol{y}^*$ πλήθους n^*, η τυπική απόκλιση $\sigma_{\log\lambda_T}$ για $T = \text{Dual}$ βασίζεται στην παρατηρούμενη πληροφορία κατά Fisher της παραμέτρου $\log\lambda_T$:

$$\sigma_{\log\lambda_T} = \left(-\frac{\partial^2}{\partial(\log\lambda_T)^2} \log f(\boldsymbol{y}^*|\,\log\lambda_T, T)^{1/n^*} \Big|_{\log\lambda_T = \log\widehat{\lambda}_D} \right)^{-\frac{1}{2}}.$$

Ας σημειωθεί ότι η παρατηρούμενη πληροφορία κατά Fisher υπολογίζεται στο σημείο $\log\widehat{\lambda}_D$ για την οικογένεια Dual. Θεωρούμε όπως και πριν ότι το $\boldsymbol{y}^*$ δηλώνει τα φανταστικά δεδομένα που έχουν ήδη μετατοπιστεί στον θετικό άξονα.

Η περιθώρια πιθανοφάνεια των μετασχηματισμένων δεδομένων που απορρέει από μια διάχυτη NIG πρότερη κατανομή για τα (μ_T, σ_T^2) έχει ήδη δωθεί προηγουμένως (βλέπε Ενότητα 2.4.1.1). Συμβολίζουμε τη νέα παράμετρο μετασχηματισμού με $\tilde{\lambda}_T = \log\lambda_T$. Η παραγώγιση γίνεται ως προς τη νέα παράμετρο $\tilde{\lambda}_T$. Συνεπώς, πρέπει να επαναπροσδιορίσουμε όλες τις συναφείς ποσότητες και εκφράσεις ως συναρτήσεις του $\tilde{\lambda}_T$. Το διάνυσμα των μετασχηματισμένων δεδομένων γίνεται:

$$z_i = (y_i^*)^{\lambda_T} = (y_i^*)^{\exp(\tilde{\lambda}_T)} = \frac{(y_i^*)^{\exp(\tilde{\lambda}_T)} - (y_i^*)^{-\exp(\tilde{\lambda}_T)}}{2\exp(\tilde{\lambda}_T)}.$$

Ο λογάριθμος της απόλυτης τιμής της Ιακωβιανής γίνεται:

$$\log\left|J\left(\boldsymbol{y}^*,\tilde{\lambda}_T|T\right)\right| = \sum_{i=1}^{n^*}\log\left(\frac{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}{2}\right).$$

Επομένως:

$$\begin{aligned}\log f\left(\boldsymbol{y}^*|\tilde{\lambda}_T,T\right)^{\frac{1}{n^*}} =& \frac{1}{n^*}\sum_{i=1}^{n^*}\log\left(\frac{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}{2}\right)\\ &-\frac{n^*}{2n^*}\log\left(\frac{(n^*-1)S_{\mathbf{z}}^2}{2}\right),\end{aligned}$$

όπου $\overline{\mathbf{z}} = \overline{(\boldsymbol{y}^*)^{(\lambda_T)}} = \frac{1}{n^*}\sum_{i=1}^{n^*} z_i$. Εν γένει, $\mathbf{x} = (x_1,\ldots,x_{n^*})^T$ και η μέση τιμή του διανύσματος $\mathbf{x}$ συμβολίζεται με $\overline{\mathbf{x}} = \frac{1}{n^*}\sum_{i=1}^{n^*} x_i$.

Αρχικά θα ασχοληθούμε με την παραγώγιση της Ιακωβιανής του μετασχηματισμού. Η πρώτη παράγωγος του λογαρίθμου της απόλυτης τιμής της Ιακωβιανής ως προς $\tilde{\lambda}_T$ είναι ίση με:

$$\begin{aligned}\frac{\partial\log\left|J\left(\boldsymbol{y}^*,\tilde{\lambda}_T|T\right)\right|}{\partial\tilde{\lambda}_T} &= \sum_{i=1}^{n^*}\frac{\partial\log\left(\frac{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}{2}\right)}{\partial\tilde{\lambda}_T}\\ &= \sum_{i=1}^{n^*}\frac{\frac{\partial(y_i^*)^{\exp(\tilde{\lambda}_T)-1}}{\partial\tilde{\lambda}_T}+\frac{\partial(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}{\partial\tilde{\lambda}_T}}{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}\\ &= \sum_{i=1}^{n^*}\log y_i^*\exp(\tilde{\lambda}_T)\frac{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}-(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}.\end{aligned}$$

Όσον αφορά την πρώτη παράγωγο της κύριας ποσότητας που μας ενδιαφέρει, έχουμε:

$$\begin{aligned}\frac{\partial\log f\left(\boldsymbol{y}^*|\tilde{\lambda}_T,T\right)^{\frac{1}{n^*}}}{\partial\tilde{\lambda}_T} =& \frac{1}{n^*}\frac{\partial\log\left|J\left(\boldsymbol{y}^*,\tilde{\lambda}_T|T\right)\right|}{\partial\tilde{\lambda}_T}-\frac{n^*}{2n^*}\frac{\frac{\partial\left(\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})^2\right)}{\partial\lambda_T}}{\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})^2}\\ =& \frac{1}{n^*}\frac{\partial\log\left|J\left(\boldsymbol{y}^*,\tilde{\lambda}_T|T\right)\right|}{\partial\tilde{\lambda}_T}-\frac{n^*}{n^*}\frac{\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})\left(\frac{\partial z_i}{\partial\tilde{\lambda}_T}-\frac{\partial\overline{\mathbf{z}}}{\partial\tilde{\lambda}_T}\right)}{\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})^2}.\end{aligned} \tag{2.18}$$

Σε όλο αυτό, θα χρειαστούμε τις ακόλουθες ποσότητες:

$$\begin{aligned}\frac{\partial z_i}{\partial\tilde{\lambda}_T} &= \frac{\frac{\partial\left((y_i^*)^{\exp(\tilde{\lambda}_T)}-(y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)}{\partial\tilde{\lambda}_T}2\exp(\tilde{\lambda}_T)-\left((y_i^*)^{\exp(\tilde{\lambda}_T)}-(y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)2\exp(\tilde{\lambda}_T)}{2^2\exp(2\tilde{\lambda}_T)}\\ &= \frac{\left((y_i^*)^{\exp(\tilde{\lambda}_T)}\log y_i^*\exp(\tilde{\lambda}_T)+(y_i^*)^{-\exp(\tilde{\lambda}_T)}\log y_i^*\exp(\tilde{\lambda}_T)\right)}{2\exp(\tilde{\lambda}_T)}\end{aligned}$$

$$
\begin{aligned}
& -\frac{\left((y_i^*)^{\exp(\tilde{\lambda}_T)} - (y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)}{2\exp(\tilde{\lambda}_T)} \\
= & \frac{\left((y_i^*)^{\exp(\tilde{\lambda}_T)} + (y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)\log y_i^* \exp(\tilde{\lambda}_T)}{2\exp(\tilde{\lambda}_T)} - \frac{(y_i^*)^{\exp(\tilde{\lambda}_T)} - (y_i^*)^{-\exp(\tilde{\lambda}_T)}}{2\exp(\tilde{\lambda}_T)} \\
= & \frac{\left((y_i^*)^{\exp(\tilde{\lambda}_T)} + (y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)\log y_i^*}{2} - \frac{(y_i^*)^{\exp(\tilde{\lambda}_T)} - (y_i^*)^{-\exp(\tilde{\lambda}_T)}}{2\exp(\tilde{\lambda}_T)} \\
= & w_i - z_i
\end{aligned}
$$

όπου $w_i = \frac{\left((y_i^*)^{\exp(\tilde{\lambda}_T)} + (y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)\log y_i^*}{2}$.

Κατά συνέπεια, έχουμε: $\frac{\partial \overline{\mathbf{z}}}{\partial \tilde{\lambda}_T} = \overline{\mathbf{w}} - \overline{\mathbf{z}}$. Η πρώτη παράγωγος του w_i είναι:

$$
\begin{aligned}
\frac{\partial w_i}{\partial \tilde{\lambda}_T} &= \frac{\partial\left((y_i^*)^{\exp(\tilde{\lambda}_T)} + (y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)}{\partial \tilde{\lambda}_T}\frac{\log y_i^*}{2} \\
&= \left((y_i^*)^{\exp(\tilde{\lambda}_T)} - (y_i^*)^{-\exp(\tilde{\lambda}_T)}\right)\exp(\tilde{\lambda}_T)\frac{\log^2 y_i^*}{2} \\
&= z_i \exp(2\tilde{\lambda}_T)\log^2 y_i^* \\
&= \phi_i.
\end{aligned}
$$

Η δεύτερη παράγωγος του z_i είναι:

$$
\begin{aligned}
\frac{\partial^2 z_i}{\partial \tilde{\lambda}_T^2} &= \frac{\partial w_i}{\partial \tilde{\lambda}_T} - \frac{\partial z_i}{\partial \tilde{\lambda}_T} \\
&= \phi_i - (w_i - z_i) \\
&= r_i
\end{aligned}
$$

και η δεύτερη παράγωγος του σχετικού διανύσματος $\mathbf{z}$ είναι:

$$
\begin{aligned}
\frac{\partial^2 \overline{\mathbf{z}}}{\partial \tilde{\lambda}_T^2} &= \overline{\boldsymbol{\phi}} - (\overline{\mathbf{w}} - \overline{\mathbf{z}}) \\
&= \overline{\mathbf{r}}.
\end{aligned}
$$

Άρα, η (2.18) γίνεται:

$$
\frac{\partial \log f\left(\boldsymbol{y}^*|\tilde{\lambda}_T, T\right)^{\frac{1}{n^*}}}{\partial \tilde{\lambda}_T} = \frac{1}{n^*}\frac{\partial \log\left|J\left(\boldsymbol{y}^*, \tilde{\lambda}_T|T\right)\right|}{\partial \tilde{\lambda}_T} - \frac{\sum_{i=1}^{n^*}(z_i - \overline{\mathbf{z}})\left(w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}})\right)}{\sum_{i=1}^{n^*}(z_i - \overline{\mathbf{z}})^2}.
$$

Η δεύτερη παράγωγος του λογαρίθμου της απόλυτης τιμής της Ιακωβιανής ως προς $\tilde{\lambda}_T$ δίνεται ως ακολούθως:

$$\frac{\partial^2 \log\left|J\left(\boldsymbol{y}^*,\tilde{\lambda}_T|T\right)\right|}{\partial\tilde{\lambda}_T^2} =$$

$$= \sum_{i=1}^{n^*} \log y_i^* \frac{\partial \exp(\tilde{\lambda}_T)\frac{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}-(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}}{\partial\tilde{\lambda}_T}$$

$$= \sum_{i=1}^{n^*} \log y_i^* \left(\frac{\frac{\partial (y_i^*)^{\exp(\tilde{\lambda}_T)-1}\exp(\tilde{\lambda}_T)}{\partial\tilde{\lambda}_T} - \frac{\partial (y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\exp(\tilde{\lambda}_T)}{\partial\tilde{\lambda}_T}}{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}} \right.$$
$$\left. - \frac{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}-(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]\exp(\tilde{\lambda}_T)\left(\frac{\partial (y_i^*)^{\exp(\tilde{\lambda}_T)-1}}{\partial\tilde{\lambda}_T}+\frac{\partial (y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}{\partial\tilde{\lambda}_T}\right)}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2} \right)$$

$$= \sum_{i=1}^{n^*} \log y_i^* \left(\frac{\exp(\tilde{\lambda}_T)\left((y_i^*)^{\exp(\tilde{\lambda}_T)-1}-(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right.}{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}} \right.$$
$$\left. + \frac{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]\log y_i^*\exp(\tilde{\lambda}_T)\right)}{(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}}$$
$$- \left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}-(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]\exp(\tilde{\lambda}_T)$$
$$\times \left. \frac{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}\log y_i^*\exp(\tilde{\lambda}_T)-(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\log y_i^*\exp(\tilde{\lambda}_T)\right]}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2} \right)$$

$$= \sum_{i=1}^{n^*} \log y_i^* \left(\frac{\exp(\tilde{\lambda}_T)(y_i^*)^{2(\exp(\tilde{\lambda}_T)-1)}+\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{2(\exp(\tilde{\lambda}_T)-1)}-\exp(\tilde{\lambda}_T)(y_i^*)^{-2}}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2} \right.$$
$$+ \frac{\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{-2}+\exp(\tilde{\lambda}_T)(y_i^*)^{-2}+\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{-2}}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2}$$
$$+ \frac{-\exp(\tilde{\lambda}_T)(y_i^*)^{-2(\exp(\tilde{\lambda}_T)+1)}+\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{-2(\exp(\tilde{\lambda}_T)+1)}}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2}$$
$$+ \frac{-\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{2(\exp(\tilde{\lambda}_T)-1)}+\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{-2}}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2}$$
$$\left. + \frac{\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{-2}-\exp(2\tilde{\lambda}_T)\log y_i^*(y_i^*)^{-2(\exp(\tilde{\lambda}_T)+1)}}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2} \right)$$

$$= \exp(\tilde{\lambda}_T)\sum_{i=1}^{n^*} \log y_i^* \frac{\left((y_i^*)^{2(\exp(\tilde{\lambda}_T)-1)}+4\log y_i^*\exp(\tilde{\lambda}_T)(y_i^*)^{-2}-(y_i^*)^{-2(\exp(\tilde{\lambda}_T)+1)}\right)}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1}+(y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2}$$

Η δεύτερη παράγωγος της κύριας ποσότητας ενδιαφέροντος διαμορφώνεται ως εξής:

$$\frac{\partial^2 \log f\left(\boldsymbol{y}^*|\tilde{\lambda}_T,T\right)^{\frac{1}{n^*}}}{\partial\tilde{\lambda}_T^2} = \frac{1}{n^*}\frac{\partial^2 \log\left|J\left(\boldsymbol{y}^*,\tilde{\lambda}_T|T\right)\right|}{\partial\tilde{\lambda}_T^2} - \frac{\partial\frac{\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})(w_i-z_i-(\overline{\mathbf{w}}-\overline{\mathbf{z}}))}{\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})^2}}{\partial\tilde{\lambda}_T}.$$

Στην παραπάνω εξίσωση, η δεύτερη παράγωγος που σχετίζεται με την Ιακωβιανή του μετασχηματισμού ως προς $\tilde{\lambda}_T$ έχει ήδη υπολογιστεί. Όσον αφορά τον δεύτερο όρο της εξίσωσης, λαμβάνοντας υπόψη τους συναφείς όρους που παράγει ο κανόνας του πηλίκου κατά την παραγώγιση, έχουμε:

$$N = \sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}}) \, (w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}})),$$

$$D = \sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}})^2 \, ,$$

$$\begin{aligned}
\frac{\partial N}{\partial \tilde{\lambda}_T} &= \sum_{i=1}^{n^*} \left(\frac{\partial (z_i - \overline{\mathbf{z}})}{\partial \tilde{\lambda}_T} (w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}})) + (z_i - \overline{\mathbf{z}}) \frac{\partial (w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}}))}{\partial \tilde{\lambda}_T} \right) \\
&= \sum_{i=1}^{n^*} \left((w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}}))^2 + (z_i - \overline{\mathbf{z}}) \frac{\partial (w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}}))}{\partial \tilde{\lambda}_T} \right) \\
&= \sum_{i=1}^{n^*} \left((w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}}))^2 + (z_i - \overline{\mathbf{z}})(\phi_i - w_i + z_i - (\overline{\boldsymbol{\phi}} - \overline{\mathbf{w}} + \overline{\mathbf{z}})) \right) \\
&= \sum_{i=1}^{n^*} \left((w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}}))^2 + (z_i - \overline{\mathbf{z}})(r_i - \overline{\mathbf{r}}) \right) ,
\end{aligned}$$

και

$$\frac{\partial D}{\partial \tilde{\lambda}_T} = 2 \sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}}) \, (w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}})) = 2N.$$

Συνεπώς:

$$\begin{aligned}
&\frac{\partial \frac{\sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}})(w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}}))}{\sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}})^2}}{\partial \tilde{\lambda}_T} = \\
&= \frac{\sum_{i=1}^{n^*} \left((w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}}))^2 + (z_i - \overline{\mathbf{z}})(r_i - \overline{\mathbf{r}}) \right) \sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}})^2}{\left[\sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}})^2 \right]^2} \\
&\quad - \frac{2 \left(\sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}}) \, (w_i - z_i - (\overline{\mathbf{w}} - \overline{\mathbf{z}})) \right)^2}{\left[\sum_{i=1}^{n^*} (z_i - \overline{\mathbf{z}})^2 \right]^2} \\
&= \frac{S^2_{\mathbf{w}-\mathbf{z}} + S_{\mathbf{zr}}}{S^2_{\mathbf{z}}} - 2(n^* - 1)^2 \frac{(S_{\mathbf{zw}} - S^2_{\mathbf{z}})^2}{[(n^* - 1) S^2_{\mathbf{z}}]^2} \\
&= \frac{S^2_{\mathbf{w}-\mathbf{z}} + S_{\mathbf{zr}}}{S^2_{\mathbf{z}}} - 2 \left(\frac{S_{\mathbf{zw}}}{S^2_{\mathbf{z}}} - 1 \right)^2 .
\end{aligned}$$

Εν τέλει, καταλήγουμε στα εξής:

$$\frac{\partial^2 \log f\left(\boldsymbol{y}^*|\tilde{\lambda}_T, T\right)^{\frac{1}{n^*}}}{\partial \tilde{\lambda}_T^2} =$$

$$\begin{aligned}
&= \frac{1}{n^*}\frac{\partial^2 \log\left|J\left(\boldsymbol{y}^*, \tilde{\lambda}_T|T\right)\right|}{\partial \tilde{\lambda}_T^2} - \frac{\partial \frac{\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})(w_i-z_i-(\overline{\mathbf{w}}-\overline{\mathbf{z}}))}{\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})^2}}{\partial \tilde{\lambda}_T} \\
&= \frac{1}{n^*}\exp(\tilde{\lambda}_T)\sum_{i=1}^{n^*}\log y_i^* \frac{\left((y_i^*)^{2(\exp(\tilde{\lambda}_T)-1)} + 4\log y_i^* \exp(\tilde{\lambda}_T)(y_i^*)^{-2} - (y_i^*)^{-2(\exp(\tilde{\lambda}_T)+1)}\right)}{\left[(y_i^*)^{\exp(\tilde{\lambda}_T)-1} + (y_i^*)^{-(\exp(\tilde{\lambda}_T)+1)}\right]^2} \\
&\quad - \left(\frac{S^2_{\mathbf{w}-\mathbf{z}} + S_{\mathbf{zr}}}{S^2_{\mathbf{z}}} - 2\left(\frac{S_{\mathbf{zw}}}{S^2_{\mathbf{z}}} - 1\right)^2\right).
\end{aligned}$$

Στην ανωτέρω τελική μορφή της δεύτερης παραγώγου, φάνηκαν χρήσιμα τα εξής:

$$\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})(w_i-z_i-\overline{\mathbf{w}}+\overline{\mathbf{z}}) = (n^*-1)\left(S_{\mathbf{wz}} - S^2_{\mathbf{z}}\right),$$

$$\sum_{i=1}^{n^*}(z_i-\overline{\mathbf{z}})(r_i-\overline{\mathbf{r}}) = (n^*-1)S_{\mathbf{zr}}$$

και

$$\sum_{i=1}^{n^*}(w_i-z_i-\overline{\mathbf{w}}+\overline{\mathbf{z}})^2 = (n^*-1)S^2_{\mathbf{w}-\mathbf{z}}$$

όπου η δειγματική (αμερόληπτη) διακύμανση του $\boldsymbol{\alpha}$ συμβολίζεται με $S^2_{\boldsymbol{\alpha}}$ και η δειγματική συνδιακύμανση μεταξύ των $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$ συμβολίζεται με $S_{\boldsymbol{\alpha\beta}}$.

Αντικαθιστώντας $\tilde{\lambda}_T = \log\widehat{\lambda}_D$ στην τελική έκφραση της δεύτερης παραγώγου της $\log f(\boldsymbol{y}^*|\lambda_T, T)^{1/n^*}$, παίρνοντας την αρνητική ποσότητα αυτής και υψώνοντάς τη στη δύναμη $-\frac{1}{2}$, λαμβάνουμε την τιμή της παραμέτρου κλίμακας $\sigma_{\tilde{\lambda}_T}$ για την οικογένεια Dual.

2.4.2 Συμπερασματολογία a posteriori

Δεδομένης της οικογένειας T και της παραμέτρου μετασχηματισμού λ_T, η a posteriori κατανομή των παραμέτρων θέσης και κλίμακας (μ_T, σ_T^2) υπό την πρότερη κατανομή (2.4) της Ενότητας 2.4.1 είναι:

$$\begin{aligned}
\pi\left(\mu_T, \sigma_T^2|\boldsymbol{y}, \lambda_T, T\right) &= NIG\left(\mu_T, \sigma_T^2; \mu_n, k_n^{-1}, \alpha_n, \beta_n\right) \\
&= N\left(\mu_T; \mu_n, \frac{\sigma_T^2}{k_n}\right) IG\left(\sigma_T^2; \alpha_n, \beta_n\right),
\end{aligned}$$

όπου

$$\begin{aligned}
\mu_n &= \frac{k_0\mu_0 + n \cdot \overline{\boldsymbol{y}^{(\lambda_T)}}}{k_0 + n} \\
k_n &= k_0 + n \\
a_n &= a_0 + \frac{n}{2} \\
\beta_n &= \beta_0 + \frac{(n-1)S_T^2}{2} + \frac{nk_0\left(\overline{\boldsymbol{y}^{(\lambda_T)}} - \mu_0\right)^2}{2(k_0+n)}
\end{aligned}$$

με S_T^2 να είναι η δειγματική διακύμανση των μετασχηματισμένων δεδομένων.

Για την εκάστοτε οικογένεια T, η a posteriori περιθώρια κατανομή του λ_T δίνεται από την ακόλουθη εξίσωση:

$$\pi\left(\lambda_T|\boldsymbol{y}, \boldsymbol{y}^*, T\right) \propto f\left(\boldsymbol{y}|\lambda_T, T\right)\pi\left(\lambda_T|\boldsymbol{y}^*, T\right). \tag{2.19}$$

Ο πρώτος όρος στο δεξί μέρος της (2.19) είναι η πιθανοφάνεια των αμετασχημάτιστων δεδομένων δεδομένου του λ_T και δίνεται από την (2.5). Η ύστερη περιθώρια κατανομή του λ_T διαμορφώνεται ως εξής:

$$\begin{aligned}
\pi\left(\lambda_T|\boldsymbol{y}, \boldsymbol{y}^*, T\right) \quad \propto \quad & \left(\beta_0 + \frac{(n-1)S_T^2}{2} + \frac{nk_0\left(\overline{\boldsymbol{y}^{(\lambda_T)}} - \mu_0\right)^2}{2(k_0+n)}\right)^{-\left(\alpha_0+\frac{n}{2}\right)} \\
& \times \left|J\left(\boldsymbol{y}, \lambda_T|T\right)\right|\pi\left(\lambda_T|\boldsymbol{y}^*, T\right).
\end{aligned} \tag{2.20}$$

Ο τρίτος όρος στο δεξί μέρος της (2.20) αφορά την πρότερη κατανομή του λ_T υπό την οικογένεια T δεδομένων των $\boldsymbol{y}^*$ και ποικίλει ανάλογα με το πλαίσιο πρότερης κατανομής που εφαρμόζεται, όπως περιγράφουμε στην Ενότητα 2.4.1, καταλήγοντας στην (2.8) για την Prior A και στις (2.10)–(2.12) για την Prior B. Ένας τυπικός αλγόριθμος τυχαίου περιπάτου Metropolis-Hastings μπορεί να εφαρμοστεί για να προσομοιώσει κανείς από την (2.20), με κανονική κατανομή εισήγησης και διακύμανση κατάλληλα επιλεγμένη ώστε να επιτευχθεί ο επιθυμητός ρυθμός αποδοχής 25%.

2.4.3 Επιλογή μετασχηματισμού

Στα πλαίσια της Μπεϋζιανής συλλογιστικής, η ανάδειξη του βέλτιστου μετασχηματισμού μεταξύ των έξι οικογενειών που έχουμε θεωρήσει είναι μια διαδικασία ισοδύναμη

με την ανεύρεση του μετασχηματισμού $T \in \mathcal{T}$ με τη μέγιστη εκ των υστέρων πιθανότητα εμφάνισης ορισμένη ως εξής:

$$\pi\left(T|\ \boldsymbol{y}, \boldsymbol{y}^*\right) = \frac{f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right)\pi(T)}{\sum_{T\in\mathcal{T}} f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right)\pi(T)}$$

όπου $f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right)$ είναι η περιθώρια πιθανοφάνεια των δεδομένων υπό την οικογένεια T και $\pi(T)$ είναι η πρότερη κατανομή της οικογένειας T η οποία δίνεται στην (2.3). Η περιθώρια πιθανοφάνεια μπορεί να επεκταθεί περαιτέρω ώστε να ενσωματώσει την επίδραση του λ_T:

$$f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right) = \int f\left(\boldsymbol{y}|\lambda_T, T\right)\pi\left(\lambda_T|\boldsymbol{y}^*, T\right)\mathrm{d}\lambda_T$$

όπου $f\left(\boldsymbol{y}|\lambda_T, T\right)$ είναι η πιθανοφάνεια του διανύσματος παρατηρήσεων $\boldsymbol{y}$ υπό το μοντέλο T περιθωριοποιημένη ως προς τις υπόλοιπες παραμέτρους εκτός του λ_T, ενώ ο όρος $\pi\left(\lambda_T|\boldsymbol{y}^*, T\right)$ αντιπροσωπεύει την πρότερη κατανομή του λ_T δεδομένου του T (βλέπε Ενότητα 2.4.1). Είναι προφανές ότι οι μετασχηματισμοί Id και Log δε συνοδεύονται από την παράμετρο λ_T, αλλά έχουμε υιοθετήσει μια ολιστική προσέγγιση συμβολισμού χάριν συνέπειας. Επομένως, η συνάρτηση $f\left(\boldsymbol{y}|\lambda_T, T\right)$ για τους δύο εν λόγω μετασχηματισμούς δίνεται από τη (2.5) όπου $\boldsymbol{y}^{(\lambda_T)}$ είναι τα αρχικά (τυποποιημένα) δεδομένα $\boldsymbol{y}$ ή ο λογάριθμος των δεδομένων $\boldsymbol{y}$ αντίστοιχα.

Στην περίπτωση της πρότερης κατανομής δύναμης (Prior A) για την παράμετρο λ_T δεδομένου του T, η περιθώρια πιθανοφάνεια, με βάση τη (2.7) και με $\delta = (n^*)^{-1}$, δίνεται από τον ακόλουθο τύπο που περιλαμβάνει δύο ολοκληρώματα:

$$f\left(\boldsymbol{y}|\ \boldsymbol{y}^*, T\right) = \frac{\int f\left(\boldsymbol{y}|\lambda_T, T\right) f\left(\boldsymbol{y}^*|\lambda_T, T\right)^{1/n^*} \pi^N(\lambda_T|T)\mathrm{d}\lambda_T}{\int f\left(\boldsymbol{y}^*|\lambda_T, T\right)^{1/n^*} \pi^N(\lambda_T|T)\mathrm{d}\lambda_T}. \tag{2.21}$$

Στην περίπτωση της εναλλακτικής πρότερης κατανομής μοναδιαίας πληροφορίας (Prior B) για την παράμετρο λ_T δεδομένου του T, ο αντίστοιχος τύπος της περιθώριας πιθανοφάνειας, με βάση τη (2.10), είναι ο ακόλουθος:

$$f(\boldsymbol{y}|\ \boldsymbol{y}^*, T) = \int f\left(\boldsymbol{y}|\widetilde{\lambda}_T, T\right) N\left(\widetilde{\lambda}_T; \mu_{\widetilde{\lambda}_T}, \sigma^2_{\widetilde{\lambda}_T}, T\right)\mathrm{d}\widetilde{\lambda}_T, \tag{2.22}$$

όπου το $\widetilde{\lambda}_T$ περιγράφεται επαρκώς από τη (2.9) και οι παράμετροι $\mu_{\widetilde{\lambda}_T}$ και $\sigma^2_{\widetilde{\lambda}_T}$ δίνονται από τις (2.11) και (2.12) αντίστοιχα.

Η εκτίμηση της περιθώριας πιθανοφάνειας (2.21) ή (2.22) καθίσταται εφικτή μέσω μιας απλής αριθμητικής ολοκλήρωσης, ή μέσω μιας επέκτασης του εκτιμητή του Chib (Chib 1995) όπως περιγράφεται στο άρθρο των Chib & Jeliazkov (2001). Στην Ενότητα 2.4.4 παρέχονται όλες οι υπολογιστικές λεπτομέρειες.

2.4.4 Υπολογισμός περιθώριας πιθανοφάνειας

Ο υπολογισμός των μη υπολογίσιμων ολοκληρωμάτων που εμπλέκονται στην περιθώρια πιθανοφάνεια $f(\boldsymbol{y}|\boldsymbol{y}^*, T)$ της (2.21) ή της (2.22) επιτυγχάνεται κατά βάση μέσω δύο διακριτών εκτιμητών. Ο βασικός εκτιμητής είναι ο εκτιμητής του Chib (βλέπε Chib & Jeliazkov (2001)) ο οποίος βασίζεται στα αποτελέσματα ενός αλγορίθμου Metropolis-Hastings (MH) που προσομοιώνει από την a posteriori κατανομή του λ_T. Παράλληλα, χρησιμοποιήθηκε και μια αριθμητική προσεγγιστική μέθοδος εκτίμησης των ολοκληρωμάτων (βλέπε Ενότητα 2.4.4.1). Η χρήση αυτών των εναλλακτικών υπολογιστικών προσεγγίσεων έγινε κυρίως για λόγους επικύρωσης των αποτελεσμάτων και εκτίμησης της ακρίβειάς τους. Τα αποτελέσματα που προκύπτουν από τις δύο αυτές τεχνικές συγκλίνουν σε ικανοποιητικό βαθμό. Ο αριθμητικός εκτιμητής ήταν κάποιες φορές πιο ασταθής σε σχέση με τον εκτιμητή του Chib για την οικογένεια Dual.

Όσον αφορά τον εκτιμητή του Chib, θεωρούμε την ακόλουθη ταυτότητα της περιθώριας πιθανοφάνειας:

$$\log f(\boldsymbol{y}|\boldsymbol{y}^*, T) = \log f\left(\boldsymbol{y}|\lambda_T^*, T\right) + \log \pi(\lambda_T^*|\boldsymbol{y}^*, T) - \log \pi(\lambda_T^*|\boldsymbol{y}, \boldsymbol{y}^*, T)$$

όπου λ_T^* είναι μια τιμή υψηλής ύστερης πυκνότητας του $\{\lambda_T\}$. Η ποσότητα $\pi\left(\lambda_T^*|\boldsymbol{y}, \boldsymbol{y}^*, T\right)$ ονομάζεται ύστερη τεταγμένη (*posterior ordinate*) και εκτιμάται μέσω του τύπου:

$$\pi\left(\lambda_T^*|\boldsymbol{y}, \boldsymbol{y}^*, T\right) = (2\pi k^*)^{-1/2} \frac{\frac{1}{M}\sum_{g=1}^{M}\left[\min\left\{1, \frac{K(\lambda_T^*)}{K\left(\lambda_T^{(g)}\right)}\right\} \exp\left\{-\frac{\left(\lambda_T^* - \lambda_T^{(g)}\right)^2}{2k^*}\right\}\right]}{\frac{1}{J}\sum_{j=1}^{J}\min\left\{1, \frac{K\left(\lambda_T^{(j)}\right)}{K(\lambda_T^*)}\right\}}$$

όπου

$$K(\lambda_T) \;=\; \left(\beta_0 + \frac{(n-1)S_T^2}{2} + \frac{nk_0\left(\overline{\boldsymbol{y}^{(\lambda_T)}} - \mu_0\right)^2}{2(k_0+n)}\right)^{-\left(\alpha_0 + \frac{n}{2}\right)} \Big| J\left(\boldsymbol{y}, \lambda_T|T\right)\Big|$$

$$\times \quad \left(\beta_0 + \frac{(n^*-1)S^{*2}_T}{2} + \frac{n^* k_0 \left(\overline{\boldsymbol{y}^{*(\lambda_T)}} - \mu_0 \right)^2}{2(k_0+n^*)} \right)^{-\left(\frac{\alpha_0}{n^*}+\frac{1}{2}\right)}$$

$$\times \quad \left| J\left(\boldsymbol{y}^*, \lambda_T | T\right) \right|^{\frac{1}{n^*}} N(\lambda_T | m_0, s_0^2) \tag{2.23}$$

για την προσέγγιση της πρότερης κατανομής δύναμης (Prior A). Για την προσέγγιση Prior B, η δεύτερη και τρίτη γραμμή της (2.23) απλά αντικαταστάθηκαν από τον πυρήνα της κανονικής πρότερης κατανομής (βλέπε Ενότητα 2.4.1.2) για όλες τις οικογένειες μετασχηματισμών T εκτός της Dual όπου χρησιμοποιήθηκε η λογαριθμοκανονική κατανομή. Επιπλέον, $\lambda_T^{(g)}$ είναι ένα τυχαίο δείγμα μεγέθους M από την ύστερη κατανομή του λ_T όπως προέκυψε από έναν αλγόριθμο τυχαίου περιπάτου MH, ενώ $\lambda_T^{(j)}$, $j = 1, \ldots, J$, είναι ένα τυχαίο δείγμα μεγέθους J όπως προσομοιώθηκε από την κατανομή εισήγησης που χρησιμοποιήθηκε στον αλγόριθμο MH, δηλαδή ένα δείγμα από την κανονική κατανομή $N\left(\lambda_T | \lambda_T^*, k^*, T\right)$ με μέση τιμή λ_T^* και διακύμανση k^* κατάλληλα επιλεγμένες ώστε να έχουμε καλή μίξη της αλυσίδας (βλέπε για παράδειγμα Ntzoufras (2009)). Γενικά, το M κυμάνθηκε γύρω στις $15000 - 18000$ επαναλήψεις επιπρόσθετα με τη *burn in* περίοδο του αλγόριθμου, ενώ για το J χρησιμοποιήσαμε μόνο 2000 καθώς αναφέρεται στο πλήθος των i.i.d. προσομοιωμένων τιμών από τη γνωστή κατανομή εισήγησης.

Ένας τρίτος εκτιμητής που δοκιμάστηκε, χωρίς ωστόσο την επιτυχία των δύο προαναφερθέντων εκτιμητών, είναι ο εκτιμητής Laplace-Metropolis (LM) (Lewis & Raftery 1997), που ονομάστηκε έτσι εξαιτίας του γεγονότος ότι κάνει χρήση ενός κατάλληλου MCMC αλγόριθμου ο οποίος παρέχει πληροφορία που ενσωματώνεται στην κλασική προσέγγιση Laplace. Ο σχετικός τύπος του εκτιμητή είναι ο εξής:

$$\log f(\boldsymbol{y}|\boldsymbol{y}^*, T) \approx \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\sigma^*_{\lambda_T}\right)^2 + \log \pi\left(\lambda_T^* | \boldsymbol{y}^*, T\right) + \log f\left(\boldsymbol{y} | \lambda_T^*, T\right).$$

Στον παραπάνω τύπο, το λ_T^* συμβολίζει την εκ των υστέρων κορυφή της αλυσίδας $\{\lambda_T\}$, η οποία θεωρητικά θέλουμε να προσεγγίζει ικανοποιητικά τον ύστερο μέσο ή την ύστερη διάμεσο και $\left(\sigma^*_{\lambda_T}\right)^2$ είναι η MCMC εκτίμηση της ύστερης διακύμανσης της αλυσίδας $\{\lambda_T\}$. Παρόλα αυτά, ο εκτιμητής LM εγκαταλείφθηκε γρήγορα εξαιτίας του υψηλού βαθμού αστάθειας που επέδειξε σε σχέση με τον εκτιμητή του Chib. Συγκεκριμένα, όταν η εκ των υστέρων κατανομή του λ_T είναι σημαντικά ασύμμετρη (π.χ. πολύ συχνό θέμα υπό την οικογένεια Dual) τότε ο εκτιμητής LM δίνει αναξιόπιστα αποτελέσματα.

2.4.4.1 Αριθμητική προσέγγιση της περιθώριας πιθανοφάνειας

Η αριθμητική μέθοδος προσέγγισης της περιθώριας πιθανοφάνειας, ή αλλιώς της ένδειξης των δεδομένων (*data evidence*), υλοποιείται μέσω της R-εντολής `integrate' ώστε να προσεγγιστεί το άλυτο ολοκλήρωμα I:

$$\begin{aligned} I &= \int f\left(\boldsymbol{y}, \lambda_T|T\right) \mathrm{d}\lambda_T \\ &= \int f\left(\boldsymbol{y}|\lambda_T, T\right) \cdot \pi(\lambda_T|\boldsymbol{y}^* T) \mathrm{d}\lambda_T \end{aligned} \tag{2.24}$$

Πρόκειται για μια αριθμητική ρουτίνα ολοκλήρωσης, βασισμένη στον κανόνα τετραγωνισμού των Gauss-Kronrod (Kronrod 1965). Ως εκ τούτου, σε περίπτωση που η προς ολοκλήρωση συνάρτηση είναι ως επί το πλείστον σταθερή (ή ακόμη και μηδενική) σχεδόν σε όλο το εύρος της, είναι πιθανό το αποτέλεσμα της μεθόδου και το εκτιμώμενο σφάλμα να είναι ιδιαιτέρως ανακριβή. Επομένως, είναι κομβικής σημασίας η προσεκτική επιλογή των ορίων ολοκλήρωσης.

Για να ξεπεραστούν τα θέματα υπερχείλισης που η R αντιμετωπίζει συχνά, εφαρμόζεται το ακόλουθο μαθηματικό τέχνασμα χρησιμοποιώντας μια κατάλληλα ορισμένη σταθερά c και δουλεύοντας με τον λογάριθμο της κύριας προς ολοκλήρωση συνάρτησης $f\left(\boldsymbol{y}, \lambda_T|T\right)$:

$$\begin{aligned} \log \mathcal{I} &= \log \int \exp\left(\log f\left(\boldsymbol{y}, \lambda_T|T\right) \pm c\right) \mathrm{d}\lambda_T \\ &= \log \left(e^{-c} \int \exp\left(\log f\left(\boldsymbol{y}, \lambda_T|T\right) + c\right) \mathrm{d}\lambda_T \right) \\ &= -c + \log \int \exp\left(\log f\left(\boldsymbol{y}, \lambda_T|T\right) + c\right) \mathrm{d}\lambda_T \\ &= -c + \log \int \left[f\left(\boldsymbol{y}, \lambda_T|T\right) + c\right] \mathrm{d}\lambda_T \\ &= -c + \log \mathcal{I}_1 \end{aligned} \tag{2.25}$$

όπου $\mathcal{I}_1 = \int [f\left(\boldsymbol{y}, \lambda_T|T\right) + c] \mathrm{d}\lambda_T$. Εναλλακτικά, θα μπορούσε κανείς να εναλλάξει τις θέσεις των $\pm c$ στην (2.25) και να υλοποιήσει την ακόλουθη ισοδύναμη διαδικασία:

$$\begin{aligned} \log \mathcal{I} &= \log \int \exp\left(\log f\left(\boldsymbol{y}, \lambda_T|T\right) \pm c\right) \mathrm{d}\lambda_T \\ &= \log \left(e^{+c} \int \exp\left(\log f\left(\boldsymbol{y}, \lambda_T|T\right) - c\right) \mathrm{d}\lambda_T \right) \\ &= +c + \log \int \exp\left(\log f\left(\boldsymbol{y}, \lambda_T|T\right) - c\right) \mathrm{d}\lambda_T \end{aligned}$$

$$
\begin{aligned}
&= +c + \log \int [f\left(\boldsymbol{y}, \lambda_T | T\right) - c] \mathrm{d}\lambda_T \\
&= +c + \log \mathcal{I}_2.
\end{aligned}
$$

όπου $\mathcal{I}_2 = \int [f\left(\boldsymbol{y}, \lambda_T | T\right) - c] \mathrm{d}\lambda_T$.

Η σταθερά c ορίζεται ως η μέγιστη (ή η ελάχιστη αν βολεύει περισσότερο) απόλυτη τιμή της βασικής προς ολοκλήρωση συνάρτησης πάνω στο εύρος της και μπορεί να υπολογιστεί μέσω μιας σύντομης και εύκολης υπο-ρουτίνας με χρήση βρόχων.

2.5 Συμπεράσματα

Στο παρόν κεφάλαιο αναπτύξαμε τη Μπεϋζιανή μεθοδολογία για την επιλογή οικογένειας μετασχηματισμών σε μια διαδικασία δύο βημάτων για μονομεταβλητά προβλήματα, έχοντας ως στόχο να φέρουμε την κατανομή των μετασχηματισμένων τιμών της μεταβλητής απόκρισης όσο το δυνατόν πλησιέστερα στην κανονικότητα. Στην προσέγγιση που παρουσιάστηκε, θεωρήσαμε τέσσερις οικογένειες μετασχηματισμών (Box-Cox, Modulus, Yeo & Johnson και Dual) μαζί με τον τετριμμένο Ταυτοτικό μετασχηματισμό και τον Λογαριθμικό μετασχηματισμό. Η προταθείσα μεθοδολογία αναδεικνύει τη βέλτιστη επιλογή μετασχηματισμού επιλέγοντας την καταλληλότερη οικογένεια T και εκτιμώντας την τιμή της συναφούς παραμέτρου λ_T χρησιμοποιώντας μεθόδους Μπεϋζιανής επιλογής μοντέλου.

Γίνεται εμφανές ότι η κατασκευή εύλογων πρότερων κατανομών για την παράμετρο μετασχηματισμού είναι θεμελιώδους σημασίας λόγω της διαφορετικής ερμηνείας του λ_T ανάμεσα στις οικογένειες. Αυτό το ζήτημα αντιμετωπίστηκε μέσω της πρότερης κατανομής δύναμης. Δοκιμάστηκε επίσης και μια εναλλακτική προσέγγιση μέσω μιας κανονικής πρότερης κατανομής μοναδιαίας πληροφορίας για το λ_T (ή λογαριθμοκανονικής κατανομής όσον αφορά την οικογένεια Dual).

Ακολουθεί το Κεφάλαιο 3 όπου η προτεινόμενη μεθοδολογία παρουσιάζεται μέσω κατάλληλων παραδειγμάτων με προσομοιωμένα σύνολα δεδομένων αλλά και παραδείγματα δεδομένων πραγματικών προβλημάτων.

Κεφάλαιο 3

Εφαρμογές σε Μονομεταβλητά Προβλήματα

3.1 Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζονται κάποια βασικά παραδείγματα εφαρμογών με βάση την προταθείσα μεθοδολογία του Κεφαλαίου 2. Οι εφαρμογές αυτές βασίζονται σε μονομεταβλητά σύνολα δεδομένων. Τα αποτελέσματα που δίνονται είναι με βάση τον εκτιμητή του Chib όσον αφορά την εκτίμηση της περιθώριας πιθανοφάνειας (βλέπε Ενότητα 2.4.4). Ας σημειωθεί ότι όλα τα αρχικά σύνολα δεδομένων έχουν υποστεί τυποποίηση πριν από τον μετασχηματισμό τους και ακόμη ότι στην περίπτωση των μετασχηματισμών Box-Cox, Dual και Log οι παρατηρήσεις (είτε δειγματικές είτε φανταστικές) έχουν μετατοπισθεί στον θετικό άξονα προσθέτοντας την απόλυτη τιμή της ελάχιστης παρατήρησης συν το ήμισυ της μικρότερης αυστηρά θετικής τιμής y_0 των μη-αρνητικών δεδομένων (i.e., $\epsilon = y_0/2$).

Περιληπτικά, η Ενότητα 3.2 περιλαμβάνει εφαρμογές σε προσομοιωμένα σύνολα δεδομένων, με βάση πληθώρα κατανομών με συγκεκριμένες ιδιότητες κατάλληλα επιλεγμένων ώστε να αποτυπώνουν γλαφυρά τα δυνατά σημεία της μεθόδου. Οι εφαρμογές αυτές βασίζονται σε μονομεταβλητά σύνολα προσομοιωμένων δεδομένων από μια πληθώρα γνωστών κατανομών μεσαίου ($n = 100$) και μεγάλου ($n = 1000$) δειγματικού μεγέθους. Παράλληλα, διενεργείται μια ανάλυση ευαισθησίας πάλι στη βάση προσομοιωμένων δε-

δομένων. Η Ενότητα 3.3 μεταχειρίζεται ένα σύνολο δεδομένων από το πεδίο του ελέγχου ποιότητας και ακολουθούν σχετικές αναλύσεις ευαισθησίας και εδώ. Το κεφάλαιο καταλήγει με κάποιες συμπερασματικές σημειώσεις (Ενότητα 3.4).

3.2 Παραδείγματα σε Προσομοιωμένα Δεδομένα

Για την πρώτη εφαρμογή, προσομοιώσαμε δεδομένα από την τυπική κανονική κατανομή, οπότε ουσιαστικά πρόκειται για μια εφαρμογή αναφοράς (Πίνακας 3.1). Ξεκινώντας με ένα μέγεθος δείγματος $n = 100$, παρατηρούμε ότι ο Ταυτοτικός μετασχηματισμός υπερισχύει σαφώς, όπως αναμενόταν, ανεξάρτητα από το πλαίσιο πρότερης κατανομής. Συγκεκριμένα, η εκ των υστέρων πιθανότητα $P(T = \text{Id}|\boldsymbol{y}, \boldsymbol{y}^*)$ είναι 77% υπό τις Prior A και Prior B. Το δεύτερο κατά σειρά προτίμησης μοντέλο είναι το Box-Cox μοντέλο με εκ των υστέρων πιθανότητα ίση με περίπου 8% και για τα δύο πλαίσια πρότερης κατανομής και ύστερη κορυφή του λ_T γύρω στο 1.03 ώστε να διορθωθεί ακόμη και η ελάχιστη απόκλιση από την κανονικότητα λόγω προσομοίωσης. Ακολουθούν με μικρές διαφορές οι οικογένειες YJ και Modulus με ύστερες πιθανότητες περίπου $6\% - 8\%$ και ύστερες κορυφές του λ_T κοντά στη μονάδα. Για το δείγμα μεγάλου μεγέθους ($n = 1000$), ο Ταυτοτικός μετασχηματισμός αναδεικνύεται και πάλι ως η βέλτιστη επιλογή, μόνο που τώρα η ύστερη πιθανότητα του μοντέλου αυτού φτάνει τα επίπεδα του 88% και για τα δύο πλαίσια πρότερης κατανομής. Ακολουθεί η οικογένεια Box-Cox με ύστερη πιθανότητα 5% και ύστερη κορυφή της παραμέτρου λ_T ίση με 0.93 (δηλαδή πολύ κοντά στη μονάδα). Η βαρύτητα της οικογένειας Box-Cox είναι σχεδόν ίδια με τη βαρύτητα των μοντέλων Modulus και YJ με όρους ύστερων πιθανοτήτων. Σε κάθε περίπτωση, ο μετασχηματισμός Log είναι ο λιγότερο κατάλληλος καθώς η μη παραμετρική του φύση τον καθιστά μη ευέλικτο συγκριτικά με τις τέσσερις παραμετρικές οικογένειες που προσαρμόζονται καλύτερα ανάλογα με το εκάστοτε σύνολο δεδομένων. Ο μετασχηματισμός Dual επίσης δείχνει ακατάλληλος για τα συγκεκριμένα σύνολα δεδομένων.

Μια σημαντική ένδειξη της σύγκλισης των αποτελεσμάτων μεταξύ των δύο προσεγγίσεων Prior A και Prior B είναι η πολύ μικρή απόκλιση των τιμών της λογαριθμημένης περιθώριας πιθανοφάνειας. Κάπως μεγαλύτερες αποκλίσεις παρατηρούνται στην περίπτωση

της οικογένειας Dual η οποία συνδέεται με μια λογαριθμοκανονική (και όχι κανονική) πρότερη κατανομή υπό το πλαίσιο της Prior B. Εν γένει, η βέλτιστη τιμή της παραμέτρου λ_T είναι πολύ κοντά στη μονάδα για κάθε παραμετρική οικογένεια εκτός της οικογένειας Dual, επιβεβαιώνοντας ότι δεν υπάρχει ανάγκη για μετασχηματισμό των δεδομένων.

Πίνακας 3.1: Εκ των υστέρων πιθανότητες και τιμές της λογαριθμημένης περιθώριας πιθανοφάνειας για κάθε οικογένεια μετασχηματισμών T, καθώς και εκτιμητές Monte Carlo της ύστερης κορυφής (sd) της παραμέτρου λ_T για προσομοιωμένα κανονικά δεδομένα.

