ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΑΡΑΓΟΝΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ, ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΠΟΙΝΙΚΟΠΟΙΗΜΕΝΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΜΜΑΝΟΥΗΛ Κ. ΑΝΔΡΟΥΛΑΚΗ

Διπλωματούχου Σχολής Ε.Μ.Φ.Ε. Ε.Μ.Π.

ΕΠΙΒΛΕΠΩΝ:

Χ. ΚΟΥΚΟΥΒΙΝΟΣ

Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, Ιούνιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΑΡΑΓΟΝΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ, ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΠΟΙΝΙΚΟΠΟΙΗΜΕΝΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΜΜΑΝΟΥΗΛ Κ. ΑΝΔΡΟΥΛΑΚΗ

Διπλωματούχου Σχολής Ε.Μ.Φ.Ε. Ε.Μ.Π.

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

1. Χ. ΚΟΥΚΟΥΒΙΝΟΣ, Καθ. Ε.Μ.Π. (Επιβλέπων)
2. Φ. ΒΟΝΤΑ, Αν. Καθ. Ε.Μ.Π.
3. Χ. ΚΑΡΩΝΗ, Καθ. Ε.Μ.Π.

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

1. Χ. ΚΟΥΚΟΥΒΙΝΟΣ, Καθ. Ε.Μ.Π. (Επιβλέπων)
2. Φ. ΒΟΝΤΑ, Αν. Καθ. Ε.Μ.Π.
3. Χ. ΚΑΡΩΝΗ, Καθ. Ε.Μ.Π.
4. Χ. ΕΥΑΓΓΕΛΑΡΑΣ, Επίκ. Καθ. Παν. Πειραιώς
5. Α. ΚΑΡΑΓΡΗΓΟΡΙΟΥ, Καθ. Παν. Αιγαίου
6. Μ. ΚΟΥΤΡΑΣ, Καθ. Παν. Πειραιώς
7. Ι. ΣΠΗΛΙΩΤΗΣ, Αν. Καθ. Ε.Μ.Π.

ΑΘΗΝΑ, Ιούνιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Παραγοντικοί Σχεδιασμοί, Γενικευμένα Γραμμικά Μοντέλα και Επιλογή Μεταβλητών με Χρήση Μεθόδων Ποινικοποιημένης Πιθανοφάνειας

Διδακτορική Διατριβή
Εμμανουήλ Κ. Ανδρουλάκη

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Χρήστος Κουκουβίνος
Καθηγητής Ε.Μ.Π. (Επιβλέπων Καθηγητής)

Φιλία Βόντα
Αναπληρώτρια Καθηγήτρια Ε.Μ.Π. (Μέλος της Τριμελούς Επιτροπής)

Χρυσηΐς Καρώνη-Ρίτσαρντσον
Καθηγήτρια Ε.Μ.Π. (Μέλος της Τριμελούς Επιτροπής)

Χαράλαμπος Ευαγγελάρας
Επίκουρος Καθηγητής Πανεπιστημίου Πειραιώς

Αλέξανδρος Καραγρηγορίου
Καθηγητής Πανεπιστημίου Αιγαίου

Μάρκος Κούτρας
Καθηγητής Πανεπιστημίου Πειραιώς

Ιωάννης Σπηλιώτης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

ΠΕΡΙΛΗΨΗ

Η διδακτορική αυτή διατριβή αποτελείται από τέσσερα μέρη και συνολικά δώδεκα κεφάλαια. Το πρώτο μέρος, αναφέρεται στη μεθοδολογία επιλογής μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας, με έμφαση στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα. Στο δεύτερο μέρος, αναπτύσσονται νέες προσεγγίσεις για την επιλογή της ρυθμιστικής παραμέτρου, η οποία διαδραματίζει σημαντικό ρόλο κατά τη χρήση των ποινικοποιημένων μεθόδων. Στο τρίτο μέρος, μελετώνται διάφορες κατηγορίες παραγοντικών σχεδιασμών και παρουσιάζονται νέες μέθοδοι εύρεσης των ενεργών επιδράσεων, υπό την παραδοχή γενικευμένων γραμμικών μοντέλων. Στο τέταρτο και τελευταίο μέρος της διατριβής, μελετώνται και αξιολογούνται κλάσεις ορθογώνιων σχηματισμών τριών επιπέδων, τόσο ως προς τη χρησιμότητά τους ως σχεδιασμοί αποκριτικών επιφανειών υπό την παρουσία συσχετισμένων παρατηρήσεων, όσο και ως προς την καταλληλότητά τους στη διάκριση μοντέλων.

ABSTRACT

The present PhD thesis is divided into four parts consisting of twelve chapters. The first part refers to the variable selection methodology via penalized likelihood, with emphasis on frailty models with clustered data. In the second part, new approaches for the selection of the tuning parameter are developed, which plays an important role in the use of penalized methods. In the third part, various categories of factorial designs are studied and new methods of identifying the active effects are presented, under the assumption of generalized linear models. In the fourth and final part of the thesis, some classes of three level orthogonal arrays are studied and evaluated with regard to their usefulness as response surface designs in the presence of correlated observations, but also with regard to their model discrimination capability.

Στους Γονείς μου,
για τη στήριξη και
την υπομονή τους.

Περιεχόμενα

Ευχαριστίες

Η εκπόνηση αυτής της διατριβής θα ήταν αδύνατη χωρίς τη συμβολή και τη συμπαράσταση πολλών ανθρώπων. Στο σημείο αυτό, θα ήθελα να τους εκφράσω τις ευχαριστίες μου, με την πεποίθηση ότι φάνηκα αντάξιος των προσδοκιών τους.

Αισθάνομαι πρωτίστως την ανάγκη να ευχαριστήσω θερμά τον Επιβλέποντα, κ. Χρήστο Κουκουβίνο, Καθηγητή της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π., για την άψογη συνεργασία που είχαμε, το ειλικρινές ενδιαφέρον του, τη συνεχή επιστημονική καθοδήγησή του, την άμεση βιβλιογραφική ενημέρωση που μου παρείχε και τις ξεκάθαρες κατευθυντήριες γραμμές που μου έδινε. Επιπλέον, μου έδωσε τη δυνατότητα κατά τη διάρκεια των διδακτορικών μου σπουδών, να έρθω σε επαφή με τον κ. Claude Brezinski, Professor Emeritus, University of Sciences and Technologies of Lille και τον κ. Aurel Galantai, Professor of Mathematics, Obuda University, στους οποίους οφείλω επίσης τις ευχαριστίες μου. Οι επιστημονικές τους γνώσεις, οι εύστοχες παρατηρήσεις τους και οι εποικοδομητικές συζητήσεις μας, διαδραμάτισαν σημαντικό ρόλο στην πορεία της έρευνάς μου.

Θα ήθελα να ευχαριστήσω και τα άλλα δύο μέλη της Τριμελούς Επιτροπής, την κα. Φιλία Βόντα, Αναπληρώτρια Καθηγήτρια της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π., για την ιδιαίτερα εποικοδομητική συνεργασία που είχαμε στα χρόνια των διδακτορικών μου σπουδών και τις πολύτιμες υποδείξεις που μου παρείχε και την κα. Χρυσηΐδα Καρώνη-Ρίτσαρντσον, Καθηγήτρια της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π., για τις χρήσιμες συμβουλές και παρατηρήσεις της καθώς και για την πολύπλευρη βοήθειά της στην πορεία μου από τα φοιτητικά μου χρόνια έως σήμερα. Ευχαριστίες οφείλω και στα υπόλοιπα μέλη της Επταμελούς Εξεταστικής Επιτροπής, τον κ. Χαράλαμπο Ευαγγελάρα, Επίκουρο Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, τον κ. Αλέξανδρο Καραγρηγορίου, Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Αιγαίου, τον κ. Μάρκο Κούτρα, Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς και τον κ. Ιωάννη Σπηλιώτη, Αναπληρωτή Καθηγητή της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π., για το χρόνο που αφιέρωσαν στη μελέτη της διατριβής, καθώς και για τις χρήσιμες υποδείξεις τους.

Ευχαριστώ τους συναδέλφους που εργαστήκαμε μαζί αυτά τα χρόνια και ιδιαίτερα τον κ. Παναγιώτη Αγγελόπουλο, Διδάκτορα Ε.Μ.Π. και την κα. Καλλιόπη Μυλωνά, Lecturer in Statistics, in Mathematical Sciences and Southampton Statistical Sciences Research Institute, University of Southampton, για τη βοήθειά τους και την αποδοτική συνεργασία που είχαμε. Ευχαριστώ επιπρόσθετα την κ. Χριστίνα Παρπούλα, Διδάκτορα Ε.Μ.Π., για την υποστήριξη και τις εποικοδομητικές συζητήσεις μας στα χρόνια των διδακτορικών μου σπουδών. Θερμές ευχαριστίες επίσης οφείλω στην αδελφή μου Μαριλένα, Φιλόλογο και στον κ. Φοίβο Ζακκάκ, Υποψήφιο Διδάκτορα του Πανεπιστημίου Κρήτης, για τη διαρκή στήριξή τους αλλά και για την προσεκτική διόρθωση της διατριβής, στη Μαρία, για τη συμπαράσταση και την υπομονή της και σε όλους τους φίλους και συγγενείς που στάθηκαν πλάι μου όλα αυτά τα χρόνια.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα τους γονείς μου, Κωνσταντίνο και Ζαχαρένια, για την αμέριστη ηθική συμπαράσταση, την οικονομική υποστήριξη και την αδιάκοπη ενθάρρυνσή τους σε όλη τη διάρκεια των σπουδών μου. Ως ένα ελάχιστο δείγμα αναγνώρισης της συνεισφοράς τους, η παρούσα διατριβή αφιερώνεται σε αυτούς.

Εμμανουήλ Κ. Ανδρουλάκης
Αθήνα, 2015

Ερευνητικό Έργο

Κατά τη διάρκεια εκπόνησης της παρούσας διδακτορικής διατριβής, προέκυψαν οι παρακάτω δημοσιευμένες ή προς δημοσίευση επιστημονικές εργασίες, στα ακόλουθα πεδία έρευνας:

Ποινικοποιημένες Μέθοδοι Επιλογής Μεταβλητών

- Estimation and variable selection via frailty models with penalized likelihood (με Χ. Κουκουβίνο και Φ. Βόντα), *Statistics in Medicine*, 31 (2012), 2223-2239.
- On the Uniform frailty model with penalized likelihood and clustered data (με Χ. Κουκουβίνο και Φ. Βόντα), *Journal of Reliability and Statistical Studies*, 5 (2012), 97-106.
- Penalized likelihood methodology and frailty models (με Χ. Κουκουβίνο και Φ. Βόντα), In: *Statistical Models and Methods for Reliability and Survival Analysis*, V. Couallier, L. Gerville-Reache, C. Huber-Carol, N. Limnios and M. Mesbah (Eds), 45-60, John Wiley and Sons, Hoboken, USA, 2014.

Επιλογή της Ρυθμιστικής Παραμέτρου στις Μεθόδους Ποινικοποιημένης Πιθανοφάνειας

- Tuning parameter estimation in penalized least squares methodology (με Χ. Κουκουβίνο και Κ. Μυλωνά), *Communications in Statistics - Simulation and Computation*, 40 (2011), 1444-1457.
- Tuning parameter selection in penalized generalized linear models for discrete data (με Χ. Κουκουβίνο και Φ. Βόντα), *Statistica Neerlandica*, 68 (2014), 276-292.
- Tuning parameter selection in penalized frailty models (με Χ. Κουκουβίνο και Φ. Βόντα), *Communications in Statistics - Simulation and Computation*, (2015), DOI: 10.1080/03610918.2014.968723.

Μέθοδοι Ανάλυσης Παραγοντικών Σχεδιασμών

- A new variable selection method for uniform designs (με Χ. Κουκουβίνο), *Journal of Applied Statistics*, 40 (2013), 2564-2578.
- A new method for the analysis of supersaturated designs with discrete data (με Χ. Κουκουβίνο και Φ. Βόντα), *Communications in Statistics - Simulation and Computation*, (2014), DOI: 10.1080/03610918.2014.884589.
- A new method for the analysis of unreplicated designs with binary response (με Κ. Δρόσου και Χ. Κουκουβίνο), (έχει υποβληθεί για δημοσίευση).

Ορθογώνιοι Σχηματισμοί τριών Επιπέδων

- A comparison of three-level orthogonal arrays in the presence of different correlation structures in observations (με Π. Αγγελόπουλο και Χ. Κουκουβίνο), *Communications in Statistics - Simulation and Computation*, 42 (2013), 552-569.
- Model discrimination criteria on model-robust designs (με Π. Αγγελόπουλο και Χ. Κουκουβίνο), *Communications in Statistics - Simulation and Computation*, 43 (2014), 1575-1582.

Περίληψη

Ένα από τα βασικά στάδια κατά τη διεξαγωγή ενός πειράματος ή μίας στατιστικής μελέτης, αποτελεί η μείωση της διάστασης των δεδομένων και η επιλογή των σημαντικών μεταβλητών. Λόγω της ύπαρξης μεγάλου αριθμού πιθανά ενεργών παραγόντων, είναι αναγκαία η δημιουργία αποδοτικών μεθόδων, κάτι που μεταξύ άλλων, αποτελεί βασικό στόχο της παρούσας διατριβής.

Η διδακτορική αυτή διατριβή επικεντρώνεται σε ορισμένα από τα πιο ενεργά και συνεχώς αναπτυσσόμενα πεδία έρευνας, με άμεση εφαρμογή σε προβλήματα επιλογής μεταβλητών και όχι μόνο, τη μεθοδολογία ποινικοποιημένης πιθανοφάνειας και τους πειραματικούς σχεδιασμούς. Αποτελείται από τέσσερα μέρη και συνολικά δώδεκα κεφάλαια. Το πρώτο μέρος, αναφέρεται στη μεθοδολογία επιλογής μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας, με έμφαση στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα. Στο δεύτερο μέρος, αναπτύσσονται νέες προσεγγίσεις για την επιλογή της ρυθμιστικής παραμέτρου, η οποία διαδραματίζει σημαντικό ρόλο κατά τη χρήση των ποινικοποιημένων μεθόδων. Στο τρίτο μέρος, μελετώνται διάφορες κατηγορίες παραγοντικών σχεδιασμών και παρουσιάζονται νέες μέθοδοι εύρεσης των ενεργών επιδράσεων, υπό την παραδοχή γενικευμένων γραμμικών μοντέλων. Στο τέταρτο και τελευταίο μέρος της διατριβής, μελετώνται και αξιολογούνται κλάσεις ορθογώνιων σχηματισμών τριών επιπέδων, ως προς τη χρήσιμότητά τους ως σχεδιασμοί αποκριτικών επιφανειών αλλά και ως προς την καταλληλότητά τους στη διάκριση μοντέλων. Η παρουσίαση των επιμέρους θεμάτων και αποτελεσμάτων, οργανώνεται ως εξής:

Στο Κεφάλαιο 1 παρέχεται μια σύντομη περιγραφή των παραδοσιακών τεχνικών επιλογής μεταβλητών που συναντώνται στη βιβλιογραφία. Ορισμένες εξ αυτών αποτέλεσαν τη βάση για την ανάπτυξη της μεθοδολογίας ποινικοποιημένης πιθανοφάνειας, την οποία και αναλύουμε στη συνέχεια.

Το Κεφάλαιο 2 και τελευταίο του πρώτου μέρους της διατριβής, αφορά τα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα όπου και προτείνεται μια γενικευμένη μορφή της ποινικοποιημένης συνάρτησης πιθανοφάνειας. Η συνάρτηση αυτή επιδέχεται τη χρήση διαφορετικών συνεχών κατανομών για την παράμετρο ευπάθειας. Επιπρόσθετα, παρουσιάζονται αναλυτικά οι περιπτώσεις της Γάμμα, της Αντίστροφης Γκαουσιανής και της Ομοιόμορφης κατανομής.

Στο Κεφάλαιο 3, προτείνεται μια νέα μέθοδος επιλογής της ρυθμιστικής παραμέτρου, στο πλαίσιο του ποινικοποιημένου γραμμικού μοντέλου. Χρησιμοποιούμε ως υπόβαθρο εργαλεία από το επιστημονικό πεδίο της Αριθμητικής Ανάλυσης και συγκεκριμένα, εκτιμήσεις σφάλματος κατά την επίλυση γραμμικών συστημάτων.

Το Κεφάλαιο 4, ασχολείται με τα γενικευμένα γραμμικά μοντέλα με διακριτά δεδομένα και προτείνονται αρχικά νέες εκτιμήσεις της νόρμας του σφάλματος, μέσω της χρήσης των Kantorovich ανισώσεων. Οι εκτιμήσεις αυτές εφαρμόζονται για τη δημιουργία μιας νέας μεθόδου επιλογής της ρυθμιστικής παραμέτρου, στην περίπτωση όπου έχουμε ποινικοποιημένη πιθανοφάνεια.

Στο Κεφάλαιο 5 και τελευταίο του δεύτερου μέρους της διατριβής, επεκτείνεται η προσέγγιση του προβλήματος επιλογής της ρυθμιστικής παραμέτρου του προηγουμένου Κεφαλαίου 4, θεωρώντας ποινικοποιημένα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα. Η μελέτη έχει πάλι ως υπόβαθρο τις Kantorovich ανισώσεις.

Στο Κεφάλαιο 6, παρατίθενται ορισμένα από τα βασικά είδη των πειραματικών σχεδιασμών, καθώς και οι σχετικοί ορισμοί και ιδιότητες αυτών. Συγκεκριμένα, αναλύονται οι παραγοντικοί και κλασματικοί παραγοντικοί σχεδιασμοί, οι μη επαναλαμβανόμενοι παραγοντικοί σχε-

διασμοί, οι υπερκορεσμένοι και οι ομοιόμορφοι σχεδιασμοί.

Στο Κεφάλαιο 7, παρουσιάζεται μια νέα μεθοδολογία επιλογής μεταβλητών στους υπερκορεσμένους σχεδιασμούς δύο επιπέδων, θεωρώντας γενικευμένα γραμμικά μοντέλα, των οποίων τα δεδομένα απόκρισης είναι διακριτά. Συγκεκριμένα, δίνεται έμφαση στις περιπτώσεις όπου η απόκριση είναι κατανομής Bernoulli και Poisson. Η προτεινόμενη μέθοδος βασίζεται στη τροποποίηση του PageRank αλγορίθμου της μηχανής αναζήτησης Google, μέσω του συνδυασμού του με κατάλληλα μέτρα από τη Θεωρία Πληροφορίας.