N(0,1)		Prior[a]	Id	Box-Cox	YJ	Modulus	Dual	Log
	$P(T\|\boldsymbol{y},\boldsymbol{y}^*)$	Prior A	0.77	0.08	0.07	0.06	< 0.01	< 0.01
		Prior B	0.77	0.08	0.08	0.07	< 0.01	< 0.01
$n = 100$	$\log f(\boldsymbol{y}\|\boldsymbol{y}^*,T)$	Prior A	-195.10	-197.33	-197.51	-197.57	-203.01	-214.21
		Prior B	-195.10	-197.33	-197.36	-197.48	-202.82	-214.21
	λ_T	Prior A	—	1.03 (0.20)	1.07 (0.13)	0.98 (0.28)	1.48 (0.21)	—
		Prior B	—	1.03 (0.20)	1.07 (0.13)	0.98 (0.28)	1.45 (0.20)	—

N(0,1)		Prior	Id	Box-Cox	Modulus	YJ	Log	Dual
	$P(T\|\boldsymbol{y},\boldsymbol{y}^*)$	Prior A	0.88	0.05	0.04	0.03	< 0.01	< 0.01
		Prior B	0.88	0.05	0.04	0.03	< 0.01	< 0.01
$n = 1000$	$\log f(\boldsymbol{y}\|\boldsymbol{y}^*,T)$	Prior A	-3106.81	-3109.63	-3109.94	-3110.23	-3436.48	-3441.49
		Prior B	-3106.81	-3109.75	-3109.90	-3110.18	-3436.48	-3440.46
	λ_T	Prior A	—	0.93 (0.06)	1.08 (0.09)	0.98 (0.04)	—	0.01 (0.01)
		Prior B	—	0.93 (0.06)	1.08 (0.09)	0.98 (0.04)	—	0.01 (0.01)

[a] Prior A: πρότερη κατανομή δύναμης (βλέπε Ενότητα 2.4.1.1), Prior B: κανονική πρότερη κατανομή μοναδιαίας πληροφορίας (βλέπε Ενότητα 2.4.1.2).

Εν συνεχεία, παρουσιάζουμε ένα παράδειγμα με προσομοιωμένα δεδομένα από μια κατανομή γάμμα$(2, 3)$ ώστε να εξετάσουμε τη συμπεριφορά της προταθείσας μεθοδολογίας σε πολύ ασύμμετρες κατανομές (Πίνακας 3.2). Οι οικογένειες που προσαρμόζονται καλύτερα στα δεδομένα είναι ξεκάθαρα η οικογένεια Box-Cox και για το μέτριο αλλά και για το μεγάλο μέγεθος δείγματος, με τον Ταυτοτικό μετασχηματισμό να καταλαμβάνει την

τελευταία θέση όπως αναμενόταν. Καθώς μετακινούμαστε από το πρώτο μέγεθος δείγματος στο δεύτερο, οι ύστερες πιθανότητες του μοντέλου Box-Cox διατηρούνται άνω του 99%, ενώ η αντίστοιχη ύστερη κορυφή του λ_T πέφτει ελάχιστα από 0.44 σε 0.35. Παρατηρήστε πώς η ύστερη τυπική απόκλιση του λ_T περίπου υποτριπλασιάζεται όταν $n = 1000$ συγκριτικά με $n = 100$. Για το μέτριο μέγεθος δείγματος, στο μοντέλο YJ με ύστερη κορυφή του λ_T ίση με 0.43 αποδίδεται ένα ελάχιστο βάρος της τάξης του 1% το οποίο γίνεται αμελητέο όταν περνάμε στο σύνολο δεδομένων μεγαλύτερου μεγέθους.

Πίνακας 3.2: Εκ των υστέρων πιθανότητες και τιμές της λογαριθμημένης περιθώριας πιθανοφάνειας για κάθε οικογένεια μετασχηματισμών T, καθώς και εκτιμητές Monte Carlo της ύστερης κορυφής (sd) της παραμέτρου λ_T για προσομοιωμένα γάμμα δεδομένα.

G(2,3)		Prior[a]	Box-Cox	YJ	Id	Modulus	Log	Dual
	$P(T\|\boldsymbol{y},\boldsymbol{y}^*)$	Prior A	0.99	0.01	< 0.01	< 0.01	< 0.01	< 0.01
		Prior B	0.99	0.01	< 0.01	< 0.01	< 0.01	< 0.01
$n = 100$	$\log f(\boldsymbol{y}\|\boldsymbol{y}^*,T)$	Prior A	-184.26	-190.48	-195.10	-197.15	-197.28	-199.97
		Prior B	-184.19	-190.35	-195.10	-197.08	-197.28	-200.38
	λ_T	Prior A	0.44 (0.09)	0.43 (0.16)	—	1.27 (0.26)	—	0.01 (0.05)
		Prior B	0.44 (0.09)	0.43 (0.16)	—	1.26 (0.26)	—	0.04 (0.04)

G(2,3)		Prior	Box-Cox	YJ	Log	Dual	Modulus	Id
	$P(T\|\boldsymbol{y},\boldsymbol{y}^*)$	Prior A	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
		Prior B	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
$n = 1000$	$\log f(\boldsymbol{y}\|\boldsymbol{y}^*,T)$	Prior A	-2957.54	-2997.08	-3014.48	-3018.28	-3105.11	-3106.81
		Prior B	-2957.62	-2996.90	-3014.48	-3018.13	-3105.01	-3106.81
	λ_T	Prior A	0.35 (0.03)	0.31 (0.05)	—	0.01 (0.03)	0.76 (0.07)	—
		Prior B	0.35 (0.03)	0.31 (0.05)	—	0.02 (0.03)	0.76 (0.07)	—

[a] Prior A: πρότερη κατανομή δύναμης (βλέπε Ενότητα 2.4.1.1), Prior B: κανονική πρότερη κατανομή μοναδιαίας πληροφορίας (βλέπε Ενότητα 2.4.1.2).

Τέλος, η κατανομή Student είναι ένα παράδειγμα ιδιαίτερα ενδιαφέρουσας κατανομής αφού συνδυάζει συμμετρία με παχιές ουρές. Σύμφωνα με τη μέχρι τώρα εμπειρία, το χα-

ρακτηριστικό αυτό συνήθως επιφέρει αποτυχία των προσπαθειών μετασχηματισμού προς την κανονικότητα για τις περισσότερες οικογένειες. Στο παράδειγμά μας, έχουμε προσομοιώσει δεδομένα από μια κατανομή Student με 2 βαθμούς ελευθερίας και παράμετρο μη κεντρικότητας ίση με -1. Κοιτώντας τον Πίνακα 3.3, παρατηρεί κανείς την αδιαμφισβήτητη υπεροχή της οικογένειας Modulus ανεξαρτήτως πρότερης προσέγγισης. Ακόμη και για το μικρότερο εκ των δύο δειγμάτων ($n = 100$), η ύστερη πιθανότητα του μοντέλου Modulus είναι γύρω στο $96\% - 97\%$ αποδίδοντας ένα βάρος περίπου 3% συνδυαστικά για τα μοντέλα Box-Cox, YJ και Id. Για $n = 1000$ το προηγούμενο βάρος υπερβαίνει το 99% για το μοντέλο Modulus. Η αντίστοιχη ύστερη κορυφή της παραμέτρου λ_T είναι περίπου 0.12 για το μέτριο μέγεθος δείγματος και -0.42 για το μεγάλο μέγεθος δείγματος, ενώ η ύστερη τυπική απόκλιση είναι ίση με 0.25 και 0.08 αντίστοιχα. Αξίζει να σημειωθεί ότι παρεμφερής συμπεριφορά και υποστήριξη της οικογένειας Modulus απαντάται και σε περιπτώσεις προσομοιωμένων δεδομένων από την κατανομή Laplace η οποία αποτελεί ακόμη ένα παράδειγμα συμμετρικής κατανομής με παχιές ουρές.

Σε γενικές γραμμές, παρατηρήθηκαν αμελητέες διαφορές στην περιθώρια πιθανοφάνεια μεταξύ των δύο πρότερων προσεγγίσεων, υποδεικνύοντας ότι η Prior A και η Prior B δίνουν συμβατά αποτελέσματα όπως στοχεύαμε. Κάπως πιο συστηματικές αποκλίσεις παρατηρήθηκαν στο μοντέλο Dual όπου σημειωτέον δεν υπάρχει θεωρητική τιμή της παραμέτρου μετασχηματισμού που να αντιστοιχεί στο κανονικό μοντέλο αναφοράς, ενώ ειδικά η κατανομή της Prior A αποκλίνει σημαντικά από την κανονικότητα.

Σαν ένα εργαλείο καλής προσαρμογής ενός μοντέλου, προσομοιώνουμε από την a posteriori κατανομή της τετραγωνικής ρίζας της μέσης τιμής των τετραγώνων των σφαλμάτων (*root mean square errors , RMSE*) με βάση διαγράμματα κανονικής πιθανότητας (*normal probability plots*) των τυποποιημένων μετασχηματισμένων δεδομένων για κάθε αλυσίδα MCMC της παραμέτρου λ_T δεδομένης της οικογένειας T. Αναφερόμαστε σε αυτές τις ποσότητες ως μέτρα QQ-RMSE ή απλώς ως QQ-RMSE. Για κάθε οικογένεια μετασχηματισμού T με δεδομένη τιμή για την παράμετρο λ_T, οι ποσότητες αυτές υπολογίζονται από τη σχέση

$$\text{QQ-RMSE}(\lambda_T, T) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[z_i^{(\lambda_T)} - Q_{i/n}\right]^2}$$

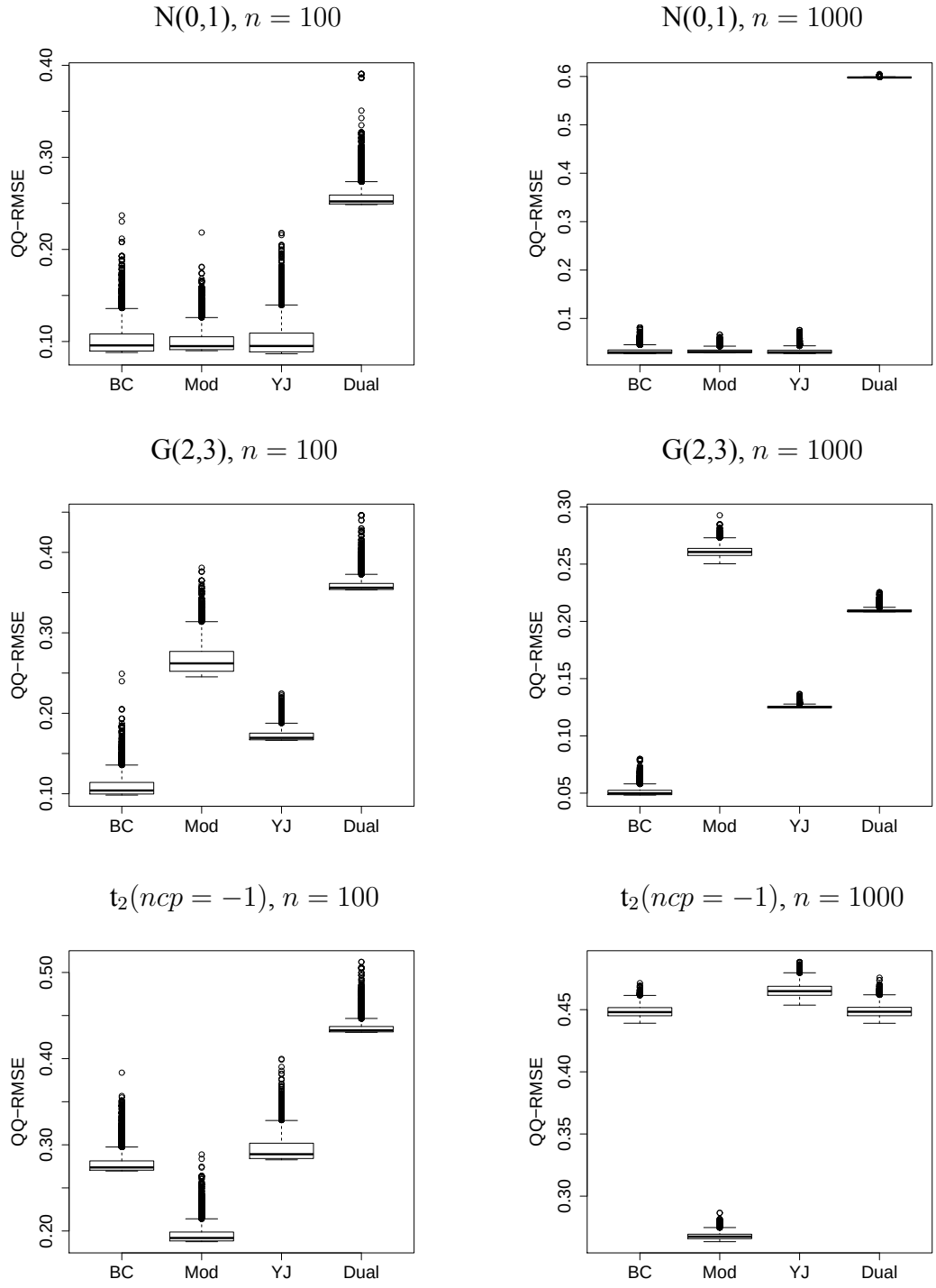
Πίνακας 3.3: Εκ των υστέρων πιθανότητες και τιμές της λογαριθμημένης περιθώριας πιθανοφάνειας για κάθε οικογένεια μετασχηματισμών T, καθώς και εκτιμητές Monte Carlo της ύστερης κορυφής (sd) της παραμέτρου λ_T για προσομοιωμένα Student δεδομένα.

$\mathrm{t}_2(ncp=-1)$		Prior[a]	Modulus	Box-Cox	YJ	Id	Dual	Log
	$P(T\|\boldsymbol{y},\boldsymbol{y}^*)$	Prior A	0.96	0.01	0.01	0.01	< 0.01	< 0.01
		Prior B	0.97	0.01	0.01	0.01	< 0.01	< 0.01
$n = 100$	$\log f(\boldsymbol{y}\|\boldsymbol{y}^*,T)$	Prior A	-190.33	-194.58	-195.05	-195.10	-204.65	-242.27
		Prior B	-190.21	-194.54	-194.93	-195.10	-204.69	-242.27
	λ_T	Prior A	0.12 (0.25)	1.43 (0.19)	1.24 (0.10)	—	1.99 (0.20)	—
		Prior B	0.12 (0.25)	1.43 (0.19)	1.24 (0.10)	—	1.98 (0.21)	—

$\mathrm{t}_2(ncp=-1)$		Prior	Modulus	YJ	Dual	Box-Cox	Id	Log
	$P(T\|\boldsymbol{y},\boldsymbol{y}^*)$	Prior A	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
		Prior B	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
$n = 1000$	$\log f(\boldsymbol{y}\|\boldsymbol{y}^*,T)$	Prior A	-2830.18	-2941.06	-2948.13	-2948.57	-3106.81	-3461.70
		Prior B	-2830.12	-2940.90	-2948.51	-2948.68	-3106.81	-3461.70
	λ_T	Prior A	-0.42 (0.08)	1.46 (0.02)	3.01 (0.12)	3.01 (0.12)	—	—
		Prior B	-0.42 (0.08)	1.46 (0.02)	3.01 (0.12)	3.01 (0.12)	—	—

[a] Prior A: πρότερη κατανομή δύναμης (βλέπε Ενότητα 2.4.1.1), Prior B: κανονική πρότερη κατανομή μοναδιαίας πληροφορίας (βλέπε Ενότητα 2.4.1.2).

όπου $z_i^{(\lambda_T)}$ είναι η τυποποιημένη τιμή του $y_i^{(\lambda_T)}$ και Q_p είναι το p-οστό ποσοστημόριο της τυπικής κανονικής κατανομής. Το Διάγραμμα 3.1 αποτελεί έναν περιεκτικό και λακωνικό τρόπο για να συνοψισθεί η ύστερη αβεβαιότητα του QQ-RMSE για κάθε προσομοιωμένο σύνολο δεδομένων και για κάθε μια από τις τέσσερις παραμετρικές οικογένειες υπό σύγκριση. Η συμπερασματολογία όπως προκύπτει από την ύστερη κατανομή του μέτρου QQ-RMSE είναι σε συμφωνία με τις ύστερες πιθανότητες των Πινάκων 3.1–3.3.



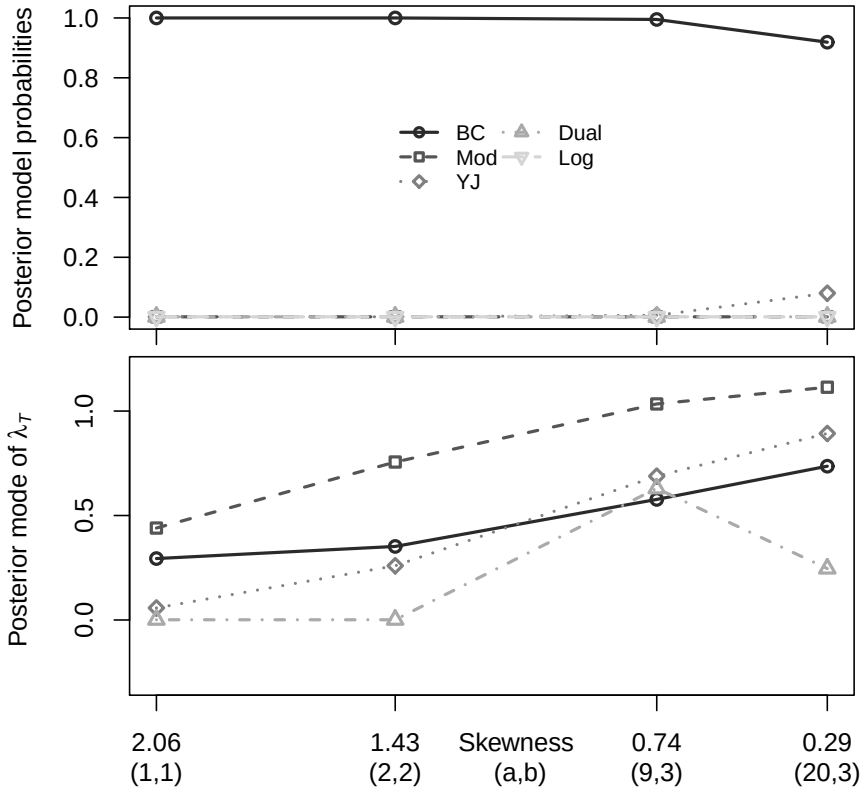
Διάγραμμα 3.1: Θηκογράμματα που συνοψίζουν την a posteriori κατανομή του QQ-RMSE κάτω από τις διάφορες οικογένειες μετασχηματισμών για τα προσομοιωμένα σύνολα δεδομένων της Ενότητας 3.2.

3.2.1 Ανάλυση ευαισθησίας με βάση τα προσομοιωμένα παραδείγματα

Σε αυτή την ενότητα, διεξάγουμε μια ανάλυση ευαισθησίας αναπαριστώντας γραφικά την επίδραση των παραμέτρων σχήματος (*shape*) ή/και ρυθμού (*rate*) των κατανομών γάμμα και Student στην ύστερη κορυφή της παραμέτρου λ_T και στην ύστερη πιθανότητα κάθε οικογένειας μετασχηματισμών.

Σε μια προσπάθεια να αποκτήσουμε μια γενικότερη εικόνα για τη συμπεριφορά των καταλληλότερων οικογενειών σχετικά με τον μετασχηματισμό δεδομένων γάμμα, εφαρμόσαμε την προτεινόμενη μεθοδολογία στα πλαίσια της προσέγγισης Prior A για διάφορους συνδυασμούς των παραμέτρων σχήματος και ρυθμού (a, b) της κατανομής κρατώντας σταθερό το μέγεθος δείγματος ($n = 1000$). Η θεωρητική ασυμμετρία της κατανομής γάμμα μειώνεται με την αύξηση της τιμής της παραμέτρου σχήματος, ενώ μια μείωση του ρυθμού της κατανομής διευρύνει τη διασπορά αυτής. Το Διάγραμμα 3.2 απεικονίζει τις εκ των υστέρων πιθανότητες $P(T|\boldsymbol{y}, \boldsymbol{y}^*)$ των πέντε καλύτερων οικογενειών μετασχηματισμών για την κατανομή γάμμα συναρτήσει του βαθμού ασυμμετρίας της κατανομής. Το κάτω γράφημα του εν λόγω διαγράμματος αποτυπώνει την ύστερη κορυφή της παραμέτρου λ_T για κάθε παραμετρική οικογένεια συναρτήσει της δειγματικής ασυμμετρίας. Ακόμη, δίνονται οι αντίστοιχοι συνδυασμοί των παραμέτρων (a, b) της κατανομής στον οριζόντιο άξονα κάτω από τις τιμές του βαθμού ασυμμετρίας. Ας σημειωθεί ότι στον πρώτο συνδυασμό τιμών (a, b) η παράμετρος σχήματος ισούται με μονάδα και άρα η γάμμα κατανομή εκφυλίζεται σε εκθετική με μέση τιμή ίση με $1/b$. Για τις υψηλότερες τιμές ασυμμετρίας του γραφήματος, δηλαδή για τις τιμές 2.0 και 1.4, παρατηρούμε ότι το μοντέλο Box-Cox υπερτερεί των υπολοίπων με ύστερη πιθανότητα άνω του 0.99. Για βαθμό ασυμμετρίας ίσο με 0.7, η ύστερη πιθανότητα του μοντέλου Box-Cox τείνει να ελαττωθεί σε αντίθεση με το μοντέλο YJ το οποίο αναδύεται για πρώτη φορά με ύστερη πιθανότητα ίση με 1%. Για χαμηλή ασυμμετρία της τάξης του 0.3, η οικογένεια Box-Cox έχει ακόμη προβάδισμα με ύστερη πιθανότητα 0.92, ενώ η οικογένεια YJ έρχεται δεύτερη με a posteriori πιθανότητα 8%. Τα υπόλοιπα μοντέλα δεν παίζουν σημαντικό ρόλο, ανεξάρτητα από το ύψος του βαθμού ασυμμετρίας. Αναφορικά με την τιμή της ύστερης κορυφής της παραμέτρου λ_T για τις διάφορες οικογένειες εκτός της οικογένειας Dual,

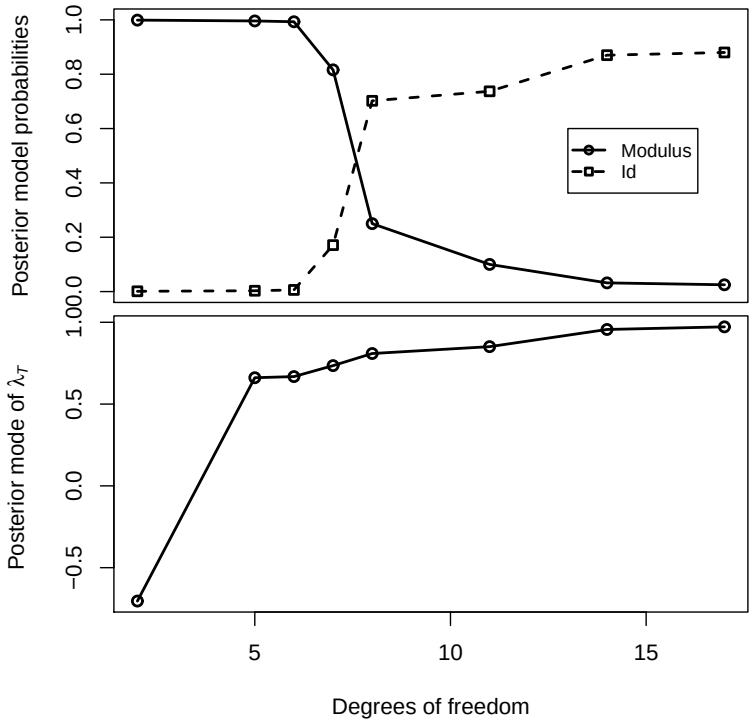
παρουσιάζει μια σταδιακή αύξηση προς τη μονάδα καθώς ο βαθμός ασυμμετρίας μειώνεται. Ειδικά για την περίπτωση του μετασχηματισμού Box-Cox, η ύστερη κορυφή της παραμέτρου λ_T είναι περίπου 0.3 για υψηλή ασυμμετρία και φτάνει κοντά στο 0.7 για χαμηλή ασυμμετρία. Οι αντίστοιχες τιμές του λ_T για την οικογένεια YJ εκτείνονται από τιμές κοντά στο μηδέν έως και 0.8.



Διάγραμμα 3.2: Ύστερες πιθανότητες $P(T|\boldsymbol{y}, \boldsymbol{y}^*)$ των μοντέλων και ύστερη κορυφή της παραμέτρου λ_T κάτω από τις οικογένειες Box-Cox, Modulus, Yeo & Johnson, Dual και Log έναντι του βαθμού δειγματικής ασυμμετρίας για προσομοιωμένα γάμμα(a,b) δεδομένα μεγέθους $n = 1000$ (οι αντίστοιχοι συνδυασμοί των (a,b) δίνονται σε παρένθεση κάτω από τις τιμές της ασυμμετρίας).

Μια παρόμοια διαδικασία ακολουθήθηκε για την κεντρική κατανομή Student (μηδενική παράμετρος μη κεντρικότητας). Το Διάγραμμα 3.3 παρέχει μια σύγκριση ανάμεσα στις ύστερες πιθανότητες των μοντέλων Modulus και Id (δηλαδή των μόνων ανταγωνιστικών μοντέλων σε αυτό το παράδειγμα) έναντι των βαθμών ελευθερίας (df) της κατα-

νομής για σταθερό μέγεθος δείγματος $n = 1000$. Για κατανομές με παχιές ουρές (δηλ. με χαμηλούς βαθμούς ελευθερίας) το μοντέλο Modulus κυριαρχεί, ενώ η εκ των υστέρων στήριξη στο μοντέλο Id αυξάνει όσο ανεβαίνουν οι βαθμοί ελευθερίας και η κατανομή Student προσεγγίζει όλο και περισσότερο την κανονική. Το κάτω γράφημα του διαγράμματος απεικονίζει τη συμπεριφορά της ύστερης κορυφής της παραμέτρου λ_T για την οικογένεια Modulus συναρτήσει των βαθμών ελευθερίας της κατανομής Student, δείχνοντας ξεκάθαρα ότι η ύστερη κορυφή του λ_T προσεγγίζει τη μονάδα καθώς οι βαθμοί ελευθερίας αυξάνουν. Αυτό είναι απολύτως εύλογο καθώς (i) η κατανομή Student προσεγγίζει την κανονική κατανομή καθώς οι βαθμοί ελευθερίας αυξάνουν και (ii) η παράμετρος μετασχηματισμού λ_T ισούται με 1 για κανονικά κατανεμημένα δεδομένα υπό τον μετασχηματισμό Modulus.



Διάγραμμα 3.3: Εκ των υστέρων πιθανότητες $P(T|\boldsymbol{y}, \boldsymbol{y}^*)$ των μοντέλων κάτω από τους μετασχηματισμούς Modulus και Id, καθώς και τιμές της ύστερης κορυφής της παραμέτρου λ_T υπό την οικογένεια Modulus συναρτήσει των βαθμών ελευθερίας (*df*) για δείγματα μεγέθους $n = 1000$ προσομοιωμένα από την κεντρική Student κατανομή.

3.3 Παράδειγμα Ελέγχου Ποιότητας: Παρακολούθηση Σφαλμάτων στη Βάση Δεδομένων του Πελάτη

Παραδοσιακές μέθοδοι στο πεδίο του ελέγχου ποιότητας, όπως τα ατομικά διαγράμματα ελέγχου (*individual control charts*), προϋποθέτουν ότι η μέτρηση κάποιας μεταβλητής απόκρισης μιας βιομηχανικής διαδικασίας είναι κανονικά κατανεμημένη. Στην πράξη, όμως, υπάρχουν ποικίλα χαρακτηριστικά σχετικά με την ποιότητα των οποίων η κατανομή είναι ουσιωδώς διαφορετική από την κανονική κατανομή. Σε ανάλογες περιπτώσεις, τα κλασικά ατομικά διαγράμματα ελέγχου δεν είναι αξιόπιστα για την παρακολούθηση της βιομηχανικής διαδικασίας καθότι τα πραγματικά ποσοστά ψευδών συναγερμών θα είναι διαφορετικά από τα υποθετικά ποσοστά ψευδών συναγερμών υπό την προϋπόθεση της κανονικότητας. Τα ποσοστά ψευδών συναγερμών είναι υψηλής σημασίας σε διαδικασίες ελέγχου ποιότητας καθώς η υπερ-εκτίμησή τους ενδέχεται να οδηγήσει σε σημαντική σπατάλη πόρων, ενώ η υπο-εκτίμησή τους θα κάνει τη μέθοδο ελέγχου ανεπαρκή και μη ικανή να εντοπίσει διαφοροποιήσεις και προβλήματα που ενδεχομένως να εμφανιστούν στη βιομηχανική διαδικασία με αποτέλεσμα ένα ανεπιθύμητο πλήθος ελαττωματικών προϊόντων (Qiu & Zhang 2014).

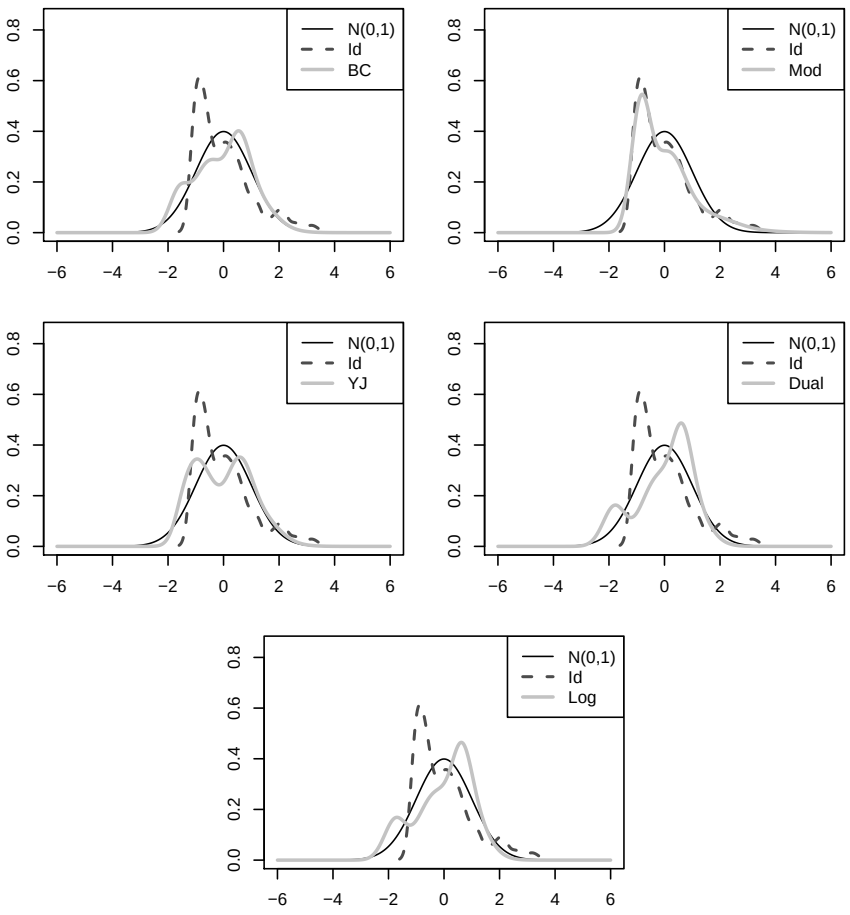
Σε αυτό το παράδειγμα, θεωρούμε 125 τιμές από τα σφάλματα που εντοπίστηκαν στα πελατειακά αρχεία της βάσης δεδομένων μιας εταιρείας μάρκετινγκ (Παράδειγμα 7.E19 στο βιβλίο του Montgomery (2009)). Τέτοια δεδομένα χρησιμοποιούνται συχνά στην παρακολούθηση της ποιότητας των πελατειακών καταλόγων σε παρόμοιες εταιρείες και ακολουθούν σοβαρά ασύμμετρες κατανομές. Επομένως, δε μπορούν να χρησιμοποιηθούν άμεσα στα σχετικά ατομικά διαγράμματα ποιότητας τα οποία προϋποθέτουν κανονική κατανομή.

Ο Πίνακας 3.4 παρουσιάζει τα a posteriori αποτελέσματα για κάθε οικογένεια μετασχηματισμών υπό την πρότερη προσέγγιση Prior A. Τα σχετικά αποτελέσματα υπό την προσέγγιση Prior B είναι απολύτως ταυτόσημα και παραλείπονται για συντομία. Σύμφωνα με τα αποτελέσματα, η οικογένεια Box-Cox υποστηρίζεται εκ των υστέρων με βάρος 57% ακολουθούμενη από τον μετασχηματισμό Log με ύστερο βάρος 34% και από την οικογένεια Dual με 9%.

Πίνακας 3.4: Εκ των υστέρων πιθανότητες (υπό την Prior A) για κάθε οικογένεια μετασχηματισμών T, καθώς και η ύστερη διάμεσος (sd) του QQ-RMSE και του QQ-MAD και η ύστερη κορυφή της παραμέτρου λ_T (sd) για τα δεδομένα ελέγχου ποιότητας.

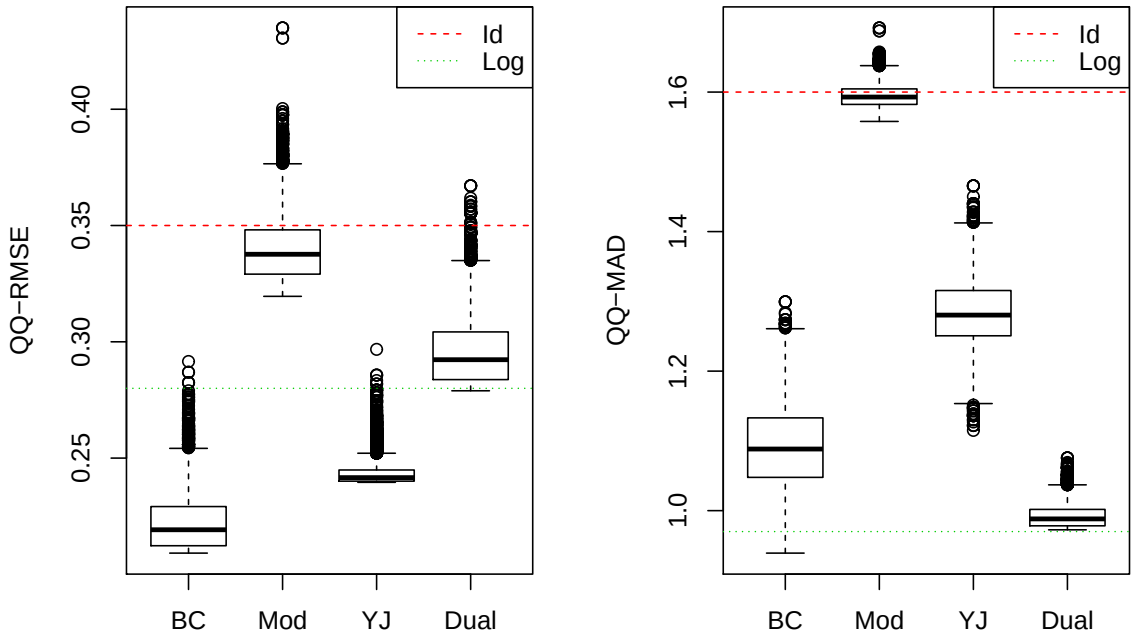
	Box-Cox	Log	Dual	YJ	Id	Modulus
$P(T\|\boldsymbol{y}, \boldsymbol{y}^*)$	0.57	0.34	0.09	< 0.01	< 0.01	< 0.01
QQ-RMSE	0.21 (0.01)	0.28 (–)	0.29 (0.01)	0.24 (0.01)	0.35 (–)	0.34 (0.01)
QQ-MAD	1.09 (0.06)	0.97 (–)	0.99 (0.02)	1.28 (0.05)	1.60 (–)	1.59 (0.02)
λ_T	0.24 (0.09)	—	0.38 (0.16)	0.15 (0.15)	—	0.83 (0.23)

Το Διάγραμμα 3.4 απεικονίζει την εκτιμώμενη πυκνότητα πιθανότητας για κάθε μετασχηματισμό υπολογισμένη στην ύστερη κορυφή της παραμέτρου λ_T (Πίνακας 3.4). Όλες



Διάγραμμα 3.4: Διαγράμματα πυκνότητας πιθανότητας των (τυποποιημένων) μετασχηματισμένων παρατηρήσεων για τα δεδομένα ελέγχου ποιότητας. Τα διαγράμματα της τυπικής κανονικής και της Id περίπτωσης συμπεριλαμβάνονται για λόγους αναφοράς.

οι κατανομές έχουν υποστεί τυποποίηση ώστε να είναι οπτικά συγκρίσιμες. Σε κάθε ένα από τα γραφήματα του Διαγράμματος 3.4 έχουν προστεθεί η τυπική κανονική κατανομή και η κατανομή του μετασχηματισμού Id ώστε να καταστεί δυνατή η άμεση σύγκριση κάθε οικογένειας με αυτές ως σημεία αναφοράς. Τα διαγράμματα πυκνότητας πιθανότητας είναι σε στενή συμφωνία με τις ύστερες πιθανότητες των μοντέλων όπως παρουσιάστηκαν στον Πίνακα 3.4 με το μοντέλο BC να έρχεται πλησιέστερα στην κανονικότητα σε σχέση με τα υπόλοιπα μοντέλα. Ο μετασχηματισμός Log είναι παρεμφερής με τον μετασχηματισμό Dual, ενώ η οικογένεια YJ εμφανίζεται ως δικόρυφη κατανομή.



Διάγραμμα 3.5: Θηκογράμματα που συνοψίζουν την ύστερη κατανομή του QQ-RMSE (αριστερά) και του QQ-MAD (δεξιά) κάτω από τις διάφορες οικογένειες μετασχηματισμών για τα δεδομένα ελέγχου ποιότητας.

Επιπρόσθετα, υπολογίσαμε τις a posteriori τιμές του QQ-RMSE (βλέπε Πίνακα 3.4) οι οποίες σε γενικές γραμμές είναι σύμφωνες με τη μεθοδολογία που αναπτύξαμε περί ύστερης σύγκρισης των μοντέλων. Ενδεικτικό, ως προς αυτό, είναι και το αριστερό γράφημα του Διαγράμματος 3.5 όπου παρατίθενται και οι αντίστοιχες τιμές για τους μετασχημα-

τισμούς Id και Log με δύο γραμμές αναφοράς. Σημειώνεται μια ιδιαιτερότητα ως προς την ύστερη κατανομή του QQ-RMSE στην περίπτωση του μοντέλου YJ η οποία, σε μια πρώτη ανάγνωση, φαίνεται να υποδεικνύει ότι έχουμε καλύτερη προσαρμογή. Παρόλα αυτά, αν εστιάσουμε στη μέγιστη απόλυτη απόσταση των ποσοστημορίων

$$\text{QQ-MAD}(\lambda_T, T) = \max_{i=1,\dots,n} \left\{ \left| z_i^{(\lambda_T)} - Q_{i/n} \right| \right\}$$

(βλέπε Πίνακα 3.4) με παρόμοιο τρόπο όπως ο έλεγχος κανονικότητας Kolmogorov - Smirnov εστιάζει στη μέγιστη απόσταση μεταξύ των αθροιστικών πιθανοτήτων, βλέπουμε καθαρά την κακή απόδοση της οικογένειας YJ ειδικά όταν συγκρίνεται με τους μετασχηματισμούς Log και Dual. Η ύστερη κατανομή του μέτρου QQ-MAD απεικονίζεται στο δεξί γράφημα του Διαγράμματος 3.5. Άρα, η προσέγγισή μας φαίνεται να είναι ευαίσθητη σε ατομικές αποκλίσεις από την κανονικότητα. Αυτό, σε συνδυασμό με το γεγονός ότι ο παράγοντας Bayes ενεργεί σαν ξυράφι του Occam (*Occam's razor*) το οποίο ποινικοποιεί μοντέλα υψηλότερης παραμετρικής διάστασης, καταλήγει σε σημαντικά υψηλότερο a posteriori βάρος για τον Λογαριθμικό μετασχηματισμό έναντι του μετασχηματισμού YJ.

3.3.1 Ενσωματώνοντας πληροφορία στην πρότερη κατανομή

Συνήθως, στην περιοχή της βιομηχανικής στατιστικής, πρότερη πληροφορία περί μετασχηματισμού των τιμών της μεταβλητής ενδιαφέροντος διατίθεται από προηγούμενα πειράματα ή δοκιμές. Παραδείγματα μπορεί κανείς να δει στο βιβλίο του Montgomery [2009, p. 323–326]. Σε τέτοιες περιπτώσεις, μπορούμε να ενσωματώσουμε τη διαθέσιμη πληροφορία στην προσέγγισή μας υλοποιώντας την ακόλουθη διαδικασία.

Ας υποθέσουμε ότι διαθέτουμε κάποιο είδος πρότερης πληροφορίας για την τιμή της παραμέτρου μετασχηματισμού μιας συγκεκριμένης οικογένειας μετασχηματισμών. Για παράδειγμα, ας θεωρήσουμε την περίπτωση όπου προηγούμενες έρευνες έχουν ταυτοποιήσει την τιμή λ_0 για μια οικογένεια μετασχηματισμών T_0 η οποία επιφέρει κανονικότητα σε ένα σύνολο δεδομένων. Έχουμε τη δυνατότητα να ενσωματώσουμε αυτή την πρότερη πληροφορία στον μηχανισμό των φανταστικών δεδομένων μέσω του αναπτύγματος Taylor της συνάρτησης $h(y) = y^{(\lambda_0)}$, χρησιμοποιώντας την οικογένεια T_0. Επομένως, με βάση το ανάπτυγμα Taylor δευτέρου βαθμού, η μέση τιμή της μετασχηματισμένης τ.μ.

μπορεί να προσεγγιστεί από τη σχέση:

$$\mathrm{E}[h(Y)] \simeq h(\mu) + \frac{h''(\mu)}{2}\,\mathrm{Var}[Y], \tag{3.1}$$

όπου $\mu = \mathrm{E}[Y]$ και $h''(y)$ είναι η δεύτερη παράγωγος της συνάρτησης $h(y)$. Παρομοίως, η διακύμανση της συνάρτησης $h(Y)$, διατηρώντας τους πρώτους δύο όρους του αναπτύγματος Taylor, μπορεί να προσεγγιστεί από τη σχέση:

$$\mathrm{Var}[h(Y)] \simeq (h'(\mu))^2\,\mathrm{Var}[Y], \tag{3.2}$$

όπου $h'(y)$ είναι η πρώτη παράγωγος της συνάρτησης $h(y)$. Επιστρέφοντας στο διάνυσμα δεδομένων $\boldsymbol{y}$, για τους μετασχηματισμούς BC και Dual αρχικά θεωρείται η τυποποιημένη μορφή των αρχικών δεδομένων τα οποία στη συνέχεια μετατοπίζονται στον θετικό άξονα κατά $\xi = |\min(\boldsymbol{y})| + \epsilon$. Συνεπώς, $\mu = \xi$ και $\mathrm{Var}[Y] = 1$. Για τις υπόλοιπες οικογένειες θεωρούμε απλά τα τυποποιημένα δεδομένα και άρα $\xi = 0$.

Η διαδικασία κατασκευής των πρότερων κατανομών με βάση διαθέσιμη πρότερη πληροφορία συνοψίζεται στα εξής βήματα:

1. Υπολόγισε προσεγγιστικά τη μέση τιμή $\mu_\xi = \mathrm{E}[h(Y)]$ και τη διακύμανση $\sigma_\xi^2 = \mathrm{Var}[h(Y)]$ των μετασχηματισμένων δεδομένων όπως δίνονται απο τις (3.1) και (3.2) αντίστοιχα για $\mu = \xi$ και $\mathrm{Var}[Y] = 1$.

2. Προσομοίωσε ένα σύνολο τυχαίων τιμών $z_i^* \sim N(\mu_\xi, \sigma_\xi^2)$ για $i = 1, \ldots, n^*$.

3. Κατασκεύασε τα φανταστικά δεδομένα με βάση το ανάπτυγμα Taylor χρησιμοποιώντας τον αντίστροφο μετασχηματισμό της οικογένειας T_0 με $\lambda_{T_0} = \lambda_0$, δηλαδή $y_i^* = h^{-1}(z_i^*)$.

4. Χρησιμοποίησε τα φανταστικά δεδομένα $\boldsymbol{y}^*$ του βήματος 3 ώστε να κατασκευάσεις την Prior A ή την Prior B για κάθε οικογένεια μετασχηματισμών υπό μελέτη.

Εδώ δείχνουμε πώς η παραπάνω διαδικασία εκτελείται στο σύνολο δεδομένων ελέγχου ποιότητας όταν διαθέτουμε πρότερη πληροφορία σχετικά με την παράμετρο μετασχηματισμού της οικογένειας Box-Cox (η οποία είναι η πιο συχνά χρησιμοποιούμενη οικογένεια σε τέτοιου είδους δεδομένα). Συγκεκριμένα για την οικογένεια Box-Cox, ο μέσος και η διακύμανση των μετασχηματισμένων δεδομένων (από το βήμα 1) προσεγγίζονται

από τις σχέσεις $\mu_\xi \simeq \frac{\xi^{\lambda_0}-1}{\lambda_0} + \frac{(\lambda_0-1)\xi^{\lambda_0-2}}{2}$ και $\sigma^2_\xi \simeq \xi^{2(\lambda_0-1)}$. Η αντίστροφη συνάρτηση δίνεται από τη σχέση $h^{-1}(z^*) = (z^*\lambda_0 + 1)^{\frac{1}{\lambda_0}}$ υπό τους περιορισμούς $(z^*\lambda_0 + 1) \geq 0$ και $\lambda_0 \neq 0$.

Τα στοιχεία του Πίνακα 3.5 περιγράφουν πώς ποικίλουν οι a posteriori πιθανότητες των έξι οικογενειών συναρτήσει της τιμής της παραμέτρου λ_0 που ενσωματώνεται στην πρότερη κατανομή μέσω των φανταστικών δεδομένων $\boldsymbol{y}^*$ μεγέθους $n^* = 100$ που κατασκευάζονται με τη μέθοδο Taylor. Τα αποτελέσματα βασίζονται σε πρότερα βάρη ενός μόνο σημείου, δηλαδή $\delta = 1/n^*$ (όπως σε όλα τα προηγούμενα παραδείγματα), και 100 σημείων, δηλαδή $\delta = 1$, ώστε να εξερευνήσει την επίδραση ισχυρότερης πρότερης πληροφορίας. Συγκρίνουμε αποτελέσματα για $\lambda_0 \in \{0.24, -0.25, -3\}$. Η τιμή 0.24 επιλέχθηκε καθώς είναι η ύστερη κορυφή της παραμέτρου μετασχηματισμού υπό την οικογένεια Box-Cox (βλέπε Πίνακα 3.4). Για αυτή την τιμή του λ_0 και με $\delta = 1/n^*$, το μοντέλο BC υποστηρίζεται εμφανώς περισσότερο με 71% ύστερη πιθανότητα συγκριτικά με το 57% (Πίνακας 3.4) και ύστερη κορυφή της παραμέτρου λ_T ίση με $\lambda_0 = 0.24$ όπως αναμενόταν (95% ΔΕ $= (0.07, 0.41)$). Κατά έναν παρεμφερή τρόπο, όταν το a posteriori βάρος αυξάνει ($\delta = 1$) η αντίστοιχη ύστερη πιθανότητα αυξάνει επίσης στο 95% καθώς περισσότερη πληροφορία υπέρ της οικογένειας BC περικλείεται. Ακόμη, η ύστερη κορυφή του λ_T είναι σχεδόν αμετάβλητη (0.23) αλλά με αρκετά πιο περιορισμένο διάστημα εμπιστοσύνης, 95% ΔΕ $= (0.11, 0.34)$.

Για τη δεύτερη τιμή $\lambda_0 = -0.25$ με $\delta = 1/n^*$, το a posteriori βάρος της οικογένειας BC είναι τώρα χαμηλότερο (66%) σε σχέση με αυτό που αντιστοιχούσε σε $\lambda_0 = 0.24$, εφόσον η οικογένεια BC υποστηρίζεται έμμεσα a priori αλλά μόνο κατά ένα σημείο, κάτι που επηρεάζει ελάχιστα τα a posteriori αποτελέσματα (η ύστερη κορυφή του λ_T είναι 0.23 με 95% ΔΕ $(0.05, 0.41)$). Από την άλλη μεριά, όταν το a priori βάρος αυξάνει σε n^* σημεία, η σειρά των οικογενειών a posteriori διαταράσσεται και το μοντέλο Box-Cox έρχεται τρίτο μετά τα μοντέλα Log και Dual με ύστερες πιθανότητες 4%, 58% και 38% αντίστοιχα. Αυτό είναι εύλογο, καθώς η ύστερη πιθανότητα του λ_T για το μοντέλο BC επηρεάζεται σημαντικά από τη διαμόρφωση της πρότερης κατανομής με αποτέλεσμα μια ύστερη κορυφή με τιμή 0.04 και 95% ΔΕ $(-0.09, 0.17)$ στο οποίο συμπεριλαμβάνεται το μηδέν και επομένως έμμεσα υποστηρίζεται ο Λογαριθμικός μετασχηματισμός.