Στο Κεφάλαιο 8, επεκτείνεται η μέθοδος κρησαρίσματος μέσω του PageRank αλγορίθμου, που αναπτύχθηκε στο προηγούμενο κεφάλαιο, στους μη επαναλαμβανόμενους παραγοντικούς σχεδιασμούς δύο επιπέδων. Τα δεδομένα απόκρισης είναι δίτιμα, συνεπώς προέρχονται από το Λογιστικό μοντέλο παλινδρόμησης.

Στο Κεφάλαιο 9 και τελευταίο του τρίτου μέρους της διατριβής, προτείνεται μια νέα διαδικασία επιλογής μεταβλητών στους ομοιόμορφους σχεδιασμούς. Για το σκοπό αυτό, συνδυάζεται η μεθοδολογία των ποινικοποιημένων ελαχίστων τετραγώνων με τη μέθοδο επιλογής της ρυθμιστικής παραμέτρου του Κεφαλαίου 3.

Στο Κεφάλαιο 10, μελετώνται οι ορθογώνιοι σχηματισμοί και παρατίθενται τα βασικά χαρακτηριστικά τους, καθώς και οι σχετικοί ορισμοί. Περιγράφονται επίσης οι έννοιες των συνδυαστικά ισόμορφων και γεωμετρικά ισόμορφων ορθογώνιων σχηματισμών.

Στο Κεφάλαιο 11, εξετάζεται και αξιολογείται η κλάση των γεωμετρικά μη ισόμορφων ορθογώνιων σχηματισμών τριών επιπέδων, με 18, 27 και 36 εκτελέσεις. Οι σχηματισμοί αυτοί συγκρίνονται αναφορικά με την αποτελεσματική τους χρήση ως σχεδιασμοί αποκριτικών επιφανειών δευτέρας τάξης. Η μελέτη υποθέτει την παρουσία διαφορετικών μορφών συσχέτισης στις παρατηρήσεις. Για το σκοπό αυτό, προτείνονται τρία νέα κριτήρια σύγκρισης τα οποία και εφαρμόζονται σε συνδυασμό με ένα κατάλληλο μέτρο εκτιμητικής ικανότητας.

Η παρουσίαση του ερευνητικού έργου για αυτή τη διατριβή ολοκληρώνεται με το Κεφάλαιο 12, όπου μελετάται μια κλάση των γεωμετρικά μη ισόμορφων, εύρωστων ως προς τα μοντέλα, ορθογώνιων σχηματισμών τριών επιπέδων. Οι σχηματισμοί αυτοί συγκρίνονται ως προς την ικανότητά τους στη διάκριση μοντέλων, μέσω της χρήσης σχετικών κριτηρίων της βιβλιογραφίας. Προτείνεται επίσης ένα νέο μέτρο το οποίο τα συνδυάζει και αξιολογεί τη συνολική ικανότητα διάκρισης μοντέλων ενός σχεδιασμού.

Η στοιχειοθεσία της παρούσας διδακτορικής διατριβής πραγματοποιήθηκε με το πρόγραμμα LaTeX (διανομές MiKTeX και BibTeX). Η συγγραφή έγινε με τη βοήθεια του προγράμματος WinEdt. Η τελική ηλεκτρονική μορφή (Portable Document Format – PDF) δημιουργήθηκε με το πρόγραμμα PDFLaTeX. Για την υλοποίηση των μεθόδων και την ανάπτυξη των αλγορίθμων και των προσομοιώσεων, χρησιμοποιήθηκε το λογισμικό Matlab. Οι γραφικές παραστάσεις έγιναν με τη βοήθεια του προγράμματος Microsoft Excel 2007 και η επεξεργασία των σχημάτων με το πρόγραμμα Adobe Photoshop.

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Μέρος Ι

Ποινικοποιημένες Μέθοδοι Επιλογής Μεταβλητών

ΚΕΦΑΛΑΙΟ **1**

Μεθοδολογία Ποινικοποιημένης Πιθανοφάνειας

Statistics is the grammar of science.

—Karl Pearson (1857–1936)

Στο πρώτο αυτό κεφάλαιο, παρουσιάζεται μια εισαγωγή στις βασικές μεθόδους επιλογής μεταβλητών που συναντώνται στη βιβλιογραφία. Ξεκινάμε με μια σύντομη περιγραφή των παραδοσιακών τεχνικών που αποτέλεσαν το έναυσμα για τη δημιουργία της μεθοδολογίας ποινικοποιημένης πιθανοφάνειας, την οποία και αναλύουμε στη συνέχεια. Ιδιαίτερη έμφαση δίνουμε στην οικογένεια των ποινικοποιημένων μεθόδων των Fan και Li [53].

1.1 Εισαγωγή

Η διαθεσιμότητα μεγάλου όγκου δεδομένων, σε συνδυασμό με τα νέα επιστημονικά προβλήματα της εποχής, έχουν αναδιαμορφώσει τη στατιστική σκέψη και ανάλυση στον ερευνητικό τομέα. Για αυτόν τον λόγο, η σωστή επιλογή των στατιστικά σημαντικών μεταβλητών είναι μείζωνος σημασίας, ειδικά σε δεδομένα υψηλής διάστασης. Τα τελευταία χρόνια, παρατηρείται ένα αυξανόμενο ενδιαφέρον για μεθόδους που βασίζονται στα ποινικοποιημένα ελάχιστα τετράγωνα και στην ποινικοποιημένη πιθανοφάνεια, με τις οποίες τα περισσότερα από τα γνωστά κριτήρια επιλογής μεταβλητών είναι στενά συνδεδεμένα. Προφανώς, τα φειδωλά μοντέλα είναι πάντα επιθυμητά. Παρέχουν απλές και εύκολα ερμηνεύσιμες σχέσεις μεταξύ των μεταβλητών. Επιπλέον, τα σφάλματα πρόβλεψης μειώνονται. Κατά συνέπεια, απαιτούνται καινοτόμες διαδικασίες επιλογής μεταβλητών που μπορούν να εφαρμοστούν σε διαφορετικούς επιστημονικούς τομείς, όπως η Ιατρική, η Βιολογία, η Δημόσια Υγεία, η Επιδημιολογία, η Εφαρμοσμένη Μηχανική, η Γεωλογία και τα Χρηματοοικονομικά μεταξύ άλλων.

Οι παραδοσιακές μέθοδοι επιλογής μεταβλητών συνίστανται στην επιλογή υποσυνόλων, όπως η επιλογή του καλύτερου υποσυνόλου (best subset selection) και η κατά βήματα επιλογή (stepwise selection). Αν και φαίνονται χρήσιμες, οι διαδικασίες αυτές αγνοούν τα στοχαστικά σφάλματα που εμφανίζονται κατά τη διαδικασία της επιλογής μεταβλητών καθώς επίσης είναι υπολογιστικά χρονοβόρες. Επιπλέον, η επιλογή του καλύτερου υποσυνόλου σε συνδυασμό με τα συνήθως χρησιμοποιημένα κριτήρια AIC [4, 5] και BIC [152], πάσχει από έλλειψη σταθερότητας, όπως σημειώνεται στον Breiman [21]. Εντούτοις, αυτά τα δύο κριτήρια ήταν τα πρώτα που πρότειναν ουσιαστικά μια ενοποιημένη προσέγγιση στην επιλογή μεταβλητών και μοντέλων, τη χρήση δηλαδή μιας μορφής ποινικοποιημένης πιθανοφάνειας, η οποία πρέπει να μεγιστοποιηθεί σε σχέση με το διάνυσμα $\boldsymbol{\beta}$ των παραμέτρων,

$$l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_0 \tag{1.1}$$

όπου $l(\boldsymbol{\beta})$ είναι η συνάρτηση της λογαριθμισμένης πιθανοφάνειας, η L_0-νόρμα μετρά τον αριθμό των μη μηδενικών συνιστωσών στο διάνυσμα $\boldsymbol{\beta}$ και $\lambda \geq 0$ είναι μια παράμετρος κανονικοποίησης (regularization parameter). Συγκεκριμένα, ας υποθέσουμε ότι τα δεδομένα είναι της μορφής $(\mathbf{x}_i^T, y_i)_{i=1}^n$ όπου $\mathbf{x}_i$ είναι $d-$διάστατο διάνυσμα των συμμεταβλητών και y_i είναι η $i-$οστή παρατήρηση του διανύσματος της απόκρισης $\mathbf{y}$. Ο Akaike [4, 5] πρότεινε ότι ένα ποσό ποινής, αναφορικά με το πλήθος των παραμέτρων του μοντέλου, πρέπει να αφαιρεθεί από τον λογάριθμο της πιθανοφάνειας. Κατά συνέπεια, πρότεινε τη χρήση της ποσότητας

$$-l(\hat{\boldsymbol{\beta}}) + \lambda dim(\hat{\boldsymbol{\beta}}). \tag{1.2}$$

Για $\lambda = 1$ προκύπτει το κριτήριο AIC. Επιπλέον, $dim(\hat{\boldsymbol{\beta}})$ είναι η διάσταση του διανύσματος των παραμέτρων στο μοντέλο και $\hat{\boldsymbol{\beta}}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας (maximum likelihood estimator) του $\boldsymbol{\beta}$. Η ανωτέρω έκφραση προφανώς ελαχιστοποιείται (και δε μεγιστοποιείται), λόγω της αλλαγής του προσήμου. Εν συνεχεία, ο Schwartz [152] πρότεινε το κριτήριο BIC που δίνεται ως

$$-l(\hat{\boldsymbol{\beta}}) + (\log n/2) dim(\hat{\boldsymbol{\beta}}). \tag{1.3}$$

Το κριτήριο BIC έχει μια παρόμοια μορφή με την (1.2), αλλά με τη χρήση του $\lambda = \log n/2$. Κατά συνέπεια, πολλές παραδοσιακές μέθοδοι μπορούν να θεωρηθούν ως ποινικοποιημένες μέθοδοι πιθανοφάνειας, με διαφορετικές επιλογές του λ. Παραδείγματος χάριν, το κριτήριο C_p του Mallows [126], και τα κριτήρια του Risannen [148] και των Hannan και Quinn [87] έχουν παρόμοιες εκφράσεις. Παρ΄ όλα αυτά, πρέπει να σημειωθεί ότι η επίλυση του ποινικοποιημένου L_0 προβλήματος (1.1) είναι ένα ιδιαίτερο πρόβλημα συνδυαστικής με NP-πολυπλοκότητα.

Προκειμένου να διατηρηθούν οι αρετές των μεθόδων επιλογής υποσυνόλου και για να αποφευχθεί η αστάθειά τους, πολλοί συγγραφείς χρησιμοποίησαν μια πιο γενική μορφή ποινικοποιημένης πιθανοφάνειας. Αυτό επιτυγχάνεται μέσω της χρήσης διαφορετικών συναρ-

τήσεων ποινής. Συγκεκριμένα, παρατηρούμε ότι η (1.1) μπορεί να γενικευθεί στη μορφή

$$l(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{1.4}$$

όπου $p_\lambda(|\beta_j|)$ είναι μια κατάλληλη συνάρτηση ποινής που βασίζεται σε μια ρυθμιστική παράμετρο (tuning parameter) λ. Ο στόχος είναι η μεγιστοποίηση της (1.4) ώστε να επιλεχθούν οι σημαντικές μεταβλητές και ταυτόχρονα να υπολογιστούν οι συντελεστές του μοντέλου. Κατά συνέπεια, οι μη σημαντικές μεταβλητές διαγράφονται αυτόματα, με τον υπολογισμό των συντελεστών τους ως μηδέν.

Ένα παράδειγμα της προαναφερθείσας προσέγγισης είναι η παλινδρόμηση γέφυρας (bridge regression) ή ισοδύναμα η ποινικοποιημένη L_p-παλινδρόμηση των Frank και Friedman [75], στην οποία $p_\lambda(|\beta_j|) = \lambda|\beta_j|^p$ για $0 < p \leq 2$. Η παλινδρόμηση bridge βρίσκεται στην ουσία μεταξύ της μεθόδου καλύτερου υποσυνόλου (L_0 ποινή) και της παλινδρόμησης κορυφογραμμής (ridge regression, L_2 ποινή). Σχετικά με την τελευταία, πρέπει να σημειωθεί ότι ο εκτιμητής κορυφογραμμής ναι μεν κανονικοποιεί και σταθεροποιεί τον εκτιμητή, αλλά παρουσιάζει μεροληψία. Επίσης δε συρρικνώνει οποιονδήποτε συντελεστή άμεσα στο μηδέν [56]. Επιπλέον, η μη αρνητική garrote μέθοδος του Breiman [20] είναι στο ίδιο πνεύμα με αυτό της παλινδρόμησης bridge. Ένα άλλο παράδειγμα είναι η ποινικοποιημένη L_1-παλινδρόμηση, γνωστή ως μέθοδος του ελάχιστου απόλυτου τελεστή συρρίκνωσης και επιλογής (Least Absolute Shrinkage and Selection Operator method, LASSO). Η μέθοδος LASSO προτάθηκε από τον Tibshirani [162], για την οποία $p_\lambda(|\beta|) = \lambda|\beta|$, με $\beta \in R$. Πρέπει επίσης να αναφέρουμε τη μέθοδο Hard [10], με την ποινή σκληρού κατωφλιού (hard thresholding penalty) $p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, όπου $I(\cdot)$ είναι μια δείκτρια συνάρτηση. Στο ίδιο πνεύμα όπως της LASSO, έχει προταθεί και η μέθοδος ποινικοποιημένης πιθανοφάνειας με μη κοίλες συναρτήσεις ποινής, η οποία έχει εφαρμοστεί σε διάφορα παραμετρικά μοντέλα, συμπεριλαμβανομένων των γενικευμένων γραμμικών μοντέλων [53], των εύρωστων γραμμικών μοντέλων [59], καθώς και κάποιων ημι-παραμετρικών, όπως το μοντέλο του Cox [28, 54] και τα μερικώς γραμμικά μοντέλα [55].

Η προηγούμενη συζήτηση δίνει την αφορμή για ένα εύλογο ερώτημα: Πώς μπορούμε να επιλέξουμε μια κατάλληλη συνάρτηση ποινής, χρήσιμη για το πρόβλημα επιλογής μεταβλητών; Για το σκοπό αυτό, οι Fan και Li [53] πρότειναν μια καινούργια μεθοδολογία, βασισμένη αρχικά στα ποινικοποιημένα ελάχιστα τετράγωνα (penalized least squares), η οποία διατηρεί τις καλές ιδιότητες της παλινδρόμησης κορυφογραμμής αλλά και της μεθόδου επιλογής καλύτερου υποσυνόλου. Η μεθοδολογία τους αυτή, επεκτείνεται και σε μοντέλα βασισμένα στην πιθανοφάνεια, όπως για παράδειγμα στην περίπτωση όπου έχουμε δίτιμη απόκριση (binary response). Στην εργασία τους [53], η διαδικασία της ποινικοποίησης συνίσταται στην εισαγωγή συναρτήσεων ποινής οι οποίες πρέπει να οδηγούν σε εκτιμητές με τις ακόλουθες ιδιότητες:

- Αμεροληψία: Ο προκύπτων εκτιμητής είναι σχεδόν αμερόληπτος, ιδίως στην περίπτωση όπου η σωστή άγνωστη παράμετρος είναι μεγάλη. Αποφεύγεται έτσι η μεροληψία του μοντέλου.

- Σποραδικότητα: Ο προκύπτων εκτιμητής αποτελεί κανόνα περιορισμού (thresholding rule), ώστε οι εκτιμηθέντες συντελεστές με μικρή τιμή, να μηδενίζονται. Έτσι, μειώνεται η πολυπλοκότητα του μοντέλου.

- Συνέχεια: Ο προκύπτων εκτιμητής είναι συνεχής. Αποφεύγεται κατά αυτόν τον τρόπο η αστάθεια στην πρόβλεψη του μοντέλου.

Στο σημείο αυτό, πρέπει να τονίσουμε ότι η ποινή L_p με $0 \leq p < 1$ δεν ικανοποιεί την προϋπόθεση της συνέχειας, ενώ για $p > 1$ δεν ικανοποιεί την προϋπόθεση της σποραδικότητας. Επιπλέον, η ποινή L_1 δεν οδηγεί σε αμερόληπτους εκτιμητές. Για αυτόν τον

λόγο, ο Fan [52] και οι Fan και Li [53] εισήγαγαν την ποινή ομαλά αποκομμένης απόλυτης απόκλισης (smoothly clipped absolute deviation-SCAD), η πρώτη παράγωγος της οποίας ορίζεται ως $p'_\lambda(\beta) = \lambda\Big\{I(\beta \leq \lambda) + \frac{(\alpha\lambda-\beta)_+}{(\alpha-1)\lambda} I(\beta > \lambda)\Big\}$, για κάποια $\beta > 0$ και $\alpha > 2$, με $p_\lambda(0) = 0$. Για την επιλογή της παραμέτρου α, σύμφωνα με τη σχετική βιβλιογραφία [53], η τιμή $\alpha \approx 3.7$ εμφανίζεται να αποδίδει αρκετά ικανοποιητικά στα πολυάριθμα προβλήματα επιλογής μεταβλητών. Η ποινή SCAD ικανοποιεί όλες τις προηγουμένως αναφερθείσες ιδιότητες. Μια άλλη ποινή παρόμοια με τη SCAD, είναι η minimax κοίλη ποινή (minimax concave penalty-MCP), που προτάθηκε από τον Zhang [182], της οποίας η παράγωγος δίνεται ως $p'_\lambda(\beta) = (\alpha\lambda - \beta)_+/\alpha$. Πρέπει επίσης να αναφέρουμε μια οικογένεια ποινών, που γεφυρώνουν την L_0 και την L_1 και μελετήθηκαν από τους Lv και Fan [125]. Επιπλέον, αρκετά χρησιμοποιούμενη ποινή είναι η elastic net που προτείνεται από τους Zou και Hastie [187] και αποτελεί ένα γραμμικό συνδυασμό των ποινών L_0 και L_1.

Η μεγιστοποίηση της (1.4) είναι ένα πρόβλημα βελτιστοποίησης, για το οποίο διάφοροι αλγόριθμοι έχουν προταθεί. Μεταξύ άλλων, οι ακόλουθοι είναι οι πιο γνωστοί. Αρχικά, οι Fan και Li [53] πρότειναν τον αλγόριθμο τοπικής τετραγωνικής προσέγγισης (local quadratic approximation-LQA) για τη βελτιστοποίηση της συνάρτησης ποινικοποιημένης πιθανοφάνειας. Εντούτοις, η μεγιστοποίησή της είναι υπολογιστικά δύσκολη. Για να ξεπεραστεί αυτό το πρόβλημα, οι Zou και Li [188] ανέπτυξαν έναν ενοποιημένο αλγόριθμο βασισμένο πλέον στην τοπική γραμμική προσέγγιση (local linear approximation-LLA), ο οποίος μειώνει το υπολογιστικό κόστος. Ένας άλλος γνωστός αλγόριθμος είναι ο αλγόριθμος παλινδρόμησης ελάχιστης γωνίας (least angle regression-LARS) που προτείνεται από τους Efron et al. [47]. Επιπλέον, ο Zhang [182] πρότεινε τον αλγόριθμο ποινικοποιημένης γραμμικής αμερόληπτης επιλογής (penalized linear unbiased selection-PLUS) που μπορεί να χρησιμοποιηθεί για τον υπολογισμό της λύσης στο πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων. Για τη βελτιστοποίηση του ίδιου προβλήματος, ένας αλγόριθμος συντονισμένης καθόδου (coordinate descent algorithm) προτάθηκε από τους Fu [76], Daubechies et al. [45] και τους Wu και Lange [180]. Παρ΄ όλα αυτά, μπορεί επίσης να χρησιμοποιηθεί και στη γενικότερη περίπτωση της ποινικοποιημένης πιθανοφάνειας. Παραδείγματος χάριν, αναφέρουμε τους Fan και Lv [58], οι οποίοι ανέπτυξαν τον επαναληπτικό αλγόριθμο συντονισμένης ανόδου (iterative coordinate ascent-ICA). Μια παρόμοια διαδικασία μελετήθηκε στους Zhang και Li [183], οι οποίοι εισήγαγαν τον επαναληπτικό αλγόριθμο δεσμευμένης μεγιστοποίησης (iterative conditional maximization-ICM).

1.2 Επιλογή Μεταβλητών Μέσω Ποινικοποιημένης Πιθανοφάνειας

Στην ενότητα αυτή θα παραθέσουμε κάποια βασικά στοιχεία της μεθοδολογίας επιλογής μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας, που προτάθηκε από τους Fan και Li [53]. Ο λόγος είναι ότι αποτέλεσαν στην ουσία τη βάση, στην οποία στηρίχθηκαν αρκετές μέθοδοι που αναπτύχθηκαν μετέπειτα στη βιβλιογραφία. Μεταξύ άλλων, αναφέρουμε ενδεικτικά τις εργασίες [54], [55], [56], [59], [114], [144], [188], καθώς και την εργασία των Fan και Lv [57] για μια επιλεκτική επισκόπηση των ποινικοποιημένων μεθόδων.

Το βασικό χαρακτηριστικό της οικογένειας μεθόδων επιλογής μεταβλητών των Fan και Li [53] είναι ότι διαφέρουν από τις παραδοσιακές τεχνικές, καθότι διαγράφουν τις μη σημαντικές μεταβλητές, εκτιμώντας τους συντελεστές τους ως μηδέν. Συνεπώς, αυτό που τελικά επιτυγχάνεται, είναι ότι ταυτόχρονα γίνεται και εκτίμηση των παραμέτρων του μοντέλου και μηδενισμός κάποιων, άρα ικανοποιείται ο σκοπός της επιλογής μεταβλητών. Να σημειωθεί επίσης, ότι οι εκτιμητές ποινικοποιημένης πιθανοφάνειας, χαρακτηρίζονται από τη λεγόμενη μαντική ιδιότητα (oracle property), δεδομένου ότι έχει επιλεγεί σωστά η ρυθμιστική παράμετρος λ. Η ιδιότητα αυτή, πρακτικά σημαίνει δύο πράγματα. Πρώτον, όταν το πραγματικό

διάνυσμα $\boldsymbol{\beta}$ του μοντέλου έχει κάποιες μηδενικές συνιστώσες, αυτές εκτιμώνται από τη μέθοδο ως μηδενικές με πιθανότητα να τείνει στη μονάδα. Δεύτερον, όσον αφορά τις μη μηδενικές συνιστώσες, αυτές εκτιμώνται τόσο καλά σαν να είναι γνωστό εκ των προτέρων το σωστό υπο-μοντέλο. Το αποτέλεσμα είναι ότι αυξάνεται η ακρίβεια εκτίμησης τόσο των μηδενικών όσο και των μη μηδενικών συνιστωσών.

Υποθέτουμε ότι τα δεδομένα $(\mathbf{x}_i, Y_i)$ έχουν συλλεχθεί ανεξάρτητα. Δεδομένου του $\mathbf{x}_i$, η απόκριση Y_i έχει συνάρτηση πιθανότητας $f_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i)$, όπου g είναι μια γνωστή συνάρτηση σύνδεσης (link function). Στην εργασία των Fan και Li [53], μια μορφή της ποινικοποιημένης πιθανοφάνειας ορίζεται ως

$$Q(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} l_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{1.5}$$

όπου $l_i = \log f_i$ είναι ο λογάριθμος της πιθανοφάνειας για τα y_i, $p_\lambda(.)$ είναι μια συνάρτηση ποινής και λ είναι η ρυθμιστική παράμετρος. Με αλλαγή του προσήμου, αντί να μεγιστοποιήσουμε την ως άνω συνάρτηση ώστε να αποκτήσουμε τον εκτιμητή του $\boldsymbol{\beta}$, ισοδύναμα ελαχιστοποιούμε την ποσότητα

$$-\sum_{i=1}^{n} l_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{1.6}$$

ως προς $\boldsymbol{\beta}$ για κάποια παράμετρο λ.

Οι Fan και Li [53] πρότειναν έναν αποδοτικό αλγόριθμο τοπικής τετραγωνικής προσέγγισης για την ελαχιστοποίηση της (1.6), χρησιμοποιώντας μια κατάλληλη αρχική τιμή για τις άγνωστες παραμέτρους. Με την τοπική τετραγωνική προσέγγιση, η λύση μπορεί να βρεθεί υπολογίζοντας αναδρομικά την ακόλουθη έκφραση, έχοντας μια αρχική τιμή $\boldsymbol{\beta}^{(0)}$:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - \{\nabla^2 l(\boldsymbol{\beta}^{(0)}) + n\sum_{\lambda}(\boldsymbol{\beta}^{(0)})\}^{-1}\{\nabla l(\boldsymbol{\beta}^{(0)}) + nU_\lambda(\boldsymbol{\beta}^{(0)})\}, \tag{1.7}$$

όπου $\nabla l(\boldsymbol{\beta}^{(0)}) = \frac{\partial\, l(\boldsymbol{\beta}^{(0)})}{\partial \boldsymbol{\beta}}$, $\nabla^2 l(\boldsymbol{\beta}^{(0)}) = \frac{\partial^2\, l(\boldsymbol{\beta}^{(0)})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}$, $U_\lambda(\boldsymbol{\beta}^{(0)}) = \sum_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(0)}$, και $\sum_\lambda(\boldsymbol{\beta}^{(0)}) = diag\{p'_\lambda(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \ldots, p'_\lambda(|\beta_d^{(0)}|)/|\beta_d^{(0)}|\}$. Το μόνο μειονέκτημα του αλγορίθμου, είναι ότι από τη στιγμή που κάποιος συντελεστής θα συρρικνωθεί στο μηδέν σε κάποιο βήμα, θα παραμείνει σε αυτήν την τιμή. Όσον αφορά τα τυπικά σφάλματα των μη μηδενικών συνιστωσών του $\hat{\boldsymbol{\beta}}^{(1)}$, υπολογίζονται με χρήση του "sandwich" τύπου $\widehat{cov}(\hat{\boldsymbol{\beta}}^{(1)}) = \{\nabla^2 l(\hat{\boldsymbol{\beta}}^{(1)}) + n\sum_\lambda(\hat{\boldsymbol{\beta}}^{(1)})\}^{-1}\widehat{cov}\{\nabla l(\widehat{\boldsymbol{\beta}}^{(1)})\}\{\nabla^2 l(\hat{\boldsymbol{\beta}}^{(1)}) + n\sum_\lambda(\hat{\boldsymbol{\beta}}^{(1)})\}^{-1}$. Ο τύπος αυτός είναι αρκετά ακριβής και για μέτρια μεγέθη δειγμάτων. Για περισσότερες λεπτομέρειες, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στην επιστημονική εργασία [53], καθώς και στην [54] για την περίπτωση του ποινικοποιημένου Cox μοντέλου και του μοντέλου ευπάθειας Γάμμα.

ΚΕΦΑΛΑΙΟ **2**

Ποινικοποιημένη Πιθανοφάνεια στα Μοντέλα Ευπάθειας με Ομαδοποιημένα Δεδομένα

The final test of a theory is its capacity
to solve the problems which originated it.

—*George Dantzig (1914–2005)*

Στο κεφάλαιο αυτό, επεκτείνουμε τη μεθοδολογία ποινικοποιημένης πιθανοφάνειας των Fan και Li [54], προτείνοντας μια γενικευμένη μορφή της ποινικοποιημένης συνάρτησης πιθανοφάνειας στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα. Η συνάρτηση αυτή επιδέχεται τη χρήση διαφορετικών συνεχών κατανομών για την παράμετρο ευπάθειας, πέραν της κατανομής Γάμμα που είναι η πιο διαδεδομένη. Αναλύονται επίσης τρεις περιπτώσεις κατανομών της μεταβλητής ευπάθειας που χρησιμοποιούνται αρκετά συχνά και συγκεκριμένα, η Γάμμα, η Αντίστροφη Γκαουσιανή και η Ομοιόμορφη κατανομή.

2.1 Ερευνητικό Πρόβλημα

Η μεθοδολογία ποινικοποιημένης πιθανοφάνειας των Fan και Li [53], έχει επεκταθεί και στο μοντέλο αναλογικής διακινδύνευσης του Cox, καθώς και στο μοντέλο ευπάθειας Γάμμα κατανομής [54], για τις περιπτώσεις όπου έχουμε δεδομένα επιβίωσης και ομαδοποιημένα δεδομένα με αποκομμένες παρατηρήσεις, αντίστοιχα. Παρ΄ όλα αυτά, οι Fan και Li στην εργασία [54] περιορίστηκαν μόνο στη Γάμμα κατανομή και η ποινικοποιημένη πιθανοφάνεια που πρότειναν δε μπορεί να χρησιμοποιηθεί για διαφορετικές κατανομές. Το γεγονός αυτό μας έδωσε το έναυσμα να επεκτείνουμε με τη σειρά μας τη μεθοδολογία τους, προτείνοντας μια γενικευμένη μορφή της συνάρτησης πιθανοφάνειας, εισάγοντας σε αυτήν και έναν όρο ποινής. Στην προτεινόμενη μορφή πιθανοφάνειας, θα χρησιμοποιήσουμε και την εκτιμήτρια του Breslow για τη σωρευτική συνάρτηση διακινδύνευσης, τεχνική η οποία χρησιμοποιείται συχνά στη βιβλιογραφία, όπως για παράδειγμα στις εργασίες [54] και [85]. Ο στόχος μας είναι να παρέχουμε στους ερευνητές διαφορετικές επιλογές, πέραν του μοντέλου ευπάθειας Γάμμα, το οποίο και έχει μελετηθεί αρκετά.

2.2 Μοντέλα Ευπάθειας

Η ανάλυση επιβίωσης, αποτελεί μια περιοχή έρευνας στη Στατιστική, η οποία επικεντρώνεται στην ανάλυση δεδομένων τα οποία δε μπορούν να επεξεργαστούν με τις συνηθισμένες στατιστικές μεθόδους για την εξαγωγή συμπερασμάτων. Στην πλειονότητα των περιπτώσεων, τα δεδομένα της ανάλυσης επιβίωσης αφορούν το λεγόμενο χρόνο επιβίωσης ή χρόνο αποτυχίας, το χρόνο δηλαδή από μια αρχική παρατήρηση, όπως για παράδειγμα την έναρξη μιας θεραπείας, έως ότου συμβεί ένα γεγονός, όπως λοίμωξη, υποτροπή νόσου ή θάνατος. Στις περιπτώσεις όπου ο χρόνος αυτός δεν είναι δυνατό να παρατηρηθεί, παραδείγματος χάριν λόγω ατελούς παρατήρησης της εξέλιξης κάποιων ασθενών ή λόγω κάποιας βίαιης διακοπής της θεραπείας σε ανύποπτη χρονική στιγμή, τότε η παρατήρηση θεωρείται αποκομμένη (censored). Τα δεδομένα που δεν είναι αποκομμένα ονομάζονται πλήρη.

Ιδιαίτερα δημοφιλές στον τομέα της ανάλυσης επιβίωσης είναι το μοντέλο παλινδρόμησης του Cox ή αλλιώς μοντέλο αναλογικής διακινδύνευσης του Cox [41], [42]. Έχει πολλές εφαρμογές κυρίως στις βιο-ιατρικές επιστήμες, καθώς και σε προβλήματα πιστωτικού κινδύνου τραπεζών. Το μοντέλο του Cox, μοντελοποιεί τη συνάρτηση διακινδύνευσης $h(t)$ και αποτελεί μια πολύ καλή τεχνική για τη διερεύνηση της σχέσης μεταξύ της επιβίωσης ενός ατόμου και αρκετών επεξηγηματικών μεταβλητών [37].

Στο σημείο αυτό, να αναφέρουμε ότι μια από τις βασικές υποθέσεις στα μοντέλα επιβίωσης, όπως και στο μοντέλο του Cox, είναι η ανεξαρτησία των παρατηρήσεων. Αυτό σημαίνει ότι ο υπό μελέτη πληθυσμός είναι ομοιογενής, αναφορικά με κάποιες μετρήσιμες συμμεταβλητές, όπως το φύλο, η ηλικία, το βάρος και το είδος της θεραπείας. Αποτέλεσμα της ομοιογένειας αυτής, είναι ότι οι πειραματικές μονάδες του πληθυσμού διατρέχουν τον ίδιο κίνδυνο (κίνδυνο επανεμφάνισης νόσου, κίνδυνο θανάτου κλπ). Σε πολλές εφαρμογές όμως, η υπόθεση αυτή είτε παραβιάζεται είτε είναι ακατάλληλη, λόγω της ύπαρξης μη παρατηρηθέντων παραγόντων. Για παράδειγμα, αρκετές φορές παρουσιάζονται διαφορές σε πανομοιότυπες πειραματικές μονάδες, λόγω διαφορετικού γενότυπου. Συνεπώς, είναι αδύνατο σε ορισμένες περιπτώσεις να συνυπολογισθούν ή να μελετηθούν όλοι οι σχετικοί παράγοντες που επηρεάζουν μια νόσο, λόγω έλλειψης σχετικής πληροφορίας, κάτι που προφανώς προκαλεί ετερογένεια στον πληθυσμό. Επιπλέον, μπορεί ο αναλυτής να μη γνωρίζει την επίδραση ενός παράγοντα κινδύνου ή ακόμα και να αγνοεί την ύπαρξή του. Να συμπληρώσουμε, ότι η αδυναμία μελέτης κάποιων παραγόντων μπορεί να οφείλεται και σε οικονομικούς λόγους, καθώς και σε έλλειψη χρόνου.

Σε τέτοιες περιπτώσεις είναι χρήσιμο να εξεταστούν δύο πηγές μεταβλητότητας στα δεδομένα επιβίωσης: η μεταβλητότητα που οφείλεται σε παρατηρήσιμους παράγοντες κινδύνου

και περιλαμβάνονται στο μοντέλο (και ως εκ τούτου θεωρητικά προβλέψιμοι) και η ετερογένεια που προκαλείται από άγνωστες συμμεταβλητές [175]. Σύμφωνα με τον Hougaard [95], υπάρχει το εξής πλεονέκτημα κατά την εξέταση χωριστά αυτών των δύο πηγών της μεταβλητότητας: Η ετερογένεια μπορεί να εξηγήσει μερικά απροσδόκητα αποτελέσματα ή να δώσει μια εναλλακτική εξήγηση, όπως στην περίπτωση όπου έχουμε υποεκτίμηση της συνάρτησης κινδύνου. Αν κάποια άτομα αντιμετωπίζουν υψηλότερο κίνδυνο αποτυχίας, τότε τα υπόλοιπα τείνουν να σχηματίσουν μια επιλεγμένη ομάδα με χαμηλότερο κίνδυνο. Εάν δε ληφθεί υπόψη η ετερογένεια αυτή, θα οδηγηθούμε σε υποεκτίμηση της συνάρτησης κινδύνου σε ολοένα και μεγαλύτερο βαθμό όσο ο χρόνος αυξάνει.

Ένας από τους τρόπους περιγραφής της ανομοιογένειας ενός πληθυσμού και μοντελοποίησης της εξάρτησης, ανάμεσα στους χρόνους των γεγονότων σε μελέτες ανάλυσης επιβίωσης, είναι μέσω της εισαγωγής στο μοντέλο, μιας θετικής τυχαίας μεταβλητής u. Η μεταβλητή καλείται ευπάθεια (frailty) και ακολουθεί μια συγκεκριμένη κατανομή. Προτάθηκε αρχικά από τους Clayton [36], Vaupel et al. [167] και Hougaard [93]. Η ευπάθεια δεν παρατηρείται και αντιπροσωπεύει παράγοντες κινδύνου μετρήσιμους ή μη, οι οποίοι διαφοροποιούν τις πειραματικές μονάδες ως προς το χρόνο επιβίωσής τους. Στην απλούστερη μορφή της, η ευπάθεια είναι ένας τυχαίος αναλογικός παράγοντας που τροποποιεί τη συνάρτηση κινδύνου του ατόμου ή των συσχετιζόμενων ατόμων, έχοντας πολλαπλασιαστικό χαρακτήρα στη βασική συνάρτηση κινδύνου. Τα μοντέλα ευπάθειας είναι στην ουσία επεκτάσεις του μοντέλου αναλογικών κινδύνων του Cox.

Είναι γνωστό ότι τα μοντέλα ευπάθειας είναι πολύ αποτελεσματικά στην περιγραφή της εξάρτησης μέσα σε μια συστάδα (cluster) των παρατηρήσεων ή της ετερογένειας μεταξύ των συστάδων. Σε πολλές εφαρμογές, συχνά συναντάμε δεδομένα που αποτελούνται από παρατηρήσεις οργανωμένες σε συστάδες - ομάδες που διαμορφώνονται είτε φυσικά είτε τεχνητά. Εκεί θα πρέπει να ληφθεί υπόψη η εξάρτηση που προκαλείται στους ομαδοποιημένους χρόνους από τον αναλυτή, όπως για παράδειγμα στις διάρκειες ζωής ασθενών στις πολυκεντρικές κλινικές δοκιμές [6]. Επίσης, στις οικογενειακές μελέτες, οι οικογένειες αποτελούν φυσικές συστάδες με εξάρτηση εντός των συστάδων που οφείλεται ενδεχομένως στα παρόμοια γενετικά προφίλ. Ένα άλλο παράδειγμα είναι και η περίπτωση όπου έχουμε δεδομένα διδύμων ατόμων. Οπότε στην περίπτωση των ομαδοποιημένων δεδομένων, προστίθεται ένας όρος ευπάθειας, συγκεκριμένης κατανομής, που δρα πολλαπλασιαστικά στις συναρτήσεις κινδύνου όλων των ατόμων στην κάθε ομάδα [54]. Ως αποτέλεσμα, συχνά υποθέτουμε ότι δεδομένης της ευπάθειας, έστω u_i, τα άτομα στην $i-$οστή ομάδα είναι ανεξάρτητα. Μια πιο ειδική περίπτωση, μπορεί να θεωρηθεί αυτή των μονομεταβλητών μοντέλων ευπάθειας (univariate frailty models), όπου κάθε άτομο έχει τη δική του ευπάθεια.