Σε μια τρίτη δοκιμή, η τιμή του λ_0 τέθηκε ίση με την ακραία τιμή -3. Εδώ, τα αποτελέσματα αλλάζουν δραστικά συγκριτικά με τις προηγούμενες περιπτώσεις είτε για χαμηλό είτε για υψηλό a priori βάρος. Το μοντέλο Log σχετίζεται με την υψηλότερη ύστερη πιθανότητα και στις δύο περιπτώσεις (62% και 99%), ενώ το μοντέλο Box-Cox έρχεται δεύτερο σε σειρά με μεγάλη μεταβλητότητα (33% έναντι 1%). Υπό το μοντέλο Box-Cox, όταν $\delta = 1/n^*$ η ύστερη κορυφή του λ_T είναι 0.24 με 95% ΔΕ $(0.06, 0.42)$, ενώ για $\delta = 1$ οι αντίστοιχες τιμές είναι 0.02 και $(-0.14, 0.19)$.

Πίνακας 3.5: Εκ των υστέρων πιθανότητες (υπό την Prior A) των κύριων ανταγωνιστικών οικογενειών μετασχηματισμών αναφορικά με τα δεδομένα ελέγχου ποιότητας, με $\boldsymbol{y}^*$ μέσω της Taylor μεθόδου για διαφορετικές τιμές της παραμέτρου λ_0.

λ_0	Prior weight (δ)	Box-Cox	Log	Dual
0.24	$1/n^*$	0.71	0.21	0.08
	1	0.95	0.03	0.02
-0.25	$1/n^*$	0.66	0.26	0.08
	1	0.04	0.58	0.38
-3	$1/n^*$	0.33	0.62	0.05
	1	0.01	0.99	< 0.01
[a]1	$1/n^*$	0.57	0.34	0.09
[b]1	1	0.03	0.96	0.01

[a] Πρότερη κατανομή μοναδιαίας πληροφορίας που βασίζεται σε φανταστικά δεδομένα $\boldsymbol{y}^*$.

[b] Πρότερη πληροφορία ζυγισμένη κατά 50% βασισμένη σε φανταστικά δεδομένα $\boldsymbol{y}^*$.

Τα στοιχεία που αντιστοιχούν στις παραλειφθείσες οικογένειες είναι < 0.001.

3.3.2 Ανάλυση ευαισθησίας με βάση το k_0

Στον Πίνακα 3.6 περιλαμβάνονται τα αποτελέσματα μιας ανάλυσης ευαισθησίας για την υπερπαράμετρο k_0 που σχετίζεται με την παράμετρο κλίμακας όπως δίνεται στην (2.4). Για ακόμη μια φορά τα στοιχεία που παρατίθενται αφορούν μόνο την προσέγγιση Prior A. Παρατηρούμε ότι όσο το k_0 αυξάνει από 10^{-3} σε 10^0, η οικογένεια BC υποστηρίζεται με παρόμοιες εκ των υστέρων πιθανότητες, αν και διακρίνεται μια μικρή αυξη-

τική τάση. Οι εκτιμώμενες τιμές της ύστερης κορυφής του λ_T διατηρούνται αμετάβλητες για τις διάφορες δοκιμές και επομένως παραλείπονται (ο αναγνώστης παραπέμπεται στις σχετικές τιμές του Πίνακα 3.4). Εφόσον, λοιπόν, η μέθοδός μας είναι αρκετά εύρωστη για διάφορες επιλογές του πολλαπλασιαστή της πρότερης διακύμανσης $1/k_0$, η τιμή που επιλέξαμε για την εν λόγω υπερπαράμετρο μπορεί να θεωρηθεί εύλογες.

Πίνακας 3.6: Εκ των υστέρων πιθανότητες (υπό την Prior A) των κυριότερων ανταγωνιστικών οικογενειών μετασχηματισμών για τα δεδομένα ελέγχου ποιότητας εξετάζοντας διαφορετικές τιμές της υπερπαραμέτρου k_0.

k_0	BC	Log	Dual
10^{-3}	0.566	0.340	0.093
10^{-2}	0.572	0.341	0.087
10^{-1}	0.591	0.325	0.084
10^{0}	0.598	0.323	0.079

Σημείωση: Τα στοιχεία που αντιστοιχούν στις παραλειφθείσες οικογένειες είναι < 0.001.

3.4 Συμπεράσματα

Το παρόν κεφάλαιο παρείχε εφαρμογές με βάση την προταθείσα Μπεϋζιανή μεθοδολογία του Κεφαλαίου 2 για την αξιολόγηση και σύγκριση διαφορετικών οικογενειών μετασχηματισμών, με σκοπό να φέρει την κατανομή της μετασχηματισμένης τυχαίας μεταβλητής ενδιαφέροντος όσο το δυνατόν πλησιέστερα στην κανονικότητα. Αυτό έχει ιδιαίτερη χρησιμότητα στην περίπτωση του ελέγχου ποιότητας, αλλά και σε πολλές άλλες ερευνητικές περιοχές όπου η κανονικότητα παίζει σημαντικό ρόλο στην αξιοπιστία των παραγόμενων αποτελεσμάτων, η οποία με τη σειρά της συνδέεται άμεσα με την επιτυχία ή αποτυχία μιας βιομηχανικής διαδικασίας.

Τα αποτελέσματα κάτω από τις δύο πρότερες προσεγγίσεις για την παράμετρο λ_T, όπως περιγράφτηκαν στο Κεφάλαιο 3, συνέκλιναν σε σημαντικό βαθμό σε όλα τα παραδείγματα που παρατέθηκαν, με εξαίρεση κάποιες αποκλίσεις στην περίπτωση της οικογένειας Dual εξαιτίας της διαφορετικής συμπεριφοράς και των διαφορετικών χαρακτηρι-

στικών της εν λόγω οικογένειας. Σύνολα δεδομένων με έντονη ασυμμετρία αντιμετωπίζονται ικανοποιητικά μέσω του μετασχηματισμού Box-Cox, ενώ σημαντική μείωση στον βαθμό ασυμμετρίας μιας κατανομής τονώνει κάπως τον ρόλο του μετασχηματισμού YJ. Συμμετρικές κατανομές με παχιές ουρές (όπως οι κατανομές Student και διπλή εκθετική) συνδέονται με την οικογένεια Modulus. Εν τω συνόλω, εμπειρικά στοιχεία δείχνουν ότι η κυριαρχία του μοντέλου Box-Cox στη βιβλιογραφία μέχρι στιγμής δεν είναι πάντα ακριβής και η επιλογή από ένα ευρύτερο σύνολο οικογενειών μετασχηματισμών θα έπρεπε να γίνει κοινή πρακτική.

Ένα ζήτημα προς προβληματισμό είναι το βέλτιστο μέγεθος της σταθεράς μετατόπισης ξ, και πιο συγκεκριμένα του ϵ (όπου $\xi = |\min(\boldsymbol{y})| + \epsilon$), η οποία στην παρούσα ερευνητική δουλειά χρησιμοποιείται στους μετασχηματισμούς Box-Cox, Dual και Log ώστε να καταστήσει τα αμετασχημάτιστα δεδομένα $\boldsymbol{y}$ αυστηρά θετικά. Για τη σταθερά μετατόπισης ϵ, δοκιμάστηκε και μια άλλη προσέγγιση: θεωρήθηκε ότι η σταθερά αυτή είναι το τετράγωνο του πρώτου τεταρτημορίου Q_1 προς το τρίτο τεταρτημόριο Q_3 των μη αρνητικών παρατηρήσεων, κατά πρόταση του Stahel (2002), δηλαδή $\epsilon = Q_1^2/Q_3$. Τα αποτελέσματα, με βάση τα προσομοιωμένα σύνολα δεδομένων, δεν κρίθηκαν τόσο αποτελεσματικά όσο η πρώτη προσέγγιση που αναφέρθηκε για τη σταθερά μετατόπισης, οπότε παραλείφθηκαν.

Το Κεφάλαιο 4 που ακολουθεί επιχειρεί να επεκτείνει τη μεθοδολογία Μπεϋζιανής επιλογής οικογένειας μετασχηματισμών σε προβλήματα με επεξηγηματικές μεταβλητές.

Κεφάλαιο 4

Επέκταση σε Προβλήματα με Επεξηγηματικές Μεταβλητές Μέσω Εναλλακτικών Παραγόντων Bayes

4.1 Εισαγωγή

Το πρόβλημα της επιλογής μετασχηματισμού υπό την παρουσία μιας μεταβλητής απόκρισης και ενός συνόλου επεξηγηματικών μεταβλητών συναντάται συχνά σε πολλά πεδία της σύγχρονης έρευνας, όπως στον έλεγχο ποιότητας, στα οικονομικά και στη γενετική. Στο παρόν κεφάλαιο, το πρόβλημα της επιλογής οικογένειας μετασχηματισμού αντιμετωπίζεται διεξοδικά, πάντα υπό το πρίσμα της Μπεϋζιανής συλλογιστικής, επιστρατεύοντας διαφορετικές μορφές εναλλακτικών παραγόντων Bayes και συγκεκριμένα τον ενδογενή και τον κλασματικό παράγοντα Bayes.

Υπάρχουν πολύ ενδιαφέρουσες εφαρμοσμένες μελέτες στην πρόσφατη βιβλιογραφία οι οποίες τονίζουν την ανάγκη να πληρείται η προϋπόθεση της κανονικότητας (εφόσον βέβαια απαιτείται κάτι τέτοιο) ούτως ώστε η σχετική συμπερασματολογία να συνοδεύεται από εγκυρότητα. Τέτοιες μελέτες είναι, μεταξύ άλλων, η δουλειά των Yang, Christensen & Sorensen (2011) πάνω στο γενετικά δομημένο μοντέλο ετεροσκεδαστικότητας, η εφαρμοσμένη έρευνα πάνω στην πολυμεταβλητή μετανάλυση μετασχηματισμένων μοντέλων από τους Kim et al. (2013), η ανάλυση χρονοσειρών των Westerberg et al. (2011) σχετικά

με την υδρολογική μοντελοποίηση και η δουλειά των Miranda, Zhu & Ibrahim (2013) πάνω σε χωρικά μετασχηματισμένα μοντέλα για νευροαπεικονιστικές μετρήσεις.

Το κεφάλαιο αυτό εστιάζει καταρχήν στην επιλογή της κατάλληλης οικογένειας μετασχηματισμών για τη μεταβλητή απόκρισης. Η μεθοδολογία που προτείνεται στις ενότητες που ακολουθούν αποτελεί επέκταση της μεθοδολογίας του Κεφαλαίου 2 συνδυάζοντας επιπρόσθετα την επιλογή επεξηγηματικών μεταβλητών και την ερευνητική δουλειά που έχει γίνει πάνω στον ενδογενή και τον κλασματικό παράγοντα Bayes. Το σύνολο των οικογενειών μετασχηματισμών που θεωρούμε για τον μετασχηματισμό προς την κανονικότητα των δειγματικών τιμών μιας απόκρισης Y υπό την παρουσία επεξηγηματικών μεταβλητών είναι το ίδιο που χρησιμοποιήθηκε και στα Κεφάλαια 2 και 3: Box-Cox, Modulus, Yeo & Johnson και Dual.

Ο κορμός του κεφαλαίου αυτού ξεκινά με την Ενότητα 4.2 δίνοντας έμφαση σε προβληματικά ζητήματα που ανέκυψαν κατά τη μετάβαση από μονομεταβλητά προβλήματα σε προβλήματα με επεξηγηματικές μεταβλητές. Η Ενότητα 4.3 είναι εισαγωγική αναφορικά με τον παράγοντα Bayes, ενώ η Ενότητα 4.4 καταπιάνεται με τη βασική βιβλιογραφία όσον αφορά τις διάφορες εναλλακτικές μορφές παραγόντων Bayes. Οι επόμενες τρεις ενότητες παρουσιάζουν βήμα προς βήμα την κατασκευή της προτεινόμενης μεθοδολογίας. Συγκεκριμένα, η Ενότητα 4.5 ξετυλίγει τη Μπεϋζιανή προσέγγιση συμπερασματολογίας και επιλογής οικογένειας μετασχηματισμών, κάνοντας ειδική μνεία στις εναλλακτικές μορφές του παράγοντα Bayes που εφαρμόζονται, δηλαδή στον κλασματικό και στον ενδογενή παράγοντα Bayes, και διερευνά κατάλληλες προσεγγίσεις πρότερων κατανομών για τις παραμέτρους των μοντέλων. Η Ενότητα 4.6 επεκτείνει τη μεθοδολογία συμπεριλαμβάνοντας πλέον και πληροφορία από επεξηγηματικές μεταβλητές στο μοντέλο. Στη συνέχεια, η Ενότητα 4.7 παρουσιάζει τον τελικό μαθηματικό συμβολισμό μαζί με ορισμένες πρόσθετες σημειώσεις πάνω στις πρότερες κατανομές του μοντελοχώρου και των παραμέτρων του μοντέλου στα πλαίσια της συνολικότερης επιλογής μοντέλου, η οποία αφορά και την επιλογή μεταβλητών. Το κεφάλαιο ολοκληρώνεται με κάποιες συμπερασματικές παρατηρήσεις στην Ενότητα 4.8.

4.2 Το Ερευνητικό Κίνητρο για Εναλλακτικές Μορφές του Παράγοντα Bayes

Αυξημένο ενδιαφέρον συγκεντρώνει ο προσδιορισμός της εμπλεκόμενης πρότερης κατανομής για το πρόβλημα του μετασχηματισμένου μοντέλου με επεξηγηματικές μεταβλητές. Με παρόμοιο τρόπο όπως και στο Κεφάλαιο 2, ο παράγοντας Bayes θα μπορούσε να υπολογιστεί στη βάση μιας πρότερης κατανομής δύναμης (Ibrahim & Chen 2000) για την παράμετρο λ_T χρησιμοποιώντας ένα διάνυσμα φανταστικών δεδομένων $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)$ προσομοιωμένο από ένα μοντέλο που να υποστηρίζει την αρχή της φειδωλότητας (*parsimony principle*). Καλούμε τον αντίστοιχο παράγοντα Bayes ως παράγοντα Bayes πρότερης κατανομής δύναμης (BF^{pp}). Στο Κεφάλαιο 2, εξηγήσαμε πώς αυτό το σύνολο φανταστικών δεδομένων επιχειρεί την επίλυση του προβλήματος πρότερης ασυμβατότητας που εμφανίζεται λόγω της σύγκρισης μεταξύ των διαφόρων οικογενειών μετασχηματισμών. Για μια πιο λεπτομερή ματιά στα θέματα ασυμβατότητας που μπορεί να προκύψουν, βλέπε για παράδειγμα τις δουλειές των Consonni & Veronese (2008) και των Dawid & Lauritzen (2000).

Λόγω ύπαρξης επεξηγηματικών μεταβλητών στο μοντέλο, η χρήση των φανταστικών δεδομένων $\boldsymbol{y}^*$ για την κατασκευή της πρότερης κατανομής δύναμης για την παράμετρο λ_T (Ενότητα 2.4.1.1) περιπλέκεται, καθώς είναι πλέον αρκετά δύσκολο να προσδιοριστεί πλήρως το μοντέλο αναφοράς από το οποίο θα γεννηθούν τα δεδομένα αυτά. Ειδικότερα, το μοντέλο αναφοράς θα μπορούσε και πάλι να θεωρηθεί το κανονικό, αλλά ο προσδιορισμός της μέσης τιμής και της διασποράς του μοντέλου αυτού είναι αρκετά δύσκολος. Στα μονομεταβλητά προβλήματα, τυποποιώντας αρχικά τα δεδομένα, ορίζαμε ως μοντέλο αναφοράς το κανονικό με μέση τιμή μηδέν και διασπορά ένα, αλλά πλέον η εν λόγω διαδικασία περιπλέκεται μιας και η ταυτόχρονη γέννηση τιμών $\boldsymbol{y}^*$ και $\mathbf{X}^*$ από το μοντέλο αναφοράς δεν είναι ξεκάθαρη.

Έτσι, στο παρόν κεφάλαιο εναλλακτικές μορφές πρότερων κατανομών χρειάζεται να αναζητηθούν για την παράμετρο λ_T οι οποίες να είναι συμβατές μεταξύ των διαφορετικών οικογενειών και συγχρόνως να είναι "αντικειμενικές", μιας και στα περισσότερα προβλήματα δε μπορεί εύκολα να υπάρξει πρότερη πληροφορία για την παράμετρο των

μετασχηματισμών, όταν έχουμε πλήρη αβεβαιότητα ως προς το ποιον μετασχηματισμό θα πρέπει να επιλέξουμε.

Για να εκφράσουμε την έλλειψη πρότερης πληροφορίας, θα μπορούσαμε να καταφύγουμε στη χρήση μίας γνήσιας πρότερης κατανομής μεγάλης διασποράς (*diffuse prior*) η οποία, όμως, ενδεχομένως να ενεργοποιούσε το παράδοξο του Lindley. Εναλλακτικά, η χρήση μη γνησίων "αντικειμενικών" πρότερων κατανομών μας απαλλάσσει από τον καθορισμό των υπερπαραμέτρων. Ο κλασικός, όμως, παράγοντας Bayes δεν επιτρέπει εν γένει τη χρήση τέτοιων μη γνήσιων πρότερων κατανομών όταν συγκρίνουμε δύο μοντέλα, λόγω της εξάρτησής του από τις άγνωστες σταθερές κανονικοποίησης των εν λόγω πρότερων κατανομών.

Σαν λύση στο ανωτέρω ζήτημα, για τη σύγκριση μεταξύ των οικογενειών μετασχηματισμών με χρήση μη γνησίων πρότερων κατανομών, επιστρατεύθηκαν εναλλακτικές μορφές παραγόντων Bayes, όπως ο ενδογενής παράγοντας Bayes των Berger & Pericchi (1996*a*) και ο κλασματικός παράγοντας Bayes που αναπτύχθηκε από τον O'Hagan (1995). Στην ουσία, υπό τις δύο αυτές προσεγγίσεις, τα φανταστικά δεδομένα που χρησιμοποιήσαμε στο Κεφάλαιο 2 αντικαθίστανται από τα δεδομένα εκπαίδευσης του ενδογενούς παράγοντα Bayes και από την κλασματική παράμετρο εκπαίδευσης b του κλασματικού παράγοντα Bayes.

4.3 Σύντομη Εισαγωγή στον Κλασικό Παράγοντα Bayes

Τα θεμέλια για τον αρχικό υπολογισμό του παράγοντα Bayes και για την μετέπειτα διαδεδομένη χρήση του τέθηκαν στο άρθρο του Jeffreys (1935) και σε ένα μετέπειτα άρθρο του ιδίου (Jeffreys 1961) όπου πλέον παρουσιάζεται η σύγχρονη όψη της εν λόγω ποσότητας, συμπεριλαμβανομένης μιας ερμηνευτικής κλίμακας του λογαρίθμου με βάση 10 των τιμών BF. Στις μέρες μας, ο παράγοντας Bayes διαδραματίζει κομβικό ρόλο στη Μπεϋζιανή στατιστική και ειδικότερα στη Μπεϋζιανή σύγκριση και επιλογή μοντέλων. Ο γενικός ορισμός του παράγοντα Bayes είναι ότι ισούται με το πηλίκο των περιθώριων πιθανοφανειών δύο υπό σύγκριση μοντέλων M_i, M_j δεδομένου ενός συνόλου παρατηρήσεων $\boldsymbol{y}$. Ισοδύναμα, ο παράγοντας Bayes αποτελεί το πηλίκο του λόγου των ύστερων

σχετικών πιθανοτήτων των δύο μοντέλων M_i, M_j προς τον λόγο των πρότερων σχετικών πιθανοτήτων τους για κάποια δεδομένα $\boldsymbol{y} = (y_1, \ldots, y_n)$:

$$B_{ij}(\boldsymbol{y}) = \frac{PO_{ij}}{PrO_{ij}} = \frac{\frac{P(M_i|\boldsymbol{y})}{P(M_j|\boldsymbol{y})}}{\frac{P(M_i)}{P(M_j)}} = \frac{f\left(\boldsymbol{y}|M_i\right)}{f\left(\boldsymbol{y}|M_j\right)}.$$

Ένα κλασικό και πολυδιαβασμένο πλέον άρθρο της πρόσφατης βιβλιογραφίας πάνω στον παράγοντα Bayes είναι αυτό των Kass & Raftery (1995) όπου εξερευνώνται σε βάθος τα χαρακτηριστικά και η χρησιμότητα του εν λόγω μεγέθους, ενώ προτείνεται και ένα νέο εναλλακτικό σύστημα ερμηνείας για τον παράγοντα Bayes. Συγκεκριμένα, ως μέτρο καθορισμού κατάλληλων κατευθυντήριων γραμμών για την ερμηνεία μιας τιμής του παράγοντα Bayes, οι συγγραφείς του άρθρου χρησιμοποιούν το διπλάσιο του φυσικού λογαρίθμου του BF αντί για τον λογάριθμο του BF με βάση το 10. Οι κατευθυντήριες γραμμές που προτείνουν παρουσιάζονται στον Πίνακα 4.1.

Πίνακας 4.1: Ερμηνεία ενός παράγοντα Bayes ($B_{ij}(\boldsymbol{y})$) μεταξύ δύο μοντέλων M_i και M_j, όπως προτείνεται από τους Kass & Raftery (1995).

$2log_e\left(B_{ij}(\boldsymbol{y})\right)$	$B_{ij}(\boldsymbol{y})$	Ένδειξη εναντίον του μοντέλου M_j
0-2	1-3	Σχεδόν αμελητέα
2-6	3-20	Θετική
6-10	20-150	Έντονη
> 10	> 150	Πολύ έντονη

Η έννοια του παράγοντα Bayes είναι θεμελιώδης για την ανάδειξη της υπεροχής της Μπεϋζιανής στατιστικής σε σχέση με την κλασική στατιστική, καθώς αναδεικνύει την ευελιξία της πρώτης μέσα από την ποσοτικοποίηση του βάρους των ενδείξεων υπέρ μιας μηδενικής υπόθεσης ή μιας ορισμένης επιστημονικής θεωρίας, σε αντίθεση με τη λειτουργία π.χ. των p-τιμών που τόσο έχουν επικριθεί. Ας σημειωθεί ότι ο παράγοντας Bayes μπορεί να υπολογιστεί για εμφωλευμένα ή μη μοντέλα. Παρόλα αυτά, όπως σε κάθε Μπεϋζιανό εργαλείο ή μέθοδο, ένα θέμα για προβληματισμό παραμένει η εμφανής ευαισθησία του παράγοντα Bayes στην επιλογή των σχετικών πρότερων κατανομών για τον υπολογισμό της περιθώριας πιθανοφάνειας των δεδομένων υπό το εκάστοτε μοντέλο. Μια λύση σε αυτό θα μπορούσε να προσφέρει ο υπολογισμός του Μπεϋζιανού κριτηρίου Schwarz

(*Schwarz Bayesian criterion*, SBC) για κάθε ένα από τα μοντέλα M_i, M_j (Schwarz 1978), το οποίο αποτελεί μια ακατέργαστη προσέγγιση του λογάριθμου του BF (επονομαζόμενος και ως βαρύτητα της ένδειξης - *weight of evidence* - από τον Good (1985)) που δεν περιλαμβάνει πρότερες κατανομές. Για λεπτομέρειες ο αναγνώστης παραπέμπεται στο άρθρο Kass & Wasserman (1995). Συχνά κανείς συναντά το ισοδύναμο Μπεϋζιανό κριτήριο πληροφορίας (*Bayesian Information Criterion*, BIC) το οποίο είναι ίσο με μείον το διπλάσιο του κριτηρίου SBC:

$$\text{BIC} = -2 \ln f\left(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_k, M_k\right) + p_k \ln n$$

όπου $f\left(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_k, M_k\right)$ είναι η μεγιστοποιημένη πιθανοφάνεια του μοντέλου M_k με διάνυσμα παραμέτρων $\boldsymbol{\theta}_k$ διάστασης p_k. Ένας περιορισμός που πρέπει να πληρείται ώστε η προσέγγιση αυτή να είναι έγκυρη είναι να έχουμε σχετικά λίγους βαθμούς ελευθερίας του προβλήματος σε σχέση με το μέγεθος δείγματος των δεδομένων.

Ακόμη ένα ζήτημα που σχετίζεται με τον παράγοντα Bayes είναι το παράδοξο του Bartlett. Σύμφωνα με αυτό, αν επιλεγεί μια διάχυτη πρότερη κατανομή μεγάλης διασποράς, δηλαδή μια σχεδόν μη πληροφοριακή πρότερη κατανομή, για την παράμετρο $\boldsymbol{\theta}_i$ του μοντέλου M_i, τότε το μοντέλο αυτό μπορεί να λάβει ποινή και άρα χαμηλότερη στήριξη με βάση την τιμή του παράγοντα Bayes, προωθώντας έτσι τα πιο φειδωλά μοντέλα ανεξάρτητα από την πληροφορία που ενέχουν τα δεδομένα (Bartlett 1957). Ένα απλό παράδειγμα που δείχνει ότι η τιμή της περιθώριας πιθανοφάνειας μειώνεται καθώς η πρότερη κατανομή των παραμέτρων διαχέεται περισσότερο είναι το ακόλουθο (De Santis & Spezzaferri 1999): θεωρώντας μια ομοιόμορφη πρότερη κατανομή για την παράμετρο θ_k, ορισμένη στο διάστημα $(-\alpha, \alpha), \alpha \in \mathbb{R}$, η περιθώρια πιθανοφάνεια $f(\boldsymbol{y}|M_k) = (2\alpha)^{-1} \int_{-\alpha}^{\alpha} f(\boldsymbol{y}|\theta_k, M_k) \mathrm{d}\boldsymbol{\theta}_k$ τείνει στο μηδέν καθώς η υπερπαράμετρος $\alpha \to \infty$, δεδομένου ότι το ολοκλήρωμα είναι πεπερασμένο.

Μη γνήσιες πρότερες κατανομές είναι γενικά αποδεκτές μόνο για παραμέτρους που είναι κοινές σε όλα τα μοντέλα του μοντελοχώρου $\mathcal{M}$. Εξαίρεση αποτελεί η πρότερη κατανομή συρρίκωσης (*shrinkage prior*) (Strachan & van Dijk 2005). Ο τρόπος επιλογής της πρότερης κατανομής οδηγεί έμμεσα και στο επόμενο ζήτημα που σχετίζεται με τον παράγοντα Bayes: αρκετά συχνά ο υπολογισμός των εμπλεκόμενων περιθώριων πιθανοφανειών δεν είναι μια τετριμμένη διαδικασία καθώς τα σχετικά ολοκληρώματα δεν

υπολογίζονται. Έχουν αναπτυχθεί διάφορες μέθοδοι προσέγγισης της περιθώριας πιθανοφάνειας, επονομαζόμενη και ως ένδειξη των δεδομένων, και μια ενδελεχής αξιολόγηση των μεθόδων αυτών συνέθεσαν οι Friel & Wyse (2012).

Μια εναλλακτική ερμηνεία του παράγοντα Bayes, ως η σχετική ισχύς του μοντέλου M_i έναντι του μοντέλου M_j στην πρόβλεψη των παρατηρούμενων τιμών παρά στην ανάδειξη ενός από τα δύο μοντέλα ως το αληθινό μοντέλο, είναι η ακόλουθη: ο λογάριθμος του BF ισούται με τη διαφορά των γενικών προβλεπτικών σκορ των δεδομένων για τα υπό σύγκριση μοντέλα, αν ως γενικό προβλεπτικό σκορ υπό το μοντέλο M_i ορίσουμε την ποσότητα $\log f(\boldsymbol{y}|M_i) = \log f\,(y_1|M_i) + \log f(y_2|y_1, M_i) + \cdots + \log f(y_n|y_{n-1}, \cdots y_1, M_i)$. Επιπλέον, η αναλογία μεταξύ του παράγοντα Bayes και του ελέγχου του λόγου των πιθανοφανειών γίνεται εμφανής αν κανείς ολοκληρώσει εκτός (αντί να μεγιστοποιήσει) τις παραμέτρους της συνάρτησης πιθανοφάνειας. Από εδώ απορρέει και η παράπλευρη ονομασία του παράγοντα Bayes ως ο λόγος ολοκληρωμένων πιθανοφανειών (*integrated likelihood ratio*).

Ένα ελκυστικό χαρακτηριστικό του παράγοντα Bayes είναι η εν δυνάμει χρησιμότητά του σε μια εξελικτική διαδικασία χτισίματος ενός μοντέλου συγκρίνοντας με σειριακό τρόπο διάφορα εναλλακτικά ή εμφωλευμένα μοντέλα. Σε συνδυασμό με το γεγονός ότι δρα σαν μια αυτόματη μορφή του ξυραφιού του Occam (*Occam's razor*), ο παράγοντας Bayes αποτελεί επίσης και ένα ισχυρό εργαλείο για πολλαπλή σύγκριση μοντέλων. Δύο αρκετά προφανείς αλλά χρήσιμες ιδιότητες του παράγοντα Bayes είναι οι εξής:

$$B_{ij}^{-1}(\boldsymbol{y}) = B_{ji}(\boldsymbol{y}) \qquad \text{και} \qquad \frac{B_{kj}(\boldsymbol{y})}{B_{lj}(\boldsymbol{y})} = B_{kl}(\boldsymbol{y}).$$

Οι άνωθεν εξισώσεις αντιπροσωπεύουν και τις βασικές συνθήκες συνοχής του παράγοντα Bayes κατά την αξιολόγηση της λειτουργικότητας και την δικαιολόγηση της λογικής πίσω από εναλλακτικές μορφές παραγόντων Bayes.

Ένα τελευταίο σχόλιο είναι ότι οι παράγοντες Bayes επιτρέπουν την αβεβαιότητα επιλογής μοντέλου, το οποίο είναι κομβικό από την άποψη της προβλεπτικής ικανότητας και έχει άμεσες συνέπειες στην λήψη αποφάσεων (π.χ. στη διαμόρφωση πολιτικής ή στην εγκληματολογική επιστήμη). Κάτι τέτοιο καθίσταται δυνατό μέσω τεχνικών Μπεϋζιανής στάθμισης μοντέλων (BMA). Θεωρώντας ένα σύνολο πιθανών μοντέλων $\{M_1, \ldots, M_K\}$ για ένα δεδομένο πρόβλημα με παρατηρηθέντα δεδομένα $\boldsymbol{D}$ και μια ποσότητα ενδιαφέ-

ροντος Δ (π.χ. μια μελλοντική παρατήρηση ή το μέγεθος της επίδρασης ενός παράγοντα), τότε η αντίστοιχη ύστερη κατανομή του Δ είναι:

$$f(\Delta|\boldsymbol{D}) = \sum_{k=1}^{K} f(\Delta|\boldsymbol{D}, M_k) f(M_k|\boldsymbol{D})$$

όπου ενσωματώνεται ύστερη πληροφορία από όλα τα πιθανά μοντέλα. Για μια ενδελεχή διερεύνηση του ζητήματος παραπέμπουμε στο άρθρο των Hoeting et al. (1999).

4.4 Εναλλακτικές Μορφές Παραγόντων Bayes

Σε περιπτώσεις μη γνήσιων πρότερων κατανομών για τις παραμέτρους ενός μοντέλου, ανακύπτει το πρόβλημα μη κλειστής μορφής του αντίστοιχου παράγοντα Bayes. Ας αναλύσουμε λίγο παραπάνω το ζήτημα αυτό. Αν i είναι ο δείκτης του μοντέλου, η μη γνήσια πρότερη κατανομή του διανύσματος παραμέτρων $\boldsymbol{\theta}_i$ μπορεί να περιγραφεί ως εξής:

$$\pi_i\left(\boldsymbol{\theta}_i\right) = c_i g_i\left(\boldsymbol{\theta}_i\right) \tag{4.1}$$

όπου c_i είναι μια άγνωστη σταθερά και $g_i\left(\boldsymbol{\theta}_i\right)$ είναι μια συνάρτηση της οποίας το ολοκλήρωμα πάνω στον παραμετρικό χώρο Θ_i αποκλίνει. Παρά το γεγονός ότι η σύνθεση του προβλήματος οδηγεί σε μια ύστερη κατανομή κλειστής μορφής υπό το μοντέλο M_i εφόσον η σταθερά c_i απαλείφεται στο σχετικό πηλίκο, δε συμβαίνει το ίδιο και με τον παράγοντα Bayes που συγκρίνει τα μοντέλα M_i, M_j όπως φαίνεται παρακάτω:

$$\begin{aligned} B_{ij}(\boldsymbol{y}) &= \frac{\int \pi_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_i}{\int \pi_j\left(\boldsymbol{\theta}_j\right) f_j\left(\boldsymbol{y}|\boldsymbol{\theta}_j\right) \mathrm{d}\boldsymbol{\theta}_j} \\ &= \frac{c_i \int g_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_i}{c_j \int g_j\left(\boldsymbol{\theta}_j\right) f_j\left(\boldsymbol{y}|\boldsymbol{\theta}_j\right) \mathrm{d}\boldsymbol{\theta}_j} \end{aligned} \tag{4.2}$$

όπου $\boldsymbol{y} = (y_1, \ldots, y_n)$ είναι οι παρατηρήσεις, $f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right)$ είναι η πιθανοφάνεια του διανύσματος $\boldsymbol{\theta}_i$ δεδομένου του $\boldsymbol{y}$ κάτω από το μοντέλο M_i και $\frac{c_i}{c_j}$ είναι ο λόγος των άγνωστων σταθερών που αντιστοιχούν στις μη γνήσιες πρότερες κατανομές $\pi_i(\boldsymbol{\theta}_i)$, $\pi_j(\boldsymbol{\theta}_j)$ των παραμέτρων των μοντέλων M_i και M_j αντίστοιχα. Σημαντικό μερίδιο της Μπεϋζιανής βιβλιογραφίας είναι αφιερωμένο στη διαμόρφωση εναλλακτικών μορφών του παράγοντα Bayes ώστε να λυθεί αυτό το πρόβλημα, μεταξύ άλλων.

4.4.1 Ο εναλλακτικός παράγοντας Bayes των Spiegelhalter και Smith

Το άρθρο των Spiegelhalter & Smith (1982) είχε ευρεία επίδραση στο Μπεϋζιανό κοινό και αναφέρεται συχνά, μαζί με ένα προγενέστερο σχετικό άρθρο των ιδίων (Smith & Spiegelhalter 1980), στα περισσότερα άρθρα που σχετίζονται με τον μερικό παράγοντα Bayes (*partial Bayes factor*), στον οποίο θα αναφερθούμε αργότερα, ή με τη σύγκριση μοντέλων κάτω από ασαφή πρότερη πληροφορία. Η γενική ιδέα ενέχει διαδοχική ενημέρωση (*sequential updating*), αρχικά με τη βοήθεια ενός εναλλακτικού ελάχιστου δείγματος εκπαίδευσης (*training sample*) που αποτελείται από φανταστικά δεδομένα $\boldsymbol{y}^*$ και ύστερα μέσω της χρήσης του πραγματικού δείγματος παρατηρήσεων $\boldsymbol{y}$. Το φανταστικό δείγμα εκπαίδευσης μπορεί να προσομοιωθεί με βάση τη μηδενική υπόθεση ή με βάση το απλούστερο υπό εξέταση μοντέλο και αποτελεί ικανή συνθήκη ώστε να λυθεί το πρόβλημα της μη υπολογισιμότητας του παράγοντα Bayes που συχνά οφείλεται σε μη φραγμένες πρότερες κατανομές. Παρόλα αυτά, ο καθορισμός του μεγέθους και των τιμών των φανταστικών δεδομένων παραμένει ένα θέμα προς διερεύνηση. Περισσότερες πληροφορίες για την τεχνική των φανταστικών δεδομένων μπορούν να αναζητηθούν στο Κεφάλαιο 2 της παρούσας διατριβής και στο άρθρο των Charitidou, Fouskakis & Ntzoufras (2015).

Εν συντομία, με σκοπό την υπερπήδηση του εμποδίου υπολογισμού των άγνωστων σταθερών στην (4.2), σχηματίζουμε αρχικά έναν παράγοντα Bayes $B_{ij}(\boldsymbol{y}^*)$ για τα δύο υπό σύγκριση μοντέλα $M_i,\ M_j$ με βάση το δείγμα φανταστικών δεδομένων $\boldsymbol{y}^*$ και με τις πρότερες κατανομές της μορφής (4.1):

$$B_{ij}(\boldsymbol{y}^*) = \frac{c_i}{c_j}\frac{\int g_i(\theta_i) f_i(\boldsymbol{y}^*|\theta_i)\mathrm{d}\theta_i}{\int g_j(\theta_j) f_j(\boldsymbol{y}^*|\theta_j)\mathrm{d}\theta_j} \Rightarrow$$

$$\frac{c_i}{c_j} = B_{ij}(\boldsymbol{y}^*)\left(\frac{\int g_i(\theta_i) f_i(\boldsymbol{y}^*|\theta_i)\mathrm{d}\theta_i}{\int g_j(\theta_j) f_j(\boldsymbol{y}^*|\theta_j)\mathrm{d}\theta_j}\right)^{-1}. \tag{4.3}$$

Χρησιμοποιώντας την έκφραση (4.3) ως μέσο για την εκτίμηση του λόγου των άγνωστων σταθερών, ο παράγοντας Bayes για τα δεδομένα $\boldsymbol{y}$ πλέον υπολογίζεται:

$$B_{ij}(\boldsymbol{y}) = B_{ij}(\boldsymbol{y}^*)\left(\frac{\int g_i(\theta_i) f_i(\boldsymbol{y}^*|\theta_i)\mathrm{d}\theta_i}{\int g_j(\theta_j) f_j(\boldsymbol{y}^*|\theta_j)\mathrm{d}\theta_j}\right)^{-1}\frac{\int g_i(\theta_i) f_i(\boldsymbol{y}|\theta_i)\mathrm{d}\theta_i}{\int g_j(\theta_j) f_j(\boldsymbol{y}|\theta_j)\mathrm{d}\theta_j}.$$

Οι Spiegelhalter και Smith αυθαίρετα εξίσωσαν την ποσότητα $B_{ij}(\boldsymbol{y}^*)$ με το 1. Αυτή όμως είναι απλώς μια βολική επιλογή που δεν δικαιολογείται θεωρητικά και η οποία αποτελεί ένα δεύτερο σημείο ένστασης για τη μέθοδο αυτή. Ένα πλεονέκτημα που απορρέει από τη

χρήση φανταστικών δεδομένων $\boldsymbol{y}^*$ αντί ενός μέρους των αρχικών παρατηρήσεων $\boldsymbol{y}$ για την εκτίμηση του λόγου των άγνωστων σταθερών είναι ότι η δεύτερη προσέγγιση θα ελάττωνε σημαντικά τη στατιστική δύναμη της ανάλυσης στην περίπτωση μικρών δειγμάτων ή πολύ περίπλοκων σχηματισμών όπου η συμμετρία και η ορθογωνιότητα του συνολικού δείγματος θα έπρεπε να διατηρηθεί. Βέβαια, για υπερβολικά μικρά μεγέθη δείγματος η χρήση των φανταστικών δεδομένων κρίνεται ως πολύ επισφαλής και συνιστάται η αποφυγή της. Η τεχνική των φανταστικών δεδομένων επιστρατεύεται στην περίπτωση σύγκρισης εμφωλευμένων μοντέλων σε πλήρως τυχαιοποιημένους σχεδιασμούς ανά μπλοκ (*randomized complete blocks*), σε δομές μίας κατεύθυνσης (*one-way*) ανάλογες του ελέγχου t, στην πολλαπλή παλινδρόμηση καθώς και σε λογαριθμοκανονικά παραδείγματα.

4.4.2 Ύστερος παράγοντας Bayes

Μια λιγότερο δημοφιλής ιδέα προτάθηκε από τον Aitkin (1991). Με αρχικό σκοπό τη διόρθωση της υπερβολικής και ενίοτε πλασματικής ενίσχυσης της ένδειξης ενάντια σε ένα μοντέλο όπως αυτή αποδίδεται μέσω των p-τιμών, ο Aitkin θεώρησε ως πρότερη κατανομή των παραμέτρων την ύστερη κατανομή τους. Έτσι, εισήγαγε τον ύστερο παράγοντα Bayes (*posterior Bayes factor*, PBF) ως ίσο με τον λόγο των ύστερων προβλεπτικών τεταγμένων (*posterion predictive ordinates*):

$$B_{ij}^{post}(\boldsymbol{y}) = \frac{\int f^2\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right)\pi\left(\boldsymbol{\theta}_i\right)\mathrm{d}\boldsymbol{\theta}_i}{\int f\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right)\pi\left(\boldsymbol{\theta}_i\right)\mathrm{d}\boldsymbol{\theta}_i} \cdot \left(\frac{\int f^2\left(\boldsymbol{y}|\boldsymbol{\theta}_j\right)\pi\left(\boldsymbol{\theta}_j\right)\mathrm{d}\boldsymbol{\theta}_j}{\int f\left(\boldsymbol{y}|\boldsymbol{\theta}_j\right)\pi\left(\boldsymbol{\theta}_j\right)\mathrm{d}\boldsymbol{\theta}_j}\right)^{-1}$$

Τα βασικά πλεονεκτήματα της μεθόδου, όπως αναφέρει ο συγγραφέας, είναι αφενός ότι αποφεύγεται το παράδοξο του Lindley και αφετέρου ότι δέχεται διάχυτες πρότερες κατανομές για τις παραμέτρους των μοντέλων. Παρά τη δημοσίευση του ερευνητικού έργου του Aitkin στο υψηλού κύρους επιστημονικό περιοδικό *Journal of the Royal Statistical Society (Series B)*, επικρίθηκε έντονα στον χώρο των στατιστικών επιστημόνων. Συγκεκριμένα, εκτός από τον όρο 'ύστερος' στο όνομά του, ο ύστερος παράγοντας Bayes δε συμμορφώνεται με κανέναν εκ των βασικών Μπεϋζιανών κανόνων. Για παράδειγμα, κάνει διπλή χρήση των δειγματικών παρατηρήσεων $\boldsymbol{y}$ στον υπολογισμό του ύστερου μέσου της πιθανοφάνειας.

4.4.3 Κλασματικός παράγοντας Bayes

Στο αμφιλεγόμενο άρθρο του, ο O'Hagan (1995) προτείνει τον κλασματικό παράγοντα Bayes (*fractional Bayes factor*, FBF) ως μια νέα εναλλακτική στις ήδη υπάρχουσες μορφές του μερικού παράγοντα Bayes. Όπως είναι φυσικό, ο κλασματικός παράγοντας Bayes φέρει την ίδια ιδιότητα όπως και όλες οι μορφές παραγόντων Bayes, ότι δηλαδή η χρήση του στοχεύει στη σύγκριση μοντέλων παρά στην επιλογή μοντέλου με την απόλυτη έννοια του όρου. Συνοπτικά, ο κλασματικός παράγοντας Bayes θεωρεί την πιθανοφάνεια των δεδομένων υψωμένη στη δύναμη b (ονομαζόμενη 'κλάσμα' ή 'κλασματική παράμετρος') έτσι ώστε να μετατρέψει μη γνήσιες πρότερες κατανομές σε γνήσιες και έτσι να δώσει λύση στο πρόβλημα της μη κλειστής μορφής του κλασικού παράγοντα Bayes. Σε αυτό το σημείο και προτού σκιαγραφήσουμε τον σχηματισμό του κλασματικού παράγοντα Bayes, θα ήταν πολύ χρήσιμο να πούμε δύο λόγια για τον μερικό παράγοντα Bayes.

Στον μερικό παράγοντα Bayes, που αρχικά παρουσιάστηκε από τον Lempers (1971), επωφελούμαστε από την προαναφερθείσα ιδιότητα γνησιότητας της ύστερης κατανομής παρά την ύπαρξη της άγνωστης σταθεράς c_i. Η διαδικασία έχει ως εξής: τα δεδομένα χωρίζονται σε δύο (ξένα) υποσύνολα, συνήθως άνισου μεγέθους: $\boldsymbol{y}_m$ μεγέθους m και $\boldsymbol{y}_{\setminus m}$ μεγέθους $(n-m)$, το πρώτο εκ των οποίων χρησιμοποιείται για να μετατρέψει τις μη γνήσιες πρότερες κατανομές σε γνήσιες (δείγμα εκπαίδευσης) και το δεύτερο χρησιμοποιείται για τη σύγκριση και διάκριση μεταξύ των μοντέλων. Επομένως, κατά κάποιον τρόπο αυτή η διαδικασία ενέχει επίσης διαδοχική ανανέωση, όπως και στο άρθρο των Spiegelhalter & Smith (1982). Η διαφορά έγκειται στο ότι τα δύο υποσύνολα προέρχονται από την ίδια αρχική πηγή παρατηρήσεων. Η τελική μορφή του μερικού παράγοντα Bayes ισούται με:

$$B_{ij}\left(\boldsymbol{y}_{\setminus m}|\boldsymbol{y}_m\right) = \frac{\int \pi_i\left(\boldsymbol{\theta}_i|\boldsymbol{y}_m\right) f_i\left(\boldsymbol{y}_{\setminus m}|\boldsymbol{\theta}_i, \boldsymbol{y}_m\right) \mathrm{d}\boldsymbol{\theta}_i}{\int \pi_j\left(\boldsymbol{\theta}_j|\boldsymbol{y}_m\right) f_j\left(\boldsymbol{y}_{\setminus m}|\boldsymbol{\theta}_j, \boldsymbol{y}_m\right) \mathrm{d}\boldsymbol{\theta}_j}$$

όπου $\pi_i\left(\boldsymbol{\theta}_i|\boldsymbol{y}_m\right)$ είναι η συνήθης κλειστής μορφής ύστερη κατανομή ενός διανύσματος παραμέτρων δεδομένου του $\boldsymbol{y}_m$ υπό την πρότερη κατανομή της μορφής (4.1):

$$\begin{aligned}
\pi_i\left(\boldsymbol{\theta}_i|\boldsymbol{y}_m\right) &= \frac{\pi_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}_m|\boldsymbol{\theta}_i\right)}{\int \pi_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}_m|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_i} \\
&= \frac{c_i g_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}_m|\boldsymbol{\theta}_i\right)}{\int c_i g_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}_m|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_i} \\
&= \frac{g_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}_m|\boldsymbol{\theta}_i\right)}{\int g_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}_m|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_m}.
\end{aligned}$$

Η επιλογή του δείγματος εκπαίδευσης $\boldsymbol{y}_m$ είναι η κύρια αιτία που διχάζει τους Μπεϋζιανούς στατιστικούς και έχει οδηγήσει στην ανάπτυξη διαφόρων μορφών μερικών ή εναλλακτικών παραγόντων Bayes. Για να αποκτήσει κανείς μια αίσθηση του μεγέθους των πιθανών συνδυασμών που προκύπτουν, αρκεί να φανταστεί ότι για ένα μέτριο μέγεθος δείγματος $n = 100$ υπάρχουν $\binom{n}{m} = \frac{100!}{3!97!} = 161700$ διαφορετικοί συνδυασμοί ή υποσύνολα παρατηρήσεων με $m = 3$ παρατηρήσεις. Όπως θα δούμε παρακάτω, οι Berger & Pericchi (1996*a*) πρότειναν τη χρήση όλων των δειγμάτων εκπαίδευσης ελάχιστου μεγέθους ($m = m_0$) και υπολόγισαν τον μέσο όρο των παραγόμενων παραγόντων Bayes, δημιουργώντας έτσι τον ενδογενή παράγοντα Bayes.

Ορισμός 4.4.1. Ένα δείγμα εκπαίδευσης $\boldsymbol{y}_m$ μεγέθους m καλείται γνήσιο αν $0 < f_i\left(\boldsymbol{y}_m|\boldsymbol{\theta}_i\right) < \infty$ για κάθε μοντέλο $M_i \in \mathcal{M}$.

Ορισμός 4.4.2. Ένα δείγμα εκπαίδευσης $\boldsymbol{y}_{m_0}$ μεγέθους m_0 καλείται ελάχιστο αν είναι γνήσιο και κανένα υποσύνολο αυτού δεν είναι γνήσιο.

Από την άλλη μεριά, ο O'Hagan αναζητούσε έναν τρόπο να απαλλαγεί από τα υπολογιστικά προβλήματα που σχετίζονται με την επιλογή όλων των πιθανών δειγμάτων εκπαίδευσης καθώς και με την επιλογή του κατάλληλου τύπου (αριθμητικού ή γεωμετρικού) για τον μέσο όρο των αντίστοιχων μερικών παραγόντων Bayes. Ο σκοπός του επιτεύχθηκε χρησιμοποιώντας ολόκληρο το δείγμα των παρατηρήσεων ως δείγμα εκπαίδευσης ($\boldsymbol{y}_m = \boldsymbol{y}$), υψώνοντας όμως την αντίστοιχη πιθανοφάνεια του δείγματος εκπαίδευσης στην δύναμη $b = \frac{m}{n}$, $m \leq n$. Μέσω αυτής της δύναμης, θεώρησε ότι πήρε μόνο ένα μέρος της πληροφορίας που παρέχουν τα δεδομένα εκπαίδευσης (ή μόνο ένα κλάσμα της συνολικής πιθανοφάνειας). Η προσέγγιση αυτή κατέληξε στον υπολογισμό της κανονικοποιημένης περιθώριας πιθανοφάνειας που αποτελεί και τον πυρήνα του κλασματικού παράγοντα Bayes:

$$\begin{aligned} m_i^F\left(\boldsymbol{y}\right) &= \frac{\int \pi_i\left(\boldsymbol{\theta}_i\right) f_i^{1-b}\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right) f_i^b\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_i}{\int \pi_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right)^b \mathrm{d}\boldsymbol{\theta}_i} \\ &= \frac{\int \pi_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_i}{\int \pi_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right)^b \mathrm{d}\boldsymbol{\theta}_i} \\ &= \frac{\int g_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right) \mathrm{d}\boldsymbol{\theta}_i}{\int g_i\left(\boldsymbol{\theta}_i\right) f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right)^b \mathrm{d}\boldsymbol{\theta}_i}. \end{aligned}$$

Η λογική πίσω από τον κλασματικό παράγοντα Bayes βασίζεται στην ασυμπτωτική θεωρία: αν οι τιμές του αριθμητή και του παρονομαστή του κλάσματος b θεωρηθούν μεγάλες, τότε η πιθανοφάνεια του δείγματος εκπαίδευσης $\boldsymbol{y}_m$ μεγέθους m προσεγγίζει την πιθανοφάνεια του ολικού δείγματος παρατηρήσεων $\boldsymbol{y}$ μεγέθους n υψωμένη στη δύναμη $b = \frac{m}{n}$. Στην πράξη, βέβαια, η παράμετρος m λαμβάνει συνήθως μια τιμή χαμηλή σε σχέση με το μέγεθος n του ολικού δείγματος. Ο κλασματικός παράγοντας Bayes διαθέτει την ελκυστική ιδιότητα του εύκολου και γρήγορου υπολογισμού ακόμη και σε περιπτώσεις μη εμφωλευμένων μοντέλων καθώς και πολλαπλών μοντέλων υπό σύγκριση.

Εν αντιθέσει με την ύστερη συμπερασματολογία με βάση ένα και μόνο μοντέλο, η ευαισθησία ενός παράγοντα Bayes αυξάνει με το n. Όσον αφορά τον FBF, η ευαισθησία του στην επιλογή πρότερης κατανομής αυξάνει με το n και μειώνεται με το b όπως θα ανέμενε κανείς, μιας και υψηλότερη τιμή για την παράμετρο b συνεπάγεται μεγαλύτερο μέγεθος δείγματος εκπαίδευσης m. Ο FBF είναι λιγότερο ευάλωτος στην πρότερη διακύμανση σε σχέση με τον κλασικό παράγοντα Bayes και με τον μερικό παράγοντα Bayes. Ακόμη, εξαιτίας του ότι στον FBF ο ρόλος του δείγματος εκπαίδευσης ανατίθεται στο ολικό δείγμα $\boldsymbol{y}$ αντί σε ένα τυχαία επιλεγμένο υποσύνολο του $\boldsymbol{y}$, παρατηρούμε ότι η προσέγγιση αυτή είναι λιγότερο ευάλωτη σε τυχόν ακραίες τιμές. Αντιθέτως, η παρουσία ακραίων τιμών ή "ακραίου" δείγματος εκπαίδευσης θα μπορούσε να επηρεάσει σοβαρά την επίδραση ενός συγκεκριμένου υποσυνόλου στον αντίστοιχο (μερικό) παράγοντα Bayes, καθώς και στον ενδογενή παράγοντα Bayes όπως θα δούμε στην επόμενη ενότητα.

Αναφορικά με το κλάσμα εκπαίδευσης b, αυτό θα μπορούσε να λάβει την τιμή m_0/n που βασίζεται στο μέγεθος του ελάχιστου δείγματος εκπαίδευσης για ένα δεδομένο πρόβλημα. Επομένως, τείνει ασυμπτωτικά προς το μηδέν εξασφαλίζοντας ότι υπάρχει συνέπεια κατά τη σύγκριση μοντέλων. Αν κανείς εκδηλώσει ενδιαφέρον για πιο εύρωστες επιλογές της παραμέτρου b, τότε θα πρέπει να στραφεί σε τιμές οι οποίες να ικανοποιούν επιπλέον και τη συνθήκη $nb \to \infty$. Κάποιες επιλογές που οδηγούν σε αυξημένη ευρωστία και ισχύ έναντι ενός ατυχούς προσδιορισμού της πρότερης κατανομής (*prior misspecification*) περιλαμβάνουν: $b = \frac{max(m_0, \log n)}{n}$ και $b = \frac{max(m_0, \sqrt{n})}{n}$ (O'Hagan 1995). Τέτοιες επιλογές ενδέχεται να οδηγήσουν σε σημαντική διακύμανση των ύστερων πιθανοτήτων των υπό

σύγκριση μοντέλων. Εκτός αυτού, θα πρέπει να τονιστεί ότι η κλασματική παράμετρος b δεν περιορίζεται στη λήψη τιμών μεγαλύτερων ή ίσων του m_0/n, αλλά μπορεί να πάρει και χαμηλότερες τιμές. Για παράδειγμα, θα μπορούσαμε να αποδώσουμε βάρος ενός σημείου ($b = n^{-1}$) κατά τους Kass & Wasserman (1992). Σύμφωνα με τον ίδιο τον O'Hagan, το ερώτημα που παραμένει ανοιχτό είναι πόσο χαμηλά μπορούμε να πάμε (βλέπε τη συζήτηση που συνοδεύει το σχετικό του άρθρο του έτους 1995), εφόσον υπάρχει ένας συμβιβασμός που πρέπει να γίνει από τον ερευνητή μεταξύ του βαθμού ευρωστίας, που αυξάνει με την τιμή του b, και της διακριτικής ισχύος, που μειώνεται με το b, τουλάχιστον όσον αφορά πεπερασμένα δείγματα.

Στη βάση του ισχυρισμού ότι ο FBF είναι λιγότερο ευαίσθητος στην επιλογή πρότερης κατανομής σε σχέση με τον κλασικό παράγοντα Bayes αλλά και με άλλες εναλλακτικές μορφές, ο O'Hagan διατυπώνει την άποψη ότι ο FBF μπορεί να προτιμηθεί ακόμη και σε περιπτώσεις γνήσιων πρότερων κατανομών, αν και η άποψη αυτή έχει κάπως ακραία χροιά.

Στη συζήτηση που ακολούθησε την παρουσίαση του άρθρου κατά τη διάρκεια μιας συνάντησης του RSS (*Royal Statistical Society*), η αίσθηση ότι ο FBF δρα ως μια *ad hoc* ψευδο-Μπεϋζιανή διαδικασία ήταν κοινή σε μια μεγάλη μερίδα των παριστάμενων στατιστικών επιστημόνων.

4.4.4 Ενδογενής παράγοντας Bayes

Μόλις ένα χρόνο μετά το άρθρο του O'Hagan, οι Berger & Pericchi (1996*b*) δημοσίευσαν ένα δεύτερο άρθρο σχετικά με τον ενδογενή παράγοντα Bayes, το οποίο ουσιαστικά συμπλήρωνε το πρώτο άρθρο τους και, μεταξύ άλλων, παρείχε απαντήσεις σε διάφορα ερωτήματα που απηύθυνε ο O'Hagan. Το άρθρο αυτό είναι εξαιρετικά καλογραμμένο και πολύ ενδιαφέρον για τον αναγνώστη. Ξεκινά απαριθμώντας μερικά από τα κύρια επιχειρήματα υπέρ της Μπεϋζιανής επιλογής μοντέλου έναντι του κλασικού ελέγχου υποθέσεων (π.χ. αδυναμίες των p-τιμών, δυσκολία στη σύγκριση μη εμφωλευμένων μοντέλων ή στην εφαρμογή του ξυραφιού του Occam, απουσία δυνατότητας εφαρμογής της μεθόδου στάθμισης Μπεϋζιανών μοντέλων για πρόβλεψη). Έτσι, έχοντας πείσει τον αναγνώστη ότι ο μόνος αξιόπιστος δρόμος είναι ο Μπεϋζιανός, οι συγγραφείς επιχειρηματολογούν υπέρ

των αυτόματων μεθόδων επιλογής μοντέλων, εφόσον στην πράξη δεν είναι πάντα εφικτό να χρησιμοποιεί κανείς υποκειμενικές πρότερες κατανομές σε όλες τις παραμέτρους του μοντέλου, ενώ η χρήση μη γνήσιων πρότερων κατανομών οδηγεί στα γνωστά πλέον προβλήματα που σχετίζονται με την αδυναμία υπολογισμού του BF.

Οι συγγραφείς ισχυρίζονται ότι ο προτεινόμενος ενδογενής παράγοντας Bayes είναι ένας εναλλακτικός παράγοντας Bayes που αντιστοιχεί σε μια τεχνική αυτόματης επιλογής μοντέλου και μπορεί κάλλιστα να εφαρμοστεί σε μεγάλο εύρος προβλημάτων (π.χ. δεν περιορίζεται σε εμφωλευμένα μοντέλα). Επομένως, η τεχνική θεωρείται ολιστική κατά μια έννοια. Υποθέτοντας ότι κάτω από το μοντέλο M_i το - αναγκαστικά ελάχιστο - δείγμα εκπαίδευσης $\boldsymbol{y}_{m_0}$ οδηγεί σε μια γνήσια ύστερη κατανομή $\pi_i\left(\boldsymbol{\theta_i}|\boldsymbol{y}_{m_0}\right)$, τότε ο μερικός παράγοντας Bayes που προκύπτει για το υπόλοιπο σύνολο $\boldsymbol{y}_{\setminus m_0}$ του αρχικού συνόλου παρατηρήσεων $\boldsymbol{y}$ δεδομένου του $\boldsymbol{y}_{m_0}$ είναι:

$$\begin{aligned} B_{ij}(\boldsymbol{y}_{\setminus m_0}|\boldsymbol{y}_{m_0}) &= B_{ij}(\boldsymbol{y}) \cdot B_{ji}(\boldsymbol{y}_{m_0}) \\ &= \frac{f_i\left(\boldsymbol{y}|M_i\right)}{f_j\left(\boldsymbol{y}|M_j\right)} \cdot \frac{f_j\left(\boldsymbol{y}_{m_0}|M_j\right)}{f_i\left(\boldsymbol{y}_{m_0}|M_i\right)} \\ &= \frac{\int f_i\left(\boldsymbol{y}|\boldsymbol{\theta}_i\right)\pi_i\left(\boldsymbol{\theta}_i\right)\mathrm{d}\boldsymbol{\theta}_i}{\int f_j\left(\boldsymbol{y}|\boldsymbol{\theta}_j\right)\pi_j\left(\boldsymbol{\theta}_j\right)\mathrm{d}\boldsymbol{\theta}_j} \cdot \frac{\int f_j\left(\boldsymbol{y}_{m_0}|\boldsymbol{\theta}_j\right)\pi_j\left(\boldsymbol{\theta}_j\right)\mathrm{d}\boldsymbol{\theta}_j}{\int f_i\left(\boldsymbol{y}_{m_0}|\boldsymbol{\theta}_i\right)\pi_i\left(\boldsymbol{\theta}_i\right)\mathrm{d}\boldsymbol{\theta}_i}. \end{aligned}$$

Το μέγεθος m_0 του ελάχιστου δείγματος εκπαίδευσης είναι συνήθως ίσο με τη μέγιστη διάσταση των διανυσμάτων παραμέτρων για τα ανταγωνιστικά μοντέλα που απαρτίζουν τον μοντελοχώρο $\mathcal{M}$. Κομβικό σημείο σε όλα αυτά αποτελεί το γεγονός ότι το σύνολο εκπαίδευσης για τον ενδογενή παράγοντα Bayes, συμβολιζόμενο ως $\boldsymbol{y}_{m_0}$, είναι μόνο ένα στιγμιότυπο από όλα τα δυνατά δείγματα εκπαίδευσης μεγέθους m_0. Ας συμβολίσουμε το σύνολο όλων των δυνατών δειγμάτων εκπαίδευσης μεγέθους m_0 με $\mathcal{Y}_{m_0}$ και τη διάστασή του με $|\mathcal{Y}_{m_0}|$. Με σκοπό να απαλείψουμε το σφάλμα που σχετίζεται με την τυχαία επιλογή ενός δείγματος εκπαίδευσης, όλα τα δείγματα εκπαίδευσης του συνόλου $\mathcal{Y}_{m_0}$ λαμβάνονται υπόψη και παίρνουμε τον μέσο όρο των παραγόμενων παραγόντων Bayes. Παρακάτω δείχουμε πώς υπολογίζεται ο IBF χρησιμοποιώντας είτε τον αριθμητικό μέσο είτε τον γεωμετρικό μέσο:

$$B_{ij}^{I,a}(\boldsymbol{y}) = B_{ij}(\boldsymbol{y}) \cdot \frac{1}{|\mathcal{Y}_{m_0}|} \sum_{\boldsymbol{y}_{m_0} \in \mathcal{Y}_{m_0}} B_{ji}\left(\boldsymbol{y}_{m_0}\right), \tag{4.4}$$

$$B_{ij}^{I,g}(\boldsymbol{y}) = B_{ij}(\boldsymbol{y}) \cdot \left(\prod_{\boldsymbol{y}_{m_0} \in \mathcal{Y}_{m_0}} B_{ji}\left(\boldsymbol{y}_{m_0}\right) \right)^{\frac{1}{|\mathcal{Y}_{m_0}|}},$$

όπου ισχύει ότι ο γεωμετρικός IBF (GIBF) είναι πάντα μικρότερος ή ίσος από τον αριθμητικό IBF (AIBF). Συνεπώς, ο AIBF προωθεί κάπως εντονότερα το απλούστερο μοντέλο M_j. Πώς, όμως, γνωρίζουμε ότι το μοντέλο του παρονομαστή είναι απλούστερο από το μοντέλο του αριθμητή; Πολύ απλά, επειδή οι Berger και Pericchi το επέβαλαν ως κανόνα. Κατά τον υπολογισμό του IBF, είναι εξ' ορισμού απαραίτητο να τοποθετούμε πρώτο το πιο πολύπλοκο μοντέλο ή ισοδύναμα το δεύτερο μοντέλο να είναι εμφωλευμένο μέσα στο πρώτο μοντέλο. Η ασυμμετρία που προκαλείται από αυτόν τον περιορισμό στη σειρά των υπό σύγκριση μοντέλων αφορά μόνο τον AIBF και όχι τον GIBF. Άρα, στην περίπτωση του αριθμητικού μέσου όρου, ο IBF του απλούστερου μοντέλου προς το πιο σύνθετο μοντέλο προσδιορίζεται από τη σχέση $B_{ji}^{I,a}(\boldsymbol{y}) = 1/B_{ij}^{I,a}(\boldsymbol{y})$ αντί να αντιστρέψει κανείς τους δείκτες της (4.4). Άλλη μια βασική ιδιότητα των παραγόντων Bayes που ο AIBF δεν ικανοποιεί εν γένει είναι η ακόλουθη:

$$\frac{B_{ij}^{I,a}(\boldsymbol{y})}{B_{kj}^{I,a}(\boldsymbol{y})} = B_{ik}^{I,a}(\boldsymbol{y}). \tag{4.5}$$

Ποιο από τα δύο μοντέλα είναι πιο πολύπλοκο από το άλλο δεν είναι πάντα ευδιάκριτο στον πραγματικό κόσμο, ιδιαίτερα όταν έχουμε πολλαπλά μοντέλα υπό σύγκριση. Μια μερική λύση θα αποτελούσε η χρήση ενός περικλείοντος μοντέλου (*encompassing model*) M_0 το οποίο να περιλαμβάνει όλα τα μοντέλα του χώρου $\mathcal{M}$ και ο υπολογισμός όλων των σχετικών IBF να γίνεται με το περικλείον μοντέλο M_0 στον αριθμητή.

Όπως αναφέρθηκε νωρίτερα, ακόμη και για χαμηλές τιμές του m_0, όπως $m_0 = 3$ ή $m_0 = 4$, το πλήθος $|\mathcal{Y}_{m_0}|$ όλων των δυνατών ελάχιστων δειγμάτων εκπαίδευσης ενδέχεται να είναι τόσο μεγάλο ώστε το υπολογιστικό κόστος να είναι ασύμφορο. Σε μια τέτοια περίπτωση, ένα τυχαία επιλεγμένο υποσύνολο του $\mathcal{Y}_{m_0}$ μπορεί να επιστρατευτεί για τον υπολογισμό του IBF. Το υποσύνολο αυτό οφείλει να είναι όσο το δυνατόν πιο ευρύ ώστε να ελαχιστοποιήσει το σφάλμα δειγματοληψίας από το $\mathcal{Y}_{m_0}$. Αν παρά ταύτα διακρίνουμε σημαντική διακύμανση μεταξύ των παραγόμενων τιμών IBF που αντιστοιχούν στα διάφορα υποσύνολα του $\mathcal{Y}_{m_0}$, προτείνεται η περικοπή (*trimming*) των αποτελεσμάτων κατά ένα ποσοστό α μέσω της αφαίρεσης $\alpha/2$ τιμών από κάθε πλευρά της δειγματικής κατανομής των μερικών παραγόντων Bayes $B_{ji}\left(\boldsymbol{y}_{m_0}\right), \boldsymbol{y}_{m_0} \in \mathcal{Y}_{m_0}$ που εμπλέκονται στην εξαγωγή του

μέσου. Μάλιστα, όσον αφορά μη εμφωλευμένα μοντέλα, συνιστάται ιδιαιτέρως η εφαρμογή περικοπής των αποτελεσμάτων κατά ένα ποσοστό 15% (Berger & Pericchi 1996*b*). Υπερβολικά ευρεία περικοπή των αποτελεσμάτων θα οδηγούσε εν τέλει στη χρήση της διαμέσου της εν λόγω κατανομής, αν και κάτι τέτοιο θα εξυπηρετούσε κυρίως όταν ο μοντελοχώρος είναι μεγέθους $|\mathcal{M}| = 2$. Κατά συνέπεια, η σειρά των μοντέλων δεν έχει πλέον επίδραση στην τελική τιμή του IBF, είτε αναφερόμαστε στον αριθμητικό είτε στον γεωμετρικό.

Σε κάποιες περιπτώσεις, ο AIBF ενδέχεται να αντιστοιχεί σε πραγματικό παράγοντα Bayes προερχόμενο από μια γνήσια πρότερη κατανομή. Αυτή η πρότερη κατανομή εύλογα ονομάζεται ενδογενής (*intrinsic prior*). Δεν θα υπεισέλθουμε σε λεπτομέρειες όσον αφορά τις ενδογενείς πρότερες κατανομές, αλλά, εκτός του βασικού άρθρου των Berger & Pericchi (1996*b*), ο αναγνώστης μπορεί να αναζητήσει λεπτομέρειες στα εξής άρθρα: Moreno (1997), Kim (2000), Cano, Kessler & Moreno (2004), Consonni & La Rocca (2008) and Torres-Ruiz, Moreno & Girón (2011).

Όπως έχει ήδη αναφερθεί, περιπτώσεις πολλαπλών μοντέλων υπό σύγκριση συγκεντρώνουν πολύ υψηλότερο ερευνητικό ενδιαφέρον. Είναι εμφανές ότι ελάχιστα δείγματα εκπαίδευσης πρέπει να οριστούν για όλα τα μοντέλα. Σε ένα τέτοιο πλαίσιο, η γεωμετρική μορφή του IBF προσφέρεται περισσότερο σε σχέση με τον αριθμητικό, παρότι οι συγγραφείς τείνουν γενικά προς τον δεύτερο εξαιτίας της αντιστοίχισής του με ενδογενείς πρότερες κατανομές, προσδίδοντας την ερμηνεία μιας πλήρως Μπεϋζιανής διαδικασίας, και στην μεγαλύτερη σταθερότητα που έχει.

Ένα επιπλέον χαρακτηριστικό του IBF είναι ότι είναι αμετάβλητος σε μονομεταβλητούς μετασχηματισμούς των δεδομένων, αλλά δε συμβαίνει το ίδιο για πολυμεταβλητούς μετασχηματισμούς. Ακόμη, η ιδιότητα επάρκειας (*sufficiency property*) ενδέχεται να μην ισχύει σε κάποιες περιπτώσεις.

4.4.4.1 Διάμεσος ενδογενής παράγοντας Bayes

Αναδρομικά, οι διάφοροι περιορισμοί του AIBF κρίθηκαν ως ιδιαίτερα σημαντικοί από τους συγγραφείς του IBF (Berger & Pericchi 1998), ειδικά στα πλαίσια μη εμφωλευμένων και πολύπλοκων μοντέλων όπου η προσέγγιση του περικλείοντος μοντέλου απο-

δεικνύεται συχνά άχρηστη και προκύπτουν θέματα έλλειψης συνοχής. Ακόμη όμως και σε απλούστερες περιπτώσεις μοντέλων, μικρό μέγεθος δείγματος ενδέχεται να φέρει μεγάλη αστάθεια στις παραγόμενες τιμές του IBF. Αυτό που προτείνουν, λοιπόν, οι συγγραφείς της μεθοδολογίας IBF είναι η χρήση του επονομαζόμενου διάμεσου IBF (MIBF):

$$B_{ij}^{I,med}(\boldsymbol{y}) = \text{Median}[B_{ij}\,(\boldsymbol{y})] = B_{ij}(\boldsymbol{y}) \cdot \text{Median}[B_{ji}(\boldsymbol{y}_1)].$$

Μια παραπλήσια προσέγγιση θα ήταν να υπολογίσουμε το πηλίκο των διάμεσων των σχετικών περιθώριων κατανομών, καταλήγοντας στον IBF του πηλίκου των διαμέσων (*Ratio of Medians IBF*, RMIBF). Αμφότερες αυτές οι προσεγγίσεις προσφέρουν μια σημαντική λύση στο θέμα της έλλειψης συνοχής του IBF. Συνεπώς, η θεμελιώδης ιδιότητα ενός BF, $B_{ij}(\boldsymbol{y}) = \frac{1}{B_{ji}(\boldsymbol{y})}$, πλέον ισχύει. Η συνοχή παρουσία πολλαπλών μοντέλων είναι έγκυρη μόνο για τον RMIBF, αν και ο MIBF πρακτικά δε φαίνεται να παραβιάζει την σχετική συνθήκη, ειδικά κάτω από πρότερες κατανομές αναφοράς (*reference priors*). Ένα μειονέκτημα που αφορά μόνο τη δεύτερη από τις δύο μορφές είναι ότι οι διάμεσες τιμές μπορεί να προέρχονται από διαφορετικά δείγματα εκπαίδευσης, κάτι που καθιστά τον RMIBF μεταβλητό σε μετασχηματισμούς του $\boldsymbol{y}$, αν και δεν αναμένεται να προκληθούν ιδιαίτερα θέματα στην πράξη. Άλλο ένα σημείο που φαίνεται να δυσαρεστεί τους συγγραφείς είναι το γεγονός ότι ο διάμεσος IBF συνήθως δεν αντιστοιχεί σε ενδογενή πρότερη κατανομή, σε αντίθεση με τον αριθμητικό IBF.

Ας σημειωθεί ότι παρότι η διάμεση μορφή του IBF δεν επηρεάζεται από αλλαγές στις πρότερες κατανομές (ακόμη και αν είναι μεγάλης διακύμανσης) σε αντίθεση με εναλλακτικές μορφές IBF, οι συγγραφείς αποθαρρύνουν τη χρήση του στην περίπτωση εμφωλευμένων μοντέλων με την πρότερη κατανομή Jeffreys. Η καλύτερη πρόταση θα ήταν να χρησιμοποιείται με μη εμφωλευμένα μοντέλα κάτω από πρότερες κατανομές αναφοράς.

Οι μεθοδολογίες επιλογής μοντέλου θα πρέπει να προσαρμόζονται ανάλογα με το πρόβλημα, όπως αναφέρουν οι Berger & Pericchi (2001) και οι Berger & Pericchi (2004). Επομένως, όπως διαισθητικά θα περίμενε κανείς, η χρήση του ελάχιστου δείγματος εκπαίδευσης δε φαίνεται να αποτελεί καθολική λύση για όλα ανεξαιρέτως τα προβλήματα.

4.4.5 Συγκρίσεις μεταξύ IBF και FBF στη βιβλιογραφία

Όπως θα δούμε στη συνέχεια, το 1997 ήταν μια ιδιαίτερα παραγωγική χρονιά αναφορικά με δημοσιεύσεις που αντιπαρέβαλλαν τις μεθοδολογίες των FBF και IBF. Αυτό εν μέρει αποδεικνύει την ξεχωριστή εντύπωση που τα αρχικά άρθρα πυροδότησαν.

Ο O'Hagan, σε επόμενο άρθρο του (O'Hagan 1997), αντιπαραβάλει και συγκρίνει άμεσα τους IBF και FBF. Το γεγονός αυτό από μόνο του καθιστά τη δουλειά αυτή άκρως ενδιαφέρουσα παρά την προφανή μεροληψία που προέρχεται από το γεγονός ότι ο συγγραφέας έχει αναπτύξει τη μία εκ των δύο αντίπαλων μεθοδολογιών υπό εξέταση. Αφού παρουσιάζονται εν συντομία οι δύο αντίπαλες μεθοδολογίες, οι IBF και FBF αντιπαραβάλλονται ως προς τέσσερις ιδιότητες συνοχής: i) συνέπεια (δηλαδή ότι ο αντίστοιχος $B_{ij}(\boldsymbol{y})$ για δύο εμφωλευμένα μοντέλα τείνει στο άπειρο καθώς $n \to \infty$ αν το μοντέλο M_i είναι αληθές), ii) εγκυρότητα της αρχής της πιθανοφάνειας (*likelihood principle*), iii) μη μεταβλητότητα σε ένα-προς-ένα μετασχηματισμούς δεδομένων ή παραμέτρων και iv) δυνατότητα διαδοχικής αναβάθμισης.

Χάριν σύνοψης των κύριων ευρημάτων που προκύπτουν από τις διάφορες συγκρίσεις που διεξάγονται, θα αναφέρουμε ότι ο FBF, σε αντίθεση με τον IBF, όχι μόνο ικανοποιεί πάντα τη βασική αρχή της πιθανοφάνειας και το κριτήριο της επάρκειας που απορρέει από αυτήν, αλλά είναι και αμετάβλητος σε μετασχηματισμούς δεδομένων εφόσον δεν επηρεάζεται από τη συνακόλουθη αλλαγή στο ελάχιστο δείγμα εκπαίδευσης. Και οι δύο μέθοδοι είναι αμετάβλητες σε νέες παραμετροποιήσεις βασισμένες σε ένα-προς-ένα μετασχηματισμούς. Η αριθμητική μορφή του IBF είναι προβληματική όσον αφορά τη σύγκριση μη εμφωλευμένων μοντέλων και παραβιάζει την γενική αρχή συνοχής ενός παράγοντα Bayes, παρότι ικανοποιεί την ιδιότητα συνέπειας που αναφέρθηκε στην προηγούμενη παράγραφο. Σχετικά με την τελευταία αυτή ιδιότητα, ο FBF την παραβιάζει εκτός και αν $b \to 0$ καθώς $n \to \infty$. Κάτι τέτοιο έρχεται σε αντίθεση με το γεγονός ότι η παράμετρος b πρέπει να θεωρείται σταθερή ώστε να ισχύει η δυνατότητα διαδοχικής αναβάθμισης. Επομένως, σε αυτό το σημείο υπάρχει ένας συμβιβασμός που πρέπει να γίνει με βάση την κρίση του εκάστοτε ερευνητή και με βάση τον σκοπό της εκάστοτε ανάλυσης.

Αναλογιζόμενος τη σύγκριση πολλαπλών μοντέλων, ο O'Hagan κατασκευάζει ένα γενικής φύσης ελάχιστο περικλείον μοντέλο M_0 ώστε να ενδυναμώσει την προσέγγιση που

προτείνεται από τους συγγραφείς του ενδογενούς παράγοντα Bayes. Αυτό το ελάχιστο περικλείον μοντέλο απαλείφει σε έναν βαθμό την αυθαίρετη επιλογή ενός μοντέλου που να περικλείει όλα τα ανταγωνιστικά μοντέλα, όμως κάποιος βαθμός αυθαιρεσίας παραμένει όσον αφορά την πρότερη κατανομή που σχετίζεται με το μοντέλο M_0. Επιπρόσθετα, η προσέγγιση του M_0 απαιτεί μεγαλύτερο δείγμα εκπαίδευσης σε σχέση με το ελάχιστο μέγεθος δείγματος που απαιτείται για τη σύγκριση πολλαπλών μη εμφωλευμένων μοντέλων παρακάμπτοντας το M_0 (De Santis & Spezzaferri 1999). Αναφορικά με τη συνοχή ενός παράγοντα Bayes παρουσία πολλαπλών μοντέλων όπως συνοψίζεται στην (4.5), υπογραμμίζεται ότι η αριθμητική εκδοχή του IBF παραβιάζει αυτή την αρχή σε αντίθεση με τον γεωμετρικό IBF και τον FBF, ο οποίος όμως απαιτεί σταθερό b σε όλους τους εμπλεκόμενους όρους.

Ο O'Hagan, για ακόμη μια φορά, δεν ικανοποιείται ούτε από την ύπαρξη των ενδογενών πρότερων κατανομών που σχετίζονται με τον (αριθμητικό) IBF εξαιτίας της θεμελίωσής τους σε ασυμπτωτική λογική. Ωστόσο, ενδογενείς πρότερες κατανομές αντιστοιχούν στον FBF αν το κλάσμα εκπαίδευσης (*training fraction*), δηλ. η κλασματική παράμετρος b, είναι της τάξης $O(n^{-1})$.

Όσον αφορά την επιλογή της κλασματικής παραμέτρου b του FBF, το κύριο συμπέρασμα είναι ότι θα πρέπει να παίρνει όσο το δυνατόν υψηλότερες τιμές ώστε να παρέχει εύρωστα και ισχυρά αποτελέσματα και ταυτόχρονα να είναι όσο το δυνατόν μικρότερη ώστε να καταστήσει δυνατή την ορθή διάκριση μεταξύ των μοντέλων. Αυτό είναι ένα θέμα για περαιτέρω διερεύνηση, καθώς το να θέσουμε απλώς την παράμετρο b ίση με το μέγεθος του ελάχιστου δείγματος εκπαίδευσης δε μπορεί να αποτελέσει κάποιο γενικό κανόνα κατάλληλα δικαιολογημένο. Ευρετικές (*heuristic*) εναλλακτικές έχουν προταθεί, οι οποίες βασίζονται στο μέγεθος δείγματος κατά έναν πιο δημιουργικό τρόπο, π.χ. $\frac{m_0}{n}$, $\frac{m_0 \log n}{n \log m_0}$ ή $\frac{m_0, \sqrt{n}}{n}$. Έχει ήδη σημειωθεί ότι ο FBF φαίνεται να έχει από τη φύση του προβλήματα συνέπειας στην περίπτωση που η παράμετρος b μένει αμετάβλητη όσο το μέγεθος δείγματος n αυξάνει.

Αξίζει να σημειωθεί ότι η συνήθης ιδιότητα της διαδοχικής αναβάθμισης δεν πληρείται από καμιά μέθοδο. Συγκεκριμένα, ο IBF θα απαιτούσε ένα i.i.d. δείγμα για να ισχύει κάτι τέτοιο, ενώ ο FBF θα απαιτούσε ένα σταθερό κλάσμα εκπαίδευσης, κάτι που θα ερ-

χόταν σε αντίθεση με ό,τι αναφέρθηκε στο τέλος της προηγούμενης παραγράφου.

Ο ίδιος ο O'Hagan αναγνωρίζει το συχνό θέμα της ανάγκης για αριθμητική ολοκλήρωση των εμπλεκόμενων ποσοτήτων και στους δύο παράγοντες, FBF και IBF, αν και το πλήθος των ολοκληρωμάτων είναι πιο περιορισμένο σε σχέση με τον κλασικό παράγοντα Bayes.

Ο Moreno (1997) ασχολήθηκε με την περίπτωση εμφωλευμένων μοντέλων και προέβη σε συγκρίσεις μεταξύ των FBF και IBF αποδίδοντας ιδιαίτερη έμφαση στην ευρωστία των αποτελεσμάτων. Ο O'Hagan (1997) παρουσιάζει την αντιστοίχιση κλασματικών πρότερων κατανομών με τον FBF, με ανάλογο τρόπο όπως οι ενδογενείς πρότερες κατανομές αντιστοιχούν στον IBF. Όπως συμβαίνει και με τις ενδογενείς πρότερες κατανομές, η σχετική κλασματική εξίσωση δεν έχει μοναδική λύση. Σύμφωνα με τον ορισμό που δίνεται, δύο γνήσιες πρότερες κατανομές που ανατίθενται σε δύο μοντέλα καλούνται κλασματικές αν οι αντίστοιχοι BF και FBF είναι ασυμπτωτικά ισοδύναμοι για κάποια ακολουθία b_n του κλάσματος εκπαίδευσης. Η απαίτηση του ορθού ορισμού της ακολουθίας b_n καθιστά αυτή την προσέγγιση λιγότερο αυτόματη σε σχέση με την προσέγγιση των ενδογενών πρότερων κατανομών. Αποδεικνύεται ότι υπάρχουν περιπτώσεις όπου η ακολουθία $b_n = \frac{m}{n}$ οδηγεί σε μια κενή κλάση κλασματικών a priori κατανομών. Παρόλα αυτά, αν η εν λόγω κλάση δεν είναι το κενό σύνολο, τότε για $m = m_0$ η ακολουθία $b_n = \frac{m_0}{n}$ φέρεται να είναι η πλέον κατάλληλη ανάμεσα σε όσες παρουσιάζονται στη βιβλιογραφία και να είναι αυτή που παράγει τιμές FBF πολύ κοντά στον παράγοντα Bayes που προκύπτει από μια ενδογενή πρότερη κατανομή.

Είναι ενδιαφέρον ότι ο Moreno προωθεί μια σχετικά ακραία άποψη προτείνοντας ότι πραγματικοί παράγοντες Bayes θα πρέπει να επιδιώκονται με κάθε μέσο και αν κάτι τέτοιο δεν είναι δυνατό, τότε ενδογενείς πρότερες κατανομές (ή παραπλήσιες αυτών) θα πρέπει να χρησιμοποιούνται ώστε να προσεγγιστούν οι εν λόγω παράγοντες Bayes, δείχνοντας παράλληλα ότι οι σχετικοί εναλλακτικοί παράγοντες είναι πράγματι παράγοντες Bayes.

Αργότερα, το άρθρο των De Santis & Spezzaferri (1999) προσέλκυσε μεγάλο ενδιαφέρον. Πρόκειται κατά κύριο λόγο για ένα άρθρο ανασκόπησης του κλασματικού παράγοντα Bayes, το οποίο δημοσιεύθηκε λίγο μετά το αρχικό άρθρο που επιθεωρεί. Εν τω συνόλω, οι συγγραφείς είναι θετικά προδιατεθειμένοι ως προς τον κατά τα λοιπά αμφιλεγόμενο

FBF, ειδικά για περιπτώσεις αδύναμης πρότερης πληροφορίας. Βρίσκουν ιδιαίτερα ελκυστική την ιδιότητα του FBF να είναι ανθεκτικός σε ατυχείς επιλογές πρότερης κατανομής. Συγκρίσεις έχουν διεξαχθεί ανάμεσα στον FBF, στον διάμεσο IBF αλλά και στις μεθόδους των Smith & Spiegelhalter (1980) και Spiegelhalter & Smith (1982). Για εύρωστες Μπεϋζιανές αναλύσεις, ο FBF ηγείται της εύρωστης επιλογής μοντέλου σε περισσότερες περιπτώσεις συγκριτικά με τον αυθεντικό παράγοντα Bayes, εφόσον σχετίζεται με πιο λογικά όρια για δεδομένες κλάσεις πρότερων κατανομών (όπως οι μονοκόρυφες και οι συμμετρικές κλάσεις πρότερων κατανομών). Υπενθυμίζεται ότι ένας παράγοντας $B_{ij}(\boldsymbol{y})$ καλείται εύρωστος πάνω σε μια δεδομένη κλάση Γ πρότερων κατανομών αν το $\sup_{\Gamma}(B_{ij}(\boldsymbol{y}))$ είναι αυστηρά μικρότερο της μονάδας υποστηρίζοντας το μοντέλο M_i ή αν το $\inf_{\Gamma}(B_{ij}(\boldsymbol{y}))$ είναι αυστηρά μεγαλύτερο της μονάδας υποστηρίζοντας το μοντέλο M_j.

Σχετικά με την ευαισθησία των διαφόρων εναλλακτικών παραγόντων Bayes σε μικρές αλλαγές των δεδομένων και των τιμών των υπερπαραμέτρων (ειδικά όταν αυτές αντιστοιχούν σε ασαφείς πρότερες κατανομές), προβάλλεται ο ισχυρισμός ότι ειδικά ο AIBF (και ο GIBF σε μικρότερο βαθμό) επηρεάζεται σημαντικά. Κάτι τέτοιο δεν ισχύει για τους MIBF και FBF. Οι ισχυρισμοί αυτοί υποστηρίζονται μέσω πολλών παραδειγμάτων (poisson έναντι αρνητικού διωνυμικού μοντέλου, εκθετικό έναντι λογαριθμοκανονικού μοντέλου) όπου μικρές διαταραχές επιβάλλονται στις υπερπαραμέτρους και στα δεδομένα. Παράλληλα, αναφέρονται αποτελέσματα της δουλειάς των Bertolino & Racugno (1996) όπου εντοπίζονται μεγάλες διαφορές στον AIBF όταν τοποθετείται εναλλάξ στον αριθμητή το απλούστερο και το πιο πολύπλοκο μοντέλο. Η εκδοχή RM του διάμεσου IBF παραμένει σταθερή στις προαναφερθείσες διαταραχές, αν και ο κλασματικός παράγοντας Bayes εμφανίζεται ακόμη καλύτερος σε ορισμένες περιπτώσεις.

Όπως έχει προαναφερθεί, η προσέγγιση του περικλείοντος μοντέλου (M_0) απαιτεί μεγαλύτερα δείγματα εκπαίδευσης από το να συγκρίνει κανείς ζεύγη των αρχικών μοντέλων απευθείας. Αυτό αποδεικνύεται εύκολα στα πλαίσια του πολλαπλού γραμμικού μοντέλου με άγνωστη διασπορά: αν κάθε μοντέλο M_k συνδέεται με έναν πίνακα σχεδιασμού $\mathbf{X}_k$ διάστασης $n \times p_k$, τότε η σύγκριση δύο μοντέλων M_{k_1}, M_{k_2} θα απαιτούσε ένα δείγμα εκπαίδευσης $\max\{p_{k_1}, p_{k_2}\} + 1$. Από την άλλη πλευρά, το περικλείον μοντέλο M_0 θα συνδεόταν με έναν πίνακα σχεδιασμού $\mathbf{X}_0$ διάστασης (n, p_0) όπου $p_0 + 1$ είναι το σύνολο

των μοναδικών συμμεταβλητών όλων των μοντέλων που απαρτίζουν τον μοντελοχώρο. Είναι ξεκάθαρο ότι $p_0 > \max\{p_{k_1}, p_{k_2}\}$ για κάθε ζεύγος (M_{k_1}, M_{k_2}) του μοντελοχώρου, γεγονός που έχει αρνητική επίδραση στην διακριτική ικανότητα της διαδικασίας, ειδικά στα πλαίσια προβλημάτων 'small n - large p'.

Αναλογικά με τις ενδογενείς πρότερες κατανομές, οι συγγραφείς έχουν αναπτύξει μια θεωρία γύρω από τις κλασματικές ενδογενείς πρότερες κατανομές (*fractional intrinsic priors*, FIPs) (De Santis & Spezzaferri 1997). Σαν μια τελευταία παρατήρηση για το συγκεκριμένο θέμα, ούτε η ύπαρξη, ούτε η μοναδικότητα ή η γνησιότητα των FIPs εγγυώνται εκ των προτέρων. Το περιεχόμενο του εν λόγω άρθρου επεκτείνεται πέρα από τα όρια μιας απλής επισκόπησης και περιλαμβάνει ταυτοποιήσεις ορισμένων προβλημάτων όπου ο FBF δεν εφαρμόζεται (όπως το πρόβλημα των Neyman & Scott). Για τέτοιες περιπτώσεις, προτείνεται μια γενίκευση του FBF με βάση μη ασυμπτωτική λογική, σε αντίθεση με το πρωτότυπο άρθρο του O'Hagan (1995), στη βάση εμφωλευμένων μοντέλων. Συγκεκριμένα, παρέχεται μια εξίσωση που συσχετίζει τον FBF και τον γεωμετρικό IBF εμφωλευμένων μοντέλων, δηλώνοντας ότι $B_{ij}^F(\boldsymbol{y}) \geq B_{ij}^{I,g}(\boldsymbol{y})$.

4.4.6 Στοιχειώδες συγκριτικό παράδειγμα

Ας θεωρήσουμε δύο εμφωλευμένα μοντέλα,

$$M_0 : Y \sim N(0, \sigma^2) \ , \qquad M_1 : Y \sim N(\mu, \sigma^2)$$

στα οποία εκχωρούνται ίσες πρότερες πιθανότητες και τυχαίο δείγμα $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $n = 1000$. Επίσης, η πρότερη πληροφορία για τις παραμέτρους του εκάστοτε μοντέλου είναι ανάλογη με $\frac{1}{\sigma^2}$. Οι αντίστοιχες περιθώριες πιθανοφάνειες δεδομένης της παραμέτρου b είναι:

$$f\left(\boldsymbol{y}|b, M_0\right) = (2\pi)^{-\frac{nb}{2}} b^{-\frac{nb}{2}} \Gamma\left(\frac{nb}{2}\right) \left(\frac{\sum_{i=1}^n y_i^2}{2}\right)^{-\frac{nb}{2}}$$

και

$$f\left(\boldsymbol{y}|b, M_1\right) = (2\pi)^{-\frac{(nb-1)}{2}} n^{-\frac{1}{2}} b^{-\frac{nb}{2}} \Gamma\left(\frac{nb-1}{2}\right) \left(\frac{(n-1)S_{\boldsymbol{y}}^2}{2}\right)^{-\frac{(nb-1)}{2}}$$

όπου $S_{\boldsymbol{y}}^2$ είναι η δειγματική διακύμανση του $\boldsymbol{y}$. Η δύναμη b ισούται με μονάδα, εκτός αν θέλουμε να υπολογίσουμε το κλασματικό μέρος του FBF. Το ελάχιστο δείγμα εκπαίδευσης

αποτελείται από $m_0 = 2$ παρατηρήσεις. Ένα δείγμα εκπαίδευσης μιας μόνο παρατήρησης απορρίπτεται, διότι διαφορετικά η περιθώρια πιθανοφάνεια υπό το μοντέλο H_1 θα ήταν ίση με μηδέν και κάτι τέτοιο θα παραβίαζε τον ορισμό του ελάχιστου δείγματος εκπαίδευσης όπως αυτός δόθηκε από τους Berger και Pericchi (1996). Επίσης, για την IBF προσέγγιση, επιλέξαμε 50000 μοναδικά διανύσματα $\boldsymbol{y}_1$ από περίπου 500000 συνολικούς συνδυασμούς (αναλογία = $1/10$).

Στην περίπτωση όπου $m = 2$, αντί μιας απλής τυχαίας δειγματοληψίας από το πλήθος των δυνατών ζευγών από το διάνυσμα $\boldsymbol{y}$, μπορεί να χρησιμοποιηθεί ένα δείγμα με κάποιους περιορισμούς που θα εξασφαλίζουν την ταυτοποίηση/εκτίμηση των παραμέτρων. Για παράδειγμα, μια τέτοια επιλογή θα μπορούσε να είναι η επιλογή τυχαίων ζευγών παρατηρήσεων με ελάχιστη απόσταση $\epsilon_d > 0$. Στις δικές μας εφαρμογές, δοκιμάσαμε διάφορες τιμές για το ϵ_d και καταλήξαμε στην τιμή 0.1. Αυτή η προσέγγιση, συγκρινόμενη με την απλή τυχαία δειγματοληψία, έδωσε αρκετά σταθερά αποτελέσματα για τους GIBF και MIBF, ενώ παράλληλα οδήγησε σε πιο ρεαλιστικά αποτελέσματα (υποδιπλασιάζοντας σχεδόν τις τιμές) για τον AIBF. Υπολογίσαμε ακόμη τιμές του ενδογενούς παράγοντα Bayes για $m = 5$, δηλαδή για κάπως μεγαλύτερο μέγεθος δείγματος εκπαίδευσης.

Το μοντέλο M_1 είναι πιο πολύπλοκο συγκριτικά με το μοντέλο M_0. Ισοδύναμα, το μοντέλο M_0 είναι εμφωλευμένο μέσα στο μοντέλο M_1. Επομένως, υπολογίσαμε την τιμή του B_{10} ώστε να είμαστε συνεπείς με τον περιορισμό που επιβάλλει ο AIBF ώστε να οδηγηθούμε σε όσο το δυνατόν πιο συγκρίσιμα αποτελέσματα με τους FBF και GIBF. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.2, με τα κυρίως συμπεράσματα να περιγράφονται ως ακολούθως.

- Για τυχαίο δείγμα $\boldsymbol{y} \sim N(0, \sigma^2)$ προσομοιωμένο από το κανονικό μοντέλο με μηδενική μέση τιμή, το μοντέλο M_0 υποστηρίζεται ξεκάθαρα, αλλά το βάρος υπέρ του μοντέλου M_0 μειώνεται καθώς το m αυξάνει. Για την ακραία τιμή $m = 0.9n$, άρα για $b = 0.9$, η αντίστοιχη τιμή του FBF είναι αποπροσανατολιστική αγγίζοντας σχεδόν τη μονάδα. Είναι ενδιαφέρον ότι εδώ φαίνεται ότι όσο πέφτει η τιμή του m, τόσο πιο αξιόπιστο είναι το αποτέλεσμα του FBF. Αναφορικά με τα αποτελέσματα του IBF, η αύξηση της τιμής του m οδηγεί σε μειωμένη στήριξη του M_0 κάτω από οποιαδήποτε εναλλακτική μορφή του IBF. Ας σημειωθεί ότι οι τιμές του FBF για

$m = 2$ είναι πολύ κοντά στις τιμές του AIBF για την ίδια τιμή του m.

- Για τυχαίο δείγμα $\boldsymbol{y} \sim N(1, \sigma^2)$ προσομοιωμένο από το κανονικό μοντέλο με μέση τιμή ίση με τη μονάδα, το μοντέλο M_1 υποστηρίζεται σθεναρά, όπως θα έπρεπε. Για πολύ υψηλές τιμές του m, η στήριξη του μοντέλου M_1 παραμένει σημαντική, αλλά κάπως εξασθενημένη σε σχέση με πριν. Ακόμη, όσο αυξάνει το μέγεθος εκπαίδευσης m που χρησιμοποιείται στον IBF, το μοντέλο M_0 προωθείται κάπως περισσότερο, αν και η διαφορά είναι αμελητέα.

- Για τυχαίο δείγμα $\boldsymbol{y} \sim N(10, \sigma^2)$ προσομοιωμένο από το κανονικό μοντέλο με μέση τιμή αρκετά μακριά από το μηδέν, όπου το M_1 θα έπρεπε να υποστηριχθεί ξεκάθαρα, έχουμε καταγράψει τα αποτελέσματα κατά τρόπο που οι διαφορές σε λογαριθμική κλίμακα να μπορούν να αξιολογηθούν, τουλάχιστον για την προσέγγιση FBF. Ας σημειωθεί ότι όσο το m αυξάνει, το $\log B_{10}(\boldsymbol{y})$ που βρίσκεται εντός της παρένθεσης μειώνεται, υποδηλώνοντας ότι η αντίστοιχη τιμή του FBF επίσης πέφτει κάπως. Τα αποτελέσματα που σχετίζονται με τον IBF δε μπορούν να εκφραστούν καλά ούτε σε λογαριθμική κλίμακα, όμως όλα τείνουν προς το άπειρο υποστηρίζοντας το M_1.

Σαν γενικό συμπέρασμα αναφορικά με τις εναλλακτικές μορφές του IBF, παρατηρούμε ότι η γεωμετρική και η διάμεση μορφή προωθούν περισσότερο το πραγματικό μοντέλο, ενώ ο AIBF συμπεριφέρεται φανερά χειρότερα. Ο γεωμετρικός IBF υποστηρίζει περισσότερο το απλούστερο μοντέλο.

4.5 Το Πρόβλημα Επιλογής Μετασχηματισμού Μέσω Εναλλακτικών Παραγόντων Bayes

Σε ένα πρώτο επίπεδο, θα δείξουμε πώς μπορούμε να ενσωματώσουμε εναλλακτικές μορφές παραγόντων Bayes στη διαδικασία επιλογής μετασχηματισμού κάτω από ασαφή πρότερη πληροφορία. Στις επόμενες ενότητες αυτού του κεφαλαίου, θα ενισχύσουμε και θα επεκτείνουμε τη διαδικασία που περιγράφεται στην τρέχουσα ενότητα θεωρώντας διαδοχικά την παρουσία συμμεταβλητών στο μοντέλο και, εν τέλει, την επιλογή μοντέλου.