Η ορθότητα των όποιων συμπερασμάτων προκύπτουν από τη χρήση των μοντέλων αυτών, εξαρτάται από την καταλληλότητα της κατανομής που υιοθετείται για τη μεταβλητή της ευπάθειας. Αρκετές κατανομές του όρου ευπάθειας αναφέρονται στις εργασίες [93], [94] και [95] όπως η Αντίστροφη Γκαουσιανή και η Positive Stable ευπάθεια. Άλλα παραδείγματα περιλαμβάνουν τη Log-Normal [131], την Power Variance [1], την Ομοιόμορφη [113] και τη Threshold [122] ευπάθεια. Επίσης, ο Parner [141], μελετά την ασυμπτωτική θεωρία για τον εκτιμητή μέγιστης πιθανοφάνειας σε ένα μοντέλο ευπάθειας με συστάδες μεγέθους μεγαλύτερου ή ίσου του 2 και όρο ευπάθειας που ακολουθεί τη Γάμμα κατανομή με άγνωστη διασπορά. Οι Vonta [168], Slud και Vonta [158] και Kosorok et al. [98], εξετάζουν την εκτίμηση των παραμέτρων σε μια κλάση μοντέλων ευπάθειας στη μονομεταβλητή περίπτωση δεδομένων.

Στην πράξη βέβαια, είναι δύσκολο να επιλεχθεί η σωστή κατανομή της ευπάθειας. Είναι προτιμότερο να εξεταστεί κάθε πρόβλημα ξεχωριστά, δεδομένου ότι η επιλογή της κατανομής εξαρτάται από το υπό θεώρηση πρόβλημα. Παραδείγματος χάριν, κατά μήκος της σειράς των χρόνων επιβίωσης, το μοντέλο ευπάθειας Γάμμα κατανομής εξηγεί την υψηλή εξάρτηση στους τελευταίους χρόνους, ενώ το μοντέλο της Αντίστροφης Γκαουσιανής κατανομής περι-

γράφει μέτρια εξάρτηση κατά μήκος ολόκληρης της σειράς των χρόνων. Επίσης, αρκετά ενδιαφέρουσα κατανομή είναι η Ομοιόμορφη. Η συγκεκριμένη κατανομή θα μπορούσε να χρησιμοποιηθεί όταν για παράδειγμα αναλύονται δεδομένα που προέρχονται από την αξιοπιστία συστημάτων δύο εξαρτημάτων, που εκτίθενται στον ίδιο περιβαλλοντικό τυχαίο παράγοντα. Όταν το λειτουργικό περιβάλλον του συστήματος διαφέρει από το εργαστηριακό περιβάλλον, μπορούμε να περιορίσουμε το στήριγμα του τυχαίου παράγοντα σε ένα σταθερό διάστημα $[a, b]$, όπως αναφέρεται στους Lee και Klein [113].

2.3 Η Προτεινόμενη Γενικευμένη Μορφή της Ποινικοποιημένης Πιθανοφάνειας

Στην ενότητα αυτή, θα επεκτείνουμε το ποινικοποιημένο μοντέλο ευπάθειας Γάμμα κατανομής των Fan και Li [54], μέσω μιας γενικής μορφής της συνάρτησης πιθανοφάνειας. Θεωρούμε το μοντέλο ευπάθειας στην περίπτωση των ομαδοποιημένων δεδομένων, όπου τα άτομα κάθε ομάδας μοιράζονται την ίδια ευπάθεια, η οποία και αποτελεί μια θετική τυχαία μεταβλητή. Έστω n το πλήθος των ομάδων, J_i το μέγεθος της ομάδας i, $i = 1, \ldots, n$ και u_i η κοινή ευπάθεια στην ομάδα i με συνάρτηση κατανομής πιθανότητας F_{u_i}. Τα δεδομένα περιγράφονται από την τριπλέτα $(z_{ij}, \delta_{ij}, \mathbf{x}_{ij})$ όπου $z_{ij} = \min\{T_{ij}, C_{ij}\}$, T_{ij} είναι ο χρόνος επιβίωσης, C_{ij} ο χρόνος αποκοπής, δ_{ij} ο δείκτης αποκοπής και $\mathbf{x}_{ij}$ είναι το $d-$διάστατο διάνυσμα των συμμεταβλητών του $j-$οστού ατόμου που ανήκει στην ομάδα i.

Πρόταση 2.1 *Η γενικευμένη μορφή του λογαρίθμου της πιθανοφάνειας δίνεται ως*

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\ln(h_0(z_{ij})) + \sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta} + \sum_{i=1}^{n}\ln\left(\left|L^{(A_i)}\Big(\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}H_0(z_{ij})\Big)\right|\right) \tag{2.1}$$

όπου $L^{(A_i)}(x) = \int_0^\infty e^{-ux}\,(-1)^{A_i}u^{A_i}\,dF_u(u)$.

Απόδειξη: Δεδομένων των συμμεταβλητών $\mathbf{x}_{ij}$ και της κοινής ευπάθειας u_i, η συνάρτηση διακινδύνευσης του $j-$οστού ατόμου στην i-οστή ομάδα είναι

$$h_{ij}(t|\mathbf{x}_{ij}, u_i) = u_i h_0(t) e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} \tag{2.2}$$

όπου $\boldsymbol{\beta} \in R^d$ είναι οι παράμετροι που μας ενδιαφέρουν και $h_0(t)$ είναι η βασική συνάρτηση διακινδύνευσης. Η συνάρτηση επιβίωσης που αντιστοιχεί προκύπτει ως

$$S_{ij}(t|\mathbf{x}_{ij}, u_i) = e^{-u_i H_0(t)e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}}, \quad i = 1, ..., n, \quad j = 1, ..., J_i \tag{2.3}$$

όπου $H_0(t)$ η βασική σωρευτική συνάρτηση διακινδύνευσης. Η συνάρτηση πυκνότητας πιθανότητας δίνεται επίσης ως

$$f_{ij}(t|\mathbf{x}_{ij}, u_i) = e^{-u_i H_0(t)e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}} u_i e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} h_0(t). \tag{2.4}$$

Δοθείσης της ευπάθειας u_i, όλα τα άτομα της ομάδας i θεωρούνται ανεξάρτητα. Συνεπώς, η από κοινού συνάρτηση επιβίωσης της ομάδας i δεδομένης της ευπάθειας u_i είναι

$$\prod_{j=1}^{J_i} S(z_{ij}|\mathbf{x}_{ij}, u_i) = e^{-u_i\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}H_0(z_{ij})}. \tag{2.5}$$

Στην περίπτωση όπου έχουμε αποκομμένα δεδομένα, η συνεισφορά της i-οστής ομάδας στην πιθανοφάνεια, ολοκληρώνοντας ως προς την ευπάθεια u_i, είναι

$$\int_0^\infty \prod_{j=1}^{J_i} f_{ij}(z_{ij}|\mathbf{x}_{ij},u_i)^{\delta_{ij}} S_{ij}(z_{ij}|\mathbf{x}_{ij},u_i)^{1-\delta_{ij}}\; dF_{u_i}(u_i) =$$

$$\int_0^\infty \prod_{j=1}^{J_i} h_{ij}(z_{ij}|\mathbf{x}_{ij},u_i)^{\delta_{ij}} S_{ij}(z_{ij}|\mathbf{x}_{ij},u_i)\; dF_{u_i}(u_i) \tag{2.6}$$

η οποία από την (2.2) και (2.3) ισούται με

$$\left(\prod_{j=1}^{J_i}(h_0(z_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}})^{\delta_{ij}}\right)\int_0^\infty \exp(-u_i\sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}H_0(z_{ij}))\, u_i^{A_i}\, dF_{u_i}(u_i) \tag{2.7}$$

όπου $A_i=\sum_{j=1}^{J_i}\delta_{ij}$.

Άρα, η συνολική πιθανοφάνεια (όλων των ομάδων) ισούται με

$$\prod_{i=1}^{n}\left\{\left(\prod_{j=1}^{J_i}(h_0(z_{ij})e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}})^{\delta_{ij}}\right)\int_0^\infty \exp(-u_i\sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}H_0(z_{ij}))\, u_i^{A_i}\, dF_{u_i}(u_i)\right\} \tag{2.8}$$

ενώ μπορεί ισοδύναμα να γραφεί ως

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\ln(h_0(z_{ij}))+\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta}+\sum_{i=1}^{n}\ln\left(\left|L^{(A_i)}\Big(\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}H_0(z_{ij})\Big)\right|\right) \tag{2.9}$$

όπου η ποσότητα

$$L^{(A_i)}(x)=\int_0^\infty e^{-ux}\,(-1)^{A_i}u^{A_i}\,dF_u(u) \tag{2.10}$$

δηλώνει την A_i-οστή παράγωγο του μετασχηματισμού Laplace $L(x)=\int_0^\infty e^{-ux}\,dF_u(u)$ της συνάρτησης κατανομής F_u της ευπάθειας. □

Πρέπει να τονιστεί εδώ ότι διαφορετικές κατανομές της μεταβλητής ευπάθειας, οδηγούν σε διαφορετικούς μετασχηματισμούς Laplace και συνεπώς προκύπτει μια νέα γενική κλάση μοντέλων ευπάθειας με λογάριθμο πιθανοφάνειας που δίνεται από την (2.9).

Πόρισμα 2.1 *Εξαλείφοντας την οχληρή παράμετρο (nuisance parameter) $h_0(\cdot)$, χρησιμοποιώντας ότι $H_0(t)=\sum_{l=1}^{N}\mu_l I(z_l\le t)$, η τελική μορφή της πιθανοφάνειας δίνεται ως*

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\sum_{l=1}^{N}\ln\mu_l+\sum_{i=1}^{n}\ln\left(\int_0^\infty\exp\Big(-u_i\sum_{j=1}^{J_i}e^{\boldsymbol{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})\Big)\,u_i^{A_i}\,dF_{u_i}(u_i)\right). \tag{2.11}$$

Απόδειξη: Για την εξάλειψη της οχληρής παραμέτρου $h_0(\cdot)$ και ακολουθώντας τη μέθοδο Breslow, θεωρούμε τη μη παραμετρική μοντελοποίηση ελάχιστης πληροφορίας για την $H_0(\cdot)$ όπου η $H_0(t)$ έχει ένα πιθανό άλμα μεγέθους μ_l στον παρατηρούμενο χρόνο διακοπής z_l. Τότε έχουμε ότι,

$$H_0(t)=\sum_{l=1}^{N}\mu_l I(z_l\le t), \tag{2.12}$$

όπου τα $z_1, ..., z_N$ είναι οι συνδυασμένοι (pooled) παρατηρούμενοι χρόνοι διακοπής. Αντικαθιστώντας την (2.12) στην (2.9) προκύπτει η ακόλουθη λογαριθμισμένη πιθανοφάνεια προφίλ (profile loglikelihood)

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta}+\sum_{l=1}^{N}\ln\mu_l+\sum_{i=1}^{n}\ln\left(\int_0^\infty \exp\Big(-u_i\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})\Big)\,u_i^{A_i}\,dF_{u_i}(u_i)\right). \tag{2.13}$$

□

Παραγωγίζοντας την παραπάνω πιθανοφάνεια ως προς μ_l, $l = 1, ..., N$ και θέτοντάς την ίση με μηδέν, οδηγούμαστε στην εξίσωση

$$\frac{1}{\mu_l}=\sum_{i=1}^{n}\frac{\left|L^{(A_i+1)}\left(\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})\right)\right|}{\left|L^{(A_i)}\left(\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})\right)\right|}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}I(z_l\le z_{ij}). \tag{2.14}$$

Η πρώτη και δεύτερη παράγωγος της (2.13) ως προς $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_d)^T$ δίνονται παρακάτω. Συγκεκριμένα, η παράγωγος ως προς β_{k_1}, $k_1 = 1, \ldots, d$ είναι

$$\frac{\vartheta l(\boldsymbol{\beta})}{\vartheta\beta_{k_1}}=\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}x_{ijk_1}-\sum_{i=1}^{n}\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})\;x_{ijk_1} \tag{2.15}$$

ενώ για τη δεύτερη παράγωγο $\frac{\vartheta^2 l(\boldsymbol{\beta})}{\vartheta\beta_{k_1}\vartheta\beta_{k_2}}$, $k_1 = 1, \ldots, d$, $k_2 = 1, \ldots, d$ έχουμε ότι ισούται με την ποσότητα

$$\sum_{i=1}^{n}\left\{\frac{|L^{(A_i+2)}(x)|}{|L^{(A_i)}(x)|}-\left(\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}\right)^2\right\}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})x_{ijk_1}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})x_{ijk_2}$$

$$-\sum_{i=1}^{n}\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})x_{ijk_1}x_{ijk_2} \tag{2.16}$$

όπου $x=\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})$, που ουσιαστικά εξαρτάται από το i.

Παρατήρηση 2.1 *Η μονομεταβλητή περίπτωση ανεξάρτητων δεδομένων (independent univariate data), όπου δηλαδή κάθε άτομο επηρεάζεται από τη δική του ευπάθεια, αποτελεί μια ειδική περίπτωση των ομαδοποιημένων δεδομένων. Η αντίστοιχη συνάρτηση του λογαρίθμου της πιθανοφάνειας μπορεί να ληφθεί από την (2.13), θέτοντας όλα τα $J_i = 1$ ενώ πλέον το n αντιπροσωπεύει τον αριθμό των ατόμων στο δείγμα. Ωστόσο, αυτή η περίπτωση πρέπει να αντιμετωπίζεται με προσοχή, καθότι η μεροληψία των εκτιμητών μέγιστης πιθανοφάνειας σε πεπερασμένου μεγέθους δείγματα μπορεί να είναι σημαντική, ιδιαίτερα όταν έχουμε μικρά ποσοστά αποκοπής [14], [85].*

Θα πρέπει επίσης να τονιστεί ότι ο λογάριθμος της πιθανοφάνειας που δίνεται στην (2.13) είναι γενικού χαρακτήρα και μπορεί να συμπεριλάβει διαφορετικές συνεχείς κατανομές της μεταβλητής ευπάθειας. Στην περίπτωση των ανεξάρτητων παρατηρήσεων, μετατρέπεται στον αντίστοιχο λογάριθμο της πιθανοφάνειας του μοντέλου Cox. Στη συνέχεια παρουσιάζουμε τρεις κατανομές της μεταβλητής ευπάθειας που χρησιμοποιούνται αρκετά συχνά και συγκεκριμένα, τη Γάμμα, η οποία έχει αναπτυχθεί και στους Fan και Li [54], την Αντίστροφη Γκαουσιανή και την Ομοιόμορφη.

2.3.1 Μοντέλο Ευπάθειας Γάμμα Κατανομής

Έστω ότι η ευπάθεια ακολουθεί την κατανομή Γάμμα$(\alpha, 1/\alpha)$ με συνάρτηση πυκνότητας πιθανότητας $f_u(u) = \frac{\alpha^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\alpha u}$, $\alpha > 0$, η οποία για λόγους προσδιορισιμότητας (identifiability) θεωρείται ότι έχει μέσο 1 και διασπορά $1/\alpha$. Ο αντίστοιχος μετασχηματισμός Laplace προκύπτει

$$L(x) = \left(\frac{\alpha}{\alpha + x}\right)^\alpha. \tag{2.17}$$

Επειδή η $n-$οστή παράγωγος του παραπάνω μετασχηματισμού Laplace είναι

$$L^{(n)}(x) = (-1)^n \frac{\Gamma(a+n)}{a^n \Gamma(a)} \Big(\frac{\alpha}{\alpha + x}\Big)^{a+n}, \tag{2.18}$$

ο λογάριθμος της πιθανοφάνειας (2.13), μαζί με την κατάλληλη συνάρτηση ποινής, γίνεται

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i} \delta_{ij} \mathbf{x}_{ij}^T \boldsymbol{\beta} + \sum_{l=1}^{N} \ln \mu_l - \sum_{i=1}^{n} (\alpha + A_i) \ln \left(\alpha + \sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} \sum_{l=1}^{N} \mu_l I(z_l \le z_{ij})\right)$$
$$+ \sum_{i=1}^{n} \alpha \ln \alpha + \sum_{i=1}^{n} \ln \left(\frac{\Gamma(A_i + \alpha)}{\Gamma(\alpha)}\right) - n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.19}$$

όπου ως $p_\lambda(.)$ θα θεωρούμε στην παρούσα διατριβή τη SCAD, τη Hard ή τη LASSO συνάρτηση ποινής. Μπορούμε εύκολα να παρατηρήσουμε ότι η νέα μορφή του λογαρίθμου της πιθανοφάνειας (2.19) διορθώνει την αντίστοιχη συνάρτηση (3.10) των Fan και Li [54], καθότι διαφέρει κατά τον επιπλέον όρο $\sum_{i=1}^{n} \ln\left(\Gamma(A_i + \alpha)\right)$. Επιπλέον, η παράγωγος ως προς μ_l, $l = 1, ..., N$ από τις (2.14) και (2.18) δίνει

$$\frac{1}{\mu_l} = \sum_{i=1}^{n} \frac{A_i + \alpha}{\alpha + \sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} \sum_{l=1}^{N} \mu_l I(z_l \le z_{ij})} \sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} I(z_l \le z_{ij}), \tag{2.20}$$

η οποία συμπίπτει με την (3.12) των Fan και Li [54].

2.3.2 Μοντέλο Ευπάθειας Αντίστροφης Γκαουσιανής Κατανομής

Έστω ότι η ευπάθεια ακολουθεί την Αντίστροφη Γκαουσιανή κατανομή $N^-(b_1, b_2)$ με συνάρτηση πυκνότητας πιθανότητας $f_u(u) = \left(\frac{b_2}{\pi u^3}\right)^{(1/2)} e^{\sqrt{4 b_1 b_2}} e^{-b_1 u - b_2 (1/u)}$, $b_1 \ge 0, b_2 > 0$, η οποία για λόγους προσδιορισιμότητας θεωρείται ότι έχει $b_1 = b_2 = b$, ώστε να έχουμε μέσο ίσο με 1 και διασπορά $1/2b$. Ο αντίστοιχος μετασχηματισμός Laplace $L(x)$ λαμβάνει τη μορφή

$$L(x) = e^{2b - \sqrt{4b(b+x)}}. \tag{2.21}$$

Η απόλυτη τιμή της A_i-οστής παραγώγου του μετασχηματισμού Laplace $L(x)$ που δίνεται στην (2.10), προκύπτει

$$|L^{(A_i)}(x)| = e^{2b - \sqrt{4(b+x)b}} E_{g_1}(U_i^{A_i}) = L(x) E_{g_1}(U_i^{A_i}), \tag{2.22}$$

όπου η αναμενόμενη τιμή λαμβάνεται ως προς μια συνάρτηση πυκνότητας πιθανότητας g_1 μίας $N^-(b + x, b)$ τυχαίας μεταβλητής. Σημειώνεται ότι η ποσότητα x αντιπροσωπεύει τον όρο $\sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}} \sum_{l=1}^{N} \mu_l I(z_l \le z_{ij})$, όπως ορίσθηκε και παραπάνω. Ως εκ τούτου, ο λογάριθμος

της πιθανοφάνειας (2.13) λαμβάνει τη μορφή

$$\sum_{l=1}^{N}\ln\mu_l+\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta}+\sum_{i=1}^{n}\ln\Big(L(x)\Big)+\sum_{i=1}^{n}\ln(E_{g_1}(U_i^{A_i})). \tag{2.23}$$

Από τη στιγμή τώρα που η $n-$οστή ροπή μίας $N^-(b_1,b_2)$ τυχαίας μεταβλητής είναι

$$e^{\sqrt{4b_1b_2}}\sqrt{4b_2/\pi}\left(\sqrt{\frac{b_2}{b_1}}\right)^{n-(1/2)}K_{1/2-n}(\sqrt{4b_1b_2}\,)$$

όπου $K_{1/2-n}(z)$ είναι η τροποποιημένη συνάρτηση Bessel δευτέρου τύπου, η (2.23) για $b_1=b+x$ και $b_2=b$ καθώς και με την κατάλληλη συνάρτηση ποινής, μετατρέπεται ως

$$\sum_{l=1}^{N}\ln\mu_l+\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta}+\sum_{i=1}^{n}\left(2b-\sqrt{4(b+x)b}\,\right)+\sum_{i=1}^{n}\frac{1}{2}(A_i-\frac{1}{2})\ln\left(\frac{b}{b+x}\right)$$

$$+\sum_{i=1}^{n}\sqrt{4(b+x)b}+\sum_{i=1}^{n}\frac{1}{2}\ln(\frac{4b}{\pi})+\sum_{i=1}^{n}\ln\left(K_{1/2-A_i}\left(\sqrt{4(b+x)b}\,\right)\right)-n\sum_{j=1}^{d}p_\lambda(|\beta_j|)=$$

$$\sum_{l=1}^{N}\ln\mu_l+\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta}+\sum_{i=1}^{n}2b+\sum_{i=1}^{n}\frac{1}{2}(A_i-\frac{1}{2})\ln\left(\frac{b}{b+x}\right)$$

$$+\sum_{i=1}^{n}\frac{1}{2}\ln(\frac{4b}{\pi})+\sum_{i=1}^{n}\ln\left(K_{1/2-A_i}\left(\sqrt{4(b+x)b}\,\right)\right)-n\sum_{j=1}^{d}p_\lambda(|\beta_j|). \tag{2.24}$$

Η παράγωγος ως προς μ_l, $l=1,...,N$ από τις (2.14) και (2.22) δίνει

$$\frac{1}{\mu_l}=\sum_{i=1}^{n}\sqrt{\frac{b}{b+x}}\,\frac{K_{1/2-A_i-1}(\sqrt{4(b+x)b}\,)}{K_{1/2-A_i}(\sqrt{4(b+x)b}\,)}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}I(z_l\le z_{ij}). \tag{2.25}$$

Παρατήρηση 2.2 *Εναλλακτικά, ο λογάριθμος της πιθανοφάνειας για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής, μπορεί να εξαχθεί λαμβάνοντας υπόψη το εξής. Επειδή η $n-$οστή ροπή μίας $N^-(b_1,b_2)$ τυχαίας μεταβλητής δίνεται και ως [153]*

$$b_1^n\sum_{j=0}^{n-1}\frac{(n-1+j)!}{j!(n-1-j)!}\left(\frac{2b_2}{b_1}\right)^{-j},$$

η αναμενόμενη τιμή στην (2.23) για $b_1=b+x$ και $b_2=b$ ισούται με

$$(b+x)^{A_i}\sum_{j=0}^{A_i-1}\frac{(A_i-1+j)!}{j!(A_i-1-j)!}\left(\frac{2b}{b+x}\right)^{-j}=$$

$$(b+x)^{A_i}\mathrm{hypergeom}\left([A_i,-A_i+1],[\,],-\frac{b+x}{2b}\right)$$

όπου η τελευταία συνάρτηση είναι η γενικευμένη υπεργεωμετρική συνάρτηση (generalized hypergeometric function). Συνεπώς, η (2.23) μαζί με την κατάλληλη συνάρτηση ποινής γίνεται

$$\sum_{l=1}^{N}\ln\mu_l+\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\sum_{i=1}^{n}\left(2b-\sqrt{4(b+x)b}\,\right)+$$

$$\sum_{i=1}^{n} A_i \ln(b+x) + \sum_{i=1}^{n} \ln\left(\mathrm{hypergeom}\left([A_i, -A_i+1], [\,], -\frac{b+x}{2b}\right)\right) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \quad (2.26)$$

Επίσης, η παράγωγος ως προς μ_l, $l = 1, ..., N$ *από τις (2.14) και (2.22) δίνει*

$$\frac{1}{\mu_l} = \sum_{i=1}^{n}(b+x)\frac{\mathrm{hypergeom}([A_i+1, -A_i], [\,], -\frac{b+x}{2b})}{\mathrm{hypergeom}([A_i, -A_i+1], [\,], -\frac{b+x}{2b})}\sum_{j=1}^{J_i} e^{\boldsymbol{x}_{ij}^T\boldsymbol{\beta}} I(z_l \le z_{ij}). \quad (2.27)$$

2.3.3 Μοντέλο Ευπάθειας Ομοιόμορφης Κατανομής

Έστω ότι η ευπάθεια ακολουθεί την Ομοιόμορφη κατανομή στο $[b, a]$ όπου $a > b > 0$. Τότε έχουμε ότι $f_u(u) = 1/(a-b), b \le u \le a$ και ο μετασχηματισμός Laplace δίνεται ως

$$L(x) = \frac{1}{(a-b)x}(e^{-bx} - e^{-ax}). \quad (2.28)$$

Από την (2.10) και έπειτα από μια σειρά ολοκληρώσεων κατά παράγοντες, έχουμε ότι η ποσότητα $L^{(A_i)}(x)$ ισούται με

$$(-1)^{A_i}\left\{e^{-bx}\left(\frac{b^{A_i}}{(a-b)x} + \frac{A_i b^{A_i-1}}{(a-b)x^2} + \frac{A_i(A_i-1)b^{A_i-2}}{(a-b)x^3} + \ldots + \frac{A_i(A_i-1)\ldots 2b^1}{(a-b)x^{A_i}} + \frac{A_i(A_i-1)\ldots 2}{(a-b)x^{A_i+1}}\right)\right.$$

$$\left. -e^{-ax}\left(\frac{a^{A_i}}{(a-b)x} + \frac{A_i a^{A_i-1}}{(a-b)x^2} + \frac{A_i(A_i-1)a^{A_i-2}}{(a-b)x^3} + \ldots + \frac{A_i(A_i-1)\ldots 2a^1}{(a-b)x^{A_i}} + \frac{A_i(A_i-1)\ldots 2}{(a-b)x^{A_i+1}}\right)\right\}$$

$$= (-1)^{A_i}\left\{\frac{e^{-bx}}{(a-b)x}\sum_{j=0}^{A_i} j!\binom{A_i}{j}\left(\frac{1}{x}\right)^j b^{A_i-j} - \frac{e^{-ax}}{(a-b)x}\sum_{j=0}^{A_i} j!\binom{A_i}{j}\left(\frac{1}{x}\right)^j a^{A_i-j}\right\}$$

$$= \frac{(-1)^{A_i}}{(a-b)x}\left\{e^{-bx}\frac{e^{bx}\left(2\,x^{-A_i} - x^{-A_i+1}\right)\Gamma\left(A_i+1, bx\right)}{-2+x} - e^{-ax}\frac{e^{ax}\left(2\,x^{-A_i} - x^{-A_i+1}\right)\Gamma\left(A_i+1, ax\right)}{-2+x}\right\}$$

η οποία απλοποιείται στην ποσότητα

$$\frac{(-1)^{A_i}}{(a-b)x}\frac{2\,x^{-A_i} - x^{-A_i+1}}{-2+x}\{\Gamma\left(A_i+1, bx\right) - \Gamma\left(A_i+1, ax\right)\} =$$

$$\frac{(-1)^{A_i+1}}{(a-b)x^{A_i+1}}\{\Gamma\left(A_i+1, bx\right) - \Gamma\left(A_i+1, ax\right)\} \quad (2.29)$$

όπου $\Gamma(a, x)$ είναι η μη-πλήρης συνάρτηση Γάμμα, η οποία ορίζεται ως $\int_x^\infty t^{a-1}e^{-t}dt$. Ο λογάριθμος της πιθανοφάνειας (2.13) μαζί με την κατάλληλη συνάρτηση ποινής προκύπτει ως

$$\sum_{l=1}^{N}\ln\mu_l + \sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta} + \sum_{l=1}^{n}\ln\left(\Gamma(A_i+1, bx) - \Gamma(A_i+1, ax)\right) -$$

$$\sum_{l=1}^{n}\ln(a-b) - \sum_{l=1}^{n}(A_i+1)\ln x - n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \quad (2.30)$$

Η παράγωγος ως προς μ_l, $l = 1, ..., N$ από τις (2.14) και (2.29) δίνει

$$\frac{1}{\mu_l} = \sum_{l=1}^{n}\left\{\frac{\Gamma\left(A_i+2, bx\right) - \Gamma\left(A_i+2, ax\right)}{\left(\Gamma\left(A_i+1, bx\right) - \Gamma\left(A_i+1, ax\right)\right)x}\right\}\sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} I(z_l \le z_{ij}). \quad (2.31)$$

2.4 Μελέτες Προσομοίωσης

Σε αυτήν την ενότητα, παρουσιάζουμε μια μελέτη προσομοιώσης που εκτελέσαμε, προκειμένου να αναδειχθεί η προτεινόμενη μεθοδολογία, δίνοντας έμφαση στις τρεις προαναφερθείσες κατανομές ευπάθειας. Να σημειώσουμε, ότι για την περίπτωση της Γάμμα κατανομής, εκτελέστηκαν προσομοιώσεις και στην επιστημονική εργασία [54], με τη διαφορά ότι εδώ παρουσιάζουμε τα αποτελέσματα βάσει της σωστής πλέον πιθανοφάνειας.

2.4.1 Σχεδιασμός Προσομοιώσεων

Χρησιμοποιήσαμε το μοτίβο προσομοιώσεων των Fan και Li [54], συνεπώς δημιουργήθηκαν 100 σύνολα δεδομένων αποτελούμενα από n ομάδες και J άτομα σε κάθε ομάδα, βάσει του μοντέλου

$$h(t|\mathbf{x}, u) = u \exp(\mathbf{x}^T \boldsymbol{\beta}), \tag{2.32}$$

όπου το διάνυσμα των πραγματικών παραμέτρων είναι $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0.6, 0)$, τα x_i παράγονται από την τυποποιημένη Κανονική κατανομή και η συσχέτιση μεταξύ των x_i και x_j είναι $\rho^{|i-j|}$ με $\rho = 0.5$. Η κατανομή του χρόνου αποκοπής είναι Εκθετική με μέσο $U \exp(\mathbf{x}^T \boldsymbol{\beta}_0)$, όπου το U παράγεται τυχαία από την Ομοιόμορφη κατανομή στο διάστημα $[1, 3]$ για κάθε προσομοιωμένο σύνολο δεδομένων, ούτως ώστε να έχουμε ποσοστό αποκοπής 30%. Το $\boldsymbol{\beta}_0$ θεωρείται ως γνωστή σταθερά ώστε η αποκοπή να είναι μη πληροφοριακή.

Αναφορικά με την ευπάθεια u, εξετάσαμε τις εξής περιπτώσεις: Αρχικά, θεωρήσαμε ευπάθεια Γάμμα κατανομής με $\alpha = 4$ και Αντίστροφης Γκαουσιανής κατανομής με $b = 2$, ώστε και στις δύο περιπτώσεις η διασπορά να είναι ίση με $1/4$. Για αυτές τις κατανομές, εξετάσαμε επίσης και την επίδραση όπου θα είχε η προσαρμογή λανθασμένης κατανομής της ευπάθειας, στην εκτίμηση των συντελεστών και στην επιλογή μεταβλητών. Τέλος, θεωρήσαμε ευπάθεια Ομοιόμορφης κατανομής, όπου βασιζόμενοι στους Lee και Klein [113], θέσαμε καταρχήν ως $b = 0.25$ και $a = 0.75$, έπειτα θέσαμε ως $b = 1.25$ και $a = 1.75$ και ως τελευταία περίπτωση θέσαμε τα b και a να έχουν τις τιμές 0.75 και 1.25 αντίστοιχα.

2.4.2 Κριτήρια Αξιολόγησης της Προτεινόμενης Μεθοδολογίας

Εξετάσαμε την απόδοση της διαδικασίας επιλογής μεταβλητών μέσω της νέας γενικευμένης μορφής ποινικοποιημένης πιθανοφάνειας, με γνώμονα τα σφάλματα μοντέλου (model errors-ME), την πολυπλοκότητα και την ακρίβεια του μοντέλου (model complexity and accuracy). Τα σφάλματα μοντέλου της προτεινόμενης μεθοδολογίας συγκρίθησαν με αυτά των εκτιμητριών της μέγιστης profile πιθανοφάνειας. Συγκεκριμένα, παρουσιάζουμε τη διάμεσο του σχετικού σφάλματος του μοντέλου (Median of Relative Model Error-MRME) των 100 προσομοιωμένων συνόλων δεδομένων, για τις προαναφερθείσες τιμές των n και J. Να σημειώσουμε, ότι στα μοντέλα ευπάθειας το σφάλμα μοντέλου ορίζεται ως [54]

$$ME = E[\exp(-\mathbf{x}^T \hat{\boldsymbol{\beta}}) - \exp(-\mathbf{x}^T \boldsymbol{\beta}_0)]^2, \tag{2.33}$$

ενώ για το σχετικό σφάλμα έχουμε ότι $RME = ME/ME_{full}$, με ME_{full} το σφάλμα μοντέλου υπολογιζόμενο προσαρμόζοντας στα δεδομένα το πλήρες μοντέλο. Αναφέρουμε επίσης το μέσο αριθμό των σωστά και λανθασμένα αναγνωρισμένων μηδενικών συντελεστών (Aver. no. of 0 coeff., στήλες correct και incorrect αντίστοιχα στους επόμενους πίνακες).

Εν συνεχεία, εξετάσαμε την ακρίβεια του τύπου εύρεσης του τυπικού σφάλματος των Fan και Li [54]. Συγκεκριμένα, η διάμεση απόλυτη απόκλιση (median absolute deviation) διαιρούμενη με 0.6745, η οποία συμβολίζεται με SD στους πίνακες, των 100 εκτιμηθέντων συντελεστών στις προσομοιώσεις, μπορεί να θεωρηθεί ως το πραγματικό τυπικό σφάλμα. Σημειώνουμε ότι, γενικά, η διάμεση απόλυτη απόκλιση (MAD) ενός συνόλου τιμών $\{t_i\}$,

$i = 1, \ldots, n$ δίνεται ως MAD=median($|t_i - \tilde{t}|$), όπου $\tilde{t}$ η διάμεση τιμή του συνόλου $\{t_i\}$. Η διάμεσος των 100 εκτιμηθέντων τυπικών αποκλίσεων (SDs), την οποία συμβολίζουμε ως SD_m και το διάμεσο απόλυτο σφάλμα απόκλισης των 100 εκτιμηθέντων τυπικών σφαλμάτων, διαιρούμενο με 0.6745, το οποίο συμβολίζεται ως SD_{mad}, καθορίζουν τη γενική συμπεριφορά και απόδοση του τύπου εύρεσης του τυπικού σφάλματος. Στις προσομοιώσεις μας, τα τυπικά σφάλματα των εκτιμηθέντων συντελεστών οι οποίοι αποκλείονται από το επιλεγμένο μοντέλο, θέτονται ίσοι με 0, συνεπώς παρουσιάζουμε μόνο τα αποτελέσματα για τους μη μηδενικούς συντελεστές.

Χρησιμοποιήσαμε τις ποινικοποιημένες μεθόδους, με τις συναρτήσεις ποινής SCAD, LASSO και Hard, σε συνδυασμό με τη γενικευμένη διασταυρωμένη επικύρωση ως τη διαδικασία επιλογής της ρυθμιστικής παραμέτρου. Για τις περιπτώσεις της Γάμμα και Αντίστροφης Γκαουσιανής κατανομής, εξετάσαμε επίσης και την απόδοση της μεθόδου επιλογής καλύτερου υποσυνόλου (Best Subset) βάσει του κριτηρίου BIC. Για τις ίδιες περιπτώσεις, παραθέτουμε για κάθε ποινικοποιημένη μέθοδο τις μέσες τιμές των 100 εκτιμηθέντων συντελεστών των πραγματικά ενεργών μεταβλητών X_1, X_4 και X_7, καθώς επίσης και τις μέσες τιμές των α και b για τις οποίες μεγιστοποιήθηκε η πιθανοφάνεια. Για όλα τα παραπάνω, θεωρήσαμε ως τιμές σύγκρισης (και στις τρεις περιπτώσεις κατανομών), τα αποτελέσματα που λάβαμε από τον Oracle εκτιμητή, τα οποία παρήχθησαν προσαρμόζοντας το ιδανικό μοντέλο που συμπεριλαμβάνει μόνο τις X_1, X_4 και X_7.