Πίνακας 4.2: Εναλλακτικές μορφές παραγόντων Bayes για προσομοιωμένα σύνολα δεδομένων μεγέθους $n = 1000$ από την κανονική κατανομή και διάφορα μεγέθη δειγμάτων εκπαίδευσης m, με έμφαση στην τιμή της παραμέτρου μ, $\mu = 0$ (M_0) έναντι $\mu \neq 0$ (M_1).

Προσέγγιση	m	$\boldsymbol{y} \sim N(0, \sigma^2)$	$\boldsymbol{y} \sim N(1, \sigma^2)$	$\boldsymbol{y} \sim N(10, \sigma^2)$
$B_{10}^{F}(\boldsymbol{y})$	2	0.028	$2.3{\cdot}10^{+142}$	$\exp(2301.2)$
$B_{10}^{F}(\boldsymbol{y})$	$\log n$	0.082	$1.3{\cdot}10^{+142}$	$\exp(2290.9)$
$B_{10}^{F}(\boldsymbol{y})$	$\frac{2\log n}{\log 2}$	0.152	$3.2{\cdot}10^{+140}$	$\exp(2261.5)$
$B_{10}^{F}(\boldsymbol{y})$	$\sqrt{n}$	0.194	$8.5{\cdot}10^{+138}$	$\exp(2234.7)$
$B_{10}^{F}(\boldsymbol{y})$	$\sqrt{2n}$	0.232	$1.3{\cdot}10^{+137}$	$\exp(2204.6)$
$B_{10}^{F}(\boldsymbol{y})$	$0.9n$	0.959	$2.5{\cdot}10^{+14}$	$\exp(230.9)$
$B_{10}^{I,a}(\boldsymbol{y})$	2	0.031	$2.8{\cdot}10^{+142}$	$> 10^{+143}$
$B_{10}^{I,g}(\boldsymbol{y})$	2	0.019	$1.4{\cdot}10^{+142}$	$> 10^{+143}$
$B_{10}^{I,med}(\boldsymbol{y})$	2	0.021	$1.6{\cdot}10^{+142}$	$> 10^{+143}$
$B_{10}^{I,a}(\boldsymbol{y})$	5	0.054	$2.6{\cdot}10^{+142}$	$> 10^{+143}$
$B_{10}^{I,g}(\boldsymbol{y})$	5	0.042	$1.1{\cdot}10^{+142}$	$> 10^{+143}$
$B_{10}^{I,med}(\boldsymbol{y})$	5	0.051	$1.4{\cdot}10^{+142}$	$> 10^{+143}$

4.5.1 Εισαγωγικοί υπολογισμοί και πρότερη πληροφορία

Η βασική προϋπόθεση του προβλήματος μετασχηματισμού είναι ότι τα μετασχηματισμένα δεδομένα $\boldsymbol{y}^{(\lambda_T)}$, για κάποια κατάλληλη τιμή της παραμέτρου λ_T κάτω από κατάλληλο μετασχηματισμό T, ακολουθούν μια κατανομή $N(\boldsymbol{y}^{(\lambda_T)}|\mu_T, \sigma_T^2)$. Η προεπιλεγμένη πρότερη κατανομή βάσης για την παράμετρο λ_T είναι $\pi^N(\lambda_T|T) \propto 1$. Χρησιμοποιώντας την πρότερη κατανομή Jeffreys $\pi^N(\mu_T, \sigma_T^2|T) \propto \sigma_T^{-2}$ για τις παραμέτρους μ_T, σ_T^2 (την οποία εδώ θεωρούμε σαν κατανομή βάσης για το εν λόγω διάνυσμα παραμέτρων, εξ' ου και ο εκθέτης N), η περιθώρια πιθανοφάνεια των δεδομένων $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ δεδομένου του λ_T και έχοντας ολοκληρώσει εκτός τις παραμέτρους (μ_T, σ_T^2) δίνεται από την ακόλουθη σχέση που εμπεριέχει και μια κατανομή Student:

$$
\begin{aligned}
f^J(\boldsymbol{y}|\lambda_T, T) &= f\left(\boldsymbol{y}^{(\lambda_T)}|T\right) \cdot |J(\boldsymbol{y}, \lambda_T|T)| \\
&= f_{St_n}\left(\boldsymbol{y}\middle| n-1, \overline{\boldsymbol{y}^{(\lambda_T)}} \cdot \mathbf{1}_n, S^2(\boldsymbol{y}) \cdot \left[\mathbf{I}_n + \frac{1}{n}\mathbf{1}_n \cdot \mathbf{1}_n'\right]\right) \cdot |J(\boldsymbol{y}, \lambda_T|T)| \cdot C
\end{aligned}
$$

$$= (2\pi)^{\frac{1-n}{2}} n^{-\frac{1}{2}} \Gamma\left(\frac{n-1}{2}\right) \left(\frac{(n-1)S_T^2}{2}\right)^{-\frac{(n-1)}{2}} \cdot |J(\boldsymbol{y}, \lambda_T|T)| \cdot C$$

όπου $\overline{\boldsymbol{y}^{(\lambda_T)}}$ και S_T^2 είναι ο δειγματικός μέσος και η δειγματική διασπορά αντίστοιχα των μετασχηματισμένων δεδομένων $\boldsymbol{y}^{(\lambda_T)}$, ενώ C είναι η άγνωστη σταθερά που προέρχεται από την πρότερη κατανομή Jeffreys. Κατ' αναλογία, η περιθώρια πιθανοφάνεια των αρχικών μη μετασχηματισμένων δεδομένων υπό τον τετριμμένο μετασχηματισμό Id δίνεται από τη σχέση:

$$f(\boldsymbol{y}|T = \text{Id}) = (2\pi)^{\frac{1-n}{2}} n^{-\frac{1}{2}} \Gamma\left(\frac{n-1}{2}\right) \left(\frac{(n-1)S_{\boldsymbol{y}}^2}{2}\right)^{-\frac{(n-1)}{2}} \cdot C$$

όπου $S_{\boldsymbol{y}}^2$ είναι η δειγματική διακύμανση των αρχικών παρατηρήσεων $\boldsymbol{y}$.

Στο γνωστό άρθρο τους πάνω στους μετασχηματισμούς μεταβλητών, οι Box και Cox (1964) θεώρησαν a priori ανεξαρτησία των παραμέτρων με τη μορφή της προαναφερθείσας πρότερης κατανομής Jeffreys για τις παραμέτρους μ_T, σ_T^2:

$$\pi^N\left(\mu_T, \sigma_T^2|T\right) \propto \frac{1}{\sigma_T^2}.$$

Επιχείρησαν επίσης να συσχετίσουν την αναμενόμενη τιμή των μετασχηματισμένων δεδομένων με την αναμενόμενη τιμή των αρχικών παρατηρήσεων υποθέτοντας ότι η εν λόγω σχέση είναι γραμμική. Έτσι, κατέληξαν στον πολλαπλασιασμό της πρότερης κατανομής βάσης $\pi^N\left(\mu_T, \sigma_T^2|T\right)$ με τον όρο $|J\left(\boldsymbol{y}, \lambda_T|T\right)|^{-\frac{1}{n}}$ που αντιπροσωπεύει μια μορφή κλίσης της συνάρτησης (*gradient*) από τα μετασχηματισμένα στα αρχικά δεδομένα. Αυτή η επιλογή, βέβαια, δεν δικαιολογήθηκε επαρκώς από τους συγγραφείς. Αναλυτικά, η πιθανοφάνεια των αρχικών δεδομένων είναι:

$$\begin{aligned} f\left(\boldsymbol{y}|\mu_T, \sigma_T^2, \lambda_T, T\right) &= f\left(\boldsymbol{y}|\mu_T, \sigma_T^2, \lambda_T, T\right) \\ &= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \cdot f\left(\boldsymbol{y}^{(\lambda_T)}|\mu_T, \sigma_T^2, \lambda_T, T\right) \\ &= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \left(2\pi\sigma_T^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_i \left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\sigma_T^2}\right). \end{aligned}$$

Μετά την ενσωμάτωση της εν λόγω πρότερης κατανομής κατά τους Box & Cox, η περιθώρια πιθανοφάνεια των αρχικών παρατηρήσεων δεδομένης της παραμέτρου λ_T γίνεται:

$f^{BC}\left(\boldsymbol{y}|\lambda_T, T\right)$

$$
\begin{aligned}
&= |J(\boldsymbol{y}, \lambda_T|T)| \cdot f^{BC}\left(\boldsymbol{y}^{(\lambda_T)}|\lambda_T, T\right) \\
&= |J(\boldsymbol{y}, \lambda_T|T)| \iint f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\mu_T, \sigma_T^2, \lambda_T, T\right) \cdot \pi^N\left(\mu_T, \sigma_T^2|T\right) \, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^{1-\frac{1}{n}} \iint (2\pi)^{-\frac{n}{2}} \left(\sigma_T^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_i \left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\sigma_T^2}\right) \frac{C}{\sigma_T^2} \, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{n-1}{n}} \int (2\pi)^{-\frac{n}{2}} \left(\sigma_T^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)S_T^2}{2\sigma_T^2}\right) \frac{1}{\sigma_T^2} \\
&\cdot \left(\int \exp\left(-\frac{n\left(\overline{\boldsymbol{y}^{(\lambda_T)}} - \mu_T\right)^2}{2\sigma_T^2}\right) \mathrm{d}\mu_T\right) \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{n-1}{n}} (2\pi)^{-\frac{n}{2}} (2\pi)^{\frac{1}{2}} n^{-\frac{1}{2}} \int \left(\sigma_T^2\right)^{-\frac{n+2}{2}} \left(\sigma_T^2\right)^{\frac{1}{2}} \exp\left(-\frac{(n-1)S_T^2}{2\sigma_T^2}\right) \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{n-1}{n}} (2\pi)^{-\frac{n-1}{2}} n^{-\frac{1}{2}} \int \left(\sigma_T^2\right)^{-\frac{n-1}{2}-1} \exp\left(-\frac{(n-1)S_T^2}{2\sigma_T^2}\right) \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{n-1}{n}} (2\pi)^{-\frac{n-1}{2}} n^{-\frac{1}{2}} \Gamma\left(\frac{n-1}{2}\right) \left(\frac{(n-1)S_T^2}{2}\right)^{-\frac{(n-1)}{2}}. \qquad (4.6)
\end{aligned}
$$

Στους ανωτέρω υπολογισμούς, οι πυρήνες μιας κανονικής και μιας αντίστροφης γάμμα κατανομής αποδείχτηκαν χρήσιμοι. Η άγνωστη σταθερά C προέρχεται από τη μη γνήσια πρότερη κατανομή.

Από την άλλη πλευρά, ο Pericchi (1981) απέρριψε την ιδέα της a priori ανεξαρτησίας των παραμέτρων ως μη ρεαλιστική και υιοθέτησε ένα ελαφρώς διαφορετικό πλαίσιο πρότερης κατανομής για τις παραμέτρους μ_T, σ_T:

$$\pi^N\left(\mu_T, \sigma_T|T\right) \propto \frac{1}{\sigma_T^2} \Leftrightarrow$$

$$\pi^N\left(\mu_T, \sigma_T^2|T\right) \propto \frac{1}{\left(\sqrt{\sigma_T^2}\right)^2} \cdot \frac{\left(\sigma_T^2\right)^{-1/2}}{2} \propto \frac{1}{\sigma_T^3}.$$

Για τη γενική περίπτωση όπου $\pi^N\left(\mu_T, \sigma_T|T\right) \propto \frac{1}{\sigma_T^{k+1}}$, παίρνουμε $\pi^N\left(\mu_T, \sigma_T^2|T\right) \propto \frac{1}{\sigma_T^{k+2}}$. Η περιθώρια πιθανοφάνεια των αρχικών παρατηρήσεων, δεδομένης της παραμέτρου λ_T, που αντιστοιχεί στην πρότερη προσέγγιση του Pericchi είναι:

$$
\begin{aligned}
&f^P\left(\boldsymbol{y}|\lambda_T, T\right) \\
&= |J(\boldsymbol{y}, \lambda_T|T)| \cdot f^P\left(\boldsymbol{y}^{(\lambda_T)}|\lambda_T, T\right)
\end{aligned}
$$

$$
\begin{aligned}
&= |J(\boldsymbol{y}, \lambda_T|T)| \iint f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\mu_T, \sigma_T^2, T\right) \cdot \pi^N\left(\mu_T, \sigma_T^2|T\right) \, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= |J(\boldsymbol{y}, \lambda_T|T)| \iint (2\pi)^{-\frac{n}{2}} \left(\sigma_T^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_i \left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\sigma_T^2}\right) \frac{C}{\sigma_T^3} \, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)| \, (2\pi)^{-\frac{n}{2}} \int \left(\sigma_T^2\right)^{-\frac{n+3}{2}} \exp\left(-\frac{(n-1)S_T^2}{2\sigma_T^2}\right) \\
&\cdot \left(\int \exp\left(-\frac{n\left(\overline{\boldsymbol{y}^{(\lambda_T)}} - \mu_T\right)^2}{2\sigma_T^2}\right) \mathrm{d}\mu_T\right) \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)| \, (2\pi)^{-\frac{n-1}{2}} \, n^{-\frac{1}{2}} \int \left(\sigma_T^2\right)^{-\frac{n+3}{2}} \left(\sigma_T^2\right)^{\frac{1}{2}} \exp\left(-\frac{(n-1)S_T^2}{2\sigma_T^2}\right) \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)| \, (2\pi)^{-\frac{n-1}{2}} \, n^{-\frac{1}{2}} \int \left(\sigma_T^2\right)^{-\frac{n}{2}-1} \exp\left(-\frac{(n-1)S_T^2}{2\sigma_T^2}\right) \mathrm{d}\sigma_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)| \, (2\pi)^{-\frac{n-1}{2}} \, n^{-\frac{1}{2}} \Gamma\left(\frac{n}{2}\right) \left(\frac{(n-1)S_T^2}{2}\right)^{-\frac{n}{2}}. \qquad (4.7)
\end{aligned}
$$

Ας σημειωθεί ότι η διαφορά της (4.7) από την (4.6) έγκειται στο ότι ο Ιακωβιανός όρος δεν είναι πλέον υψωμένος σε καμία δύναμη (πέραν της μονάδας προφανώς) και ότι ο όρος της γάμμα συνάρτησης καθώς και ο εκθέτης του S_T^2 όρου περιέχουν $-\frac{n}{2}$ αντί για $-\frac{n-1}{2}$.

4.5.2 Ο κλασματικός παράγοντας Bayes για το πρόβλημα της επιλογής μετασχηματισμού

Για τον κλασματικό παράγοντα Bayes, χρειάζεται επίσης να υπολογισθεί η περιθώρια πιθανοφάνεια $f\left(\boldsymbol{y}|\lambda_T, b, T\right)$ δεδομένου ότι η αρχική πιθανοφάνεια υψώνεται στη δύναμη $b = \frac{m}{n}$, για την οποία συχνή επιλογή αποτελεί το μέγεθος του ελάχιστου δείγματος εκπαίδευσης $m = m_0$. Για το μονομεταβλητό πρόβλημα μετασχηματισμού που διερευνούμε σε αυτό το σημείο, έχουμε $b = 3/n$. Άλλες επιλογές για το m που προτείνονται στα άρθρα του O'Hagan (1995) και του O'Hagan (1997) είναι οι εξής:

$\log n$	$\sqrt{n}$	$\frac{m_0 \log n}{\log m_0}$	$\sqrt{m_0 n}$

Με βάση τον O'Hagan (1995), η εξίσωση για τη λ_T-περιθώρια πιθανοφάνεια που αντιστοιχεί στην πιθανοφάνεια υψωμένη στη δύναμη b (για την απλή περίπτωση όπου δεν

συμπεριλαμβάνονται επεξηγηματικές μεταβλητές), και με βάση την πρότερη κατανομή Jeffreys για το διάνυσμα (μ_T, σ_T^2), είναι:

$$f\left(\boldsymbol{y}|\lambda_T, b, T\right) \propto \left(\sum_{i=1}^{n}\left(y_i^{(\lambda_T)} - \overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2\right)^{-\frac{(nb-1)}{2}} \cdot |J(\boldsymbol{y}, \lambda_T|T)|^b.$$

Ως ακολούθως, αποδεικνύουμε αναλυτικά τα άνωθεν αποτελέσματα ενσωματώνοντας σε ένα πρώτο κομμάτι την προσέγγιση κατά Box & Cox και ύστερα την προσέγγιση κατά Pericchi, όπως αυτές παρουσιάστηκαν στην Ενότητα 4.5.1, αναφορικά με την πρότερη κατανομή για το διάνυσμα (μ_T, σ_T^2). Η πιθανοφάνεια των δεδομένων υψωμένη στη δύναμη b είναι εμφανώς η ίδια και για τις δύο προσεγγίσεις:

$$\begin{aligned}
f\left(\boldsymbol{y}|\mu_T, \sigma_T^2, \lambda_T, b, T\right) &= f\left(\boldsymbol{y}|\mu_T, \sigma_T^2, \lambda_T, T\right)^b \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^b \cdot f\left(\boldsymbol{y}^{(\lambda_T)}|\mu_T, \sigma_T^2, \lambda_T, T\right)^b \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^b \left(2\pi\sigma_T^2\right)^{-\frac{nb}{2}} \exp\left(-\frac{\sum_i\left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\sigma_T^2/b}\right) \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^b \left(2\pi\sigma_T^2\right)^{-\frac{nb}{2}} \exp\left(-\frac{\sum_i\left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right) \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^b (2\pi)^{-\frac{nb}{2}} \left(b\cdot\tilde{\sigma}_T^2\right)^{-\frac{nb}{2}} \exp\left(-\frac{\sum_i\left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right)
\end{aligned}$$

όπου $\tilde{\sigma}_T^2 = \frac{\sigma_T^2}{b}$.

Η περιθώρια πιθανοφάνεια των αρχικών μη μετασχηματισμένων παρατηρήσεων δεδομένης της παραμέτρου λ_T που αντιστοιχεί στη πιθανοφάνεια δύναμης και στην πρότερη προσέγγιση κατά Box & Cox είναι:

$$\begin{aligned}
&f^{BC}\left(\boldsymbol{y}|\lambda_T, b, T\right) \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^b \cdot f^{BC}\left(\boldsymbol{y}^{(\lambda_T)}|b, \lambda_T, T\right) \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^b \iint f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\mu_T, \sigma_T^2, b, T\right) \cdot \pi^N\left(\mu_T, \sigma_T^2|T\right)\, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^{b-\frac{1}{n}} \iint (2\pi)^{-\frac{nb}{2}} \left(b\cdot\tilde{\sigma}_T^2\right)^{-\frac{nb}{2}} \exp\left(-\frac{\sum_i\left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right) \frac{C}{\sigma_T^2}\, \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= C\cdot|J(\boldsymbol{y}, \lambda_T|T)|^{\frac{nb-1}{n}} \iint (2\pi)^{-\frac{nb}{2}} \frac{1}{b} \left(b\cdot\tilde{\sigma}_T^2\right)^{-\frac{nb}{2}} \exp\left(-\frac{\sum_i\left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right)
\end{aligned}$$

$$
\begin{aligned}
&\cdot \frac{1}{\tilde{\sigma}_T^2}\,\mathrm{d}\mu_T \cdot b \cdot \mathrm{d}\tilde{\sigma}_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{nb-1}{n}} (2\pi)^{-\frac{nb}{2}} (b)^{-\frac{nb}{2}} \int \left(\tilde{\sigma}_T^2\right)^{-\frac{nb+2}{2}} \exp\left(-\frac{\sum_i \left(y_i^{(\lambda_T)} - \overline{\boldsymbol{y}^{(\lambda_T)}}\right)^2}{2\tilde{\sigma}_T^2}\right) \\
&\cdot \left(\int \exp\left(-\frac{n\left(\overline{y^{(\lambda_T)}} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right) \mathrm{d}\mu_T\right) \mathrm{d}\tilde{\sigma}_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{nb-1}{n}} \cdot (2\pi)^{-\frac{nb}{2}} b^{-\frac{nb}{2}} (2\pi)^{\frac{1}{2}} n^{-\frac{1}{2}} \int \left(\tilde{\sigma}_T^2\right)^{-\frac{nb+2}{2}} \left(\tilde{\sigma}_T^2\right)^{\frac{1}{2}} \exp\left(-\frac{(n-1)S_T^2}{2\tilde{\sigma}_T^2}\right) \mathrm{d}\tilde{\sigma}_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{nb-1}{n}} \cdot (2\pi)^{-\frac{nb-1}{2}} b^{-\frac{nb}{2}} n^{-\frac{1}{2}} \int \left(\tilde{\sigma}_T^2\right)^{-\frac{nb-1}{2}-1} \exp\left(-\frac{(n-1)S_T^2}{2\tilde{\sigma}_T^2}\right) \mathrm{d}\tilde{\sigma}_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{\frac{nb-1}{n}} \cdot (2\pi)^{-\frac{nb-1}{2}} b^{-\frac{nb}{2}} n^{-\frac{1}{2}} \Gamma\left(\frac{nb-1}{2}\right) \left(\frac{(n-1)S_T^2}{2}\right)^{-\frac{nb-1}{2}}. \qquad (4.8)
\end{aligned}
$$

Παρατηρήστε ότι η παράμετρος b δε μπορεί να πάρει την τιμή $\frac{1}{n}$ εφόσον ο όρος $(nb-1)$, που είναι παρών στον αριθμητή των δύο όρων δύναμης της περιθώριας πιθανοφάνειας, θα μηδενιστεί.

Ομοίως, η περιθώρια πιθανοφάνεια των αρχικών παρατηρήσεων δεδομένου του λ_T που αντιστοιχεί στη πιθανοφάνεια δύναμης και στην πρότερη προσέγγιση κατά Pericchi είναι:

$$
\begin{aligned}
&f^P\left(\boldsymbol{y}|\lambda_T, b, T\right) \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^b \cdot f^P\left(\boldsymbol{y}^{(\lambda_T)}|\lambda_T, b, T\right) \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^b \iint f\left(\boldsymbol{y}^{(\lambda_T)}\middle|\mu_T, \sigma_T^2, b, T\right) \cdot \pi^N\left(\mu_T, \sigma_T^2|T\right) \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^b \iint (2\pi)^{-\frac{nb}{2}} \left(b\tilde{\sigma}_T^2\right)^{-\frac{nb}{2}} \exp\left(-\frac{\sum_i \left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right) \frac{C}{\sigma_T^3} \mathrm{d}\mu_T \mathrm{d}\sigma_T^2 \\
&= |J(\boldsymbol{y}, \lambda_T|T)|^b \iint (2\pi)^{-\frac{nb}{2}} b^{-\frac{nb+3}{2}} \left(\tilde{\sigma}_T^2\right)^{-\frac{nb+3}{2}} \exp\left(-\frac{\sum_i \left(y_i^{(\lambda_T)} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right) \mathrm{d}\mu_T \cdot b \cdot \mathrm{d}\tilde{\sigma}_T^2 \\
&= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^b (2\pi)^{-\frac{nb}{2}} b^{-\frac{nb+1}{2}} \int \left(\tilde{\sigma}_T^2\right)^{-\frac{nb+3}{2}} \exp\left(-\frac{(n-1)S_T^2}{2\tilde{\sigma}_T^2}\right) \\
&\cdot \left(\int \exp\left(-\frac{n\left(\overline{\boldsymbol{y}^{(\lambda_T)}} - \mu_T\right)^2}{2\tilde{\sigma}_T^2}\right) \mathrm{d}\mu_T\right) \mathrm{d}\tilde{\sigma}_T^2
\end{aligned}
$$

$$= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^b \, (2\pi)^{-\frac{nb-1}{2}} \, b^{-\frac{nb+1}{2}} n^{-\frac{1}{2}} \int \left(\tilde{\sigma}_T^2\right)^{-\frac{nb+3}{2}} \left(\tilde{\sigma}_T^2\right)^{\frac{1}{2}} \exp\left(-\frac{(n-1)S_T^2}{2\tilde{\sigma}_T^2}\right) \mathrm{d}\tilde{\sigma}_T^2$$

$$= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^b \, (2\pi)^{-\frac{nb-1}{2}} \, b^{-\frac{nb+1}{2}} n^{-\frac{1}{2}} \int \left(\tilde{\sigma}_T^2\right)^{-\frac{nb}{2}-1} \exp\left(-\frac{(n-1)S_T^2}{2\tilde{\sigma}_T^2}\right) \mathrm{d}\tilde{\sigma}_T^2$$

$$= C \cdot |J(\boldsymbol{y}, \lambda_T|T)|^b \, (2\pi)^{-\frac{nb-1}{2}} \, b^{-\frac{nb+1}{2}} n^{-\frac{1}{2}} \Gamma\left(\frac{nb}{2}\right) \left(\frac{(n-1)S_T^2}{2}\right)^{-\frac{nb}{2}}. \tag{4.9}$$

Όσον αφορά τον μοντελοχώρο $\mathcal{T}$ για το συγκεκριμένο πρόβλημα, ορίζεται όπως στη (2.2). Άρα, η διάστασή του ισούται με $|\mathcal{T}| = 6$ και μειώνεται υποχρεωτικά σε $|\mathcal{T}| = 5$ για προβλήματα αυστηρά θετικών δεδομένων, δηλαδή όταν δεν απαιτείται μετατόπιση και ο μετασχηματισμός Modulus είναι ισοδύναμος με τον μετασχηματισμό YJ. Συμβολίζοντας την προεπιλεγμένη πρότερη κατανομή του μοντέλου M_i με $\pi_i = \frac{1}{|\mathcal{T}|}$, η ύστερη πιθανότητα του μοντέλου M_i δίνεται από τη γενική σχέση που ακολουθεί:

$$\begin{aligned} P\left(M_i|\boldsymbol{y}\right) &= \left(\sum_{j=1}^{|\mathcal{T}|} \frac{\pi_j}{\pi_i} B_{ji}^F(\boldsymbol{y})\right)^{-1} \\ &= \left(1 + \sum_{\substack{j=1 \\ j\neq i}}^{|\mathcal{T}|} B_{ji}^F(\boldsymbol{y})\right)^{-1}, \end{aligned} \tag{4.10}$$

όπου ο FBF μπορεί να αντικατασταθεί από οποιαδήποτε άλλη μορφή παράγοντα Bayes, όπως ο ενδογενής παράγοντας Bayes.

Ο κλασματικός παράγοντας Bayes $B_{ij}^F(\boldsymbol{y})$ μεταξύ δύο μοντέλων M_i, M_j υπολογίζεται ως εξής:

$$\begin{aligned} B_{ij}^F(\boldsymbol{y}) &= B_{ij}(\boldsymbol{y}) \cdot B_{ji}^b(\boldsymbol{y}) \\ &= \frac{m_i(\boldsymbol{y})}{m_j(\boldsymbol{y})} \cdot \frac{m_j(\boldsymbol{y}, b)}{m_i(\boldsymbol{y}, b)} \\ &= \frac{\int f_i(\boldsymbol{y}|\lambda_T, T)\pi_i^N(\lambda_T|T)\mathrm{d}\lambda_T}{\int f_j(\boldsymbol{y}|\lambda_T, T)\pi_j^N(\lambda_T|T)\mathrm{d}\lambda_T} \cdot \frac{\int f_j(\boldsymbol{y}|\lambda_T, b, T)\pi_j^N(\lambda_T|T)\mathrm{d}\lambda_T}{\int f_i(\boldsymbol{y}|\lambda_T, b, T)\pi_i^N(\lambda_T|T)\mathrm{d}\lambda_T} \end{aligned}$$

καταλήγοντας στο πηλίκο των περιθώριων πιθανοφανειών $m_i(\boldsymbol{y}), m_j(\boldsymbol{y})$ κάτω από τα μοντέλα M_i, M_j επί το πηλίκο των περιθώριων πιθανοφανειών δύναμης $m_j(\boldsymbol{y}, b), m_i(\boldsymbol{y}, b)$ υπό τα μοντέλα M_j, M_i. Η προεπιλεγμένη πρότερη κατανομή βάσης για την παράμετρο λ_T είναι

$$\pi_i^N(\lambda_T|T) \propto 1$$

για όλα τα μοντέλα στο σύνολο $\mathcal{T}$.

Εξαιτίας του γεγονότος ότι οι αλγόριθμοι που έχουμε υλοποιήσει για τις εν λόγω διαδικασίες υπολογίζουν τον λογάριθμο των σχετικών περιθώριων πιθανοφανειών για το μοντέλο M_i, συμβολιζόμενων ως $\log m_i(\boldsymbol{y})$ και $\log m_i(\boldsymbol{y}, b)$, χρησιμοποιούμε την εξής σχέση για τον προσδιορισμό των τιμών του FBF:

$$B_{ij}^F(\boldsymbol{y}) = \exp\left(\log m_i(\boldsymbol{y}) - \log m_j(\boldsymbol{y})\right) \cdot \exp\left(\log m_j(\boldsymbol{y}, b) - \log m_i(\boldsymbol{y}, b)\right).$$

4.5.3 Ο ενδογενής παράγοντας Bayes για το πρόβλημα επιλογής μετασχηματισμού

Ο αριθμητικός ενδογενής παράγοντας Bayes $B_{ij}^{I,a}(\boldsymbol{y})$ μεταξύ δύο μοντέλων M_i, M_j υπολογίζεται ως εξής, συμβολίζοντας με $|\mathcal{Y}_{m_0}|$ το σύνολο όλων των δυνατών δειγμάτων εκπαίδευσης μεγέθους m_0:

$$\begin{aligned} B_{ij}^{I,a}(\boldsymbol{y}) &= B_{ij}(\boldsymbol{y}) \cdot \frac{1}{|\mathcal{Y}_{m_0}|} \sum_{l=1}^{|\mathcal{Y}_{m_0}|} B_{ji}(\boldsymbol{y}_{m_0 l}) \\ &= \frac{m_i(\boldsymbol{y})}{m_j(\boldsymbol{y})} \cdot \frac{1}{|\mathcal{Y}_{m_0}|} \sum_{l=1}^{|\mathcal{Y}_{m_0}|} \frac{m_j(\boldsymbol{y}_{m_0 l})}{m_i(\boldsymbol{y}_{m_0 l})} \\ &= \frac{\int f_i(\boldsymbol{y}|\lambda_T, T)\pi_i^N(\lambda_T|T)\mathrm{d}\lambda_T}{\int f_j(\boldsymbol{y}|\lambda_T, T)\pi_j^N(\lambda_T|T)\mathrm{d}\lambda_T} \cdot \frac{1}{|\mathcal{Y}_{m_0}|} \sum_{l=1}^{|\mathcal{Y}_{m_0}|} \frac{\int f_j(\boldsymbol{y}_{m_0 l}|\lambda_T, T)\pi_j^N(\lambda_T|T)\mathrm{d}\lambda_T}{\int f_i(\boldsymbol{y}_{m_0 l}|\lambda_T, T)\pi_i^N(\lambda_T|T)\mathrm{d}\lambda_T} \end{aligned}$$

καταλήγοντας στο πηλίκο των περιθώριων πιθανοφανειών του συνολικού δείγματος $\boldsymbol{y}$ υπό τα μοντέλα M_i, M_j επί το πηλίκο των περιθώριων πιθανοφανειών του δείγματος εκπαίδευσης $\boldsymbol{y}_{m_0 l}$ υπό τα μοντέλα M_j, M_i. Όπως και στην περίπτωση του FBF προηγουμένως, η προεπιλεγμένη πρότερη κατανομή βάσης για το λ_T είναι $\pi_i^N(\lambda_T|T) = \pi^N(\lambda_T|T) \propto 1$ για όλα τα μοντέλα στο σύνολο $\mathcal{T}$.

Ο υπολογισμός του γεωμετρικού ενδογενούς παράγοντα Bayes $B_{ij}^{I,g}(\boldsymbol{y})$ είναι άμεσος:

$$B_{ij}^{I,g}(\boldsymbol{y}) = B_{ij}(\boldsymbol{y}) \cdot \left(\prod_{l=1}^{|\mathcal{Y}_{m_0}|} B_{ji}(\boldsymbol{y}_{m_0 l}) \right)^{\frac{1}{|\mathcal{Y}_{m_0}|}}.$$

Το ελάχιστο δείγμα εκπαίδευσης έχει μέγεθος $m = 3$ για μονομεταβλητά δεδομένα. Στις εφαρμογές του Κεφαλαίου 5 που θα ακολουθήσουν, χρησιμοποιούμε 1000 διαφορετικά διανύσματα εκπαίδευσης $\boldsymbol{y}_1$, δηλαδή 1000 μοναδικούς συνδυασμούς στοιχείων

του διανύσματος $\boldsymbol{y}$. Οι ύστερες πιθανότητες των μοντέλων υπολογίζονται από μια σχέση ανάλογη με την (4.10) προσαρμοσμένη για τον IBF αντί για τον FBF.

Όπως ήδη αναφέρθηκε, οι αλγόριθμοι που υλοποιήσαμε υπολογίζουν τον λογάριθμο των περιθώριων πιθανοφανειών, συμβολιζόμενων ως $\log m_i(\boldsymbol{y})$ και $\log m_i(\boldsymbol{y}_{m_0})$ για το συνολικό δείγμα και το δείγμα εκπαίδευσης αντίστοιχα, και έτσι χρησιμοποιούμε τις εξής σχέσεις για τον IBF:

$$B_{ij}^{I,a}(\boldsymbol{y}) = \exp\left(\log m_i(\boldsymbol{y}) - \log m_j(\boldsymbol{y})\right) \cdot \frac{1}{|\mathcal{Y}_{m_0}|} \sum_{l=1}^{|\mathcal{Y}_{m_0}|} \exp\left(\log m_j(\boldsymbol{y}_{m_0 l}) - \log m_i(\boldsymbol{y}_{m_0 l})\right)$$

$$B_{ij}^{I,g}(\boldsymbol{y}) = \exp\left(\log m_i(\boldsymbol{y}) - \log m_j(\boldsymbol{y})\right) \cdot \left(\prod_{l=1}^{|\mathcal{Y}_{m_0}|} \exp\left(\log m_j(\boldsymbol{y}_{m_0 l}) - \log m_i(\boldsymbol{y}_{m_0 l})\right) \right)^{\frac{1}{|\mathcal{Y}_{m_0}|}}.$$

4.6 Ενσωμάτωση Συμμεταβλητών στο Μοντέλο

Παρουσία επεξηγηματικών μεταβλητών στο μοντέλο, υποθέτουμε ότι τα μετασχηματισμένα δεδομένα της μεταβλητής απόκρισης για κάποια τιμή της παραμέτρου λ_T και κάποια οικογένεια T προέρχονται από:

$$\boldsymbol{y}^{(\lambda_T)} | \boldsymbol{\beta}_T, \sigma_T^2, \mathbf{X}_T, \lambda_T, T \sim N_n\left(\mathbf{X}_T \boldsymbol{\beta}_T, \sigma_T^2 \mathbf{I}_n\right),$$

όπου $\mathbf{X}_T$ είναι ένας πίνακας σχεδιασμού διάστασης $n \times (p+1)$, $\boldsymbol{\beta}_T$ είναι ένα διάνυσμα $(p+1) \times 1$ και $\boldsymbol{y}^{(\lambda_T)}$ είναι ένα διάνυσμα $n \times 1$ όπως και πριν. Στη συνέχεια, θα προσαρμόσουμε τις σχέσεις (4.9) και (4.8) ώστε να ενσωματωθεί η πληροφορία που παρέχουν οι επεξηγηματικές μεταβλητές. Προς αυτόν τον σκοπό, οι επόμενες αρκετά γνωστές εκφράσεις θα φανούν χρήσιμες:

$$\hat{\boldsymbol{\beta}_T} = \left(\mathbf{X}_T^\top \mathbf{X}_T\right)^{-1} \mathbf{X}_T \cdot \boldsymbol{y}^{(\lambda_T)},$$

$$RSS_T = \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \left(\mathbf{I}_n - \mathbf{X}_T \left(\mathbf{X}_T^\top \mathbf{X}_T\right)^{-1} \mathbf{X}_T\right) \boldsymbol{y}^{(\lambda_T)} = \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}_T}.$$

Ας ξεκινήσουμε από τη διαμόρφωση της πιθανοφάνειας των μη μετασχηματισμένων δεδομένων:

$$\begin{aligned} & f\left(\boldsymbol{y} | \boldsymbol{\beta}_T, \sigma_T^2, \mathbf{X}_T, \lambda_T, T\right) \\ & = \left| J\left(\boldsymbol{y}, \lambda_T | T\right) \right| f\left(\boldsymbol{y}^{(\lambda_T)} | \boldsymbol{\beta}_T, \sigma_T^2, \mathbf{X}_T, \lambda_T, T\right) \end{aligned}$$

$$= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \left(2\pi\sigma_T^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_T^2}\left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)^\top \left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)\right\}$$

όπου η έκφραση $\left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)^\top \left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)$ αναλύεται περαιτέρω:

$$\begin{aligned}
&\left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)^\top \left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right) \\
&= \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \mathbf{X}_T\boldsymbol{\beta}_T - \boldsymbol{\beta}_T^\top\mathbf{X}_T^\top\boldsymbol{y}^{(\lambda_T)} + \boldsymbol{\beta}_T^\top\mathbf{X}_T^\top\mathbf{X}_T\boldsymbol{\beta}_T \\
&= \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - 2\boldsymbol{\beta}_T^\top\mathbf{X}_T^\top\boldsymbol{y}^{(\lambda_T)} + \boldsymbol{\beta}_T^\top\mathbf{X}_T^\top\mathbf{X}_T\boldsymbol{\beta}_T \\
&= \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - 2\boldsymbol{\beta}_T^\top \left(\mathbf{X}_T{}^\top\mathbf{X}_T\right)\left(\mathbf{X}_T{}^\top\mathbf{X}_T\right)^{-1}\mathbf{X}_T^\top\boldsymbol{y}^{(\lambda_T)} + \boldsymbol{\beta}_T^\top\mathbf{X}_T^\top\mathbf{X}_T\boldsymbol{\beta}_T \pm \hat{\boldsymbol{\beta}}_T{}^\top\mathbf{X}_T^\top\mathbf{X}_T\hat{\boldsymbol{\beta}}_T \\
&= \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} + \boldsymbol{\beta}_T^\top\mathbf{X}_T^\top\mathbf{X}_T\boldsymbol{\beta}_T - 2\boldsymbol{\beta}_T^\top\mathbf{X}_T{}^\top\mathbf{X}_T\hat{\boldsymbol{\beta}}_T \pm \hat{\boldsymbol{\beta}}_T{}^\top\mathbf{X}_T^\top\mathbf{X}_T\hat{\boldsymbol{\beta}}_T \\
&= \left\{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}}_T{}^\top\mathbf{X}_T^\top\mathbf{X}_T\hat{\boldsymbol{\beta}}_T\right\} + \left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right)^\top \mathbf{X}_T{}^\top\mathbf{X}_T\left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right).
\end{aligned}$$

Τώρα, χρησιμοποιώντας την πρότερη κατανομή κατά Box & Cox της μορφής

$$\pi\left(\boldsymbol{\beta}_T, \sigma_T^2|T\right) \propto \frac{1}{\sigma_T^2}|J\left(\boldsymbol{y}, \lambda_T|T\right)|^{-\frac{p+1}{n}},$$

περιθωριοποιούμε μερικώς την πιθανοφάνεια $f\left(\boldsymbol{y}|\boldsymbol{\beta}_T, \sigma_T^2, \mathbf{X}_T, \lambda_T, T\right)$ ολοκληρώνοντας εκτός τις παραμέτρους $\boldsymbol{\beta}_T$ και σ_T^2:

$$\begin{aligned}
&f^{BC}\left(\boldsymbol{y}|\mathbf{X}_T, \lambda_T, T\right) \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} \int\int f\left(\boldsymbol{y}|\boldsymbol{\beta}_T, \sigma_T^2, \mathbf{X}_T, \lambda_T, T\right)\cdot\sigma_T^{-2}\mathrm{d}\boldsymbol{\beta}_T\mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} \int\int \frac{1}{\sigma_T^2}\left(2\pi\sigma_T^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_T^2}\left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)^\top \left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)\right\} \\
&\quad \mathrm{d}\boldsymbol{\beta}_T\mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} (2\pi)^{-\frac{n}{2}} \int \left(\sigma_T^2\right)^{-\frac{n+2}{2}} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}}_T{}^\top\mathbf{X}_T^\top\mathbf{X}_T\hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\} \\
&\quad \cdot \left(\int \exp\left\{-\frac{\left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right)^\top \mathbf{X}_T{}^\top\mathbf{X}_T\left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right)}{2\sigma_T^2}\right\}\mathrm{d}\boldsymbol{\beta}_T\right)\mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} (2\pi)^{-\frac{n}{2}} \int \left(\sigma_T^2\right)^{-\frac{n+2}{2}} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}}_T{}^\top\mathbf{X}_T^\top\mathbf{X}_T\hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\} \\
&\quad \cdot \left(2\pi\sigma_T^2\right)^{\frac{p+1}{2}} |\mathbf{X}_T{}^\top\mathbf{X}_T|^{-\frac{1}{2}}\mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} (2\pi)^{-\frac{n-p-1}{2}} |\mathbf{X}_T{}^\top\mathbf{X}_T|^{-\frac{1}{2}} \\
&\quad \cdot \int \left(\sigma_T^2\right)^{-\frac{n-p+1}{2}} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}}_T{}^\top\mathbf{X}_T^\top\mathbf{X}_T\hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\}\mathrm{d}\sigma_T^2
\end{aligned}$$

$$
\begin{aligned}
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} (2\pi)^{-\frac{n-p-1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \\
&\quad \cdot \int \left(\sigma_T^2\right)^{-\frac{n-p-1}{2}-1} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\} \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} (2\pi)^{-\frac{n-p-1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \Gamma\left(\frac{n-p-1}{2}\right) \\
&\quad \cdot \left(\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T}{2}\right)^{-\frac{n-p-1}{2}} \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{1-\frac{p+1}{n}} \pi^{-\frac{n-p-1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \Gamma\left(\frac{n-p-1}{2}\right) {RSS_T}^{-\frac{n-p-1}{2}}. \qquad (4.11)
\end{aligned}
$$

Θεωρώντας τη μεθοδολογία του Pericchi, υποθέτουμε την εξαρτημένη πρότερη κατανομή Jeffreys της μορφής $\pi\left(\boldsymbol{\beta}_T, \sigma_T^2|T\right) \propto \left(\sigma_T^2\right)^{-\frac{(p+1)+2}{2}} = \left(\sigma_T^2\right)^{-\frac{p+3}{2}}$. Ακολουθεί το αποτέλεσμα της ολοκλήρωσης των παραμέτρων $\boldsymbol{\beta}_T$ και σ_T^2:

$$
\begin{aligned}
&f^P\left(\boldsymbol{y}|\mathbf{X}_T, \lambda_T, T\right) \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \int\int f\left(\boldsymbol{y}|\boldsymbol{\beta}_T, \sigma_T^2, \mathbf{X}_T, \lambda_T, T\right) \cdot \sigma_T^{-(p+3)} \mathrm{d}\boldsymbol{\beta}_T \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \int\int \frac{\exp\left\{-\frac{1}{2\sigma_T^2}\left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)^\top \left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)\right\}}{\left(\sigma_T^2\right)^{\frac{p+3}{2}} \left(2\pi\sigma_T^2\right)^{\frac{n}{2}}} \mathrm{d}\boldsymbol{\beta}_T \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \left(2\pi\right)^{-\frac{n}{2}} \int \left(\sigma_T^2\right)^{-\frac{n+p+3}{2}} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\} \\
&\quad \cdot \left(\int \exp\left\{-\frac{\left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right)^\top {\mathbf{X}_T}^\top \mathbf{X}_T \left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right)}{2\sigma_T^2}\right\} \mathrm{d}\boldsymbol{\beta}_T\right) \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \left(2\pi\right)^{-\frac{n}{2}} \int \left(\sigma_T^2\right)^{-\frac{n+p+3}{2}} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\} \\
&\quad \cdot \left(2\pi\sigma_T^2\right)^{\frac{p+1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \left(2\pi\right)^{-\frac{n-p-1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \\
&\quad \cdot \int \left(\sigma_T^2\right)^{-\frac{n+2}{2}} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\} \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \left(2\pi\right)^{-\frac{n-p-1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \\
&\quad \cdot \int \left(\sigma_T^2\right)^{-\frac{n}{2}-1} \exp\left\{-\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T}{2\sigma_T^2}\right\} \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \left(2\pi\right)^{-\frac{n-p-1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \Gamma\left(\frac{n}{2}\right) \left(\frac{\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T}{2}\right)^{-\frac{n}{2}}
\end{aligned}
$$

$$= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \pi^{-\frac{n-p-1}{2}} 2^{\frac{p+1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \Gamma\left(\frac{n}{2}\right) {RSS_T}^{-\frac{n}{2}}. \tag{4.12}$$

Βασιζόμενοι είτε στην (4.11) είτε στην (4.12), η διαμόρφωση του IBF είναι άμεση δεδομένης της διαδικασίας που περιγράφτηκε στην Ενότητα 4.5.3 προσαρμοσμένη για την παρουσία επεξηγηματικών μεταβλητών. Από την άλλη πλευρά, ο υπολογισμός του κλασματικού παράγοντα Bayes απαιτεί τον προσδιορισμό της ποσότητας $f\left(\boldsymbol{y}|\mathbf{X}_T, \lambda_T, b, T\right)$ δεδομένου ότι η πιθανοφάνεια υψώνεται στη δύναμη $b = \frac{m}{n}$. Θα παρουσιάσουμε τους υπολογισμούς για τη γενική περίπτωση όπου η πρότερη κατανομή των παραμέτρων παίρνει την ακόλουθη μορφή:

$$\pi\left(\boldsymbol{\beta}_T, \sigma_T^2\right) \propto C(t) \cdot \left(\sigma_T^2\right)^{-t},$$

όπου $C(t)$ είναι μια ποσότητα που σχετίζεται με την εκάστοτε πρότερη προσέγγιση που επιλέγεται:

$$C(t) = \begin{cases} 1, & \text{για την προσέγγιση κατά Pericchi όπου } t = 1 \\ |J(\boldsymbol{y}, \lambda_T|T)|^{-\frac{p+1}{n}}, & \text{για την προσέγγιση κατά Box \& Cox όπου } t = \frac{p+3}{2} \end{cases} \tag{4.13}$$

και η παράμετρος t επιλέγεται κατάλληλα όπως φαίνεται. Η γενική (*generic*) διαδικασία ολοκλήρωσης των παραμέτρων $(\boldsymbol{\beta}_T, \sigma_T^2)$ για τυχαία τιμή της παραμέτρου t είναι η εξής:

$$\begin{aligned}
&f^G\left(\boldsymbol{y}|\mathbf{X}_T, \lambda_T, b, T\right) \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| \int\int f\left(\boldsymbol{y}|\boldsymbol{\beta}_T, \sigma_T^2, \mathbf{X}_T, \lambda_T, b, T\right) \cdot C(t) \left(\sigma_T^2\right)^{-t} \mathrm{d}\boldsymbol{\beta}_T \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| C(t) \int\int \frac{\left(\sigma_T^2\right)^{-t} \left(2\pi\sigma_T^2\right)^{-\frac{nb}{2}}}{\exp\left\{\frac{b\left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)^\top \left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_T\boldsymbol{\beta}_T\right)}{2\sigma_T^2}\right\}} \mathrm{d}\boldsymbol{\beta}_T \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| C(t) \left(2\pi\right)^{-\frac{nb}{2}} \int \left(\sigma_T^2\right)^{-\frac{nb+2t}{2}} \exp\left\{-\frac{b\left(\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T\right)}{2\sigma_T^2}\right\} \\
&\quad \cdot \left(\int \exp\left\{-\frac{b\left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right)^\top {\mathbf{X}_T}^\top \mathbf{X}_T \left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right)}{2\sigma_T^2}\right\} \mathrm{d}\boldsymbol{\beta}_T\right) \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| C(t) \left(2\pi\right)^{-\frac{nb}{2}} \int \left(\sigma_T^2\right)^{-\frac{nb+2t}{2}} \exp\left\{-\frac{b\left(\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - {\hat{\boldsymbol{\beta}}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}}_T\right)}{2\sigma_T^2}\right\} \\
&\quad \cdot \left(2\pi\sigma_T^2\right)^{\frac{p+1}{2}} b^{-\frac{p+1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| C(t) \left(2\pi\right)^{-\frac{nb-p-1}{2}} b^{-\frac{p+1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&\cdot \int \left(\sigma_T^2\right)^{-\frac{nb+2t-p-1}{2}} \exp\left\{-\frac{b\left(\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}_T}\right)}{2\sigma_T^2}\right\} \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| C(t) \left(2\pi\right)^{-\frac{nb-p-1}{2}} b^{-\frac{p+1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \\
&\cdot \int \left(\sigma_T^2\right)^{-\frac{nb+2t-p-3}{2}-1} \exp\left\{-\frac{b\left(\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}_T}\right)}{2\sigma_T^2}\right\} \mathrm{d}\sigma_T^2 \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| C(t) \left(2\pi\right)^{-\frac{nb-p-1}{2}} b^{-\frac{p+1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \Gamma\left(\frac{nb+2t-p-3}{2}\right) \\
&\cdot \left(\frac{b\left(\left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \boldsymbol{y}^{(\lambda_T)} - \hat{\boldsymbol{\beta}_T}^\top \mathbf{X}_T^\top \mathbf{X}_T \hat{\boldsymbol{\beta}_T}\right)}{2}\right)^{-\frac{nb+2t-p-3}{2}} \\
&= |J\left(\boldsymbol{y}, \lambda_T|T\right)| C(t) \frac{\pi^{-\frac{nb-p-1}{2}} 2^{t-1}}{b^{\frac{nb+2t-2}{2}}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{-\frac{1}{2}} \Gamma\left(\frac{nb+2t-p-3}{2}\right) {RSS_T}^{-\frac{nb+2t-p-3}{2}}.
\end{aligned}
\tag{4.14}
$$

Κατά συνέπεια, θεωρώντας την προσέγγιση κατά Box & Cox με $t = 1$ και $C(t)$ όπως στην (4.13), η (4.14) γίνεται:

$$
f^{BC}\left(\boldsymbol{y}|\mathbf{X}_T, \lambda_T, b, T\right) = |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{\frac{nb-p-1}{n}} \frac{\pi^{-\frac{nb-p-1}{2}} \Gamma\left(\frac{nb-p-1}{2}\right)}{b^{\frac{nb}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{\frac{1}{2}}} {RSS_T}^{-\frac{nb-p-1}{2}}.
$$

Τέλος, υπό την προσέγγιση κατά Pericchi με $t = \frac{p+3}{2}$ και $C(t)$ όπως στην (4.13), η (4.14) γίνεται:

$$
f^{P}\left(\boldsymbol{y}|\mathbf{X}_T, \lambda_T, b, T\right) = |J\left(\boldsymbol{y}, \lambda_T|T\right)|^{b} \frac{\pi^{-\frac{nb-p-1}{2}} 2^{\frac{p+1}{2}} \Gamma\left(\frac{nb}{2}\right)}{b^{\frac{nb+p+1}{2}} |{\mathbf{X}_T}^\top \mathbf{X}_T|^{\frac{1}{2}}} {RSS_T}^{-\frac{nb}{2}}.
$$

4.7 Ενσωμάτωση της Επιλογής Επεξηγηματικών Μεταβλητών

Η τρέχουσα ενότητα θα ξεκινήσει παρουσιάζοντας τον βασικό συμβολισμό που ενσωματώνει τις έννοιες που σχετίζονται με την επιλογή μοντέλου. Ως συνήθως, θεωρούμε ότι $\mathcal{T}$ είναι το σύνολο όλων των οικογενειών μετασχηματισμών υπό εξέταση. Κάθε οικογένεια διακρίνεται με βάση τον δείκτη $T \in \mathcal{T}$ και περιλαμβάνει μια παράμετρο μετασχηματισμού λ_T. Οι παρατηρήσεις συμβολίζονται με $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ και το διάνυσμα των μετασχηματισμένων δεδομένων κάτω από κάποια συγκεκριμένη τιμή της παραμέτρου λ_T για δεδομένη οικογένεια T συμβολίζεται με $\boldsymbol{y}^{(\lambda_T)} = \left(y_1^{(\lambda_T)}, \ldots, y_n^{(\lambda_T)}\right)^\top$. Χρησιμοποιώντας τον συνήθη συμβολισμό για την επιλογή επεξηγηματικών μεταβλητών, ο δείκτης $\boldsymbol{\gamma}$

είναι ένα διάνυσμα δίτιμων δεικτών μήκους p, με στοιχεία στο σύνολο $\mathcal{G} = \{0, 1\}^p$. Ειδικότερα, κάθε στοιχείο του διανύσματος $\boldsymbol{\gamma}$ ισούται είτε με ένα είτε με μηδέν ανάλογα με το αν η αντίστοιχη επεξηγηματική μεταβλητή είναι απούσα ή παρούσα στο μοντέλο. Ο σταθερός όρος συμπεριλαμβάνεται πάντα σε όλα τα μοντέλα. Ο κύριος στόχος είναι να αποκαλύψουμε την ύστερη πιθανότητα κάθε συνδυασμού $T \times \boldsymbol{\gamma}$, κάτω από κάποια κατάλληλη τιμή της παραμέτρου μετασχηματισμού λ_T, να αντιστοιχεί στο ακόλουθο κανονικό γραμμικό μοντέλο ($M_{\boldsymbol{\gamma},T}$):

$$\boldsymbol{y}^{(\lambda_T)}|\mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \lambda_T, T \sim N_n\left(\boldsymbol{y}^{(\lambda_T)}|\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma},T}, \sigma_T^2\mathbf{I}_n\right),$$

με $p + 1$ παραμέτρους, δηλαδή p συμμεταβλητές, στο πλήρες μοντέλο και μια άγνωστη παράμετρο διασποράς σ_T^2 και:

$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top$ είναι ο δείκτης των επεξηγηματικών μεταβλητών

$\boldsymbol{\beta}_{\boldsymbol{\gamma},T} : (p_{\boldsymbol{\gamma}} + 1) \times 1$ διάνυσμα παραμέτρων του μοντέλου $M_{\boldsymbol{\gamma},T}$

$\mathbf{X}_{\boldsymbol{\gamma}} : n \times (p_{\boldsymbol{\gamma}} + 1)$ πίνακας σχεδιασμού του μοντέλου $M_{\boldsymbol{\gamma},T}$

$\mathbf{I}_n : n \times n$ ταυτοτικός πίνακας

$\boldsymbol{\theta}_{\boldsymbol{\gamma},T} = \left(\boldsymbol{\beta}_{\boldsymbol{\gamma},T}, \sigma_T^2, \lambda_T\right)^\top$ είναι το διάνυσμα παραμέτρων διάστασης $(p_{\boldsymbol{\gamma}} + 3)$.

Επισημαίνεται ότι $\mathbf{X}_{\boldsymbol{\gamma}}$ είναι ένας υποπίνακας του πίνακα σχεδιασμού του πλήρους μοντέλου $\mathbf{X}$. Εν γένει, όταν ο δείκτης $\boldsymbol{\gamma}$ παραλείπεται από μια ποσότητα, όπως λόγου χάρη στις ποσότητες $\mathbf{X}$ και $\boldsymbol{\beta}_T$, αυτό υπονοεί ότι η αντίστοιχη ποσότητα αναφέρεται στο πλήρες μοντέλο με όλες τις p επεξηγηματικές μεταβλητές. Τέλος, θεωρούμε ότι η παράμετρος διασποράς σ_T^2 δεν εξαρτάται από το πλήθος των συμμεταβλητών του εκάστοτε μοντέλου και επομένως δε χαρακτηρίζεται από δείκτη $\boldsymbol{\gamma}$.

Για οποιαδήποτε οικογένεια T η πιθανοφάνεια των αρχικών μη μετασχηματισμένων δεδομένων $\boldsymbol{y}$ προσδιορίζεται πλήρως μέσω του αντίστροφου μετασχηματισμού $\boldsymbol{y}^{(\lambda_T)} \to \boldsymbol{y}$. Έτσι, η πιθανοφάνεια του $\mathbf{y}$, κάτω από το μοντέλο $M_{\boldsymbol{\gamma},T}$, συμπεριλαμβάνει την πιθανοφάνεια των μετασχηματισμένων παρατηρήσεων πολλαπλασιασμένη επί την απόλυτη τιμή της ορίζουσας του σχετικού Ιακωβιανού όρου του μετασχηματισμού $\left|J(\boldsymbol{y}, \lambda_T|T)\right| = \prod_{i=1}^n \left|\frac{\partial y_i^{(\lambda_T)}}{\partial y_i}\right|$. Η σχετική εξίσωση είναι η εξής:

$$f\left(\boldsymbol{y}|\mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma},T}, \sigma_T^2, \lambda_T, \boldsymbol{\gamma}, T\right) =$$

$$\left(2\pi\sigma_T^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_T^2}\left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma},T}\right)^\top \left(\boldsymbol{y}^{(\lambda_T)} - \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma},T}\right)\right) \times \left|J(\boldsymbol{y}, \lambda_T|T)\right|.$$

4.7.1 Ζητήματα πρότερων κατανομών

Ο μοντελοχώρος $\mathcal{M}$ διάστασης $|\mathcal{M}|$ για το υπό διερεύνηση πρόβλημα ορίζεται ως το σύνολο των μοντέλων για κάθε συνδυασμό οικογένειας μετασχηματισμών $\mathcal{T}$ και δείκτη επεξηγηματικών μεταβλητών $\boldsymbol{\gamma}$. Η πρότερη κατανομή για τον μοντελοχώρο είναι:

$$\pi(M_{\boldsymbol{\gamma},T}) = \pi(T, \boldsymbol{\gamma}) = \pi(T) \cdot \pi(\boldsymbol{\gamma})$$

για κάθε $M_{\boldsymbol{\gamma},T} \in \mathcal{M}$, άρα για κάθε συνδυασμό $T \in \mathcal{T}$ και $\boldsymbol{\gamma} \in \mathcal{G}$.

Αναφορικά με την πρότερη πιθανότητα σε επίπεδο οικογένειας μετασχηματισμών, μια διακριτή ομοιόμορφη κατανομή στο σύνολο $\mathcal{T}$ θα ήταν μια τυπική επιλογή για να εκφράσουμε την a priori άγνοιά μας. Αντίθετα, ποινικοποιούμε τις παραμετρικές οικογένειες ($\mathcal{T}_p = \{\text{BC}, \text{Mod}, \text{YJ}, \text{Dual}\}$) περισσότερο συγκριτικά με τους μη παραμετρικούς μετασχηματισμούς ($\mathcal{T}_{np} = \{\text{Id}, \text{Log}\}$) θέτοντας τις πρότερες πιθανότητές τους ως ακολούθως:

$$\pi(T) = \begin{cases} \frac{1}{2\cdot|\mathcal{T}_p|}, & T \in \mathcal{T}_p \\ \frac{1}{2\cdot|\mathcal{T}_{np}|}, & T \in \mathcal{T}_{np} \end{cases} \tag{4.15}$$

διασφαλίζοντας προφανώς ότι το συνολικό άθροισμα είναι μονάδα. Στην περίπτωση μιας αυστηρά θετικής απόκρισης (κάτι που είναι αρκετά σύνηθες), η οικογένεια Yeo-Johnson είναι ισοδύναμη με την οικογένεια Modulus και επομένως $|\mathcal{T}_p| = 3$, δηλαδή η πρότερη πιθανότητα μιας παραμετρικής οικογένειας γίνεται ίση με $1/6$. Κατά παρόμοιο τρόπο, αντί να επιλέξουμε μια διακριτή ομοιόμορφη πρότερη κατανομή για την παράμετρο $\boldsymbol{\gamma}$ για κάθε ένα από τα 2^p δυνατά μοντέλα, χρησιμοποιούμε μια βήτα-διωνυμική ιεραρχική πρότερη κατανομή που διαφοροποιεί τα μοντέλα με βάση το μέγεθός τους (Ley & Steel 2009). Ειδικότερα, για οποιοδήποτε μοντέλο $\boldsymbol{\gamma}$ στον μοντελοχώρο μας, η πρότερη κατανομή παίρνει τη μορφή

$$\begin{aligned} \pi(\boldsymbol{\gamma}) &= Bin(p, \tilde{\pi}), \\ \tilde{\pi} &\sim Beta\left(\alpha_{\tilde{\pi}}, \beta_{\tilde{\pi}}\right) \end{aligned} \tag{4.16}$$

όπου επιλέγουμε τις ακόλουθες τιμές για τις δύο υπερπαραμέτρους: $\alpha_{\tilde{\pi}} = 1$ και $\beta_{\tilde{\pi}} = (p - \tilde{m})/\tilde{m}$ με $\tilde{m}$ να είναι το πρότερο μέσο μέγεθος των μοντέλων. Η εν λόγω πρότερη

κατανομή εκφυλίζεται σε διακριτή ομοιόμορφη σχετικά με το μέγεθος των μοντέλων για $\tilde{m} = p/2$. Για $\alpha_{\tilde{\pi}} = \beta_{\tilde{\pi}} = 1$ όλα τα μοντέλα με ίδιο πλήθος συμμεταβλητών θεωρούνται a priori ισοπίθανα.

Η πρότερη κατανομή του πλήρους διανύσματος παραμέτρων $\boldsymbol{\theta}_{\boldsymbol{\gamma},T} = \left(\boldsymbol{\beta}_{\boldsymbol{\gamma},T}, \sigma_T^2, \lambda_T\right)^\top$ έχει την ακόλουθη ιεραρχική μορφή: $\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma},T}|\boldsymbol{\gamma}, T) = \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma},T}, \sigma_T^2|\lambda_T, \boldsymbol{\gamma}, T)\pi^N(\lambda_T|T)$. Η παράμετρος κύριου ενδιαφέροντος μέσα σε κάθε οικογένεια είναι το λ_T, ενώ οι υπόλοιπες παράμετροι του $\boldsymbol{\theta}_{\boldsymbol{\gamma},T}$ θεωρούνται παράμετροι δευτερεύουσας σημασίας για τη σύγκριση μοντέλων και επομένως μια ανεξάρτητη πρότερη κατανομή Jeffreys θα μπορούσε να εκχωρηθεί σε αυτές. Σχετικά με το πλαίσιο πρότερης πληροφορίας του $\boldsymbol{\theta}_{\boldsymbol{\gamma},T}$, μια επιλογή θα μπορούσε να βασιστεί στην πρότερη κατανομή των Box & Cox, προσαρμοσμένη για την επιλογή μοντέλου:

$$\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma},T}|\boldsymbol{\gamma}, T) \propto \pi^N(\lambda_T|T)\left(\sigma_T^2\right)^{-1} \cdot |J(\boldsymbol{y}, \lambda_T|T)|^{-\frac{p_{\boldsymbol{\gamma}}+1}{n}} \tag{4.17}$$

όπου $\pi^N(\lambda_T|T)$ είναι μια μη πληροφοριακή πρότερη κατανομή βάσης για την παράμετρο λ_T. Ένα αυθαίρετο σημείο στην κατασκευή αυτή είναι η ανάμειξη του Ιακωβιανού όρου. Από την άλλη πλευρά, η προσαρμοσμένη κατά Pericchi πρότερη κατανομή είναι:

$$\pi(\boldsymbol{\theta}_{\boldsymbol{\gamma},T}|\boldsymbol{\gamma}, T) \propto \pi^N(\lambda_T|T)\left(\sigma_T^2\right)^{-\frac{p_{\boldsymbol{\gamma}}+3}{2}} \tag{4.18}$$

βασιζόμενη στην πρότερη κατανομή Jeffreys για τις παραμέτρους $\left(\boldsymbol{\beta}_{\boldsymbol{\gamma},T}, \sigma_T^2\right)$ η οποία έχει την επιθυμητή ιδιότητα να μην εξαρτάται από τις τιμές της μεταβλητής απόκρισης σε αντίθεση με την (4.17). Ο πλήρης προσδιορισμός των IBF και FBF που ακολουθεί βασίζεται στη χρήση των εν λόγω μη πληροφοριακών πρότερων κατανομών βάσης για την παράμετρο λ_T, συγκεκριμένα $\pi^N(\lambda_T|T) \propto 1$, όπως επιλέγεται συχνά για το πρόβλημα επιλογής μετασχηματισμού (π.χ. βλέπε Sweeting (1984) και Box & Cox (1964)).

4.7.2 Διαμόρφωση των εναλλακτικών παραγόντων Bayes

Στα πλαίσια της Μπεϋζιανής συλλογιστικής, η ταυτοποίηση του βέλτιστου μοντέλου $M_{\boldsymbol{\gamma},T}$ είναι ισοδύναμη με την εύρεση του συνδυασμού οικογένειας μετασχηματισμών $T \in \mathcal{T}$ και δείκτη μεταβλητών $\boldsymbol{\gamma} \in \mathcal{G} = \{0,1\}^p$ που συγκεντρώνει την υψηλότερη ύστερη πιθανότητα $\pi\left(T, \boldsymbol{\gamma}|\boldsymbol{y}, \mathbf{X}_{\boldsymbol{\gamma}}\right)$.

Καθ' όλη την τρέχουσα ενότητα, θεωρούμε ένα τυχαίο μοντέλο που αντιστοιχεί στον συνδυασμό $\{T, \boldsymbol{\gamma}\}$ για τις εκφράσεις που ακολουθούν. Σχετικά με την πρώτη προσέγγιση που κάνει χρήση του διάμεσου IBF, θεωρούμε το σύνολο $\mathcal{D}_l$ που περιλαμβάνει όλα τα δυνατά (L) πολυμεταβλητά υποσύνολα $\boldsymbol{D}_l$ ελάχιστου μεγέθους m_0 από τα δεδομένα $\boldsymbol{D} = (\boldsymbol{y}, \mathbf{X})$. Με $B_{ij}(\boldsymbol{D})$ συμβολίζουμε τον παράγοντα Bayes μεταξύ των μοντέλων $i, j \in \mathcal{M}$ κάτω από την μη γνήσια πρότερη κατανομή $\pi^N(\lambda_T|T)$. Ο αντίστοιχος MIBF δίνεται από:

$$B_{ij}^{I,med}(\boldsymbol{D}) = B_{ij}(\boldsymbol{D}) \cdot \text{Median}\left(B_{ji}(\boldsymbol{D}_l)\right).$$

Η κατασκευή του FBF απαιτεί ένα κλάσμα b της λ_T-περιθώριας πιθανοφάνειας όπως δείξαμε στην (4.14), όπου $b = \frac{m}{n}$:

$$B_{ij}^{F}(\boldsymbol{D}) = B_{ij}(\boldsymbol{D}) \frac{\int f(\boldsymbol{y}|\mathbf{X}_{\boldsymbol{\gamma}_j}, \lambda_{T_j}, b, \boldsymbol{\gamma}_j, T_j)\pi^N(\lambda_{T_j}|T_j)\mathrm{d}\lambda_{T_j}}{\int f(\boldsymbol{y}|\mathbf{X}_{\boldsymbol{\gamma}_i}, \lambda_{T_i}, b, \boldsymbol{\gamma}_i, T_i)\pi^N(\lambda_{T_i}|T_i)\mathrm{d}\lambda_{T_i}}.$$

Σε όλα τα παραπάνω, χρησιμοποιούμε την ακόλουθη γενική μορφή της λ_T-περιθώριας πιθανοφάνειας των μη μετασχηματισμένων δεδομένων υπό το μοντέλο $M_{\boldsymbol{\gamma},T}$, αφήνοντας περιθώρια για την κλασματική παράμετρο b (χαμηλότερη του 1 για την περίπτωση του FBF, διαφορετικά ίση με 1):

$$\begin{aligned} f^{BC}\left(\boldsymbol{y}|\mathbf{X}_{\boldsymbol{\gamma}}, \lambda_T, b, \boldsymbol{\gamma}, T\right) \propto \quad & |J(\boldsymbol{y}, \lambda_T|T)|^{b-\frac{p_{\boldsymbol{\gamma}}+1}{n}} \cdot \pi^{-\frac{nb-p_{\boldsymbol{\gamma}}-1}{2}} |\mathbf{X}_{\boldsymbol{\gamma}}^\top \mathbf{X}_{\boldsymbol{\gamma}}|^{-\frac{1}{2}} b^{-\frac{nb}{2}} \\ & \cdot \left(S_T^2\right)^{-\frac{nb-p_{\boldsymbol{\gamma}}-1}{2}} \Gamma\left(\frac{nb-p_{\boldsymbol{\gamma}}-1}{2}\right) \end{aligned}$$

υπό την (4.17) και

$$\begin{aligned} f^{P}\left(\boldsymbol{y}|\mathbf{X}_{\boldsymbol{\gamma}}, \lambda_T, b, \boldsymbol{\gamma}, T\right) \propto \quad & |J(\boldsymbol{y}, \lambda_T|T)|^{b} \cdot \pi^{-\frac{nb-p_{\boldsymbol{\gamma}}-1}{2}} 2^{\frac{p_{\boldsymbol{\gamma}}+1}{2}} |\mathbf{X}_{\boldsymbol{\gamma}}^\top \mathbf{X}_{\boldsymbol{\gamma}}|^{-\frac{1}{2}} b^{-\frac{nb+p_{\boldsymbol{\gamma}}+1}{2}} \\ & \cdot \left(S_T^2\right)^{-\frac{nb}{2}} \Gamma\left(\frac{nb}{2}\right) \end{aligned}$$

υπό την (4.18), όπου $S_T^2 = \left(\boldsymbol{y}^{(\lambda_T)}\right)^\top \{\mathbf{I}_n - \mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}_{\boldsymbol{\gamma}}^\top \mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^\top\}\boldsymbol{y}^{(\lambda_T)}$ είναι το άθροισμα τετραγώνων των υπολοίπων για το κανονικό μοντέλο παλινδρόμησης.

Εν ολίγοις, η προσέγγιση AIBF είναι δύσκολο να χρησιμοποιηθεί στην περίπτωσή μας εξαιτίας της σύγκρισης πολλαπλών μη εμφωλευμένων μοντέλων που αντιμετωπίζουμε, ενώ η διάμεση μορφή του IBF προτιμάται έναντι της γεωμετρικής μορφής σύμφωνα και με τη σύσταση των Berger & Pericchi (1998). Ο FBF αποτελεί μια εναλλακτική του κλασικού παράγοντα Bayes στη βάση ενός κλάσματος b της ολικής πιθανοφάνειας προς εκπαίδευση. Παρά ταύτα, ο κλασματικός παράγοντας Bayes έχει την ελκυστική ιδιότητα

να είναι εύκολα υπολογίσιμος ακόμη και σε περιπτώσεις μη εμφωλευμένων μοντέλων ή πολλαπλών υπό σύγκριση μοντέλων.

Από υπολογιστικής απόψεως, απαιτείται προσοχή όσον αφορά τα όρια ολοκλήρωσης της παραμέτρου λ_T κατά τη διαδικασία υπολογισμού της περιθώριας πιθανοφάνειας. Στην πράξη, τιμές του λ_T χαμηλότερες από -2.4 οδηγούν σε μετασχηματισμένες τιμές της μεταβλητής απόκρισης που είναι σχεδόν ή εντελώς ταυτόσημες, εξαλείφοντας κάθε μεταβλητότητα ανάμεσα στα στοιχεία του $\boldsymbol{y}^{(\lambda_T)}$. Το γεγονός αυτό σε συνδυασμό με μια μεγάλη τιμή ορίζουσας του $\mathbf{X}_\gamma^\top \mathbf{X}_\gamma$ επιφέρει υπολογιστικά προβλήματα και σε σχέση με το ολικό δείγμα αλλά και σε σχέση με τα αντίστοιχα υποσύνολα εκπαίδευσης.

Μια τελευταία σημείωση είναι ότι η πρότερη κατανομή κατά Box-Cox είναι αρκετά περιοριστική όσον αφορά αποδεκτές τιμές για την κλασματική παράμετρο b. Μόνο τιμές μεγαλύτερες από $\frac{p_\gamma}{n}$ επιτρέπονται ώστε η βασική συνάρτηση προς ολοκλήρωση να ορίζεται. Αυτός είναι ένας ακόμη λόγος υπέρ της προσέγγισης κατά Pericchi. Έτσι, τα αποτελέσματα του επόμενου κεφαλαίου βασίζονται αποκλειστικά στη χρήση της πρότερης κατανομής του Pericchi όπως ορίζεται στην (4.18), καθώς χαρακτηρίζεται από μεγαλύτερη ευελιξία αλλά και αξιοπιστία όπως έχουμε επαληθεύσει σε προκαταρκτικές δοκιμές των σχετικών αλγόριθμων.

4.8 Συμπεράσματα

Στο παρόν κεφάλαιο αναπτύχθηκε μεθοδολογία με σκοπό να ξεπεραστούν τα μεθοδολογικά προβλήματα που ανακύπτουν στην επιλογή οικογένειας μετασχηματισμών σε μοντέλα παλινδρόμησης. Μια αξιόπιστη λύση περιγράφτηκε με τη βοήθεια εναλλακτικών μορφών παραγόντων Bayes, αν και κάποια δύσκολα σημεία παραμένουν. Ένα σημαντικό ζήτημα που αφορά όλες τις μορφές του IBF, συγκεκριμένα, έγκειται στο ότι οι ποιοτικές μεταβλητές ενδέχεται να μην ενσωματώνονται σωστά από τη διαδικασία, εξαιτίας του γεγονότος ότι εύκολα προκύπτουν ιδιάζοντες (*singular*) πίνακες αν το δείγμα εκπαίδευσης δεν επιλεγεί προσεκτικά. Ο FBF φαίνεται γενικά πιο ευέλικτος από πολλές απόψεις. Σχετικά με την επιλογή μοντέλου, κατάλληλες επιλογές για την a priori κατανομή του μοντελοχώρου συζητήθηκαν.

Στη συνέχεια, περνάμε στο Κεφάλαιο 5 για μια λεπτομερή εφαρμογή της μεθοδολογίας σε προβλήματα με επεξηγηματικές μεταβλητές, συμπεριλαμβανομένων ελέγχων εγκυρότητας.

Κεφάλαιο 5

Εφαρμογές σε Προβλήματα με Επεξηγηματικές Μεταβλητές

5.1 Εισαγωγή

Οι επόμενες ενότητες περιλαμβάνουν εφαρμογές της μεθοδολογίας του Κεφαλαίου 4. Πρόκειται για σύνολα δεδομένων με επεξηγηματικές μεταβλητές τα οποία, σε αντίθεση με τη διαδικασία που ακολουθήθηκε στο Κεφάλαιο 3, δεν υπόκεινται σε κανενός είδους τυποποίηση, αλλά ούτε και σε μετατόπιση κατά ξ καθώς οι τιμές των αντίστοιχων μεταβλητών απόκρισης είναι θετικές. Φυσικά, οι αντίστοιχες διαδικασίες μπορούν να ενεργοποιηθούν σε περίπτωση που άλλα σύνολα δεδομένων το απαιτήσουν.

Συνοπτικά, η Ενότητα 5.2 περιλαμβάνει τρεις πρώτες υποενότητες, μια για κάθε σύνολο δεδομένων υπό εξέταση. Το πρώτο σύνολο δεδομένων είναι προσομοιωμένο με βάση ένα μοντέλο γραμμικής παλινδρόμησης έχοντας ως στόχο μια πρώτη επικύρωση της σωστής λειτουργίας της μεθοδολογίας μας. Τα επόμενα δύο σύνολα δεδομένων είναι γνωστά στη διεθνή βιβλιογραφία και έχουν αποδειχθεί χρήσιμα για μεθοδολογίες επιλογής μεταβλητών. Πρόκειται για τα δεδομένα Hald των Woods, Steinour & Starke (1932), εξαιρετικά μικρού δειγματικού μεγέθους, που αποτελούνται από δύο ζεύγη υψηλά συσχετισμένων επεξηγηματικών μεταβλητών και για τα δεδομένα Highway, μεσαίου δειγματικού μεγέθους, που περιλαμβάνουν εννέα επεξηγηματικές μεταβλητές. Τα δεδομένα Highway προέρχονται από μια αδημοσίευτη εργασία του Carl Hoffstedt και αναφέρονται

από τον Weisberg (2014). Μια τέταρτη υποενότητα (Ενότητα 5.2.4) προσθέτει κάποιες ακόμη πρακτικές λεπτομέρειες και διευκρινίσεις όσον αφορά τις εφαρμογές που προηγήθηκαν (ή άλλες που παραλείφθηκαν). Κάποια βασικά συμπεράσματα υπογραμμίζονται στην Ενότητα 5.3.

5.2 Παραδείγματα Εφαρμογών

Με στόχο τη διερεύνηση της συμπεριφοράς των προσεγγίσεων που παρουσιάστηκαν στο Κεφάλαιο 4, εφαρμόσαμε τη μεθοδολογία αυτή σε μονομεταβλητά προβλήματα καθώς και σε προβλήματα με επεξηγηματικές μεταβλητές, αν και τα αποτελέσματα του παρόντος κεφαλαίου αφορούν μόνο τη δεύτερη περίπτωση προβλημάτων. Υπενθυμίζεται ότι το ενδιαφέρον μας στρέφεται στις μονοπαραμετρικές οικογένειες μετασχηματισμών Box-Cox, Modulus, Yeo & Johnson και Dual. Για τους αντίστοιχους μαθηματικούς τύπους μπορεί να ανατρέξει κανείς στον Πίνακα 2.1 του Κεφαλαίου 2. Οι μετασχηματισμοί Id και Log συμπεριλαμβάνονται επίσης στη διαδικασία επιλογής μοντέλου, όπως και προηγουμένως.

Όσον αφορά την πρότερη κατανομή του μοντελοχώρου, χρησιμοποιούνται τρεις εναλλακτικές με βάση όσα περιγράφτηκαν στην Ενότητα 4.7.1: μια διακριτή ομοιόμορφη κατανομή για την παράμετρο $\mathcal{T}$ και για την παράμετρο $\boldsymbol{\gamma}$ (Prior 1), μια διακριτή ομοιόμορφη κατανομή για την παράμετρο $\boldsymbol{\gamma}$ με την πρότερη κατανομή για το $\mathcal{T}$ να υποδεικνύεται από την (4.15) (Prior 2) και τέλος μια βήτα-διωνυμική πρότερη κατανομή στο $\boldsymbol{\gamma}$, όπως φαίνεται στην (4.16), σε συνδυασμό με την (4.15) για το $\mathcal{T}$ (Prior 3). Μια σύνοψη των τριών αυτών πρότερων προσεγγίσεων για τον μοντελοχώρο δίνεται στον Πίνακα 5.1.Κατά συνέπεια, το εκάστοτε σύνολο δεδομένων συνοδεύεται από τρεις κύριους πίνακες αποτελεσμάτων που αντιστοιχούν στις τρεις πρότερες προσεγγίσεις.

Σαν πρόσθετη πληροφορία στο κομμάτι της επιλογής μεταβλητών, παρέχονται εκτιμήσεις των ύστερων πιθανοτήτων ένταξης, $P\left(\gamma_j = 1|\boldsymbol{y}, \mathbf{X}\right)$, των επεξηγηματικών μεταβλητών $\{X_j\}, j = 1, \ldots, p$ για κάθε ένα από τα παραδείγματα που ακολουθούν και για κάθε πρότερη προσέγγιση. Στην περίπτωση που διαθέταμε αποτελέσματα με τη μορφή μιας MCMC αλυσίδας, τότε αυτές οι προηγούμενες εκτιμήσεις θα παράγονταν μέσω των

Πίνακας 5.1: Οι τρεις εναλλακτικές πρότερες κατανομές αναφορικά με τον μοντελοχώρο $\mathcal{M} = \{\mathcal{T} \times \mathcal{G}\}$ που χρησιμοποιούνται στις εφαρμογές του παρόντος κεφαλαίου.

Prior 1	Prior 2	Prior 3
$\pi(T) = \frac{1}{\lvert\mathcal{T}\rvert}$	$\pi(T) = \begin{cases} \frac{1}{2\cdot\lvert\mathcal{T}_p\rvert}, T \in \mathcal{T}_p \\ \frac{1}{2\cdot\lvert\mathcal{T}_{np}\rvert}, T \in \mathcal{T}_{np} \end{cases}$	$\pi(T) = \begin{cases} \frac{1}{2\cdot\lvert\mathcal{T}_p\rvert}, T \in \mathcal{T}_p \\ \frac{1}{2\cdot\lvert\mathcal{T}_{np}\rvert}, T \in \mathcal{T}_{np} \end{cases}$
$\pi(\boldsymbol{\gamma}) = \frac{1}{\lvert\mathcal{G}\rvert} = \frac{1}{2^p}$	$\pi(\boldsymbol{\gamma}) = \frac{1}{\lvert\mathcal{G}\rvert} = \frac{1}{2^p}$	$\pi(\boldsymbol{\gamma}) = \text{Bin}(p, \tilde{\pi})$, $\tilde{\pi} \sim \text{Beta}\,(\alpha_{\tilde{\pi}} = 1, \beta_{\tilde{\pi}} = 1)$

Σημείωση: $\mathcal{T}_p$ είναι το σύνολο των παραμετρικών οικογενειών μετασχηματισμών, ενώ $\mathcal{T}_{np}$ είναι το σύνολο των μη παραμετρικών μετασχηματισμών {Id, Log}.

δείκτριων μεταβλητών $I\left(\gamma_j^{(k)} = 1\right)$ για κάθε επανάληψη $k = 1, \ldots, K$ του MCMC (μετά την αφαίρεση του *burn in* τμήματος αυτής):

$$P\left(\gamma_j = 1|\boldsymbol{y}, \mathbf{X}\right) = (K)^{-1} \sum_{k=1}^{K} I\left(\gamma_j^{(k)} = 1\right).$$

Στην περίπτωσή μας, όπου λαμβάνει χώρα πλήρης απαρίθμηση (*full enumeration*), η εκτιμώμενη ύστερη πιθανότητα ένταξης $P\left(\gamma_j = 1|\boldsymbol{y}, \mathbf{X}\right)$ εξάγεται από το αντίστοιχο αλγοριθμικό αποτέλεσμα σύμφωνα με την ακόλουθη εξίσωση που περιλαμβάνει τις ύστερες πιθανότητες των μοντέλων $P(\boldsymbol{\gamma}, T|\boldsymbol{y}, \mathbf{X}_\gamma)$:

$$P\left(\gamma_j = 1|\boldsymbol{y}, \mathbf{X}\right) = \sum_{T \in \mathcal{T}} \sum_{\boldsymbol{\gamma} \in \mathcal{G}} P(\boldsymbol{\gamma}, T|\boldsymbol{y}, \mathbf{X}_\gamma) \cdot I(\gamma_j = 1).$$

Παρομοίως, μπορούμε να υπολογίσουμε την ύστερη πιθανότητα μιας συγκεκριμένης οικογένειας μετασχηματισμών $T_f \in \mathcal{T}$ μέσω της σχέσης:

$$P\left(T_f|\boldsymbol{y}, \mathbf{X}\right) = \sum_{\boldsymbol{\gamma} \in \mathcal{G}} P(\boldsymbol{\gamma}, T|\boldsymbol{y}, \mathbf{X}) \cdot I(T = T_f).$$

Ας σημειωθεί ότι στον υπολογισμό του FBF χρησιμοποιείται η τιμή $b = \frac{1}{n}$ και επίσης η τιμή $b = \frac{m_0}{n}$ όπου m_0 είναι το μέγεθος του ελάχιστου δείγματος εκπαίδευσης το οποίο χρησιμοποιείται αναγκαστικά για τον υπολογισμό του IBF. Επίσης, δίνεται μόνο η διάμεση μορφή του IBF εδώ (MIBF), καθώς παρουσιάζει την πιο αξιόπιστη συμπεριφορά σε σχέση με τις μορφές AIBF και GIBF στα πλαίσια του προβλήματος επιλογής μετασχηματισμού.

5.2.1 Προσομοιωμένα δεδομένα από το γραμμικό μοντέλο παλινδρόμησης

Σαν παράδειγμα αναφοράς για την μεθοδολογία του Κεφαλαίου 4, προσομοιώσαμε δεδομένα από ένα κανονικό γραμμικό μοντέλο παλινδρόμησης. Συγκεκριμένα, το σύνολο δεδομένων αποτελείται από $n = 30$ παρατηρήσεις με $p = 3$ επεξηγηματικές μεταβλητές προσομοιωμένες από μια κανονική κατανομή με μέση τιμή 5 και διακύμανση 1:

$$X_{ij} \sim N(5, 1),\ j = 1, \ldots, 3,\ i = 1, \ldots, 30$$

ενώ η μεταβλητή απόκρισης προσομοιώθηκε από την κατανομή:

$$Y_i \sim N(1 + 8X_{i1} + 4X_{i2}, 1),\ i = 1, \ldots, 30.$$

Οι προσομοιωμένες επεξηγηματικές μεταβλητές συσχετίζονται πολύ ελαφρά, με τον μέγιστο βαθμό συσχέτισης να είναι ίσος με 15% για το ζεύγος των μεταβλητών X_1, X_2. Επομένως, το ζήτημα της πολυσυγγραμικότητας δεν θα μας απασχολήσει σε αυτό το παράδειγμα.

Τα αποτελέσματα που προέκυψαν από την εφαρμογή της προτεινόμενης μεθοδολογίας για την επιλογή μοντέλου, με έμφαση αφενός στην επιλογή οικογένειας μετασχηματισμών και αφετέρου στην επιλογή μεταβλητών, εμπεριέχονται στους Πίνακες 5.2–5.4 οι οποίοι αντιστοιχούν στις πρότερες προσεγγίσεις Prior 1–3. Το μέγεθος του ελάχιστου δείγματος εκπαίδευσης είναι ίσο με 6.

Στον Πίνακα 5.2, παρατηρούμε ότι το σύνολο των δέκα καλύτερων μοντέλων, με όρους των ύστερων πιθανοτήτων (PMP) του κάθε μοντέλου, διαφοροποιείται ανάμεσα στις τρεις προσεγγίσεις των παραγόντων BF. Η σειρά κατάταξης των οικογενειών μετασχηματισμών δε φαίνεται να αλλάζει μεταξύ των δύο μορφών του FBF, δηλαδή για $b = 1/n$ και για $b = m_0/n$. Παρόλα αυτά, η δεύτερη επιλογή για την τιμή της κλασματικής παραμέτρου b οδηγεί εμφανώς σε λιγότερο αξιόπιστα αποτελέσματα, εφόσον το επικρατέστερο μοντέλο (με PMP $= 22.3\%$) λανθασμένα εμπεριέχει και τις τρεις επεξηγηματικές μεταβλητές, είναι δηλαδή το πλήρες μοντέλο. Εν αντιθέσει, το βέλτιστο μοντέλο έτσι όπως αναδεικνύεται από τον FBF με κλασματική παράμετρο $b = 1/n$, συνδεόμενο με διπλάσια ύστερη πιθανότητα PMP ίση με 45.2%, αντιστοιχεί στο πραγματικό μοντέλο

Πίνακας 5.2: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο προσομοιωμένων δεδομένων από το γραμμικό μοντέλο παλινδρόμησης υπό την Prior 1 ($m_0 = 6$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Id	$X_1 + X_2$	0.452	Id	$X_1 + X_2 + X_3$	0.223	Id	$X_1 + X_2$	0.632
Id	$X_1 + X_2 + X_3$	0.221	Id	$X_1 + X_2$	0.221	Id	$X_1 + X_2 + X_3$	0.121
Dual	$X_1 + X_2$	0.087	Dual	$X_1 + X_2$	0.095	Dual	$X_1 + X_2$	0.091
Mod	$X_1 + X_2$	0.069	Mod	$X_1 + X_2$	0.094	Mod	$X_1 + X_2$	0.069
BC	$X_1 + X_2$	0.068	BC	$X_1 + X_2$	0.094	BC	$X_1 + X_2$	0.068
Dual	$X_1 + X_2 + X_3$	0.039	Dual	$X_1 + X_2 + X_3$	0.091	Dual	$X_1 + X_2 + X_3$	0.010
Mod	$X_1 + X_2 + X_3$	0.032	Mod	$X_1 + X_2 + X_3$	0.090	Mod	$X_1 + X_2 + X_3$	0.005
BC	$X_1 + X_2 + X_3$	0.031	BC	$X_1 + X_2 + X_3$	0.090	BC	$X_1 + X_2 + X_3$	0.005
BC	$-$	< 0.001	BC	$-$	< 0.001	BC	$-$	< 0.001
BC	X_3	< 0.001	BC	X_3	< 0.001	BC	X_3	< 0.001

το οποίο δε χρειάζεται μετασχηματισμό των τιμών της μεταβλητής απόκρισης (οικογένεια Id) και εμπεριέχει μόνο τις πρώτες δύο επεξηγηματικές μεταβλητές. Το δεύτερο καλύτερο μοντέλο της προσέγγισης FBF με $b = 1/n$ επίσης δεν υποδεικνύει μετασχηματισμό αλλά εμπεριέχει και τις τρεις επεξηγηματικές μεταβλητές. Το γεγονός αυτό δείχνει ότι η κανονικότητα του μηχανισμού παραγωγής δεδομένων αναγνωρίζεται εύκολα. Επιπρόσθετα, η προσέγγιση MIBF παράγει την ίδια ακριβώς σειρά οικογενειών μετασχηματισμών και συνόλων επεξηγηματικών μεταβλητών για τα δέκα καλύτερα μοντέλα, αν και η ύστερη πιθανότητα του βέλτιστου μοντέλου (που είναι και το πραγματικό μοντέλο) είναι υψηλότερη και ίση με 63%. Ας σημειωθεί ακόμη ότι τα δύο τελευταία μοντέλα σε κάθε προσέγγιση υπό την Prior 1 έχουν ύστερη πιθανότητα μικρότερη του 0.1% και επομένως είναι αμελητέας ύστερης βαρύτητας. Ένα εξ αυτών, μάλιστα, δεν περιέχει καμιά επεξηγηματική μεταβλητή παρά μόνο τον σταθερό όρο.

Τα αποτελέσματα υπό την Prior 2 (βλέπε Πίνακα 5.3) δε διαφοροποιούνται σημαντικά σε σχέση με τα αποτελέσματα του Πίνακα 5.2. Η σειρά των αποτελεσμάτων σχετικά με την επιλογή μεταβλητών παραμένει απαράλλαχτη, ενώ η σειρά των οικογενειών μετασχηματισμών των δέκα καλύτερων μοντέλων αλλάζει ελάχιστα στην προσέγγιση FBF με

Πίνακας 5.3: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο προσομοιωμένων δεδομένων από το γραμμικό μοντέλο παλινδρόμησης υπό την Prior 2 ($m_0 = 6$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Id	$X_1 + X_2$	0.507	Id	$X_1 + X_2 + X_3$	0.274	Id	$X_1 + X_2$	0.689
Id	$X_1 + X_2 + X_3$	0.248	Id	$X_1 + X_2$	0.271	Id	$X_1 + X_2 + X_3$	0.132
Dual	$X_1 + X_2$	0.065	Dual	$X_1 + X_2$	0.078	Dual	$X_1 + X_2$	0.066
Mod	$X_1 + X_2$	0.052	BC	$X_1 + X_2$	0.077	Mod	$X_1 + X_2$	0.050
BC	$X_1 + X_2$	0.051	Mod	$X_1 + X_2$	0.077	BC	$X_1 + X_2$	0.049
Dual	$X_1 + X_2 + X_3$	0.030	Dual	$X_1 + X_2 + X_3$	0.075	Dual	$X_1 + X_2 + X_3$	0.007
Mod	$X_1 + X_2 + X_3$	0.024	BC	$X_1 + X_2 + X_3$	0.074	Mod	$X_1 + X_2 + X_3$	0.004
BC	$X_1 + X_2 + X_3$	0.023	Mod	$X_1 + X_2 + X_3$	0.074	BC	$X_1 + X_2 + X_3$	0.003
BC	$-$	< 0.001	BC	$-$	< 0.001	BC	$-$	< 0.001
BC	X_3	< 0.001	BC	X_3	< 0.001	BC	X_3	< 0.001

$b = m_0/n$. Επιπλέον, το πραγματικό μοντέλο αναδεικνύεται κάτω από την προσέγγιση FBF με $b = 1/n$, αλλά και κάτω από την προσέγγιση MIBF με ακόμη υψηλότερη ύστερη πιθανότητα από πριν (50.7% και 68.9% αντίστοιχα).

Τέλος, τα αποτελέσματα που αντιστοιχούν στην πρότερη προσέγγιση Prior 3 παρουσιάζονται στον Πίνακα 5.4. Παρά το γεγονός ότι ο Ταυτοτικός μετασχηματισμός Id διατηρεί τη θέση του στην κορυφή των καλύτερων μοντέλων για κάθε προσέγγιση, η διαδικασία επιλογής μεταβλητών αποτυγχάνει στις περισσότερες περιπτώσεις. Η προσέγγιση FBF με την κλασματική παράμετρο b υπολογισμένη στη βάση του m_0 αποδίδει μόνο 13.6% ύστερη πιθανότητα στο σωστό μοντέλο. Αν ρίξουμε μια πιο προσεκτική ματιά, θα παρατηρήσουμε ότι μόνο ο MIBF προτείνει το σωστό μοντέλο ως βέλτιστο, αν και με μειωμένη ύστερη πιθανότητα (PMP $= 53.3\%$) σε σχέση με τα μοντέλα που απορρέουν υπό τις πρότερες προσεγγίσεις Prior 1 και Prior 2.

Για μια πιο συνολική οπτική των ύστερων πιθανοτήτων ένταξης των επεξηγηματικών μεταβλητών στο μοντέλο, ο αναγνώστης παραπέμπεται στον Πίνακα 5.5. Οι μεταβλητές X_1 και X_2 επιλέγονται σίγουρα προς ένταξη στο τελικό μοντέλο από όλες τις προσεγγίσεις, εφόσον οι ύστερες πιθανότητες ένταξής τους είναι ίσες ή σχεδόν ίσες με 1. Από την

Πίνακας 5.4: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο προσομοιωμένων δεδομένων από το γραμμικό μοντέλο παλινδρόμησης υπό την Prior 3 ($m_0 = 6$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Id	$X_1 + X_2 + X_3$	0.451	Id	$X_1 + X_2 + X_3$	0.412	Id	$X_1 + X_2$	0.533
Id	$X_1 + X_2$	0.308	Id	$X_1 + X_2$	0.136	Id	$X_1 + X_2 + X_3$	0.307
Dual	$X_1 + X_2 + X_3$	0.054	Dual	$X_1 + X_2 + X_3$	0.112	Dual	$X_1 + X_2$	0.051
Mod	$X_1 + X_2 + X_3$	0.043	Mod	$X_1 + X_2 + X_3$	0.111	Mod	$X_1 + X_2$	0.039
BC	$X_1 + X_2 + X_3$	0.042	BC	$X_1 + X_2 + X_3$	0.111	BC	$X_1 + X_2$	0.038
Dual	$X_1 + X_2$	0.040	Dual	$X_1 + X_2$	0.039	Dual	$X_1 + X_2 + X_3$	0.016
Mod	$X_1 + X_2$	0.031	BC	$X_1 + X_2$	0.038	Mod	$X_1 + X_2 + X_3$	0.008
BC	$X_1 + X_2$	0.031	Mod	$X_1 + X_2$	0.038	BC	$X_1 + X_2 + X_3$	0.008
BC	$-$	< 0.001	BC	$-$	< 0.001	BC	$-$	< 0.001
BC	X_3	< 0.001	BC	X_3	< 0.001	BC	X_3	< 0.001

Πίνακας 5.5: Εκτιμήσεις των ύστερων πιθανοτήτων ένταξης των μεταβλητών στο τελικό μοντέλο με βάση τον ενδογενή και τον κλασματικό παράγοντα Bayes για το σύνολο προσομοιωμένων δεδομένων από το γραμμικό μοντέλο παλινδρόμησης ($m_0 = 6$).

Προσέγγιση	X_1	X_2	X_3
		Prior 1	
FBF, $b = 1/n$	1	1	0.323
FBF, $b = m_0/n$	0.999	0.999	0.495
MIBF	0.999	0.999	0.141
		Prior 2	
FBF, $b = 1/n$	1	1	0.324
FBF, $b = m_0/n$	1	1	0.497
MIBF	1	1	0.146
		Prior 3	
FBF, $b = 1/n$	0.999	0.999	0.590
FBF, $b = m_0/n$	0.999	0.999	0.748
MIBF	1	1	0.339

άλλη μεριά, στη μεταβλητή X_3 αποδίδεται το χαμηλότερο ύστερο βάρος (14.1%) κάτω από τον MIBF και την προσέγγιση Prior 1, αν και το αντίστοιχο αποτέλεσμα υπό την Prior 2 είναι μόνο 0.05 μονάδες υψηλότερο. Ο FBF με $b = 1/n$ συμπεριφέρεται επίσης ορθά υπό τις Prior 1 και Prior 2, με ύστερη πιθανότητα ένταξης της X_3 γύρω στο 32%. Η πρότερη προσέγγιση Prior 3 δεν παρέχει αποδεκτά αποτελέσματα, καθώς η ύστερη πιθανότητα ένταξης της μεταβλητής X_3 υπερβαίνει το 50% στις δύο από τις τρεις περιπτώσεις.

Ο Πίνακας 5.6 παρουσιάζει τις ύστερες περιθώριες πιθανότητες των οικογενειών μετασχηματισμών για αυτό το πρόβλημα, όπου όλες οι προσεγγίσεις αποδίδουν το μεγαλύτερο ύστερο βάρος στον Ταυτοτικό μετασχηματισμό. Ο διάμεσος ενδογενής παράγοντας Bayes δίνει υψηλότερες ύστερες πιθανότητες στον Ταυτοτικό μετασχηματισμό σε σχέση με τον κλασματικό παράγοντα Bayes με την καλύτερη απόδοση υπό την Prior 2 (83.9%). Ο FBF με τη χαμηλή τιμή της παραμέτρου b είναι πίσω από τον MIBF κατά περίπου 8 ποσοστιαίες μονάδες για κάθε πρότερη προσέγγιση κυμαινόμενος από 67.3% μέχρι 75.9%, ενώ η χαμηλή τιμή της παραμέτρου b συνδέεται με τιμές FBF σχεδόν 30 ποσοστιαίες μονάδες κάτω από τον MIBF για κάθε πρότερη προσέγγιση, με εύρος από 44.3% έως 54.7%.