Τέλος, να αναφέρουμε λίγα στοιχεία για τον αλγόριθμο μεγιστοποίησης που εφαρμόσαμε στις προσομοιώσεις, όπου βασιστήκαμε στην εργασία των Fan και Li [54]. Καταρχήν, η περίπτωση της Ομοιόμορφης κατανομής $U(b, a)$ δεν παρουσιάζει ιδιαίτερες δυσκολίες, καθότι οι τιμές των b και a είναι προκαθορισμένες. Για τις υπόλοιπες κατανομές, υιοθετήσαμε την ιδέα της χρήσης ενός συνόλου από πιθανές τιμές της παραμέτρου α (ή b) που εμφανίζεται στην κανανομή της ευπάθειας, ώστε να αποφύγουμε τυχόν δυσκολίες στη μεγιστοποίηση ως προς α (ή b). Συγκεκριμένα, για την περίπτωση της Γάμμα (ή της Αντίστροφης Γκαουσιανής) κατανομής του ποινικοποιημένου μοντέλου ευπάθειας, ξεκινάμε δίνοντας αρχικές τιμές για τα α (ή b αντίστοιχα), $\boldsymbol{\beta}$ και μ_l, $l = 1, \ldots, N$. Ανανεώνουμε έπειτα την ποσότητα μ_l από την (2.20) (ή την (2.25)) και προκύπτει η ποινικοποιημένη πιθανοφάνεια (2.19) (ή (2.24)). Εν συνεχεία, μεγιστοποιούμε την (2.19) (ή την (2.24)) ως προς $(\alpha, \boldsymbol{\beta})$ (ή $(b, \boldsymbol{\beta})$) και επαναλαμβάνουμε αυτά τα βήματα. Ένας αρχικός εκτιμητής για το $\boldsymbol{\beta}$ μπορεί να αποκτηθεί αγνοώντας την πιθανή εξάρτηση μεταξύ των ομάδων. Συνεπώς, χρησιμοποιήθηκε ο εκτιμητής μέγιστης ψευδο-μερικής πιθανοφάνειας του $\boldsymbol{\beta}$. Επιπλέον, ως αρχικός εκτιμητής για το μ_l, $l = 1, ..., N$, ορίστηκε η ποσότητα $\hat{\mu}_l = \left(\sum_{i \epsilon R_l} \exp(\mathbf{x}_i^T \boldsymbol{\beta})\right)^{-1}$, όπου το R_l δηλώνει το σύνολο κινδύνου (risk set) ακριβώς πριν τον παρατηρούμενο χρόνο αποτυχίας z_l.

2.4.3 Αποτελέσματα Προσομοιώσεων

Στο σημείο αυτό παραθέτουμε στους παρακάτω Πίνακες 2.1-2.16 όλα τα αποτελέσματα των προσομοιώσεων που εκτελέσαμε, τα οποία και σχολιάζουμε στη συνέχεια.

Πίνακας 2.1: Αποτελέσματα προσομοίωσης για δεδομένα παραγόμενα βάσει του μοντέλου ευπάθειας Γάμμα κατανομής: Η εκτίμηση των παραμέτρων γίνεται μέσω του μοντέλου ευπάθειας Γάμμα καθώς και του Αντίστροφου Γκαουσιανού μοντέλου - περίπτωση λανθασμένης κατανομής (σε παρένθεση)

Μέθοδος	**MRME**	**Aver. no. of 0 coeff.**	
		correct	incorrect
n=50, J=2			
SCAD	0.5470(0.5433)	4.5600(4.4500)	0.0300(0.0100)
LASSO	0.5590(0.8049)	4.1200(3.9600)	0.0200(0.0100)
Hard	0.7255(0.9045)	4.2800(4.1900)	0.0200(0.0200)
Best Subset	0.6068(0.7721)	4.7200(4.7100)	0.1100(0.1000)
Oracle	0.4210(0.4671)	5(5)	0(0)
n=50, J=5			
SCAD	0.6290(0.6523)	4.6200(4.4600)	0(0)
LASSO	0.7750(0.8899)	4.2000(4.0300)	0(0)
Hard	0.9579(0.9508)	4.3600(4.1400)	0(0)
Best Subset	0.5517(0.6026)	4.9100(4.9000)	0(0.0300)
Oracle	0.5799(0.5789)	5(5)	0(0)
n=100, J=2			
SCAD	0.6432(0.5277)	4.6900(4.4000)	0(0)
LASSO	1.0945(1.3630)	4.1000(4.0500)	0(0)
Hard	0.9470(0.9389)	4.1700(4.1000)	0(0)
Best Subset	0.5825(0.6428)	4.8600(4.8500)	0.0100(0.0100)
Oracle	0.6272(0.4844)	5(5)	0(0)
n=100, J=5			
SCAD	0.6215(0.6410)	4.3500(4.3500)	0(0)
LASSO	1.0648(0.6629)	4.0600(4.1700)	0(0)
Hard	0.9584(0.9880)	4.1600(4.2800)	0(0)
Best Subset	0.5321(0.5900)	4.9000(4.9000)	0(0.0300)
Oracle	0.5980(0.5805)	5(5)	0(0)

Πίνακας 2.2: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Γάμμα κατανομής

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD	0.0923	0.1235	0.0139	0.1107	0.1324	0.0183	0.1157	0.1203	0.0176
LASSO	0.0853	0.1119	0.0100	0.1132	0.1208	0.0134	0.1262	0.1053	0.0121
Hard	0.1173	0.1352	0.0168	0.1508	0.1507	0.0221	0.1459	0.1387	0.0214
Best Subset	0.1106	0.1345	0.0119	0.1191	0.1309	0.0173	0.1324	0.1271	0.0194
Oracle	0.0924	0.1257	0.0130	0.0986	0.1322	0.0161	0.1058	0.1281	0.0138
n=50, J=5									
SCAD	0.0625	0.0857	0.0087	0.0824	0.0854	0.0083	0.0482	0.0835	0.0070
LASSO	0.0583	0.0808	0.0083	0.0733	0.0829	0.0080	0.0576	0.0781	0.0064
Hard	0.0666	0.0916	0.0117	0.0817	0.0934	0.0101	0.0694	0.0935	0.0111
Best Subset	0.0568	0.0823	0.0063	0.0740	0.0826	0.0074	0.0738	0.0816	0.0076
Oracle	0.0580	0.0848	0.0084	0.0827	0.0854	0.0070	0.0471	0.0837	0.0064
n=100, J=2									
SCAD	0.0887	0.0963	0.0070	0.0772	0.0949	0.0070	0.0729	0.0926	0.0083
LASSO	0.0723	0.0867	0.0064	0.0886	0.0883	0.0059	0.0705	0.0829	0.0062
Hard	0.0793	0.0998	0.0088	0.1065	0.1053	0.0117	0.0978	0.1046	0.0113
Best Subset	0.0780	0.0954	0.0079	0.0721	0.0930	0.0070	0.0779	0.0944	0.0072
Oracle	0.0866	0.0963	0.0072	0.0744	0.0959	0.0072	0.0742	0.0925	0.0079
n=100, J=5									
SCAD	0.0390	0.0605	0.0040	0.0463	0.0608	0.0039	0.0440	0.0605	0.0037
LASSO	0.0427	0.0578	0.0040	0.0382	0.0590	0.0038	0.0477	0.0576	0.0034
Hard	0.0425	0.0638	0.0053	0.0515	0.0661	0.0055	0.0529	0.0663	0.0057
Best Subset	0.0345	0.0582	0.0034	0.0557	0.0589	0.0043	0.0526	0.0603	0.0038
Oracle	0.0385	0.0604	0.0039	0.0453	0.0604	0.0039	0.0464	0.0605	0.0038

Πίνακας 2.3: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Γάμμα κατανομής - περίπτωση λανθασμένης κατανομής: Η εκτίμηση των παραμέτρων γίνεται μέσω του Αντίστροφου Γκαουσιανού μοντέλου ευπάθειας

	$\hat{\boldsymbol{\beta}}_1$			$\hat{\boldsymbol{\beta}}_4$			$\hat{\boldsymbol{\beta}}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD	0.1138	0.1235	0.0154	0.1116	0.1210	0.0129	0.0873	0.1238	0.0154
LASSO	0.1082	0.1121	0.0112	0.1219	0.1103	0.0111	0.0955	0.1109	0.0111
Hard	0.1276	0.1371	0.0192	0.1152	0.1332	0.0163	0.1162	0.1422	0.0197
Best Subset	0.1136	0.1321	0.0117	0.1196	0.1282	0.0189	0.1320	0.1234	0.0177
Oracle	0.1101	0.1255	0.0159	0.1027	0.1230	0.0134	0.0877	0.1276	0.0122
n=50, J=5									
SCAD	0.0829	0.0837	0.0085	0.0709	0.0847	0.0076	0.0638	0.0864	0.0093
LASSO	0.0870	0.0779	0.0078	0.0769	0.0813	0.0076	0.0727	0.0786	0.0074
Hard	0.0942	0.0900	0.0096	0.0870	0.0955	0.0110	0.0728	0.0940	0.0085
Best Subset	0.0562	0.0807	0.0118	0.0617	0.0838	0.0155	0.0695	0.0815	0.0115
Oracle	0.08087	0.0838	0.0083	0.0661	0.0853	0.0071	0.0648	0.0852	0.0091
n=100, J=2									
SCAD	0.0632	0.0939	0.0088	0.0784	0.0938	0.0066	0.0690	0.0943	0.0073
LASSO	0.0567	0.0848	0.0074	0.0728	0.0887	0.0063	0.0648	0.0848	0.0056
Hard	0.0668	0.1001	0.0113	0.0770	0.1075	0.0105	0.0788	0.1029	0.0103
Best Subset	0.0778	0.0931	0.0077	0.0744	0.0909	0.0078	0.0795	0.0931	0.0083
Oracle	0.0569	0.0943	0.0087	0.0781	0.0947	0.0071	0.0643	0.0945	0.0065
n=100, J=5									
SCAD	0.0468	0.0590	0.0033	0.0495	0.0587	0.0036	0.0396	0.0606	0.0039
LASSO	0.0513	0.0565	0.0029	0.0543	0.0579	0.0036	0.0474	0.0577	0.0035
Hard	0.0472	0.0619	0.0046	0.0544	0.0645	0.0053	0.0480	0.0652	0.0051
Best Subset	0.0444	0.0596	0.0045	0.0671	0.0607	0.0054	0.0682	0.0610	0.0046
Oracle	0.0474	0.0588	0.0032	0.0490	0.0587	0.0033	0.0388	0.0608	0.0040

Πίνακας 2.4: Μέσες τιμές των μη μηδενικών συντελεστών και των τιμών α για το μοντέλο ευπάθειας Γάμμα κατανομής

Μέθοδος	β_1	β_4	β_7	α
n=50, J=2				
SCAD	0.8266	1.0138	0.5608	4.3144
LASSO	0.6907	0.8502	0.4662	4.6011
Hard	0.8424	1.0276	0.6048	4.9217
Best Subset	0.7840	1.0145	0.6798	3.9027
Oracle	0.8175	1.0099	0.5905	4.1605
n=50, J=5				
SCAD	0.8931	1.1025	0.6428	4.2618
LASSO	0.7947	0.9885	0.5504	4.3608
Hard	0.9008	1.1142	0.6436	4.1292
Best Subset	0.8411	0.9901	0.5509	4.2840
Oracle	0.8878	1.1014	0.6462	4.2187
n=100, J=2				
SCAD	0.8092	1.0062	0.5884	4.1143
LASSO	0.7111	0.8914	0.4948	4.3271
Hard	0.8220	1.0217	0.5986	3.6090
Best Subset	0.7776	0.9926	0.6011	3.9102
Oracle	0.8105	1.0077	0.5896	3.8727
n=100, J=5				
SCAD	0.8480	1.0371	0.6161	3.5645
LASSO	0.7847	0.9609	0.5567	3.4804
Hard	0.8512	1.0354	0.6172	3.5809
Best Subset	0.7822	0.9892	0.5928	3.8275
Oracle	0.8464	1.0369	0.6151	3.6071

Πίνακας 2.5: Μέσες τιμές των μη μηδενικών συντελεστών και των τιμών α για το μοντέλο ευπάθειας Γάμμα κατανομής - περίπτωση λανθασμένης κατανομής: Η εκτίμηση των παραμέτρων γίνεται μέσω του Αντίστροφου Γκαουσιανού μοντέλου ευπάθειας

Μέθοδος	β_1	β_4	β_7	α
n=50, J=2				
SCAD	0.7693	0.9975	0.5937	4.2901
LASSO	0.6562	0.8691	0.4954	4.7124
Hard	0.7918	1.0343	0.6273	4.0549
Best Subset	0.8203	0.9713	0.5899	4.4186
Oracle	0.7828	0.9950	0.6079	3.9637
n=50, J=5				
SCAD	0.8352	1.1050	0.6540	3.1664
LASSO	0.7769	0.9915	0.5695	4.0940
Hard	0.8583	1.1025	0.6666	3.1661
Best Subset	0.7662	0.9613	0.5856	3.7867
Oracle	0.8381	1.0575	0.6509	3.3048
n=100, J=2				
SCAD	0.7772	1.0013	0.6101	3.4800
LASSO	0.6807	0.8823	0.5161	3.8980
Hard	0.7849	1.0076	0.6142	3.3873
Best Subset	0.8094	0.9813	0.5092	3.9602
Oracle	0.7846	1.0050	0.6110	3.1052
n=100, J=5				
SCAD	0.8357	1.0529	0.6223	3.6326
LASSO	0.7720	0.9776	0.5658	3.6478
Hard	0.8379	1.0546	0.6265	3.6204
Best Subset	0.7784	0.9741	0.5792	3.9745
Oracle	0.8309	1.0485	0.6224	3.6208

Πίνακας 2.6: Αποτελέσματα προσομοίωσης για δεδομένα παραγόμενα βάσει του μοντέλου ευπάθειας Αντίστροφης Γκαουσιανής κατανομής: Η εκτίμηση των παραμέτρων γίνεται μέσω του Αντίστροφου Γκαουσιανού μοντέλου ευπάθειας καθώς και του Γάμμα μοντέλου - περίπτωση λανθασμένης κατανομής (σε παρένθεση)

Μέθοδος	**MRME**	**Aver. no. of 0 coeff.**	
		correct	incorrect
n=50, J=2			
SCAD	0.4544(0.4364)	4.5000(4.6900)	0(0.0600)
LASSO	0.5253(0.8207)	3.8800(4.0800)	0(0.0400)
Hard	0.9188(0.9907)	4.0400(3.9700)	0(0.0400)
Best Subset	0.4310(0.4035)	4.7800(4.7900)	0.1000(0.0600)
Oracle	0.4452(0.3657)	5(5)	0(0)
n=50, J=5			
SCAD	0.5741(0.5626)	4.5100(4.3000)	0(0)
LASSO	0.7874(0.6746)	4.0100(4.0300)	0(0)
Hard	0.9599(0.9430)	4.1500(4.1000)	0(0)
Best Subset	0.6943(0.4961)	4.9200(4.9200)	0(0)
Oracle	0.4991(0.4516)	5(5)	0(0)
n=100, J=2			
SCAD	0.5592(0.5711)	4.4800(4.7600)	0(0)
LASSO	1.0795(0.6610)	4.0800(4.3200)	0(0)
Hard	0.8918(0.7378)	4.2200(4.3300)	0(0)
Best Subset	0.8572(0.5956)	4.8500(4.7700)	0.0200(0.0600)
Oracle	0.5275(0.5183)	5(5)	0(0)
n=100, J=5			
SCAD	0.7056(0.7159)	4.4200(4.3400)	0(0)
LASSO	0.8014(0.7617)	4.1400(4.1800)	0(0)
Hard	0.9549(0.9710)	4.1900(4.2300)	0(0)
Best Subset	0.8021(0.8267)	4.8400(4.8500)	0.0100(0.0600)
Oracle	0.7199(0.6890)	5(5)	0(0)

Πίνακας 2.7: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD	0.1121	0.1252	0.0150	0.1271	0.1264	0.0154	0.0937	0.1140	0.0143
LASSO	0.1193	0.1088	0.0112	0.1265	0.1137	0.0091	0.0885	0.1091	0.0089
Hard	0.1461	0.1329	0.0161	0.1513	0.1441	0.0178	0.1216	0.1406	0.0158
Best Subset	0.1337	0.1271	0.0130	0.1333	0.1238	0.0155	0.1182	0.1231	0.0116
Oracle	0.1095	0.1255	0.0151	0.1206	0.1262	0.0139	0.0914	0.1249	0.0124
n=50, J=5									
SCAD	0.0538	0.0827	0.0080	0.0593	0.0819	0.0072	0.0567	0.0815	0.0069
LASSO	0.0615	0.0779	0.0067	0.0606	0.0798	0.0066	0.0587	0.0768	0.0064
Hard	0.0711	0.0892	0.0098	0.0729	0.0914	0.0087	0.0619	0.0883	0.0083
Best Subset	0.0537	0.0808	0.0108	0.0732	0.0824	0.0084	0.0788	0.0838	0.0090
Oracle	0.0543	0.0827	0.0078	0.0582	0.0817	0.0066	0.0544	0.0804	0.0065
n=100, J=2									
SCAD	0.0736	0.0920	0.0077	0.0729	0.0903	0.0085	0.0813	0.0919	0.0072
LASSO	0.0717	0.0843	0.0066	0.0756	0.0857	0.0066	0.0796	0.0839	0.0056
Hard	0.0915	0.0986	0.0094	0.0875	0.1016	0.0108	0.0938	0.1041	0.0104
Best Subset	0.0730	0.0915	0.0078	0.1195	0.0927	0.0075	0.0686	0.0926	0.0088
Oracle	0.0721	0.0933	0.0072	0.0728	0.0904	0.0082	0.0800	0.0927	0.0069
n=100, J=5									
SCAD	0.0445	0.0587	0.0043	0.0537	0.0590	0.0038	0.0461	0.0593	0.0038
LASSO	0.0516	0.0556	0.0036	0.0579	0.0583	0.0035	0.0468	0.0568	0.0035
Hard	0.0495	0.0619	0.0048	0.0629	0.0660	0.0063	0.0563	0.0652	0.0055
Best Subset	0.0897	0.0623	0.0094	0.0980	0.0644	0.0116	0.0759	0.0658	0.0126
Oracle	0.0439	0.0587	0.0040	0.0556	0.0587	0.0040	0.0414	0.0590	0.0042

Πίνακας 2.8: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής - περίπτωση λανθασμένης κατανομής: Η εκτίμηση των παραμέτρων γίνεται μέσω του Γάμμα μοντέλου ευπάθειας