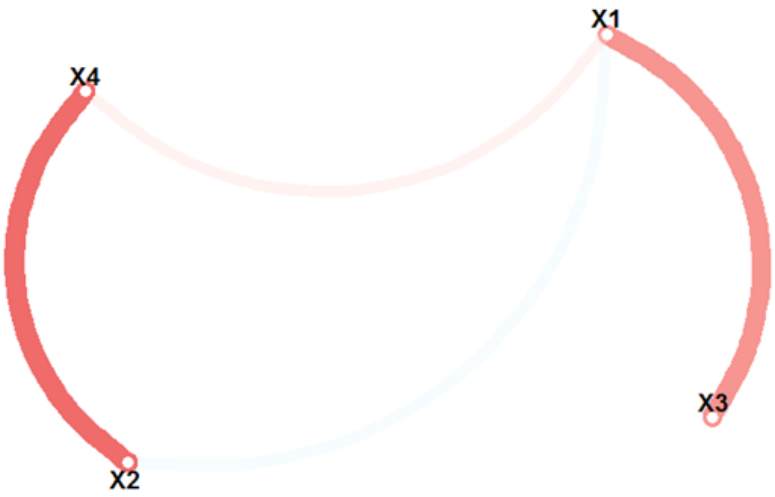
Πίνακας 5.6: Εκτιμήσεις των ύστερων περιθώριων πιθανοτήτων των οικογενειών μετασχηματισμών με βάση τον ενδογενή και τον κλασματικό παράγοντα Bayes για το σύνολο προσομοιωμένων δεδομένων από το γραμμικό μοντέλο παλινδρόμησης ($m_0 = 6$).

Προσέγγιση	BC	Mod	YJ	Log	Id
			Prior 1		
FBF, $b = 1/n$	0.099	0.101	0.130	< 0.001	0.673
FBF, $b = m_0/n$	0.185	0.185	0.186	< 0.001	0.443
MIBF	0.072	0.073	0.101	< 0.001	0.754
			Prior 2		
FBF, $b = 1/n$	0.073	0.074	0.093	< 0.001	0.759
FBF, $b = m_0/n$	0.150	0.151	0.152	< 0.001	0.547
MIBF	0.046	0.047	0.068	< 0.001	0.839
			Prior 3		
FBF, $b = 1/n$	0.075	0.075	0.095	< 0.001	0.755
FBF, $b = m_0/n$	0.152	0.152	0.153	< 0.001	0.544
MIBF	0.053	0.053	0.073	< 0.001	0.821

Συνολικά, όσον αφορά αυτό το παράδειγμα αναφοράς, ο MIBF φαίνεται να δίνει τα πιο αξιόπιστα αποτελέσματα, ενώ ακολουθεί στενά ο FBF με τη χαμηλή τιμή της κλασματικής παραμέτρου b. Είναι αρκετά εμφανές ότι η πρόταση που συνηθέστερα ανευρίσκεται στη βιβλιογραφία περί κατασκευής της κλασματικής παραμέτρου b του FBF με βάση το μέγεθος του ελάχιστου δείγματος εκπαίδευσης m_0 αποδεικνύεται όχι απόλυτα ακριβής, καθώς υπό καμιά πρότερη προσέγγιση για τον μοντελοχώρο δεν αναδεικνύεται το σωστό μοντέλο ως βέλτιστο σε αυτό το παράδειγμα αναφοράς. Ανάμεσα στις διάφορες πρότερες προσεγγίσεις που δοκιμάστηκαν για τον μοντελοχώρο, η προσέγγιση Prior 3 κρίνεται ως αρκετά ακραία, οδηγώντας σε μη ικανοποιητικά αποτελέσματα ως επί το πλείστον. Η προσέγγιση Prior 2 αποδεικνύεται ευέλικτη ώστε να αντιπροσωπεύσει καλύτερα την πρότερη διακύμανση του μοντελοχώρου και άρα να δώσει συνεπή αποτελέσματα.

5.2.2 Σύνολο δεδομένων Hald

Το σύνολο δεδομένων Hald (Woods, Steinour & Starke 1932, Hald 1952), το οποίο περιέχεται στη βιβλιοθήκη `mombf` της R, έχει χρησιμοποιηθεί συχνά για συγκρίσεις σε προβλήματα επιλογής μεταβλητών. Ο λόγος που προτιμάται σε τέτοιου είδους προβλήματα έχει να κάνει με την πολύ υψηλή συσχέτιση μεταξύ κάποιων ζευγών επεξηγηματικών μεταβλητών. Αποτελείται από 13 παρατηρήσεις μιας μεταβλητής απόκρισης (θερμότητα που αναπτύσσεται κατά τη διάρκεια 180 ημερών σκλήρυνσης του τσιμέντου, σε θερμίδες ανά γραμμάριο τσιμέντου) και τεσσάρων επεξηγηματικών μεταβλητών $X_j,\ j = 1, \ldots, 4$ που εκφράζουν το ποσοστό τεσσάρων διαφορετικών υλικών στο τελικό μείγμα τσιμέντου (X_1: αργυλικό ασβέστιο, X_2: πυριτικό ασβέστιο, X_3: αργυλοσιδηρικό τετρασβέστιο, X_4: πυριτικό διασβέστιο). Η τιμή του δειγματικού συντελεστή συσχέτισης ανάμεσα στις X_1 και X_3 είναι 82.4%, ενώ η αντίστοιχη τιμή μεταξύ των X_2 και X_4 είναι 97.3%, με τον δείκτη VIF (*variance inflation factor*) να κυμαίνεται από 38 έως 254 μονάδες για τις τέσσερις μεταβλητές. Το Διάγραμμα 5.1 απεικονίζει το διάγραμμα δικτύου των επεξηγηματικών μεταβλητών με βάση τις τιμές δειγματικής συσχέτισης, όπου το κόκκινο και το μπλε χρώμα δηλώνουν συσχετίσεις με θετικό και αρνητικό πρόσημο αντίστοιχα, ενώ χρωματικά πιο έντονες τροχιές και πιο κοντινοί κόμβοι δηλώνουν πιο έντονη συσχέτιση. Το μέγεθος του ελάχιστου δείγματος εκπαίδευσης είναι ίσο με 7 παρατηρήσεις ($m_0 = 7$).



Διάγραμμα 5.1: Διάγραμμα δικτύου (network plot) των επεξηγηματικών μεταβλητών $X_j,\ j = 1, \ldots, 4$, των δεδομένων Hald με βάση τις τιμές δειγματικής συσχέτισης.

Πριν περάσουμε στην ανάλυση των αποτελεσμάτων, να αναφέρουμε ότι, μεταξύ άλλων, οι Villa & Lee (2015) εκτίμησαν ότι το βέλτιστο πλήθος μεταβλητών για το μοντέλο των δεδομένων Hald είναι δύο και, πιο συγκεκριμένα, ανέδειξαν το μοντέλο με τις X_1 και X_2 ως σταθερά επικρατέστερο δοκιμάζοντας διάφορες πρότερες κατανομές. Πριν από αυτούς, οι Berger & Pericchi (1996*a*) είχαν επίσης εστιάσει σε μοντέλα δύο επεξηγηματικών μεταβλητών συμπεραίνοντας ότι το σύνολο $\{X_1, X_2\}$ είναι σχετικά προτιμότερο του συνόλου $\{X_1, X_4\}$ και κατά πολύ προτιμότερο του συνόλου $\{X_3, X_4\}$. Με βάση τις ύστερες πιθανότητες των μοντέλων στο άρθρο των Casella & Moreno (2006) και τα αποτελέσματα του εκτιμητή Gibbs των Kuo & Mallick (1998), το μοντέλο με τις $\{X_1, X_2\}$ επιλέγεται ως βέλτιστο.

Τα αποτελέσματα που αντιστοιχούν στις τρεις διαφορετικές πρότερες προσεγγίσεις για τον μοντελοχώρο ενέχονται στους Πίνακες 5.7–5.9, όπου με μια πρώτη ματιά δε διακρίνουμε κανένα μοντέλο με μόνο μία επεξηγηματική μεταβλητή. Αυτό είναι ένα πρώτο θετικό στοιχείο, καθώς το μείγμα τσιμέντου απαιτείται να περιέχει τουλάχιστον δύο υλικά.

Τα μοντέλα Id και Dual υπερισχύουν στα αποτελέσματα που σχετίζονται με την προσέγγιση Prior 1 (Πίνακας 5.7). Όσον αφορά τις ύστερες πιθανότητες των μοντέλων υπό τον FBF με $b = 1/n$, η οικογένεια BC και η οικογένεια Modulus (ως ισοδύναμη με την οικογένεια YJ για θετικά δεδομένα) συνδέονται με πιθανότητα 4.1% για το μοντέλο που

εμπεριέχει τις μεταβλητές X_1 και X_2. Στις άλλες δύο BF προσεγγίσεις, παρατηρούμε επίσης ότι αυτές οι δύο οικογένειες έχουν ίσες ύστερες πιθανότητες όταν αναφέρονται στο ίδιο διάνυσμα $\boldsymbol{\gamma}$. Καθώς η κλασματική παράμετρος b αυξάνει από $1/n$ σε $m_0/n = 7/13$, παρατηρείται μια ελαφριά τάση ομοιογενοποίησης μεταξύ των ύστερων πιθανοτήτων των μοντέλων. Αυτό ήταν αναμενόμενο αν σκεφτεί κανείς την κατασκευή του FBF, όπως περιγράφτηκε νωρίτερα. Επομένως, η επιλογή της τιμής $b = m_0/n$ μάλλον μας προβληματίζει, εφόσον το μοντέλο που φαίνεται να αποφεύγει ενδεχομένως ζητήματα πολυσυγγραμμικότητας, καθώς δεν περιλαμβάνει ταυτόχρονα τις X_2 και X_4 για παράδειγμα, έρχεται δέκατο κατά σειρά φθίνουσας ύστερης πιθανότητας ($PMP = 3.8\%$), ενώ θα αναμέναμε να έρθει ψηλότερα στην κατάταξη. Αμφότεροι οι FBF (χαμηλής τιμής b) και MIBF αναδεικνύουν πρώτο το μοντέλο Id με τις μεταβλητές X_1 και X_2, όπως αναμενόταν, παρότι η πρώτη εκ των δύο αυτών προσεγγίσεων υποστηρίζει το μοντέλο αυτό περισσότερο με ύστερη πιθανότητα 8.7% συγκρινόμενη με ύστερη πιθανότητα 8.2% της MIBF προσέγγισης. Εν γένει, παρατηρώντας τις ύστερες πιθανότητες των 2-3 επικρατέστερων μοντέλων οδηγούμαστε στο συμπέρασμα ότι η ανάδειξη του μετασχηματισμού Id είναι πιο σημαντική από το σύνολο συμμεταβλητών που θα αναδειχθεί.

Πίνακας 5.7: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο δεδομένων Hald υπό την Prior 1 ($m_0 = 7$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Id	$X_1 + X_2$	0.087	Id	$X_1 + X_2 + X_4$	0.049	Id	$X_1 + X_2$	0.082
Id	$X_1 + X_2 + X_4$	0.074	Id	$X_1 + X_2 + X_3$	0.048	Id	$X_1 + X_2 + X_4$	0.059
Id	$X_1 + X_2 + X_3$	0.073	Dual	$X_1 + X_2 + X_3$	0.043	Dual	$X_1 + X_2$	0.056
Dual	$X_1 + X_2$	0.065	Dual	$X_1 + X_2 + X_4$	0.042	Id	$X_1 + X_2 + X_3$	0.056
Dual	$X_1 + X_2 + X_3$	0.056	Id	$X_1 + X_3 + X_4$	0.041	Mod	$X_1 + X_2$	0.050
Dual	$X_1 + X_2 + X_4$	0.054	BC	$X_1 + X_2 + X_3$	0.040	Log	$X_1 + X_2$	0.050
Id	$X_1 + X_3 + X_4$	0.053	Mod	$X_1 + X_2 + X_3$	0.040	BC	$X_1 + X_2$	0.049
Mod	$X_1 + X_2$	0.041	BC	$X_1 + X_2 + X_4$	0.039	Id	$X_1 + X_3 + X_4$	0.047
BC	$X_1 + X_2$	0.041	Mod	$X_1 + X_2 + X_4$	0.039	Id	$X_1 + X_4$	0.041
Dual	$X_1 + X_3 + X_4$	0.039	Id	$X_1 + X_2$	0.038	Dual	$X_1 + X_2 + X_3$	0.041

Περνώντας στην πρότερη προσέγγιση Prior 2 (Πίνακας 5.8), βλέπουμε το ίδιο μοτίβο

όπως και στην Ενότητα 5.2.1, ότι δηλαδή το μοντέλο που αναμέναμε να αναδειχθεί στις πρώτες θέσεις υποστηρίζεται όντως περισσότερο σε σχέση με τα αποτελέσματα υπό την προσέγγιση Prior 1. Το μοντέλο Id με τις μεταβλητές X_1 και X_2 λαμβάνει ύστερο βάρος ίσο με 10.7%, 4.8% και 9.9% υπό τους FBF χαμηλού b ($b = 1/n$), FBF υψηλού b ($b = m_0/n$) και MIBF αντίστοιχα. Οι προηγούμενες αντίστοιχες τιμές από τον Πίνακα 5.7 ισούνται με 8.7%, 3.8% και 8.2%. Τα Mod και BC μετασχηματισμένα μοντέλα που περιέχουν τις μεταβλητές X_1 και X_2 εμφανίζονται προς το τέλος της λίστας των μοντέλων με περίπου 3.3% και 4% για τους FBF χαμηλού b και MIBF αντίστοιχα. Ο FBF υψηλού b αναδεικνύει μοντέλα που διακρίνονται από πολυπλοκότητα και αφήνουν υπόνοιες για μη ορθή αντιμετώπιση ζητημάτων πολυσυγγραμμικότητας.

Πίνακας 5.8: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο δεδομένων Hald υπό την Prior 2 ($m_0 = 7$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Id	$X_1 + X_2$	0.107	Id	$X_1 + X_2 + X_4$	0.062	Id	$X_1 + X_2$	0.099
Id	$X_1 + X_2 + X_4$	0.092	Id	$X_1 + X_2 + X_3$	0.061	Id	$X_1 + X_2 + X_4$	0.072
Id	$X_1 + X_2 + X_3$	0.090	Id	$X_1 + X_3 + X_4$	0.052	Id	$X_1 + X_2 + X_3$	0.068
Id	$X_1 + X_3 + X_4$	0.065	Id	$X_1 + X_2$	0.048	Log	$X_1 + X_2$	0.061
Dual	$X_1 + X_2$	0.053	Id	$X_1 + X_2 + X_3 + X_4$	0.046	Id	$X_1 + X_3 + X_4$	0.058
Dual	$X_1 + X_2 + X_3$	0.046	Dual	$X_1 + X_2 + X_3$	0.036	Id	$X_1 + X_4$	0.050
Dual	$X_1 + X_2 + X_4$	0.045	Dual	$X_1 + X_2 + X_4$	0.036	Dual	$X_1 + X_2$	0.045
Log	$X_1 + X_2$	0.044	BC	$X_1 + X_2 + X_3$	0.033	Mod	$X_1 + X_2$	0.041
Mod	$X_1 + X_2$	0.034	Mod	$X_1 + X_2 + X_3$	0.033	BC	$X_1 + X_2$	0.040
BC	$X_1 + X_2$	0.033	Mod	$X_1 + X_2 + X_4$	0.032	Dual	$X_1 + X_2 + X_3$	0.033

Η προσέγγιση υπό την Prior 3 δίνει αρκετά διαφορετικά αποτελέσματα (βλέπε Πίνακα 5.9), με τις προσεγγίσεις FBF να αναδεικνύουν το πλήρες μοντέλο Id ως βέλτιστο με αντίστοιχες ύστερες πιθανότητες 9.6% και 13.4%. Μάλιστα, η προσέγγιση MIBF παρουσιάζεται ως ελαφρά πιο εύρωστη συγκριτικά με τις άλλες δύο προσεγγίσεις, αν και το καλύτερο μοντέλο κατ' αυτήν συμπεριλαμβάνει τις μεταβλητές X_1 και X_2 μαζί με τη μεταβλητή X_4. Όλα τα FBF μοντέλα υψηλού b δε φαίνεται να αναδεικνύουν φειδωλά μοντέλα καθώς περιλαμβάνουν τουλάχιστον τρεις επεξηγηματικές μεταβλητές σε κάθε

περίπτωση, οπότε για ακόμη μια φορά η χαμηλή τιμή της κλασματικής παραμέτρου b φαίνεται να δρα καλύτερα.

Πίνακας 5.9: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο δεδομένων Hald υπό την Prior 3 ($m_0 = 7$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Id	$X_1 + X_2 + X_3 + X_4$	0.096	Id	$X_1 + X_2 + X_3 + X_4$	0.134	Id	$X_1 + X_2 + X_4$	0.076
Id	$X_1 + X_2 + X_4$	0.085	Dual	$X_1 + X_2 + X_3 + X_4$	0.080	Id	$X_1 + X_2 + X_3$	0.071
Id	$X_1 + X_2 + X_3$	0.084	Mod	$X_1 + X_2 + X_3 + X_4$	0.073	Id	$X_1 + X_2$	0.070
Id	$X_1 + X_2$	0.066	BC	$X_1 + X_2 + X_3 + X_4$	0.073	Id	$X_1 + X_3 + X_4$	0.061
Id	$X_1 + X_3 + X_4$	0.060	Log	$X_1 + X_2 + X_3 + X_4$	0.069	Id	$X_1 + X_2 + X_3 + X_4$	0.057
Dual	$X_1 + X_2 + X_3 + X_4$	0.050	Id	$X_1 + X_2 + X_4$	0.046	Log	$X_1 + X_2$	0.043
Dual	$X_1 + X_2 + X_3$	0.043	Id	$X_1 + X_2 + X_3$	0.045	Id	$X_1 + X_4$	0.035
Dual	$X_1 + X_2 + X_4$	0.041	Id	$X_1 + X_3 + X_4$	0.038	Dual	$X_1 + X_2 + X_3$	0.035
Dual	$X_1 + X_2$	0.033	Dual	$X_1 + X_2 + X_3$	0.027	Dual	$X_1 + X_2$	0.032
Mod	$X_1 + X_2 + X_3 + X_4$	0.031	Dual	$X_1 + X_2 + X_4$	0.026	Log	$X_1 + X_2 + X_4$	0.032

Από τον Πίνακα 5.10 αποδεικνύεται εμφανώς ότι οι μεταβλητές X_1 και X_2 έχουν κερδίσει τη θέση τους στο τελικό μοντέλο, καθώς οι ύστερες πιθανότητές τους είναι υψηλότερες από 92% και 70% αντίστοιχα λαμβάνοντας υπόψη όλες τις προσεγγίσεις. Οι υπόλοιπες δύο επεξηγηματικές μεταβλητές συνοδεύονται από χαμηλότερες ύστερες πιθανότητες ένταξης, μολονότι αυτές δεν πέφτουν κάτω του 40% ως επί το πλείστον. Για παράδειγμα, η μεταβλητή X_3 λαμβάνει τη χαμηλότερη ύστερη πιθανότητα ένταξης (περίπου 40%) για τον MIBF υπό τις πρότερες προσεγγίσεις Prior 1 και Prior 2. Από την άλλη μεριά, η ύστερη πιθανότητα ένταξης της μεταβλητής X_4 δεν πέφτει κάτω του 50% σε καμιά περίπτωση. Τις πιο χαμηλές πιθανότητες (γύρω στο 52%) τις δίνει ο FBF με $b = 1/n$ υπό τις προσεγγίσεις Prior 1 and Prior 2. Οι υψηλότερες τιμές των ύστερων πιθανοτήτων ένταξης των μεταβλητών X_3 και X_4 εμφανίζονται υπό την προσέγγιση Prior 3.

Τέλος, ο Πίνακας 5.11 με τις ύστερες περιθώριες πιθανότητες των οικογενειών μετασχηματισμών κάνει φανερό ότι ο Ταυτοτικός μετασχηματισμός κρίνεται ως ο πιο κατάλληλος κάτω από όλες τις προσεγγίσεις ακολουθούμενος από την οικογένεια YJ. Πιο αναλυτικά, βλέπουμε ότι πλέον τα αποτελέσματα ανάμεσα στις προσεγγίσεις Prior 2 και

Πίνακας 5.10: Εκτιμήσεις των ύστερων πιθανοτήτων ένταξης των μεταβλητών στο τελικό μοντέλο με βάση τον ενδογενή και τον κλασματικό παράγοντα Bayes για το σύνολο δεδομένων Hald ($m_0 = 7$).

Προσέγγιση	X_1	X_2	X_3	X_4
		Prior 1		
FBF, $b = 1/n$	0.976	0.781	0.458	0.519
FBF, $b = m_0/n$	0.928	0.751	0.577	0.651
MIBF	0.939	0.708	0.397	0.544
		Prior 2		
FBF, $b = 1/n$	0.976	0.782	0.458	0.518
FBF, $b = m_0/n$	0.927	0.752	0.578	0.651
MIBF	0.936	0.710	0.401	0.546
		Prior 3		
FBF, $b = 1/n$	0.978	0.815	0.599	0.637
FBF, $b = m_0/n$	0.948	0.837	0.745	0.781
MIBF	0.937	0.746	0.517	0.621

Prior 3 συγκλίνουν πολύ για κάθε εναλλακτικό παράγοντα Bayes και για κάθε οικογένεια, ενώ η ύστερη πιθανότητα του μετασχηματισμού Id κυμαίνεται από 31.1% (για τον FBF με χαμηλή τιμή του b υπό την Prior 2) έως 41.1% (για τον FBF με υψηλή τιμή του b υπό την Prior 2). Ο μετασχηματισμός YJ αναδεικνύεται κάπως περισσότερο υπό την προσέγγιση Prior 1 σε σχέση με τις άλλες δύο προσεγγίσεις με ύστερη περιθώρια πιθανότητα από 21.1% έως 25.2% για τον MIBF και τον FBF χαμηλού b αντίστοιχα, ενώ οι σχετικές τιμές για τον μετασχηματισμό Id είναι αρκετά κοντά και κυμαίνονται από 24.5% έως 33.2% για τον FBF υψηλού και χαμηλού b αντίστοιχα. Οι ύστερες πιθανότητες των οικογενειών BC και Modulus έχουν σχεδόν ταυτόσημες τιμές για κάθε προσέγγιση εναλλακτικού παράγοντα Bayes και για κάθε πρότερη προσέγγιση (Prior 1 - Prior 3), ενώ ο μετασχηματισμός Log τις ακολουθεί από κοντά.

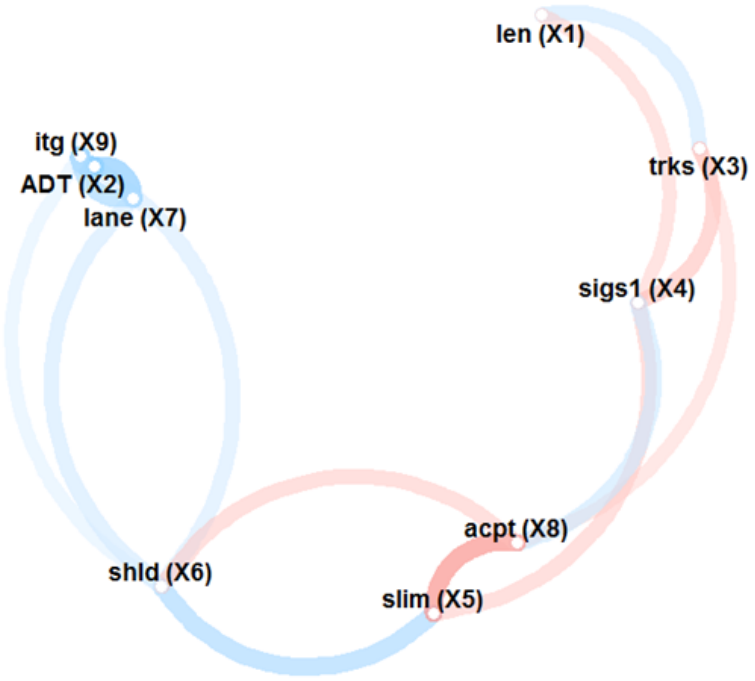
Πίνακας 5.11: Εκτιμήσεις των ύστερων περιθώριων πιθανοτήτων των οικογενειών μετασχηματισμών με βάση τον ενδογενή και τον κλασματικό παράγοντα Bayes για το σύνολο δεδομένων Hald ($m_0 = 7$).

Προσέγγιση	BC	Mod	YJ	Log	Id
			Prior 1		
FBF, $b = 1/n$	0.156	0.157	0.252	0.103	0.332
FBF, $b = m_0/n$	0.198	0.199	0.219	0.137	0.245
MIBF	0.159	0.161	0.211	0.156	0.312
			Prior 2		
FBF, $b = 1/n$	0.129	0.130	0.208	0.122	0.411
FBF, $b = m_0/n$	0.167	0.167	0.185	0.169	0.311
MIBF	0.123	0.125	0.170	0.189	0.379
			Prior 3		
FBF, $b = 1/n$	0.128	0.129	0.207	0.127	0.409
FBF, $b = m_0/n$	0.166	0.166	0.184	0.173	0.309
MIBF	0.129	0.130	0.171	0.189	0.379

5.2.3 Σύνολο δεδομένων Highway

Το τελευταίο παράδειγμα αναφέρεται επίσης σε ένα συχνά χρησιμοποιούμενο σύνολο δεδομένων, το σύνολο δεδομένων Highway (Weisberg 2014), αποτελούμενο από 39 παρατηρήσεις και 9 επεξηγηματικές μεταβλητές. Η μεταβλητή απόκρισης είναι το ποσοστό ατυχημάτων ανά 1000 μίλια όπως καταγράφηκε το έτος 1973 σε μεγάλες λεωφόρους της πολιτείας της Μινεσότα (Η.Π.Α.). Τα δεδομένα περιέχονται στη βιβλιοθήκη `car` της R, ενώ ο συμβολισμός και η περιγραφή των επεξηγηματικών μεταβλητών δίνεται στον Πίνακα 5.12. Κατά την ανάλυση του συγκεκριμένου συνόλου δεδομένων, πρωταρχικός στόχος είναι η κατανόηση της επίδρασης των μεταβλητών `sigs` (X_4), `slim` (X_5), `shld` (X_6) και `acpt` (X_8), που θεωρητικά τελούν υπό τον έλεγχο της διοίκησης του αυτοκινητόδρομου, στα ατυχήματα. Οι πιο έντονες συσχετίσεις ανευρίσκονται ανάμεσα στα ακόλουθα ζεύγη μεταβλητών: X_9 και X_2 ($r = 90\%$), X_7 και X_2 ($r = 82\%$), X_7 και X_9 ($r = 70\%$), X_5 και X_6 ($r = 69\%$), X_8 και X_5 ($r = -68\%$), X_4 και X_8 ($r = 51\%$), X_3 και X_1 ($r = 50\%$). Το Διάγραμμα 5.2 απεικονίζει το διάγραμμα δικτύου των επεξηγηματικών μεταβλητών με βάση τις τιμές δειγματικής συσχέτισης, όπου το κόκκινο και το μπλε χρώμα

δηλώνουν συσχετίσεις με θετικό και αρνητικό πρόσημο αντίστοιχα, ενώ χρωματικά πιο έντονες τροχιές και πιο κοντινοί κόμβοι δηλώνουν πιο έντονη συσχέτιση.



Διάγραμμα 5.2: Διάγραμμα δικτύου (network plot) των επεξηγηματικών μεταβλητών $X_j,\ j = 1, \ldots, 9$, των δεδομένων Highway με βάση τις τιμές δειγματικής συσχέτισης.

Σύμφωνα με τον Weisberg (2005), οι πιο σημαντικές επεξηγηματικές μεταβλητές για το σύνολο δεδομένων Highway είναι οι X_1, X_5, X_8, ενώ σύμφωνα με τους Hoeting, Raftery & Madigan (2002) οι επεξηγηματικές μεταβλητές που φαίνεται να έχουν σημασία για τη μοντελοποίηση της μεταβλητής απόκρισης (με βάση τον αλγόριθμο MC^3) περιλαμβάνονται στο σύνολο $\{X_1, X_4, X_5, X_8, X_9\}$. Τέλος, μικρή είναι η διαφοροποίηση μεταξύ του συνόλου $\{X_1, X_4, X_5, X_8\}$ και οποιουδήποτε υποσυνόλου αυτού κατά τους Thall, Russell & Simon (1997).

Τα αποτελέσματα υπό τις προσεγγίσεις Prior 1, Prior 2 και Prior 3 δίνονται στους Πίνακες 5.13–5.15. Καθώς η τιμή m αυξάνει από $m = 1$ σε $m = m_0 = 12$, τα ύστερα βάρη των κορυφαίων μοντέλων μειώνονται και όλα τα δέκα μοντέλα που παρουσιάζονται μοιάζουν σχεδόν ισοδύναμα, ιδιαίτερα αν κοιτάξει κανείς τα αποτελέσματα των δύο πρώτων προαναφερθέντων πινάκων. Αυτό είναι απόλυτα λογικό εξαιτίας του γεγονότος ότι υψηλότερη τιμή του m συνεπάγεται λιγότερη εναπομείνασα πληροφορία για τη διάκριση μεταξύ των υπό σύγκριση μοντέλων.

Πίνακας 5.12: Συμβολισμός, ονομασία και περιγραφή των επεξηγηματικών μεταβλητών που περιλαμβάνονται στο σύνολο δεδομένων Highway.

Σύμβολο	Όνομα	Περιγραφή
X_1	`len`	Μήκος του τμήματος του αυτοκινητόδρομου Highway1 σε μίλια
X_2	`ADT`	Μέση μέτρηση της ημερήσιας κυκλοφορίας σε χιλιάδες
X_3	`trks`	Όγκος φορτηγών ως ποσοστό του συνολικού όγκου
X_4	`sigs`	Πλήθος των σημάτων ανά μίλι αυτοκινητόδρομου (όχι μηδενικές τιμές): $\frac{\#\text{signalized interchanges per mile} \times \text{len} + 1}{\text{len}}$
X_5	`slim`	Όριο ταχύτητας για το έτος 1973
X_6	`shld`	Πλάτος, σε πόδια, της λωρίδας έκτακτης ανάγκης (ΛΕΑ) του δρόμου
X_7	`lane`	Συνολικό πλήθος λωρίδων κυκλοφορίας
X_8	`acpt`	Πλήθος σημείων πρόσβασης ανά μίλι
X_9	`itg`	Πλήθος κόμβων αυτοκινητόδρομου ανά μίλι

Υπό την Prior 1 (βλέπε Πίνακα 5.13), οι μετασχηματισμοί που υπερισχύουν είναι οι Log και Dual, με την οικογένεια BC να εμφανίζεται στην τελευταία θέση (με PMP γύρω στο 1.2%). Το βέλτιστο μοντέλο για όλες τις προσεγγίσεις BF είναι το λογαριθμικό, αν και υπάρχει διαφοροποίηση ως προς τις επιλεγόμενες συμμεταβλητές. Ο FBF χαμηλού b προτείνει ένα σύνολο τριών συμμεταβλητών: X_1, X_5 και X_8 με ύστερη πιθανότητα του μοντέλου ίση με 4.8%. Ο MIBF επίσης επιλέγει τρεις συμμεταβλητές, με τη διαφορά ότι έχει αντικαταστήσει τη μεταβλητή X_8 με τη μεταβλητή X_3 και το ύστερο βάρος του μοντέλου πέφτει σε σχέση με πριν στο 2.7%. Το βέλτιστο μοντέλο που παράγεται από τον FBF υψηλού b περιέχει την ένωση των δύο προηγούμενων συνόλων επεξηγηματικών μεταβλητών. Αξίζει να παρατηρηθεί ότι οι τιμές των ύστερων πιθανοτήτων των μοντέλων είναι πολύ μικρότερες από αυτές του συνόλου δεδομένων Hald, κάτι απολύτως αναμενόμενο καθώς το μέγεθος του μοντελοχώρου έχει διογκωθεί λόγω του διπλασιασμού του πλήθους των επεξηγηματικών του πλήρους μοντέλου. Ακόμη, οι ύστερες πιθανότητες των μοντέλων για τον FBF με $b = m_0/n$ είναι πολύ πιο κοντά στο μηδέν σε σχέση με τις άλλες δύο εναλλακτικές μορφές παραγόντων Bayes.

Ο Πίνακας 5.14 παρουσιάζει τα αποτελέσματα της επιλογής μοντέλου υπό την Prior 2. Η οικογένεια BC δεν εμφανίζεται πλέον στη λίστα των δέκα καλύτερων μοντέλων και, πέρα από τον μετασχηματισμό Log, η οικογένεια Dual παίζει ένα μικρό ρόλο στην επι-

Πίνακας 5.13: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο δεδομένων Highway υπό την Prior 1 ($m_0 = 12$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Log	$X_1 + X_5 + X_8$	0.048	Log	$X_1 + X_3 + X_5 + X_8$	0.007	Log	$X_1 + X_3 + X_5$	0.027
Log	$X_1 + X_4 + X_5$	0.035	Log	$X_1 + X_4 + X_5 + X_8$	0.006	Log	$X_1 + X_5$	0.025
Log	$X_1 + X_5$	0.032	Log	$X_1 + X_3 + X_4 + X_5 + X_8$	0.006	Log	$X_1 + X_4 + X_5$	0.025
Log	$X_1 + X_3 + X_5$	0.025	Log	$X_1 + X_5 + X_8$	0.005	Log	$X_1 + X_5 + X_8$	0.017
Log	$X_1 + X_3 + X_5 + X_8$	0.024	Dual	$X_1 + X_3 + X_5 + X_8$	0.005	Dual	$X_1 + X_4 + X_5$	0.015
Log	$X_1 + X_4 + X_5 + X_8$	0.022	Dual	$X_1 + X_4 + X_5 + X_8$	0.005	Dual	$X_1 + X_3 + X_5$	0.015
Dual	$X_1 + X_5 + X_8$	0.018	Log	$X_1 + X_2 + X_3 + X_5 + X_8$	0.005	Log	$X_1 + X_3 + X_4 + X_5$	0.014
BC	$X_1 + X_5$	0.015	Log	$X_1 + X_3 + X_4 + X_5$	0.004	Log	$X_1 + X_3 + X_5 + X_8$	0.014
Log	$X_1 + X_3 + X_4 + X_5$	0.014	Dual	$X_1 + X_3 + X_4 + X_5 + X_8$	0.004	Dual	$X_1 + X_5$	0.014
BC	$X_1 + X_5 + X_8$	0.012	Log	$X_1 + X_4 + X_5$	0.004	BC	$X_1 + X_5$	0.013

λογή μετασχηματισμού. Οι συμμεταβλητές και η οικογένεια των τεσσάρων καλύτερων μοντέλων για τις διάφορες προσεγγίσεις BF διατηρούνται αμετάβλητες συγκριτικά με τα αντίστοιχα αποτελέσματα του Πίνακα 5.13, μόνο που τώρα οι ύστερες πιθανότητές τους έχουν αυξηθεί. Για παράδειγμα, το βέλτιστο μοντέλο του FBF χαμηλού b έχει ύστερη πιθανότητα αυξημένη κατά σχεδόν 1% σε σχέση με πριν και το πρώτο μοντέλο που προτείνει ο MIBF έχει ύστερη πιθανότητα αυξημένη κατά 0.6%. Οι συμμεταβλητές που έχουν την πιο έντονη παρουσία σε αυτά τα αποτελέσματα είναι οι X_1, X_3, X_5, X_8 και ίσως και η X_4.

Στον Πίνακα 5.15, διαφαίνεται ένα διαφορετικό μοτίβο σε σχέση με τα αποτελέσματα των Ενοτήτων 5.2.2 και 5.2.1. Τουλάχιστον για τον FBF χαμηλού b και τον MIBF, τα καλύτερα μοντέλα είναι πιο φειδωλά συγκριτικά με τα αντίστοιχα της προσέγγισης Prior 2 και συνοδεύονται από υψηλότερες ύστερες πιθανότητες. Και οι δύο αυτοί παράγοντες προτείνουν το λογαριθμικό μοντέλο με τις μεταβλητές X_1 και X_5, με ύστερες πιθανότητες ίσες με 8.9% και 7.0% αντίστοιχα. Εν αντιθέσει, ο FBF υψηλού b είναι υπερβολικά ευέλικτος προτείνοντας ξανά το πλήρες μοντέλο ως βέλτιστο με ύστερο βάρος 6.6%, όπως συνέβει και στις προηγούμενες ενότητες.

Πίνακας 5.14: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο δεδομένων Highway υπό την Prior 2 ($m_0 = 12$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Log	$X_1 + X_5 + X_8$	0.057	Log	$X_1 + X_3 + X_5 + X_8$	0.008	Log	$X_1 + X_3 + X_5$	0.033
Log	$X_1 + X_4 + X_5$	0.042	Log	$X_1 + X_4 + X_5 + X_8$	0.008	Log	$X_1 + X_5$	0.031
Log	$X_1 + X_5$	0.038	Log	$X_1 + X_3 + X_4 + X_5 + X_8$	0.007	Log	$X_1 + X_4 + X_5$	0.030
Log	$X_1 + X_3 + X_5$	0.030	Log	$X_1 + X_5 + X_8$	0.007	Log	$X_1 + X_5 + X_8$	0.021
Log	$X_1 + X_3 + X_5 + X_8$	0.029	Log	$X_1 + X_2 + X_3 + X_5 + X_8$	0.006	Log	$X_1 + X_3 + X_4 + X_5$	0.018
Log	$X_1 + X_4 + X_5 + X_8$	0.026	Log	$X_1 + X_3 + X_4 + X_5$	0.006	Log	$X_1 + X_3 + X_5 + X_8$	0.017
Log	$X_1 + X_3 + X_4 + X_5$	0.016	Log	$X_1 + X_4 + X_5$	0.006	Dual	$X_1 + X_4 + X_5$	0.012
Dual	$X_1 + X_5 + X_8$	0.014	Log	$X_1 + X_3 + X_5 + X_7 + X_8$	0.005	Dual	$X_1 + X_3 + X_5$	0.012
Log	$X_1 + X_5 + X_7 + X_8$	0.013	Log	$X_1 + X_3 + X_5 + X_6 + X_8$	0.005	Dual	$X_1 + X_5$	0.011
Log	$X_1 + X_2 + X_5 + X_8$	0.012	Log	$X_1 + X_3 + X_5 + X_8 + X_9$	0.005	Log	$X_3 + X_4 + X_5$	0.011

Πίνακας 5.15: Τα δέκα εκ των υστέρων επικρατέστερα μοντέλα όπως αναδείχθηκαν από κάθε προσέγγιση (FBF, MIBF) συμπεριλαμβανομένης της οικογένειας μετασχηματισμών, των επιλεγμένων συμμεταβλητών και των ύστερων πιθανοτήτων των μοντέλων (PMP) για το σύνολο δεδομένων Highway υπό την Prior 3 ($m_0 = 12$).

FBF, $b = 1/n$			FBF, $b = m_0/n$			MIBF		
T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP	T	Συμμεταβλητές	PMP
Log	$X_1 + X_5$	0.089	Log	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$	0.066	Log	$X_1 + X_5$	0.070
Log	$X_1 + X_5 + X_8$	0.057	Dual	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$	0.031	Log	$X_1 + X_3 + X_5$	0.032
Log	$X_1 + X_4 + X_5$	0.043	BC	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$	0.025	Log	$X_1 + X_4 + X_5$	0.030
Log	$X_1 + X_3 + X_5$	0.030	Mod	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$	0.018	Dual	$X_1 + X_5$	0.026
BC	$X_1 + X_5$	0.028	Log	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_8 + X_9$	0.013	BC	$X_1 + X_5$	0.025
Log	$X_3 + X_5$	0.022	Log	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8$	0.013	Log	$X_3 + X_5$	0.024
Mod	$X_1 + X_5$	0.021	Log	$X_1 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$	0.012	Mod	$X_1 + X_5$	0.022
Log	$X_1 + X_3 + X_5 + X_8$	0.019	Log	$X_1 + X_2 + X_3 + X_4 + X_5 + X_7 + X_8 + X_9$	0.010	Log	$X_1 + X_5$	0.021
Log	$X_1 + X_4 + X_5 + X_8$	0.018	Log	$X_1 + X_2 + X_3 + X_5 + X_6 + X_7 + X_8 + X_9$	0.009	Log	X_5	0.015
Dual	$X_1 + X_5 + X_8$	0.014	Log	$X_1 + X_2 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9$	0.007	Log	$X_4 + X_5$	0.015

Σύμφωνα με τις εκτιμήσεις των ύστερων πιθανοτήτων ένταξης των μεταβλητών για το σύνολο δεδομένων Highway του Πίνακα 5.16, σημειώνεται ότι τα αποτελέσματα βασισμένα στον FBF ή MIBF μεταβάλλονται ελάχιστα καθώς αλλάζουμε την πρότερη κατανομή του μοντελοχώρου. Επομένως, υπάρχει ένας αξιοσημείωτος βαθμός σταθερότητας σχετικά με το βέλτιστο σύνολο μεταβλητών που αναδεικνύεται. Επιπρόσθετα, διακρίνονται μικρότερες ή μεγαλύτερες διαφορές μεταξύ των FBF και MIBF προσεγγίσεων και μάλιστα ως επί το πλείστον η απόκλιση μεταξύ του FBF με $b = 1/n$ και του MIBF είναι μικρότερη από την απόκλιση μεταξύ των δύο μορφών FBF (με τη χαμηλή και την υψηλή τιμή του b). Αυτό πιθανότατα οφείλεται στη μειωμένη αξιοπιστία που φέρουν τα αποτελέσματα τα σχετικά με τον FBF με κλασματική παράμετρο $b = \frac{m_0}{n}$. Ξεχωρίζοντας

μεταβλητές με ύστερη πιθανότητα ένταξης πάνω από 50%, το βέλτιστο σύνολο συμμεταβλητών για τα δεδομένα Highway φαίνεται να είναι το σύνολο $\{X_5, X_1, X_8\}$ σύμφωνα με τον FBF ($b = 1/n$) υπό τις Prior 1 και Prior 2 και το σύνολο $\{X_5, X_1\}$ σύμφωνα με τον MIBF, κάτω και από τις τρεις εναλλακτικές πρότερες κατανομές για τον μοντελοχώρο.

Πίνακας 5.16: Εκτιμήσεις των ύστερων πιθανοτήτων ένταξης των μεταβλητών στο τελικό μοντέλο με βάση τον ενδογενή και τον κλασματικό παράγοντα Bayes για το σύνολο δεδομένων Highway ($m_0 = 12$).

Προσέγγιση	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
					Prior 1				
FBF, $b = 1/n$	0.829	0.163	0.402	0.372	0.921	0.189	0.155	0.514	0.148
FBF, $b = m_0/n$	0.826	0.387	0.581	0.535	0.875	0.440	0.371	0.662	0.366
MIBF	0.765	0.207	0.475	0.336	0.924	0.200	0.173	0.289	0.190
					Prior 2				
FBF, $b = 1/n$	0.827	0.163	0.401	0.372	0.920	0.188	0.153	0.516	0.148
FBF, $b = m_0/n$	0.826	0.386	0.581	0.536	0.874	0.441	0.372	0.667	0.368
MIBF	0.766	0.208	0.477	0.338	0.925	0.201	0.174	0.298	0.192
					Prior 3				
FBF, $b = 1/n$	0.795	0.136	0.357	0.322	0.919	0.156	0.129	0.441	0.124
FBF, $b = m_0/n$	0.877	0.570	0.709	0.672	0.905	0.613	0.557	0.766	0.556
MIBF	0.701	0.158	0.398	0.281	0.919	0.152	0.132	0.243	0.144

Τέλος, με βάση τις ύστερες περιθώριες πιθανότητες των οικογενειών μετασχηματισμών (Πίνακας 5.17), ο μετασχηματισμός Log συγκεντρώνει την υψηλότερη ύστερη προτίμηση με τιμές να κυμαίνονται από 35% έως 62% και ακολουθούν οι οικογένειες YJ, BC και Mod με μικρές διαφορές στα αντίστοιχα ύστερα βάρη μεταξύ τους. Όπως και για τα αντίστοιχα αποτελέσματα της Ενότητας 5.2.2, τα αποτελέσματα για κάθε συγκεκριμένη οικογένεια και για κάθε συγκεκριμένη εναλλακτική μορφή παράγοντα Bayes διαφοροποιούνται ελάχιστα ανάμεσα στις προσεγγίσεις Prior 2 και Prior 3. Πιο λεπτομερώς, η μικρότερη απόσταση ανάμεσα στην οικογένεια YJ και στον μετασχηματισμό Log ($\sim$10 ποσοστιαίες μονάδες) παρατηρείται υπό την προσέγγιση Prior 1 και συγκεκριμένα για τον FBF υψηλού b, εκεί όπου η ύστερη περιθώρια πιθανότητα της οικογένειας YJ λαμβάνει την υψηλότερη τιμή της (24.8%). Η υψηλότερη ύστερη περιθώρια βαρύτητα για τον

μετασχηματισμό Log προκύπτει υπό την Prior 3 και τον FBF χαμηλού b και ισούται με 62%. Η ύστερη βαρύτητα του Ταυτοτικού μετασχηματισμού είναι κάτω από 2% σε κάθε περίπτωση.

Πίνακας 5.17: Εκτιμήσεις των ύστερων περιθώριων πιθανοτήτων των οικογενειών μετασχηματισμών με βάση τον ενδογενή και τον κλασματικό παράγοντα Bayes για το σύνολο δεδομένων Highway ($m_0 = 12$).

Προσέγγιση	BC	Mod	YJ	Log	Id
			Prior 1		
FBF, $b = 1/n$	0.163	0.127	0.175	0.523	0.002
FBF, $b = m_0/n$	0.221	0.164	0.248	0.350	0.008
MIBF	0.176	0.144	0.219	0.445	0.003
			Prior 2		
FBF, $b = 1/n$	0.134	0.105	0.131	0.617	0.003
FBF, $b = m_0/n$	0.183	0.133	0.213	0.449	0.011
MIBF	0.147	0.123	0.179	0.531	0.005
			Prior 3		
FBF, $b = 1/n$	0.128	0.099	0.138	0.620	0.003
FBF, $b = m_0/n$	0.185	0.138	0.209	0.446	0.012
MIBF	0.142	0.118	0.179	0.544	0.005

5.2.4 Μερικές πρόσθετες παρατηρήσεις

Αναφορικά με την τιμή της κλασματικής παραμέτρου b, δοκιμάστηκαν και κάποιες άλλες επιλογές, όπως $\frac{\log(n)}{n}$, $\frac{\sqrt{n}}{n}$, $\frac{m_0 \cdot log(n)}{log(m_0)n}$, $\frac{\sqrt{m_0 \cdot n}}{n}$, αλλά δεδομένων των περιορισμένων δειγματικών μεγεθών απέτυχαν, όπως βέβαια αναμενόταν αν κανείς είχε παρατηρήσει την εξέλιξη των αποτελεσμάτων περνώντας από την τιμή $b = 1/n$ στην τιμή $b = m_0/n$.

Σχετικά με τα παραδείγματα μονομεταβλητών μοντέλων, ένα βασικό παράδειγμα αναφοράς, το οποίο έχει παραλειφθεί στην παρούσα διατριβή, περιλαμβάνει την προσομοίωση 1000 σημείων από μια κατανομή $N(0, 1)$. Όλες οι προσεγγίσεις ορθώς αναγνώρισαν το μοντέλο Id ως βέλτιστο, ενεργώντας ως Μπεϋζιανοί έλεγχοι κανονικότητας. Ο MIBF παρήγαγε τα καλύτερα αποτελέσματα. Για τα δεδομένα της κατανομής Student με παχιές

ουρές, αναδείχθηκε η ανωτερότητα της οικογένειας Modulus κάτω από όλες τις προσεγγίσεις. Τα αποτελέσματα ενός συνόλου βιομηχανικών δεδομένων που αφορούν τον χρόνο μεταξύ βλαβών (*time between failures* - TBF, βλέπε Montgomery (2009)) δεν ήταν τόσο εύρωστα για διαφορετικές προσεγγίσεις όσο στα προηγούμενα παραδείγματα, αν και η σειρά των οικογενειών μετασχηματισμών διατηρείται και το μοντέλο Box-Cox αναδεικνύεται συνολικά ως το μοντέλο μέγιστης a posteriori πιθανότητας.