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD	0.1271	0.1340	0.0109	0.1410	0.1278	0.0143	0.1185	0.1249	0.0170
LASSO	0.1045	0.1156	0.0089	0.1461	0.1175	0.0101	0.1203	0.1107	0.0092
Hard	0.1366	0.1433	0.0140	0.1778	0.1490	0.0188	0.1597	0.1508	0.0216
Best Subset	0.0874	0.0889	0.0209	0.0891	0.0980	0.0212	0.1210	0.0915	0.0175
Oracle	0.1245	0.1345	0.0099	0.1439	0.1296	0.0144	0.1188	0.1364	0.0130
n=50, J=5									
SCAD	0.0591	0.0822	0.0075	0.0705	0.0864	0.0101	0.0694	0.0849	0.0070
LASSO	0.0598	0.0761	0.0055	0.0620	0.0832	0.0077	0.0741	0.0782	0.0075
Hard	0.0706	0.0863	0.0067	0.0779	0.0960	0.0101	0.0965	0.0905	0.0126
Best Subset	0.0650	0.0809	0.0089	0.0684	0.0821	0.0083	0.0836	0.0827	0.0083
Oracle	0.0560	0.0822	0.0076	0.0652	0.0862	0.0094	0.0668	0.0842	0.0074
n=100, J=2									
SCAD	0.0835	0.0995	0.0084	0.0869	0.0961	0.0086	0.0840	0.0967	0.0091
LASSO	0.0781	0.0895	0.0061	0.0835	0.0912	0.0073	0.0976	0.0862	0.0072
Hard	0.0792	0.1071	0.0095	0.0869	0.1108	0.0117	0.0933	0.1094	0.0111
Best Subset	0.0946	0.1034	0.0170	0.1175	0.1055	0.0135	0.1043	0.1007	0.0120
Oracle	0.0835	0.0994	0.0083	0.0867	0.0973	0.0082	0.0880	0.0965	0.0086
n=100, J=5									
SCAD	0.0503	0.0592	0.0039	0.0536	0.0590	0.0033	0.0425	0.0592	0.0033
LASSO	0.0496	0.0568	0.0032	0.0582	0.0585	0.0036	0.0487	0.0560	0.0037
Hard	0.0563	0.0609	0.0045	0.0612	0.0648	0.0048	0.0509	0.0654	0.0060
Best Subset	0.0584	0.0635	0.0100	0.0623	0.0641	0.0098	0.0686	0.0664	0.0116
Oracle	0.0509	0.0592	0.0039	0.0546	0.0589	0.0032	0.0460	0.0590	0.0034

Πίνακας 2.9: Μέσες τιμές των μη μηδενικών συντελεστών και των τιμών b για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής

Μέθοδος	β_1	β_4	β_7	b
n=50, J=2				
SCAD	0.8512	1.0668	0.6277	2.3499
LASSO	0.7635	0.9480	0.5720	2.7565
Hard	0.8493	1.0568	0.6424	2.4584
Best Subset	0.8100	1.0235	0.6199	2.3259
Oracle	0.8317	1.0531	0.6314	2.3743
n=50, J=5				
SCAD	0.8380	1.0116	0.6051	2.2826
LASSO	0.7545	0.9170	0.5269	2.2831
Hard	0.8439	1.0242	0.6064	2.4730
Best Subset	0.8051	1.0304	0.6071	2.5682
Oracle	0.8347	1.0101	0.6042	2.3407
n=100, J=2				
SCAD	0.8057	0.9974	0.5956	2.1514
LASSO	0.7050	0.8871	0.5091	2.4742
Hard	0.8157	1.0264	0.6160	2.5848
Best Subset	0.8288	0.9891	0.6120	2.6646
Oracle	0.8013	0.9936	0.5964	2.3843
n=100, J=5				
SCAD	0.7905	0.9895	0.5786	2.2237
LASSO	0.7551	0.9392	0.5574	2.3298
Hard	0.7644	0.9707	0.5737	2.5580
Best Subset	0.8123	1.0337	0.6196	2.4640
Oracle	0.7881	0.9863	0.5674	2.1810

Πίνακας 2.10: Μέσες τιμές των μη μηδενικών συντελεστών και των τιμών b για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής - περίπτωση λανθασμένης κατανομής: Η εκτίμηση των παραμέτρων γίνεται μέσω του μοντέλου ευπάθειας Γάμμα

Μέθοδος	β_1	β_4	β_7	b
n=50, J=2				
SCAD	0.8349	1.0483	0.5906	2.6912
LASSO	0.6989	0.9092	0.5495	2.8535
Hard	0.8775	1.1148	0.6779	2.6584
Best Subset	0.8087	1.0343	0.6086	2.6183
Oracle	0.8465	1.0824	0.6338	2.4315
n=50, J=5				
SCAD	0.8352	1.0319	0.6251	2.6172
LASSO	0.7528	0.9339	0.5456	2.7168
Hard	0.8425	1.0434	0.6394	2.7317
Best Subset	0.8184	1.0446	0.6140	2.4253
Oracle	0.8313	1.0267	0.6212	2.6372
n=100, J=2				
SCAD	0.8126	1.0067	0.5992	2.4465
LASSO	0.7088	0.8938	0.5099	2.7263
Hard	0.8162	1.0261	0.6152	2.8374
Best Subset	0.8162	1.0261	0.6152	2.5030
Oracle	0.8049	0.9978	0.5983	2.4632
n=100, J=5				
SCAD	0.8221	1.0303	0.6179	2.3404
LASSO	0.7639	0.9616	0.5646	2.6171
Hard	0.8289	1.0363	0.6284	2.5291
Best Subset	0.8238	1.0475	0.6280	2.3801
Oracle	0.8192	1.0255	0.6152	2.3087

Πίνακας 2.11: Αποτελέσματα για το μοντέλο ευπάθειας Ομοιόμορφης κατανομής - Περίπτωση 1: $b = 0.25$ και $a = 0.75$

Μέθοδος	**MRME**	**Aver. no. of 0 coeff.**	
		correct	incorrect
n=50, J=2			
SCAD	0.2723	4.4100	0.1300
LASSO	0.5858	4.0400	0.0200
Hard	0.4425	4.5700	0.1800
Oracle	0.3114	5	0
n=50, J=5			
SCAD	0.6909	4.8000	0
LASSO	1.1094	4.0200	0
Hard	0.6308	4.8400	0
Oracle	0.6318	5	0
n=100, J=2			
SCAD	0.6193	4.7200	0
LASSO	1.1136	3.8900	0
Hard	0.6737	4.8700	0.0200
Oracle	0.5720	5	0
n=100, J=5			
SCAD	0.8584	4.4600	0
LASSO	1.2529	3.9700	0
Hard	0.8928	4.8700	0
Oracle	0.8763	5	0

Πίνακας 2.12: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Ομοιόμορφης κατανομής - Περίπτωση 1: $b = 0.25$ και $a = 0.75$

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD	0.1346	0.1188	0.0066	0.1275	0.1040	0.0084	0.1149	0.0916	0.0106
LASSO	0.1106	0.0821	0.0120	0.0985	0.0659	0.0127	0.1018	0.0725	0.0082
Hard	0.1359	0.1062	0.0084	0.1365	0.1145	0.0081	0.1158	0.0902	0.0095
Oracle	0.1256	0.1091	0.0078	0.1298	0.0901	0.0085	0.1002	0.0809	0.0068
n=50, J=5									
SCAD	0.0597	0.0568	0.0043	0.0731	0.0561	0.0050	0.0595	0.0537	0.0041
LASSO	0.0632	0.0402	0.0034	0.0612	0.0445	0.0036	0.0615	0.0454	0.0043
Hard	0.0632	0.0570	0.0042	0.0729	0.0563	0.0051	0.0599	0.0550	0.0047
Oracle	0.0635	0.0569	0.0041	0.0756	0.0567	0.0051	0.0581	0.0552	0.0045
n=100, J=2									
SCAD	0.0744	0.0603	0.0043	0.0810	0.0712	0.0012	0.0730	0.0657	0.0057
LASSO	0.0810	0.0479	0.0050	0.0963	0.0506	0.0060	0.0996	0.0446	0.0065
Hard	0.0649	0.0560	0.0040	0.0815	0.0575	0.0041	0.0715	0.0549	0.0044
Oracle	0.0683	0.0565	0.0038	0.0785	0.0582	0.0036	0.0685	0.0552	0.0040
n=100, J=5									
SCAD	0.0435	0.0411	0.0020	0.0531	0.0404	0.0020	0.0445	0.0401	0.0027
LASSO	0.0480	0.0339	0.0018	0.0584	0.0345	0.0028	0.0405	0.0309	0.0025
Hard	0.0428	0.0411	0.0020	0.0535	0.0404	0.0023	0.0411	0.0406	0.0026
Oracle	0.0412	0.0412	0.0019	0.0504	0.0405	0.0020	0.0398	0.0406	0.0026

Πίνακας 2.13: Αποτελέσματα για το μοντέλο ευπάθειας Ομοιόμορφης κατανομής - Περίπτωση 2: $b = 1.25$ και $a = 1.75$

Μέθοδος	**MRME**	**Aver. no. of 0 coeff.**	
		correct	incorrect
n=50, J=2			
SCAD	0.4762	4.4100	0.1500
LASSO	1.0312	3.9600	0.0600
Hard	0.4494	4.8700	0.1900
Oracle	0.3112	5	0
n=50, J=5			
SCAD	0.5304	4.2800	0.0300
LASSO	0.8233	4.0600	0
Hard	0.6513	4.9300	0
Oracle	0.5767	5	0
n=100, J=2			
SCAD	0.5435	4.5700	0.0100
LASSO	1.0674	3.8200	0.0100
Hard	0.4759	4.7400	0.0600
Oracle	0.4614	5	0
n=100, J=5			
SCAD	0.5802	4.6500	0
LASSO	0.8838	3.9900	0
Hard	0.5738	4.4800	0
Oracle	0.5242	5	0

Πίνακας 2.14: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Ομοιόμορφης κατανομής - Περίπτωση 2: $b = 1.25$ και $a = 1.75$

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD	0.1344	0.1046	0.0086	0.1261	0.0859	0.0126	0.1215	0.1013	0.0113
LASSO	0.1144	0.0721	0.0093	0.1038	0.0755	0.0180	0.1497	0.0816	0.0137
Hard	0.1383	0.1045	0.0119	0.1127	0.0861	0.0061	0.1355	0.1071	0.0194
Oracle	0.1267	0.0972	0.0089	0.1234	0.0965	0.0076	0.1289	0.0972	0.0093
n=50, J=5									
SCAD	0.0968	0.0896	0.0103	0.1096	0.0684	0.0089	0.0810	0.0777	0.0077
LASSO	0.0858	0.0606	0.0138	0.0954	0.0630	0.0093	0.0889	0.0560	0.0123
Hard	0.0711	0.0722	0.0092	0.0890	0.0717	0.0081	0.0754	0.0716	0.0073
Oracle	0.0720	0.0741	0.0063	0.0817	0.0738	0.0051	0.0673	0.0731	0.0049
n=100, J=2									
SCAD	0.0774	0.0702	0.0096	0.0853	0.0701	0.0083	0.0702	0.0671	0.0092
LASSO	0.0687	0.0612	0.0045	0.0774	0.0636	0.0047	0.0805	0.0552	0.0048
Hard	0.0717	0.0715	0.0077	0.0853	0.0717	0.0074	0.0785	0.0711	0.0049
Oracle	0.0727	0.0748	0.0051	0.0767	0.0737	0.0059	0.0757	0.0733	0.0058
n=100, J=5									
SCAD	0.0479	0.0475	0.0050	0.0632	0.0483	0.0043	0.0752	0.0466	0.0059
LASSO	0.0491	0.0440	0.0021	0.0598	0.0459	0.0022	0.0527	0.0415	0.0020
Hard	0.0498	0.0498	0.0027	0.0553	0.0505	0.0028	0.0477	0.0495	0.0030
Oracle	0.0478	0.0500	0.0026	0.0558	0.0507	0.0026	0.0482	0.0496	0.0027

Πίνακας 2.15: Αποτελέσματα για το μοντέλο ευπάθειας Ομοιόμορφης κατανομής - Περίπτωση 3: $b = 0.75$ και $a = 1.25$

Μέθοδος	**MRME**	**Aver. no. of 0 coeff.**	
		correct	incorrect
n=50, J=2			
SCAD	0.3542	4.6100	0.0900
LASSO	1.0597	3.8600	0.0400
Hard	0.3773	4.3300	0.1500
Oracle	0.2797	5	0
n=50, J=5			
SCAD	0.6394	4.1900	0
LASSO	1.0649	4.0500	0
Hard	0.7948	4.1300	0
Oracle	0.5157	5	0
n=100, J=2			
SCAD	0.5767	4.9800	0.0100
LASSO	0.8588	4.1400	0
Hard	0.6065	4.9200	0.0100
Oracle	0.5662	5	0
n=100, J=5			
SCAD	0.6835	4.7100	0
LASSO	0.8057	4.0600	0
Hard	0.7770	4.1800	0
Oracle	0.6558	5	0

Πίνακας 2.16: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Ομοιόμορφης κατανομής - Περίπτωση 3: $b = 0.75$ και $a = 1.25$

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD	0.1050	0.0801	0.0074	0.1136	0.0904	0.0123	0.1117	0.0981	0.0123
LASSO	0.1245	0.0755	0.0075	0.1143	0.0804	0.0062	0.1263	0.0885	0.0101
Hard	0.1080	0.0771	0.0091	0.1093	0.0884	0.0120	0.1152	0.0895	0.0115
Oracle	0.1061	0.0855	0.0072	0.1119	0.0952	0.0058	0.1053	0.0958	0.0088
n=50, J=5									
SCAD	0.0643	0.0626	0.0057	0.0841	0.0608	0.0055	0.0811	0.0612	0.0066
LASSO	0.0569	0.0524	0.0031	0.0750	0.0524	0.0037	0.0752	0.0687	0.0035
Hard	0.0677	0.0666	0.0058	0.0864	0.0644	0.0067	0.0773	0.0675	0.0063
Oracle	0.0618	0.0645	0.0059	0.0845	0.0616	0.0053	0.0725	0.0647	0.0048
n=100, J=2									
SCAD	0.0831	0.0600	0.0038	0.0847	0.0616	0.0038	0.0935	0.0613	0.0088
LASSO	0.0800	0.0557	0.0040	0.0858	0.0592	0.0034	0.0993	0.0617	0.0048
Hard	0.0709	0.0658	0.0057	0.0867	0.0660	0.0055	0.0933	0.0639	0.0051
Oracle	0.0702	0.0673	0.0047	0.0830	0.0670	0.0049	0.0914	0.0648	0.0043
n=100, J=5									
SCAD	0.0623	0.0454	0.0026	0.0575	0.0456	0.0027	0.0503	0.0459	0.0026
LASSO	0.0490	0.0397	0.0020	0.0520	0.0410	0.0021	0.0472	0.0383	0.0021
Hard	0.0645	0.0456	0.0024	0.0600	0.0476	0.0036	0.0554	0.0476	0.0035
Oracle	0.0612	0.0453	0.0026	0.0583	0.0456	0.0027	0.0480	0.0459	0.0026

Από τους Πίνακες 2.1-2.16, παρατηρούμε ότι όλες οι μέθοδοι επιλέγουν σχεδόν τον ίδιο αριθμό των πραγματικά ενεργών μεταβλητών. Η SCAD και η Best Subset υπερέχουν στις περιπτώσεις των κατανομών Γάμμα και Αντίστροφης Γκαουσιανής, ενώ για την Ομοιόμορφη κατανομή τα καλύτερα αποτελέσματα τα δίνουν η SCAD και η Hard. Επιπλέον, ο τύπος εύρεσης του τυπικού σφάλματος των Fan και Li [54], συμπεριφέρεται πολύ αποδοτικά και για τις τρεις κατανομές της ευπάθειας.

Πιο συγκεκριμένα, για τη Γάμμα και Αντίστροφη Γκαουσιανή κατανομή, η SCAD υπερέχει γενικά της LASSO και της Hard και η απόδοσή της πλησιάζει τον Oracle εκτιμητή, ως προς τις τιμές για το MRME. Η Best Subset ναι μεν έχει παρόμοια απόδοση με τη SCAD, είναι όμως αρκετά πιο χρονοβόρα διαδικασία συγκριτικά με τις υπόλοιπες ποινικοποιημένες μεθόδους, καθώς επίσης συμπεριφέρεται ελαφρώς χειρότερα ως προς το σφάλμα της λανθασμένης εκτίμησης συντελεστών ίσων με το μηδέν.

Σχετικά με την Ομοιόμορφη κατανομή, οι διαφορετικές τιμές των b και a που χρησιμοποιήθηκαν, δεν επηρέασαν τα αποτελέσματά μας. Οι μέθοδοι SCAD και Hard παρουσιάζουν καλύτερη απόδοση από τη LASSO και συμπεριφέρονται το ίδιο καλά με τον Oracle εκτιμητή.

Όσον αφορά στην προσαρμογή λανθασμένης κατανομής για τις περιπτώσεις της Γάμμα και Αντίστροφης Γκαουσιανής κατανομής, κάτι τέτοιο δε φαίνεται να επηρεάζει σημαντικά την επιλογή των πραγματικά ενεργών μεταβλητών ή την ακρίβεια εκτίμησης των συντελεστών τους. Φαίνεται όμως να έχει ως αποτέλεσμα την ελαφρά υπερεκτίμηση ή υποεκτίμηση των παραμέτρων α και b. Αναφορικά με τις μέσες τιμές των μη μηδενικών συντελεστών, μπορούμε να παρατηρήσουμε ότι σε όλες τις περιπτώσεις, χρησιμοποιώντας τόσο τη σωστή όσο και τη λανθασμένη κατανομή ευπάθειας, οι εκτιμηθείσες και οι πραγματικές τιμές βρίσκονται πολυ κοντά.

2.5 Συμπεράσματα

Σε αυτό το κεφάλαιο, επεκτείναμε το μοντέλο ευπάθειας Γάμμα των Fan και Li [54] σε μια μεγάλη κλάση μοντέλων και προτείναμε μια γενικευμένη μορφή της συνάρτησης πιθανοφάνειας, η οποία επιτρέπει τη χρήση διαφορετικών συνεχών κατανομών της ευπάθειας. Παρουσιάσαμε τρία διαφορετικά παραδείγματα κατανομών, παραθέτοντας τα αντίστοιχα θεωρητικά αποτελέσματα για τα μοντέλα ευπάθειας Γάμμα, Αντίστροφης Γκαουσιανής και Ομοιόμορφης κατανομής, με την ποινικοποιημένη πιθανοφάνεια κατασκευασμένη για ομαδοποιημένα δεδομένα. Να σημειώσουμε ότι η μονομεταβλητή περίπτωση ανεξάρτητων δεδομένων αποτελεί μια ειδική περίπτωση της προτεινόμενης μορφής της πιθανοφάνειας. Η μεθοδολογία που αναπτύξαμε, μπορεί να θεωρηθεί αρκετά χρήσιμη στην πράξη, καθότι με τη χρήση ενός εύκολα υλοποιήσιμου αλγορίθμου, αφενός μεν αποκτάμε εκτιμητές των συντελεστών του μοντέλου και της διασποράς της ευπάθειας και εφετέρου επιλέγουμε τις σημαντικές μεταβλητές για διάφορες κατανομές. Συνεπώς, μπορούμε να οδηγηθούμε σε σημαντικά συμπεράσματα σχετικά με το πραγματικό μοντέλο και την έκταση της εξάρτησης που κυβερνά τα δεδομένα.