5.3 Συμπεράσματα

Η μεθοδολογία που εφαρμόσθηκε, σύμφωνα με το Κεφάλαιο 4, αποτελεί μια συγχώνευση της μεθοδολογίας Μπεϋζιανής επιλογής οικογένειας μετασχηματισμών με τη θεωρία του ενδογενούς και του κλασματικού παράγοντα Bayes με στόχο την κανονικότητα της κατανομής της μεταβλητής απόκρισης υπό την παρουσία επεξηγηματικών μεταβλητών στο μοντέλο. Όπως και στα Κεφάλαια 2 και 3, η επιλογή οικογένειας μετασχηματισμών αφορά τέσσερις παραμετρικές οικογένειες μετασχηματισμών (Box-Cox, Modulus, Yeo & Johnson και Dual) μαζί με τον Ταυτοτικό και τον Λογαριθμικό μετασχηματισμό.

Τα μέχρι τώρα συμπεράσματα συνιστούν ότι το ζήτημα πρότερης συμβατότητας στην επιλογή μετασχηματισμού αντιμετωπίζεται επαρκώς μέσω του παράγοντα Bayes πρότερης κατανομής δύναμης. Επιπρόσθετα, οι IBF και FBF προσφέρουν σημαντικά πλεονεκτήματα, όπως η δυνατότητα χρήσης μη γνήσιων πρότερων κατανομών που συχνά είναι αναπόφευκτες στο πρόβλημα επιλογής μετασχηματισμού με το οποίο ασχολούμαστε, εφόσον η εκμαίευση υποκειμενικής πρότερης πληροφορίας είναι μη τετριμμένη ή και αδύνατη.

Παρατηρώντας τα αποτελέσματα του παρόντος κεφαλαίου, ο FBF με κλασματική παράμετρο $b = \frac{1}{n}$ φαίνεται να συμπεριφέρεται καλύτερα από τον MIBF στα πλαίσια μη τετριμμένων μοντέλων και εξαιρετικά καλύτερα σε σχέση με τον FBF με κλασματική παράμετρο $b = \frac{m_0}{n}$.

Ο παράγοντας Bayes που βασίζεται στην προσέγγιση πρότερης κατανομής δύναμης του Κεφαλαίου 2 δεν είναι εύκολο να επεκταθεί σε πλαίσια με επεξηγηματικές μεταβλητές. Ως εκ τούτου, όταν $p \in \mathbb{N}^+$ χρησιμοποιούμε είτε τον ενδογενή είτε τον κλασματικό

παράγοντα Bayes. Τέλος, η σύγκριση πολλαπλών (μη εμφωλευμένων) μοντέλων καθίσταται εξαιρετικά πολύπλοκη μέσω του IBF, ενώ το υπολογιστικό κόστος είναι σημαντικά αυξημένο σε σχέση με την περίπτωση εφαρμογής του FBF.

Κεφάλαιο 6

Συζήτηση

6.1 Εισαγωγή

Η Μπεϋζιανή θεωρία επιλογής μοντέλου εν γένει θεωρείται περιοχή αιχμής στη σύγχρονη στατιστική έρευνα και ο αριθμός των ερευνητών που δραστηριοποιούνται σε αυτήν διαρκώς αυξάνει. Το γεγονός αυτό αποδεικνύει τη σπουδαιότητα που έχουν οι μεθοδολογίες που αναπτύσσονται στον τομέα αυτόν, καθώς πέρα από το ακαδημαϊκό ενδιαφέρον που συγκεντρώνουν, έχουν και πολλές εφαρμογές σε πραγματικά δεδομένα από ποικίλες επιστημονικές περιοχές και αντικείμενα. Η παρούσα διατριβή επιχείρησε να καλύψει σημαντικά κενά της σύγχρονης Μπεϋζιανής θεωρίας αναφορικά με την επιλογή μοντέλων και να επιλύσει ανοιχτά προβλήματα που έχουν τεθεί από συναφείς δημοσιεύσεις. Φιλοδοξούμε να ανοίξουμε νέες κατευθύνσεις στην έρευνα μέσα από τα αποτελέσματα που προέκυψαν τα οποία κρίθηκαν να έχουν σημαντική άξια για ανακοινώσεις σε διεθνή συνέδρια, αλλά και για δημοσιεύσεις σε υψηλού κύρους διεθνή επιστημονικά περιοδικά στατιστικής μεθοδολογίας.

6.2 Συζήτηση και Συμπεράσματα

Ο βασικός στόχος της έρευνας που παρουσιάστηκε στην παρούσα διατριβή ήταν να παρέχει μια συνεκτική μεθοδολογία Μπεϋζιανής συλλογιστικής με σκοπό τη συμπερασματολογία, σύγκριση και αξιολόγηση διαφορετικών οικογενειών μετασχηματισμών T που μετασχηματίζουν ένα σύνολο δεδομένων προς την κανονικότητα. Στην προτεινό-

μενη ενοποιημένη προσέγγιση θεωρούμε τέσσερις παραμετρικές οικογένειες μετασχηματισμών (Box-Cox, Modulus, Yeo & Johnson and Dual) και ακόμη τον Ταυτοτικό και τον Λογαριθμικό μετασχηματισμό. Η προτεινόμενη μεθοδολογία αναδεικνύει τη βέλτιστη επιλογή οικογένειας T και τη βέλτιστη επιλογή για την τιμή της παραμέτρου μετασχηματισμού λ_T μέσα από τη Μπεϋζιανή επιλογή μοντέλου με χρήση κατάλληλων MCMC αλγόριθμων ή αριθμητικών μεθόδων κατά περίπτωση.

Προς αυτή την κατεύθυνση, έχει γίνει ήδη φανερό ότι η κατασκευή λογικών πρότερων κατανομών για τις υπό μελέτη οικογένειες μετασχηματισμών είναι ένα θεμελιώδες ζήτημα εξαιτίας της διαφορετικής ερμηνείας της παραμέτρου λ_T μεταξύ των οικογενειών. Στην περίπτωση των μονομεταβλητών προβλημάτων (χωρίς επεξηγηματικές μεταβλητές), τα θέματα συμβατότητας που σχετίζονται με την επιλογή μετασχηματισμού αντιμετωπίστηκαν μέσω της πρότερης κατανομής δύναμης, αλλά και μέσω της χρήσης κοινών φανταστικών δεδομένων που προσομοιώθηκαν από το μοντέλο αναφοράς του Ταυτοτικού μετασχηματισμού. Ένα δεύτερο πλαίσιο πρότερης κατανομής, με τη μορφή κανονικής πρότερης κατανομής μοναδιαίας πληροφορίας για την παράμετρο λ_T (ή λογαριθμοκανονικής πρότερης κατανομής στην περίπτωση του Dual) χρησιμοποιήθηκε σαν μια εναλλακτική του πρώτου πλαισίου πρότερης κατανομής. Όποιο πλαίσιο πρότερης πληροφορίας και αν επιλέξει κανείς εκ των δύο προτεινόμενων, η όποια πρότερη πληροφορία υπάρχει σχετικά με την τιμή της παραμέτρου μετασχηματισμού λ_T (και της οικογένειας μετασχηματισμών) μπορεί να ενσωματωθεί στην προτεινόμενη μεθοδολογία μέσω της τεχνικής των φανταστικών δεδομένων όπως παρουσιάστηκε στην Ενότητα 3.3.1.

Η προτεινόμενη προσέγγιση αποφεύγει την εξάρτηση από τις παραμέτρους θέσης και κλίμακας επιλέγοντας έναν απλό τρόπο: την τυποποίηση του συνόλου των δεδομένων. Εναλλακτικά, θα μπορούσαμε να τοποθετήσουμε πρότερες κατανομές απευθείας στις παραμέτρους θέσης και κλίμακας των αρχικών μη μετασχηματισμένων δεδομένων, κάτι που στα πλαίσια της Μπεϋζιανης φιλοσοφίας είναι επιθυμητό. Παρόλα αυτά, μια τέτοια προσθήκη συνεπάγεται ότι ο μέσος και η διακύμανση των μετασχηματισμένων δεδομένων θα πρέπει με κάποιον τρόπο να προσεγγιστούν, π.χ. με χρήση της μεθόδου Taylor. Στη συνέχεια, MCMC αλγόριθμοι για το πλήρες διάνυσμα παραμέτρων του μοντέλου θα πρέπει να αναπτυχθούν για την εκτίμηση εκ των υστέρων κατανομών και περιθώριων κατανομών.

Επιπρόσθετα, το τι είδους πληροφορία χρειάζεται να εισάγει κανείς αναφορικά με τον μέσο και τη διακύμανση των μετασχηματισμένων δεδομένων είναι κάθε άλλο παρά αυτονόητο. Ακόμη και αν γινόταν εφικτό να αποσπάσουμε πληροφορία για τις υπερπαραμέτρους της προταθείσας NIG πρότερης κατανομής των μ_T και σ_T^2 για μια δεδομένη οικογένεια μετασχηματισμών T και της αντίστοιχης παραμέτρου λ_T, οι παραγόμενες NIG πρότερες κατανομές δε θα είναι αναγκαστικά συμβατές για διαφορετικές τιμές του ζεύγους (λ_T, T). Τονίζουμε, πάντως, ότι ο πυρήνας αυτής της ερευνητικής δουλειάς περιλαμβάνει το θέμα της επιλογής οικογένειας μετασχηματισμού στην περίπτωση όπου πρότερη πληροφορία για τα μ_T και σ_T^2 δεν είναι διαθέσιμη. Με βάση αυτή την παρατήρηση, θεωρείται ότι η επίδραση της συγκεκριμένης πρότερης κατανομής για το πρόβλημα της επιλογής μοντέλου είναι δευτερεύουσας σημασίας, καθώς οι παράμετροι μ_T και σ_T^2 αντιμετωπίζονται ως οχληρές παράμετροι και ολοκληρώνονται εκτός. Στην περίπτωση εφαρμογής μιας μη-υποκειμενικής πρότερης κατανομής, η εκ των υστέρων κατανομή του ζεύγους παραμέτρων μ_T και σ_T^2 δεδομένης της οικογένειας και της παραμέτρου μετασχηματισμού (T, λ_T) θα βασίζεται αποκλειστικά στην πιθανοφάνεια και θα επηρεάζεται ελάχιστα από τη συγκεκριμένη επιλογή πρότερης κατανομής και από τις τιμές των υπερπαραμέτρων αυτής. Συνεπώς, η προαναφερθείσα πρότερη ασυμβατότητα θα έχει αμελητέα επίδραση στην a posteriori επιλογή οικογένειας μετασχηματισμών.

Τα αποτελέσματα των δύο προσεγγίσεων οσον αφορά τις πρότερες κατανομές της παραμέτρου λ_T έτσι όπως παρουσιάζονται στο Κεφάλαιο 2 δείχνουν να συγκλίνουν σε πολύ μεγάλο βαθμό. Κάποιες αποκλίσεις που τυχόν παρατηρήθηκαν στα αποτελέσματα ανάμεσα στα δυο πλαίσια πρότερων κατανομών αφορούν κυρίως την περίπτωση της οικογένειας Dual και οφείλονται, τουλάχιστον εν μέρει, στην ιδιαίτερη φύση του εν λόγω μετασχηματισμού και των σημαντικά διαφορετικών χαρακτηριστικών του σε σχέση με τις υπόλοιπες οικογένειες υπό εξέταση, όπως π.χ. η έλλειψη ουδέτερης τιμής της παραμέτρου μετασχηματισμού και η συμμετρία γύρω από την τιμή $\lambda_T = 0$. Ασύμμετρα σύνολα δεδομένων, προερχόμενων από τη γάμμα κατανομή, φαίνεται ότι διορθώνονται καλύτερα μέσω της οικογένειας μετασχηματισμών Box-Cox, ενώ καθώς ο βαθμός ασυμμετρίας της κατανομής μειώνεται ο ρόλος της οικογένειας YJ αναβαθμίζεται όσον αφορά τον βέλτιστο μετασχηματισμό των δεδομένων. Κατανομές με παχιές ουρές, όπως είναι η Student

και η διπλή εκθετική κατανομή (ή αλλιώς κατανομή Laplace), μετασχηματίζονται αποδοτικά προσεγγίζοντας την κανονικότητα μέσω της οικογένειας Modulus. Συνοψίζοντας, είναι αναγκαίο να τονιστεί πως τα μέχρι τώρα εμπειρικά στοιχεία υποδεικνύουν ότι η επικράτηση του μετασχηματισμού Box-Cox στη σχετική βιβλιογραφία δεν είναι πάντα η ορθότερη επιλογή και ότι η επιλογή από ένα ευρύτερο σύνολο οικογενειών θα έπρεπε να γίνει κοινή πρακτική σε τέτοιου είδους προβλήματα.

Μια ακόμη πολύ ενδιαφέρουσα εφαρμογή της σχετικής μεθοδολογίας είναι η ακόλουθη: η προτεινόμενη διαδικασία της επιλογής μετασχηματισμού μπορεί να εξυπηρετήσει και τον σκοπό ενός ολοκληρωμένου Μπεϋζιανού ελέγχου κανονικότητας παρέχοντας μάλιστα μεγαλύτερη ευελιξία σε σχέση με τους κλασικούς ελέγχους της κατηγορίας αυτής (όπως οι έλεγχοι κανονικότητας Kolmogorov-Smirnov και Shapiro-Wilk). Αυτή η εφαρμογή γίνεται φανερή και μέσα από το παράδειγμα των κανονικά κατανεμημένων δεδομένων (Πίνακας 3.1) και τη συνακόλουθη υπεροχή του Ταυτοτικού μοντέλου Id, αλλά και από την ανάδειξη του εν λόγω μοντέλου στην περίπτωση που οι βαθμοί ελευθερίας μιας κεντρικής κατανομής Student αυξηθούν αρκετά ώστε αυτή να προσεγγίσει την κανονική κατανομή.

Όσον αφορά προβλήματα με επεξηγηματικές μεταβλητές, τα πράγματα διαφοροποιούνται αρκετά σε σχέση με τα μονομεταβλητά προβλήματα. Αφενός, τα φανταστικά δεδομένα, που έδωσαν σημαντική λύση στο θέμα της συμβατότητας των πρότερων κατανομών μεταξύ των οικογενειών σε περιπτώσεις μονομεταβλητών μοντέλων, δεν είναι άμεσης εκμετάλλευσης πλέον. Ο λόγος είναι ότι η παρουσία επεξηγηματικών μεταβλητών καθιστά εξαιρετικά δύσκολο τον πλήρη προσδιορισμό ενός μοντέλου αναφοράς από το οποίο να προσομοιώσουμε τα φανταστικά δεδομένα, με τρόπο ανάλογο της μονομεταβλητής περίπτωσης. Αφετέρου, όπως και στα μονομεταβλητά προβλήματα, το ζητούμενο είναι να προταθούν λύσεις στην περίπτωση όπου η πρότερη πληροφορία σχετικά με τους μετασχηματισμούς ουσιαστικά απουσιάζει. Εκ των πραγμάτων, ο κλασικός παράγοντας Bayes αδυνατεί να ανταπεξέλθει σε πλαίσια πρότερων κατανομών ανοικτού τύπου, όπως είναι οι μη γνήσιες μη πληροφοριακές πρότερες κατανομές, καθώς οι άγνωστες σταθερές που μεταφέρονται στον τελικό τύπο καθιστούν αδύνατο τον υπολογισμό του. Αναζητήθηκαν, επομένως, εναλλακτικές μορφές παραγόντων Bayes στη βιβλιογραφία και

αφού αναλύθηκαν τα πλεονεκτήματα και τα μειονεκτήματα κάθε προσέγγισης ξεχωριστά αλλά και ειδικότερα σε σχέση με το πλαίσιο του υπό μελέτη προβλήματος, δηλαδή της σύγκρισης πολλαπλών μη εμφωλευμένων μοντέλων μετασχηματισμών, καταλήξαμε στον διάμεσο ενδογενή παράγοντα Bayes και στον κλασματικό παράγοντα Bayes. Έτσι, τα φανταστικά δεδομένα έδωσαν τη θέση τους στα δεδομένα εκπαίδευσης του ενδογενούς παράγοντα Bayes και στην κλασματική παράμετρο εκπαίδευσης b του κλασματικού παράγοντα Bayes. Επιπλέον, το ερευνητικό πλαίσιο εμπλουτίστηκε ενσωματώνοντας τη διαδικασία επιλογής επεξηγηματικών μεταβλητών παράλληλα με την επιλογή οικογένειας μετασχηματισμών.

Σαν παράδειγμα αναφοράς, προσομοιώθηκαν δεδομένα από ένα κανονικό μοντέλο πολλαπλής γραμμικής παλινδρόμησης. Όλες οι προσεγγίσεις που εφαρμόστηκαν ανέδειξαν πρώτο τον Ταυτοτικό μετασχηματισμό, αλλά δεν επέδειξαν το ίδιο καλή συμπεριφορά όσον αφορά την επιλογή μεταβλητών. Ο ενδογενής παράγοντας Bayes φάνηκε να έχει την πιο σταθερή συμπεριφορά ακολουθούμενος από τον κλασματικό παράγοντα Bayes με την χαμηλή τιμή της παραμέτρου b. Στην περίπτωση των δεδομένων Hald, το θέμα της πολυσυγγραμικότητας είναι παρόν και τα δεδομένα σχεδόν αγγίζουν το 'small n - large p' πρόβλημα λόγω πολύ περιορισμένου μεγέθους δείγματος (και όχι λόγω μεγάλου αριθμού συμμεταβλητών). Εδώ, παρατηρήθηκε ότι ο κλασματικός παράγοντας Bayes υπό τη χαμηλή τιμή της παραμέτρου b υπερείχε κάπως σε σχέση με τον διάμεσο IBF, αν και οι δύο επέδειξαν ορθή συμπεριφορά στο θέμα της επιλογής μεταβλητών παρά την έντονη πολυσυγγραμικότητα του πλήρους μοντέλου. Τέλος, στο παράδειγμα των δεδομένων Highway με το αυξημένο πλήθος επεξηγηματικών μεταβλητών, όλες οι προσεγγίσεις ανέδειξαν την ίδια οικογένεια μετασχηματισμών σαν βέλτιστη, αν και όχι απαραίτητα το ίδιο σύνολο επεξηγηματικών μεταβλητών.

Σε όλες τις εφαρμογές, η υψηλή τιμή που δοκιμάστηκε για την κλασματική παράμετρο b του κλασματικού παράγοντα Bayes, σχετιζόμενη και με το ελάχιστο μέγεθος του δείγματος εκπαίδευσης, έχει την τάση να αναδεικνύει πολύπλοκα μοντέλα με μεγάλο πλήθος επεξηγηματικών μεταβλητών ενάντια στην αρχή της φειδωλότητας. Συνεπώς, το αν οδηγεί σε σωστή διαφοροποίηση των υπό σύγκριση μοντέλων, ιδιαιτέρως υπό την παρουσία πολυσυγγραμικότητας, παραμένει ένα ερώτημα.

Όσον αφορά στον μοντελοχώρο, η πρότερη κατανομή Prior 2 με ομοιόμορφη πρότερη πιθανότητα στον χώρο των μεταβλητών και διαφορετική πρότερη πιθανότητα στον χώρο των οικογενειών μετασχηματισμών, με βάση το αν αυτές είναι παραμετρικές ή όχι, φάνηκε να δίνει γενικά τα καλύτερα αποτελέσματα. Η προσέγγιση Prior 3 με τη βήτα-διωνυμική πρότερη κατανομή για τον χώρο των μεταβλητών μάλλον απορρίπτεται σε κάθε περίπτωση, αν και όσο μεγαλώνει το μέγεθος του δείγματος τόσο φαίνεται να μειώνεται η επίδραση της πρότερης κατανομής του μοντελοχώρου και να παρατηρείται μια σχετική σύγκλιση των αποτελεσμάτων των διαφορετικών προσεγγίσεων.

Μέσα από τη διερεύνηση και την εφαρμογή των εναλλακτικών παραγόντων Bayes, κατέστη δυνατό να εξάγουμε σημαντικά συμπεράσματα τα οποία δεν ήταν εμφανή μέσα από τη βιβλιογραφική ανασκόπηση, αλλά παρόλα αυτά επηρεάζουν άμεσα την καταλληλότητα και την αξιοπιστία κάθε εναλλακτικής μορφής BF ανάλογα με το εκάστοτε ερευνητικό πλαίσιο. Αρχικά, παρατηρήθηκε ότι ο αριθμητικός ενδογενής παράγοντας Bayes είναι αρκετά δύστροπος από πολλές απόψεις. Αν και στη βιβλιογραφία προτείνεται συχνά έναντι άλλων μορφών BF λόγω της άμεσης σχέσης του με τις ενδογενείς πρότερες κατανομές, εντούτοις η εφαρμογή του μπορεί ανά περίπτωση να παράγει παραπλανητικά αποτελέσματα, πέραν του ότι είναι υπολογιστικά ασύμφορος για μέτριο ή μεγάλο πλήθος επεξηγηματικών μεταβλητών. Μια περίπτωση στην οποία τα αποτελέσματα του AIBF μπορεί να μην είναι αξιόπιστα είναι όταν έχουμε πολλαπλά μη εμφωλευμένα μοντέλα υπό σύγκριση, καθώς η κατάταξη των μοντέλων από το πιο πολύπλοκο στο λιγότερο πολύπλοκο δεν είναι συχνά εφικτή και, επομένως, δε μπορεί να εφαρμοστεί ο βασικός κανόνας κατασκευής του AIBF.

Ο γεωμετρικός ενδογενής παράγοντας Bayes είναι σημαντικά βελτιωμένος σε σχέση με τον αριθμητικό στην περίπτωση πολλών ακραίων τιμών με βάση τα δείγματα εκπαίδευσης, αλλά στην περίπτωσή μας και πάλι υστερούσε. Μόνο ο διάμεσος ενδογενής παράγοντας Bayes έδειξε να συμπεριφέρεται σωστά και για αυτόν τον λόγο προτιμήθηκε στις εφαρμογές που παρουσιάστηκαν. Οι Berger και Pericchi στα τελευταία τους επιστημονικά άρθρα πάνω στο ζήτημα δείχνουν να το αναγνωρίζουν εν μέρει, ειδικά για τις περιπτώσεις όπου έχουμε μικρό δείγμα και πολλά υπό σύγκριση μοντέλα.

Τέλος, ένα ζήτημα που αφορά όλες τις μορφές ενδογενών παραγόντων Bayes είναι

ο τρόπος επιλογής των δειγμάτων εκπαίδευσης. Στην περίπτωση μικρού μεγέθους του ολικού δείγματος, το πλήθος των δυνατών δειγμάτων εκπαίδευσης (συγκεκριμένου μεγέθους) είναι περιορισμένο και άρα η τιμή του IBF ενδέχεται να επηρεαστεί σημαντικά από ακραίες ενδιάμεσες τιμές. Αντίθετα, όταν τα δυνατά δείγματα εκπαίδευσης είναι πολύ μεγάλου πλήθους, τότε το υπολογιστικό κόστος αυξάνει σημαντικά και είναι συνήθης πρακτική να χρησιμοποιείται τελικά μόνο ένα υποσύνολο των δειγμάτων εκπαίδευσης. Ακόμη όμως και η επιλογή αυτού του υποσυνόλου πρέπει να γίνεται με προσοχή. Με τα ανωτέρω ζητήματα συνδέεται άμεσα και η ύπαρξη ποιοτικών επεξηγηματικών μεταβλητών/παραγόντων. Όσο πιο λίγες κατηγορίες έχει μια ποιοτική μεταβλητή και όσο πιο λίγες ποσοτικές συμμεταβλητές συμπεριλαμβάνονται στο μοντέλο, τόσο πιο πιθανό είναι να προκύψουν ιδιάζοντες πίνακες σχετιζόμενοι με κάποια δείγματα εκπαίδευσης, οι οποίοι εμποδίζουν τον υπολογισμό του IBF και άρα θα πρέπει να εντοπιστούν και να αποκλειστούν. Το συγκεκριμένο θέμα της διαχείρισης των ποιοτικών μεταβλητών αξίζει σίγουρα περαιτέρω συστηματική διερεύνηση.

Από την άλλη μεριά, ο κλασματικός παράγοντας Bayes έχει χαμηλό υπολογιστικό κόστος ακόμα και όταν συμπεριλαμβάνει αρκετές δεκάδες επεξηγηματικών μεταβλητών, χαρακτηριστικό που τον καθιστά άκρως ελκυστικό σε περιπτώσεις μεγάλου p. Ακόμη, το γεγονός ότι δε χρησιμοποιεί δείγματα εκπαίδευσης τον κάνει ανθεκτικό σε ακραίες περιπτώσεις και δεν έχει ανάγκη από τόσο μεγάλο δείγμα για να δώσει ικανοποιητικά αποτελέσματα. Παρόλα αυτά, ακόμη και ο κλασματικός παράγοντας Bayes χρειάζεται κάποιο ελάχιστο μέγεθος δείγματος για να λειτουργήσει σωστά, καθώς η εμπλεκόμενη κλασματική παράμετρος b, που εν μέρει αντικαθιστά και τη λειτουργία των δειγμάτων εκπαίδευσης, ουσιαστικά χωρίζει τη δειγματική πληροφορία σε δύο μέρη: ένα πρώτο μέρος για την εκπαίδευση του μοντέλου (b) και ένα δεύτερο για τη σύγκριση και διαφοροποίηση μεταξύ των μοντέλων ($1-b$). Η ίδια η επιλογή της τιμής της παραμέτρου b δεν ακολουθεί συγκεκριμένες κατευθυντήριες γραμμές και συνιστάται να δοκιμάζονται διάφορες επιλογές πριν καταλήξει κανείς στο ποια είναι η ιδανική για το εκάστοτε ερευνητικό πλαίσιο στο οποίο δρα. Μια καλή πρακτική είναι να επιλέγει κανείς μικρές τιμές της κλασματικής παραμέτρου b, αλλά να αυξάνει την τιμή με την αύξηση του μεγέθους του δείγματος. Δεν πρέπει βεβαίως να ξεχνάμε ότι ο κλασματικός παράγοντας Bayes δε στηρίζεται σε αμιγώς

Μπεϋζιανή λογική και έτσι ο τρόπος κατασκευής του κρίνεται αμφίβολος. Πολλοί είναι οι ερευνητές εκείνοι που τον απορρίπτουν εξολοκλήρου ως μη Μπεϋζιανό εργαλείο. Εν γένει, η Μπεϋζιανή κοινότητα δείχνει προτίμηση στον ενδογενή παράγοντα Bayes καθώς στηρίζεται στη λογική της τεχνικής διασταυρούμενης επικύρωσης, αλλά κυρίως λόγω της σύνδεσής του με καθαρά Μπεϋζιανές μεθόδους όπως είναι οι ενδογενείς πρότερες κατανομές.

6.3 Μελλοντική Έρευνα

Η έρευνα της παρούσας διατριβής οδήγησε σε πρωτότυπα αποτελέσματα τα οποία, με τη σειρά τους, συμβάλλουν στην καλύτερη κατανόηση του προβλήματος της Μπεϋζιανής επιλογής μετασχηματισμού και ανοίγουν νέες κατευθύνσεις έρευνας στην περιοχή.

Αναμφισβήτητα, η έρευνα στο συγκεκριμένο πεδίο έχει αρκετό μέλλον. Ιδιαίτερη έμφαση μπορεί να δοθεί στην ταυτόχρονη διαχείριση των προβλημάτων της Μπεϋζιανής επιλογής μετασχηματισμού, επιλογής μεταβλητών και ανίχνευσης ακραίων τιμών, με βάση την πολύ ενδιαφέρουσα δουλειά των Hoeting, Raftery & Madigan (2002) και των Gottardo & Raftery (2009) η οποία αναφέρθηκε εκτενώς στο Κεφάλαιο 1.

Ένα θέμα που έχει αποσιωπηθεί αρκετά, με την έννοια ότι η έλλειψη σχετικών ερευνητικών άρθρων στη διεθνή βιβλιογραφία είναι εκκωφαντική, αλλά τελευταία αρχίζει να συγκεντρώνει αυξημένο ενδιαφέρον για πολλούς ερευνητές είναι η βέλτιστη εκτίμηση της παραμέτρου μετατόπισης ξ, και πιο συγκεκριμένα του ϵ (όπου $\xi = |\min(\boldsymbol{y})| + \epsilon$). Η χρήση της παραμέτρου ϵ είναι αναγκαία για τους μετασχηματισμούς Box-Cox, Dual και Log στην περίπτωση μη θετικών παρατηρήσεων στο δείγμα ώστε το σύνολο των μη μετασχηματισμένων παρατηρήσεων τελικά να υπάγεται στον αυστηρά θετικό άξονα. Μια απλή ανάλυση ευαισθησίας θα συνίστατο στο να θεωρήσει κανείς ένα πλήθος δυνατών τιμών $\epsilon_k, k \in \mathbb{N}$ από μια αυστηρά θετική ομοιόμορφη κατανομή μεγάλης διασποράς και να καταγράψει πώς διαφοροποιούνται οι τιμές του λ_T σύμφωνα με τις τιμές ϵ_k. Αυτή η διαδικασία εφαρμόστηκε ως ένα βαθμό δείχνοντας ότι σημαντική αύξηση της παραμέτρου μετατόπισης έχει ως αποτέλεσμα την αύξηση της τιμής του λ_T. Παρά ταύτα, είναι σημαντικό να γίνει μια εις βάθος διερεύνηση του συγκεκριμένου ζητήματος, ιδιαίτερα αν

τα δεδομένα του προβλήματος παίρνουν τιμές πολύ κοντά στο μηδέν.

Μια ενδιαφέρουσα προσέγγιση για να επιλυθεί, μεταξύ άλλων, το θέμα της ασυμβατότητας των πρότερων κατανομών μεταξύ των οικογενειών μετασχηματισμών θα ήταν, αντί του ενδογενούς παράγοντα Bayes, να χρησιμοποιηθεί η ιδέα των Pérez & Berger (2002) αναφορικά με τη μεταγενέστερα-αναμενόμενη κατανομή πρότερης πληροφορίας (*expected posterior prior*) για την κατασκευή του παράγοντα Bayes. Η μέθοδος αυτή υποδεικνύει έναν νέο τρόπο κατασκευής πρότερων κατανομών για τις παραμέτρους ενός μοντέλου σε μια παρεμφερή λογική με αυτή της ιεραρχικής δόμησης. Στη βάση εκπαίδευσης των μοντέλων με ένα κοινό σύνολο φανταστικών δεδομένων, αποφεύγεται η προβληματική διαδικασία του κατακερματισμού του συνόλου δεδομένων σε πολλαπλά υποσύνολα εκπαίδευσης. Επιπρόσθετα, η πρότερη κατανομή που δημιουργείται έχει εύκολη ερμηνεία σε αντίθεση με τις ενδογενείς και τις κλασματικές ενδογενείς πρότερες κατανομές. Επιγραμματικά, η χρήση της μεταγενέστερα-αναμενόμενης κατανομής πρότερης πληροφορίας αποτελεί μια πλήρως Μπεϋζιανή μέθοδο η οποία έχει αποδειχθεί ότι ασυμπτωτικά τείνει να δώσει ίδια αποτελέσματα με τον ενδογενή παράγοντα Bayes. Παρόλα αυτά, αποτελεί μια μέθοδο που στηρίζεται σε πολύπλοκους υπολογισμούς και ίσως αυτός είναι και ο λόγος που δεν έχει χρησιμοποιηθεί όσο οι δυνατότητές της θα επέτρεπαν.

Σαν τελική παρατήρηση, έχοντας ως στόχο να ελαφρύνουμε το υπολογιστικό κόστος που συνδέεται με τον ενδογενή παράγοντα Bayes παρουσία πολλών επεξηγηματικών μεταβλητών στο μοντέλο, θα ήταν ιδιαίτερα επωφελές να επιστρατευτούν πιο ευέλικτες αλγοριθμικές τεχνικές εξερεύνησης του μοντελοχώρου, όπως ο αλγόριθμος MC^3 (Madigan & York 1995). Ακόμη πιο προηγμένοι αλγόριθμοι θα πρέπει να εφαρμοστούν για προβλήματα μεγάλης διάστασης, πιθανότατα αξιοποιώντας την έννοια του παράλληλου προγραμματισμού του οποίου η υλοποίηση ευνοείται στο περιβάλλον της R. Προς αυτή την κατεύθυνση, ένας αλγόριθμος στοχαστικής αναζήτησης ενδεχομένως να είχε θεμελιώδη επίδραση στη δραστική μείωση του χρόνου επεξεργασίας, όπως η προσέγγιση *shotgun stochastic search* (SSS) που ανέπτυξαν οι Hans, Dobra & West (2007).

Βιβλιογραφικές Αναφορές

Aitkin, M. (1991), 'Posterior Bayes factors', *Journal of the Royal Statistical Society B,* **53**, 111–142.

Attkinson, A. C. (1982*a*), 'Diagnostic regression analysis and shifted power transformation', *Technometrics,* **25**, 23–33.

Attkinson, A. C. (1982*b*), 'Regression diagnostics, transformation and constructed variables', *Journal of the American Statistical AssociationSeries B,* **44**, 1–36.

Attkinson, A. C. (1985), *Plots, Transformation and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford: Clarendon Press.

Attkinson, A. C. (1986), 'Diagnostic tests for transformation', *Technometrics,* **28**, 29–37.

Bartlett, M. S. (1947), 'The use of transformations', *Biometrics,* **3**, 39–52.

Bartlett, M. S. (1957), 'Comment on D.V. Lindley's statistical paradox', *Biometrika,* **44**, 533–534.

Berger, J. O. & Pericchi, L. R. (1996*a*), The intrinsic Bayes factor for linear models, In J. Bernardo, J. Berger, A. Dawid, and A. Smith (eds.), *Bayesian Statistics*, Vol. 5, Oxford University Press, pp. 25–44.

Berger, J. O. & Pericchi, L. R. (1996*b*), 'The intrinsic Bayes factor for model selection and prediction', *Journal of the American Statistical Association,* **91**, 109–122.

Berger, J. O. & Pericchi, L. R. (1998), 'Accurate and stable Bayesian model selection: The median intrinsic Bayes factor', *Sankhȳa B,* **60**, 1–18.

Berger, J. O. & Pericchi, L. R. (2001), Objective Bayesian methods for model selection: Introduction and comparison, *in* P. Lahiri, ed., 'Model selection', Vol. 38 of *Lecture Notes–Monograph Series*, Institute of Mathematical Statistics, pp. 135–207.

Berger, J. O. & Pericchi, L. R. (2004), 'Training samples in objective model selection', *Annals of Statistics,* **32**, 841–869.

Bertolino, F. & Racugno, W. (1996), 'Is the intrinsic Bayes factor instrinsic?', *Metron,* **56**, 5–15.

Bickel, P. J. & Docksum, A. (1981), 'An analysis of transformations revisited', *Journal of the American Statistical Association,* **76**, 296–311.

Box, G. E. P. & Cox, D. R. (1964), 'An analysis of transformations (with discussion)', *Journal of the Royal Statistical Society Series B,* **26**, 211–252.

Box, G. E. P. & Cox, D. R. (1982), 'An analysis of transformation revisited, rebutted', *Journal of the American Statistical Association,* **77**, 209–210.

Box, G. E. P. & Tidwell, P. W. (1962), 'Transformation of the independent variables', *Technometrics,* **4**, 531–550.

Cano, J. A., Kessler, M. & Moreno, E. (2004), 'On intrinsic priors for nonnested models', *Test,* **13**, 445–463.

Carroll, R. J. & Ruppert, D. (1981), 'On prediction and the power transformation family', *Biometrika,* **79**, 321–328.

Casella, G. & Moreno, E. (2006), 'Objective Bayesian variable selection', *Journal of the American Statistical Association,* **101**, 157–167.

Casella, G., Moreno, E. & Girón, F. J. (2014), 'Cluster analysis, model selection, and prior distributions on models', *Bayesian Analysis,* **9**, 613–658.

Charitidou, E., Fouskakis, D. & Ntzoufras, I. (2015), 'Bayesian transformation family selection: Moving toward a transformed Gaussian universe', *The Canadian Journal of Statistics,* **43**, 600–623.

Chib, S. (1995), 'Marginal likelihood from the Gibbs output', *Journal of the American Statistical Association,* **90**, 1313–1321.

Chib, S. & Jeliazkov, I. (2001), 'Marginal likelihood from the Metropolis-Hastings output', *Journal of the American Statistical Association,* **96**, 270–281.

Consonni, G. & La Rocca, L. (2008), 'Tests based on intrinsic priors for the equality of two correlated proportions', *Journal of the American Statistical Association,* **103**, 1260–1269.

Consonni, G. & Veronese, P. (2008), 'Compatibility of prior specifications accross linear models', *Statistical Science,* **23**, 332–363.

Cook, R. D. & Wang, P. C. (1983), 'Transformation and influential cases in regression', *Technometrics,* **25**, 337–343.

Cui, W. & George, E. I. (2008), 'Empirical Bayes vs. fully Bayes variable selection', *Journal of Statistical Planning and Inference,* **138**, 888 – 900.

Dawid, A. P. & Lauritzen, S. L. (2000), Compatible prior distributions, *in* George, E. I., ed., 'Bayesian methods with applications to science, policy and official statistics, ISBA 2000', Eurostat (Statistical Office of the European Communities), pp. 109–118.

De Oliveira, V., Kedem, B. & Short, D. A. (1997), 'Bayesian prediction of transformed Gaussian random fields', *Journal of the American Statistical Association,* **92**, 1422–1433.

De Santis, F. & Spezzaferri, F. (1997), 'Alternative Bayes factors for model selection', *Canadian Journal of Statistics,* **25**, 503–515.

De Santis, F. & Spezzaferri, F. (1999), 'Methods for default and robust Bayesian model comparison: The fractional Bayes factor approach', *International Statistical Review,* **67**, 267–286.

Dellaportas, P., Forster, J. J. & Ntzoufras, I. (2012), 'Joint specification of model space and parameter space prior distributions', *Statistical Science,* **27**, 232–246.

Fan, T. H., Wang, W. L. & Balakrishnan, N. (2008), 'Exponential progressive step-stress life-testing with link function based on Box–Cox transformation', *Journal of statistical planning and inference,* **138**, 2340–2354.

Freni, G. & Mannina, G. (2010), 'Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution', *Journal of Hydrology,* **392**, 31–39.

Friel, N. & Wyse, J. (2012), "Estimating the model evidence: A review", *Statistica Neerlandica,* **66**, 288–308.

Good, I. J. (1985), Weight of evidence: A brief survey, In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds.), *Bayesian statistics 2*, Amsterdam: Elsevier Science Publishers, pp. 249–270.

Gottardo, R. & Raftery, A. E. (2007), *Bayesian robust variable and transformation selection: A unified approach*, Technical Report no. 508, Department of Statistics, University of Washington.

Gottardo, R. & Raftery, A. E. (2009), 'Bayesian robust variable and transformation selection: A unified approach', *Canadian Journal of Statistics,* **37**, 1–20.

Hald, A. (1952), *Statistical theory with engineering applications*, New York: Wiley.

Hans, C., Dobra, A. & West, M. (2007), 'Shotgun stochastic search for "large p" regression', *Journal of the American Statistical Association,* **102**, 507–516.

Hinkley, D. V. & Runger, G. (1984), 'The analysis of transformed data (with comments)', *Journal of the American Statistical Association,* **79**, 302–320.

Hoeting, J. A. & Ibrahim, J. G. (1998), 'Bayesian predictive simultaneously variable and transformation selection in the linear model', *Journal of Computational Statistics and Data Analysis,* **28**, 87–103.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999), 'Bayesian model averaging: A tutorial', *Statistical Science,* **14**, 382–417.

Hoeting, J. A., Raftery, A. E. & Madigan, D. (2002), 'A method for simultaneous variable and transformation selection in linear regression', *Journal of Computational and Graphical Statistics,* **11**, 485–507.

Ibrahim, J. G. & Chen, M. H. (2000), 'Power prior distributions for regression models', *Statistical Science,* **15**, 46–60.

Ibrahim, J. G., Chen, M. H. & Sinha, D. (2003), 'On optimality properties of the power prior', *Journal of the American Statistical Association,* **98**, 204–213.

Jeffreys, H. (1935), 'Some tests of significance, treated by the theory of probability', *Proccedings of the Cambridge Philosophy society,* **31**, 203–222.

Jeffreys, H. (1961), *Theory of probability (3rd ed.)*, Oxford: Oxford University Press.

John, J. A. & Draper, N. R. (1980), 'An alternative family of transformations', *Applied Statistics,* **29**, 190–197.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association,* **90**, 773–795.

Kass, R. E. & Wasserman, L. (1992), *A reference Bayesian Test for Nested Hypotheses with large samples*, Technical Report 567, Department of Statistics, Carnegie Mellon University, Pittsburgh.

Kass, R. E. & Wasserman, L. (1995), 'A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion', *Journal of the American Statistical Association,* **90**, 928–934.

Kim, S., Chen, M. H., Ibrahim, J. G., Shah, A. K. & Lin, J. (2013), 'Bayesian inference for multivariate meta-analysis Box-Cox transformation models for individual patient data with applications to evaluation of cholesterol-lowering drugs', *Statistics in Medicine,* **32**, 3972–3990.

Kim, S. W. (2000), 'Intrinsic priors for testing exponential means', *Statistics & Probability Letters,* **46**, 195 – 201.

Klein Entink, R. H., van der Linden, W. J. & Fox, J. P. (2009), 'A Box-Cox normal model for response times', *British journal of mathematical and statistical psychology,* **62**, 621–640.

Kronrod, A. S. (1965), *Nodes and weights of quadrature formulas. Sixteen-place tables*, Consultants Bureau (Authorized translation from the Russian), New York.

Kuo, L. & Mallick, B. (1998), "Variable selection for regression models", *Sankhyā B,* **60**, 65–81.

Laud, P. W. & Ibrahim, J. G. (1995), 'Predictive model selection', *Journal of the Royal Statistical Society B,* **57**, 247–262.

Lee, J. C., Lin, T. I., Lee, K. J. & Hsu, Y. L. (2005), 'Bayesian analysis of Box–Cox transformed linear mixed models with ARMA (p, q) dependence', *Journal of statistical Planning and Inference,* **133**, 435–451.

Lempers, F. B. (1971), *Posterior probabilities of alternative linear models*, University Press, Rotterdam.

Lewis, S. M. & Raftery, A. E. (1997), 'Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator', *Journal of the American Statistical Association,* **92**, 648–655.

Ley, E. & Steel, M. F. J. (2009), 'On the effect of prior assumptions in Bayesian model averaging with applications to growth regression', *Journal of Applied Econometrics,* **24**, 651–674.

Li, P. (2005), 'Box-Cox transformations: An overview', *www.stat.uconn.edu/~studentjournal/index_files/pengfi_s05.pdf* .

Lindley, D. V. (1957), 'A statistical paradox', *Biometrika,* **44**, 187–192.

Madigan, D. & York, J. (1995), 'Bayesian graphical models for discrete data', *International Statistical Review,* **63**, 215–232.

Manly, B. F. (1976), 'Exponential data transformation', *The Statistician,* **25**, 37–42.

Miranda, M. F., Zhu, H. & Ibrahim, J. G. (2013), 'Bayesian spatial transformation models with applications in neuroimaging data', *Biometrics,* **69**, 1074–1083.

Montgomery, D. C. (2009), *Statistical Quality Control*, New Jersey: Wiley & Sons.

Moreno, E. (1997), Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness, *IMS Lecture Notes*, Monograph Series, vol. 31, pp. 257–270.

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*, Wiley Series in Computational Statistics, Hoboken, NJ.

Ntzoufras, I. & Tarantola, C. (2013), 'Conjugate and conditional conjugate Bayesian analysis of discrete graphical models of marginal independence', *Computational Statistics and Data Analysis,* **66**, 161–177.

O'Hagan, A. (1995), 'Fractional Bayes factors for model comparison', *Journal of the Royal Statistical Society B,* **57**, 99–138.

O'Hagan, A. (1997), 'Properties of intrinsic and fractional Bayes factors', *Test* **6**, 101–118.

Pérez, J. M. & Berger, J. O. (2002), 'Expected-posterior prior distributions for model selection', *Biometrika,* **89**, 491–511.

Pericchi, L. R. (1981), 'A Bayesian approach to transformations to normality', *Biometrika,* **68**, 35–43.

Qiu, P. & Zhang, J. (2014), 'On phase II SPC in cases when normality is invalid', *Quality and Reliability Engineering International,* **31**, 27–35.

Razali, N. M. & Wah, Y. B. (2011), 'Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests', *Journal of Statistical Modelling and Analytics,* **2**, 21–33.

Sakia, R. M. (1992), 'The Box-Cox transformation technique: A review', *Journal of the Royal Statistical Society, Series D (The Statistician),* **41**, 169–178.

Sarkar, N. (1985), 'Box-Cox transformation and the problem of heteroscedasticity', *Communications in Statistics-Theory and Methods,* **14**, 363–379.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics,* **6**, 461–464.

Scott, J. G. & Berger, J. O. (2006), 'An exploration of aspects of Bayesian multiple testing', *Journal of Statistical Planning and Inference,* **136**, 2144–2162.

Scott, J. G. & Berger, J. O. (2010), 'Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem', *The Annals of Statistics,* **38**, 2587–2619.

Smith, A. F. M. & Spiegelhalter, D. J. (1980), 'Bayes factors and choice criteria for linear models', *Journal of the Royal Statistical Society B,* **42**, 213–220.

Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Statistics in Practice, Wiley, Chichester, UK.

Spiegelhalter, D. J. & Smith, A. F. M. (1982), 'Bayes factors for linear and log-linear models with vague prior information', *Journal of the Royal Statistical Society B,* **44**, 377–387.

Stahel, W. A. (2002), *Statistische Datenanalyse, Eine Einführung für Naturwissenschaftler*, Vieweg, Braunschweig, DE.

Stow, C. A., Reckhow, K. H. & Qian, S. S. (2006), 'A Bayesian approach to retransformation bias in transformed regression', *Ecology,* **87**, 1472–1477.

Strachan, R. W. & van Dijk, H. K. (2005), *Improper priors with well defined Bayes factors*, Technical Report 05-4, Department of Economics, University of Leicester.

Sweeting, T. J. (1984), 'On the choice of the prior distribution for the Box-Cox transformed linear model', *Biometrika,* **71**, 127–134.

Sweeting, T. J. (1985), Consistent prior distributions for transformed models, In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M. (eds.), *Bayesian Statistics 2*, Amsterdam: Elsevier Science Publishers, pp. 755–762.

Taylor, J. M. G. (1985), 'Measures of location of skew distributions obtained through Box-Cox transformations', *Journal of the American Statistical Association,* **80**, 427–432.

Taylor, J. M. G., Cumberland, W. G. & Meng, X. (1996), 'Components of variance models with transformations', *Australian Journal of Statistics,* **38**, 183–191.

Thall, P. F., Russell, K. E. & Simon, R. M. (1997), 'Variable selection in regression via repeated data splitting', *Journal of the American Statistical Association,* **6**, 416–434.

Thode, H. C. (2002), *Testing for Normality*, New York: Marcel Dekker.

Torres-Ruiz, F., Moreno, E. & Girón, F. J. (2011), 'Intrinsic priors for model comparison in multivariate normal regression', *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas,* **105**, 273–289.

Tukey, J. W. (1957), 'On the comparative anatomy of transformations', *Annals of Mathematical Statistics,* **28**, 602–632.

van Zwet, W. R. (1964), *Convex Transformations of Random Variables*, Amsterdam: Mathematisch Centrum.

Villa, C. & Lee, J. E. (2015), 'Model prior distribution for variable selection in linear regression models', *arXiv:1512.08077,* available at `http://arxiv.org/abs/1104.0861` .

Wang, Q. J. (2008), 'A Bayesian method for multi-site stochastic data generation: Dealing with non-concurrent and missing data, variable transformation and parameter uncertainty', *Environmental Modelling & Software,* **23**, 412–421.

Weisberg, S. (2005), *Applied Linear Regression*, 3rd ed., Wiley-Interscience, New Jersey.

Weisberg, S. (2014), *Applied Linear Regression*, 4th ed., Wiley, New Jersey.

Westerberg, I., Guerrero, J. L., Seibert, J., Beven, K. J. & Halldin, S. (2011), 'Stage-discharge uncertaintly derived with a non-sttionary rating curve in the Choluteca river, honduras', *Hydrological processes,* **25**, 603–613.

Woods, H., Steinour, H. & Starke, H. (1932), 'Effect of composition of Portland cement on heat evolved during hardening', *Industrial and Engineering Chemistry Research* **24**, 1207–1214.

Yang, Y., Christensen, O. F. & Sorensen, D. (2011), 'Analysis of a genetically structured variance heterogeneity model using the Box-Cox transformation', *Genetics Reseach,* **93**, 33–46.

Yang, Z. (2006), 'A modified family of power transformations', *Economic Letters,* **92**, 14–19.

Yeo, I. K. & Johnson, R. A. (2000), 'A new family of power transformations to improve normality or symmetry', *Biometrika,* **87**, 954–959.

Zamba, K. D., Tsiamyrtzis, P. & Hawkins, D. M. (2008), 'A sequential Bayesian control model for influenza-like illnesses and early detection of intentional outbreaks', *Quality Engineering,* **20**, 495–507.

Zellner, A. (1986), On assessing prior distributions and Bayesian regression analysis using g-prior distributions, In: P. K. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Amsterdam: Elsevier Science Publishers, pp. 233–243.