Μέρος II

Επιλογή της Ρυθμιστικής Παραμέτρου στις Μεθόδους Ποινικοποιημένης Πιθανοφάνειας

ΚΕΦΑΛΑΙΟ **3**

Ποινικοποιημένο Γενικό Γραμμικό Μοντέλο

A big computer, a complex algorithm
and a long time does not equal science.

—*Robert Gentleman (2003)*

Όπως αναφέραμε και σε προηγούμενο κεφάλαιο, η επιλογή της ρυθμιστικής παραμέτρου διαδραματίζει καθοριστικό ρόλο στην απόδοση των ποινικοποιημένων μεθόδων. Συνεπώς, πρέπει να επιλέγεται κατάλληλα. Ο στόχος του παρόντος κεφαλαίου είναι η ανάπτυξη μιας νέας διαδικασίας επιλογής της παραμέτρου αυτής, στο πλαίσιο της παλινδρόμησης ποινικοποιημένων ελαχίστων τετραγώνων. Η μέθοδος βασίζεται σε εκτιμήσεις της νόρμας του σφάλματος κατά την επίλυση συστημάτων γραμμικών εξισώσεων των Brezinski et al. [22, 23], τις οποίες και τροποποίησαμε κατάλληλα.

3.1 Ερευνητικό Πρόβλημα

Η συνάρτηση ποινής που εισάγεται στην ποινικοποιημένη πιθανοφάνεια, εξαρτάται από μια ρυθμιστική παράμετρο λ. Η παράμετρος αυτή έχει ουσιαστικό ρόλο, καθότι καθορίζει το μέγεθος της ποινικοποίησης που θα επιβληθεί στους συντελεστές του μοντέλου και συνεπώς έχει άμεση σχέση με την αποδοτικότητα των ποινικοποιημένων μεθόδων και με την πολυπλοκότητα του προκύπτοντος μοντέλου. Ως εκ τούτου, θα πρέπει να επιλέγεται κατάλληλα. Αρκετές μέθοδοι έχουν πλέον προταθεί στη βιβλιογραφία για το σκοπό αυτό. Μεταξύ άλλων, οι Fan και Li [53] προτείνουν τη χρήση της διασταυρωμένης επικύρωσης (cross-validation-CV) και της γενικευμένης διασταυρωμένης επικύρωσης (generalized cross-validation-GCV) [43], οι οποίες και χρησιμοποιούνται αρκετά συχνά. Επιπλέον, οι Wang et al. [172] προτείνουν τη χρήση του BIC κριτηρίου στην περίπτωση των ποινικοποιημένων ελαχίστων τετραγώνων με τη ποινή SCAD. Οι Zhang et al. [184] υιοθέτησαν το κριτήριο γενικευμένης πληροφορίας (generalized information criterion-GIC) του Nishii [137], για την επιλογή της παραμέτρου αυτής στην περίπτωση μεθόδων μη κοίλης ποινικοποιημένης πιθανοφάνειας. Πρόσφατα, οι Fan και Tang [60] ερεύνησαν την επιλογή της ρυθμιστικής παραμέτρου στα ποινικοποιημένα γενικευμένα γραμμικά μοντέλα, όπου η διάσταση των συμμεταβλητών αυξάνει εκθετικά ως προς το μέγεθος του δείγματος, καθώς επίσης μελέτησαν διεξοδικά το κριτήριο GIC.

Οι προαναφερθείσες μέθοδοι, βασίζονται είτε στην ελαχιστοποίηση ενός κατάλληλου κριτηρίου, όπως τα (BIC, GIC, GCV), είτε σε μεθόδους αναδειγματοληψίας που μπορούν όμως να γίνουν αρκετά χρονοβόρες [140], όπως η διασταυρωμένη επικύρωση. Παρ΄ όλα αυτά, η ανάπτυξη μεθόδων που θα βασίζονται σε εκτιμήσεις σφάλματος, οι οποίες και θα μπορούν να χρησιμοποιηθούν για την επιλογή της παραμέτρου αυτής, είναι ένα πολύ σημαντικό θέμα που δεν έχει αναπτυχεί στη βιβλιογραφία. Αυτό μας έδωσε το έναυσμα να αναπτύξουμε νέες μεθόδους, οι οποίες θα στηρίζονται στην ελαχιστοποίηση του σφάλματος εκτίμησης των ποινικοποιημένων συντελεστών και θα οδηγούν ταυτόχρονα στην επιλογή της βέλτιστης τιμής της ρυθμιστικής παραμέτρου. Ξεκινάμε με το πλαίσιο των ποινικοποιημένων ελαχίστων τετραγώνων, βασιζόμενοι στις εκτιμήσεις της νόρμας του σφάλματος για την επίλυση συστημάτων γραμμικών εξισώσεων των Brezinski et al. [22], [23], τις οποίες και θα τροποποιήσουμε στη συνέχεια.

3.2 Εκτίμηση Σφάλματος στα Γραμμικά Συστήματα

Οι Brezinski et al. [22] μελέτησαν διάφορες εκτιμήσεις της νόρμας του σφάλματος στην επίλυση συστημάτων γραμμικών εξισώσεων. Οι εκτιμήσεις αυτές επεκτάθηκαν στην εργασία [23], στο πρόβλημα ελαχίστων τετραγώνων

$$\min_{x\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|, \tag{3.1}$$

όπου ο πίνακας $\mathbf{A}\in\mathbb{R}^{m\times n}$ είναι κακής κατάστασης (ill-conditioned) και $\mathbf{b}\in\mathbb{R}^m$ είναι το διάνυσμα των δεδομένων στο οποίο θεωρείται ότι υπάρχει τυχαίος θόρυβος (random noise) με μέσο 0, συνεπώς η χρήση μιας μεθόδου κανονικοποίησης (regularization) είναι απαραίτητη ώστε να αποκτήσουμε τα επιθυμητά αποτελέσματα. Να σημειωθεί ότι δεν υπάρχουν περιορισμοί στα m και n, ούτως ώστε μπορούμε να έχουμε είτε $m\geq n$ είτε $m<n$.

Θεωρούμε το πρόβλημα ελαχίστων τετραγώνων (3.1), όπου $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\text{rank}(\mathbf{A})=r\leq \min(m,n)$ και έστω $\mathbf{x}^*$ η κανονική λύση της (3.1). Ορίζουμε επίσης ως $\mathbf{x}$ κάθε προσεγγιστική λύση της (3.1) και θεωρούμε το διάνυσμα σφάλματος $\mathbf{e}=\mathbf{x}^*-\mathbf{x}$, το αντίστοιχο υπόλοιπο $\mathbf{r}=\mathbf{b}-\mathbf{A}\mathbf{x}$ και το κανονικό υπόλοιπο $\boldsymbol{\rho}=\mathbf{A}^T\mathbf{r}$. Έστω επίσης ότι $\mathbf{A}=U\Sigma V^T$ είναι η παραγοντοποίηση ιδιαζουσών τιμών (singular value decomposition-SVD) του $\mathbf{A}$ [83], με

$$U=[\mathbf{u}_1,\ldots,\mathbf{u}_m],\ \mathbf{u}_i\in\mathbb{R}^m,\ U^TU=I_m,$$

$$V = [\mathbf{v}_1, \ldots, \mathbf{v}_n],\ \mathbf{v}_i \in \mathbb{R}^n,\ V^T V = I_n,$$

$$\Sigma = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m \times n}, \Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r),$$

και $\sigma_1 \geq \ldots \geq \sigma_r > 0$. Οι Brezinksi et al. [23] πρότειναν την ακόλουθη εκτίμηση της νόρμας του σφάλματος

$$\|\mathbf{e}\|^2 \simeq \eta_\nu^2 = d_0^{\nu-1} d_1^{5-2\nu} d_2^{\nu-3}, \nu \in \mathbb{R} \tag{3.2}$$

όπου

$$d_0 = \|\mathbf{r}\|^2 = \|U^T \mathbf{r}\|^2 = \sum_{i=1}^{m} \alpha_i^2$$

$$d_1 = \|\boldsymbol{\rho}\|^2 = \|\mathbf{A}^T \mathbf{r}\|^2 = \sum_{i=1}^{r} \sigma_i^2 \alpha_i^2$$

$$d_2 = \|\mathbf{A}\boldsymbol{\rho}\|^2 = \|\mathbf{A}\mathbf{A}^T \mathbf{r}\|^2 = \sum_{i=1}^{r} \sigma_i^4 \alpha_i^2 \tag{3.3}$$

και $\alpha_i = \mathbf{u}_i^T \mathbf{r}$.

Η εκτιμήτρια της νόρμας του σφάλματος (3.2) τροποποιήθηκε από τους Brezinski et al. [23] ώστε να χρησιμοποιηθεί κατάλληλα στην Tikhonov κανονικοποίηση, όπου έχουμε το ποινικοποιημένο πρόβλημα

$$\min_{x \in \mathbb{R}^n} \{\|A\mathbf{x} - \mathbf{b}\|^2 + \lambda^2 \|H\mathbf{x}\|^2\} \tag{3.4}$$

όπου $H \in \mathbb{R}^{p \times n}$ είναι ένας πίνακας κανονικοποίησης. Στην ίδια εργασία, η (3.2) χρησιμοποιήθηκε για τον εύρεση της παραμέτρου λ της Tikhonov. Για περισσότερες λεπτομέρειες σχετικά με την δημιουργία της εκτιμήτριας (3.2), καθώς και για εφαρμογές αυτής σε διάφορα προβλήματα κανονικοποίησης, παραπέμπουμε τον αναγνώστη στην επιστημονική εργασία [23].

3.3 Η Προτεινόμενη Μέθοδος Επιλογής της Ρυθμιστικής Παραμέτρου

Θεωρούμε την περίπτωση του γραμμικού μοντέλου παλινδρόμησης $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, όπου $\mathbf{y}$ είναι ένα $n \times 1$ διάνυσμα, $\mathbf{X}$ είναι ένας $n \times d$ πίνακας και $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ είναι το $n \times 1$ διάνυσμα από iid τυχαία σφάλματα, όπου $\varepsilon_i \sim N(0, \sigma^2)$ για όλα τα $i = 1, 2, \ldots, n$. Η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}$ επιτυγχάνεται μέσω της ελαχιστοποίησης της ποσότητας $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$.

Υπενθυμίζουμε ότι, οι Fan και Li [53], πρότειναν μια οικογένεια μεθόδων επιλογής μεταβλητών, μέσω ποινικοποιημένων ελαχίστων τετραγώνων και μη-κοίλης ποινικοποιημένης πιθανοφάνειας. Στην εργασία τους, η μορφή των ποινικοποιημένων ελαχίστων τετραγώνων έχει ως εξής:

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{3.5}$$

όπου $p_\lambda(.)$ είναι μια συνάρτηση ποινής και λ είναι η άγνωστη ρυθμιστική παράμετρος. Η λύση για το πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων, μπορεί να βρεθεί υπολογίζοντας αναδρομικά την ακόλουθη ποσότητα, με μια αρχική τιμή $\boldsymbol{\beta}^{(0)}$:

$$\boldsymbol{\beta}^{(1)} = [\mathbf{X}^T \mathbf{X} + n \textstyle\sum_\lambda (\boldsymbol{\beta}^{(0)})]^{-1} \mathbf{X}^T \mathbf{y}, \tag{3.6}$$

όπου $\sum_\lambda(\boldsymbol{\beta}^{(0)}) = \text{diag}\{p'_\lambda(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \ldots, p'_\lambda(|\beta_{d^*}^{(0)}|)/|\beta_{d^*}^{(0)}|\}$. Να τονίσουμε ότι d^* είναι ο αριθμός των στατιστικά σημαντικών παραγόντων, ενώ για τους υπόλοιπους παράγοντες η αντίστοιχη ποινή είναι 0.

Βασιζόμενοι στις εργασίες των Brezinski et al. [22, 23], προτείνουμε τώρα μια καινούργια μέθοδο για την επιλογή της ρυθμιστικής παραμέτρου λ, κατασκευάζοντας αρχικά μια νέα εκτιμήτρια του σφάλματος στο πλαίσιο των ποινικοποιημένων ελαχίστων τετραγώνων των Fan και Li [53].

Πρόταση 3.1 *Μια εκτίμηση του σφάλματος στα ποινικοποιημένα ελάχιστα τετράγωνα των Fan και Li (2001) δίνεται ως*

$$\eta_\nu(\lambda) = \|\boldsymbol{r}_\lambda\|^{\nu-1} \cdot \|n\sum\nolimits_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}\|^{5-2\nu} \cdot \|\boldsymbol{X}n\sum\nolimits_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}\|^{\nu-3}, \nu \in \mathbb{R}. \tag{3.7}$$

Απόδειξη: Η λύση της (3.5) προσδιορίζεται από τη σχέση

$$[\mathbf{X}^T\mathbf{X} + n\sum\nolimits_\lambda(\boldsymbol{\beta}^{(0)})]\boldsymbol{\beta}^{(1)} = \mathbf{X}^T\mathbf{y}, \tag{3.8}$$

η οποία οδηγεί στη σχέση

$$\mathbf{X}^T\mathbf{r}_\lambda = n\sum\nolimits_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}, \tag{3.9}$$

όπου $\mathbf{r}_\lambda = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(1)}$. Επιπλέον, χρησιμοποιώντας τη σημειογραφία αυτής της ενότητας καθώς και τις σχέσεις (3.3), η εκτιμήτρια της νόρμας του σφάλματος (3.2) τροποποιείται ως

$$\eta_\nu(\lambda) = \|\mathbf{r}_\lambda\|^{\nu-1} \cdot \|\mathbf{X}^T\mathbf{r}_\lambda\|^{5-2\nu} \cdot \|\mathbf{X}\mathbf{X}^T\mathbf{r}_\lambda\|^{\nu-3}, \nu \in \mathbb{R}. \tag{3.10}$$

Συνεπώς, αντικαθιστώντας την ποσότητα $\mathbf{X}^T\mathbf{r}_\lambda$ με την ποσότητα $n\sum_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}$ στην (3.10), έχουμε την ακόλουθη σχέση

$$\eta_\nu(\lambda) = \|\mathbf{r}_\lambda\|^{\nu-1} \cdot \|n\sum\nolimits_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}\|^{5-2\nu} \cdot \|\mathbf{X}n\sum\nolimits_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}\|^{\nu-3}, \nu \in \mathbb{R}. \tag{3.11}$$

□

Η κατάλληλη ελαχιστοποίηση της (3.11) μπορεί να οδηγήσει σε μια αρκετά αποδοτική επιλογή της παραμέτρου λ. Συγκεκριμένα, αρχίζουμε με ένα επιλεγμένο διάστημα που περιέχει τις αρχικές τιμές για την παράμετρο λ. Για κάθε λ, χρησιμοποιούμε τον αλγόριθμο των Fan και Li [53] ώστε να υπολογίσουμε τη λύση $\boldsymbol{\beta}^{(1)}$ και εν συνεχεία υπολογίζουμε τη νόρμα του σφάλματος από τη σχέση (3.11). Η παράμετρος η οποία ελαχιστοποιεί την (3.11) είναι η επιλεχθείσα παράμετρος. Τονίζουμε επίσης ότι, στην περίπτωση όπου οι προκύπτουσες ποινές είναι όλες μηδέν, αντί της (3.11), ελαχιστοποιούμε την αρχική έκφραση (3.10). Διαφορετικά, ο διαγώνιος πίνακας $\sum_\lambda(\boldsymbol{\beta}^{(0)})$ θα περιέχει μόνο μηδενικά στοιχεία και η σχέση (3.11) δε μπορεί να χρησιμοποιηθεί.

3.4 Συγκριτική Μελέτη Προσομοίωσης

Για την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου, εκτελέσαμε προσομοιώσεις για ένα ευρύ φάσμα μοντέλων, με χρήση πέντε διαφορετικών στατιστικών σχεδιασμών, μαζί με δύο προσομοιωμένα σύνολα δεδομένων. Επίσης, χρησιμοποιήσαμε τρεις διαφορετικές συναρτήσεις ποινής, τη Hard, την L_1 (LASSO) και τη SCAD, εφαρμόζοντας για σκοπούς σύγκρισης, δύο διαφορετικές μεθόδους για την εκτίμηση της ρυθμιστικής παραμέτρου λ: Τη γενικευμένη διασταυρωμένη επικύρωση (GCV) και τη νέα μέθοδο $\eta_\nu(\lambda)$ ως εναλλακτική.

3.4.1 Σχεδιασμός Προσομοιώσεων

Ένας 2^5 παραγοντικός σχεδιασμός (τον οποίο και θα συμβολίζουμε ως FD), ένας 2^{8-2} και ένας 2^{10-3} κλασματικός παραγοντικός σχεδιασμός αναλυτικής τάξης V (FFD), ένας OA(18,7,3,2) και ένας OA(27,13,3,2) περιλαμβάνονται στις προσομοιώσεις. Περισσότερες λεπτομέρειες σχετικά με τα είδη αυτών των σχεδιασμών, τις βασικές έννοιες και ιδιότητές τους, αναφέρουμε στα Κεφάλαια 6 και 10. Επιπλέον, το πρώτο προσομοιωμένο σύνολο δεδομένων (υποδηλώνεται ως dataset 1) αποτελείται από 100 γραμμές και 10 στήλες. Τα στοιχεία των πρώτων έξι στηλών προέκυψαν από την τυποποιημένη Κανονική κατανομή. Τα στοιχεία των τελευταίων τεσσάρων στηλών είναι ανεξάρτητα και ισόνομα κατανεμημένα από την κατανομή Bernoulli με πιθανότητα επιτυχίας 0.5. Το δεύτερο σύνολο δεδομένων (υποδηλώνεται ως dataset 2) αποτελείται επίσης από 100 γραμμές και 10 στήλες και όλα τα στοιχεία του είναι ανεξάρτητα και ισόνομα κατανεμημένα από την U(0,20).

Για τα πειράματα προσομοίωσης, δημιουργήσαμε γραμμικά μοντέλα με συντελεστές τυχαία επιλεγμένους από -5 έως 5. Όταν ένας παραγόμενος συντελεστής ήταν 'σχεδόν μηδέν' τότε έγινε αντικατάστασή του από το 50% του μέγιστου συντελεστή. Ένα τυχαίο σφάλμα $\varepsilon_i \sim N(0,1)$ για όλα τα $i = 1, 2, \ldots, n$ προστέθηκε σε κάθε αντίστοιχη παρατήρηση y_i. Μόνο μοντέλα κυρίων επιδράσων θεωρήθηκαν στις προσομοιώσεις. Επιπλέον, οι πραγματικά ενεργές επιδράσεις επιλέχθηκαν τυχαία, σύμφωνα με την Ομοιόμορφη κατανομή, από το σύνολο των $1, \ldots, d$ δυνητικά ενεργών παραγόντων, αναφορικά με τον ήδη καθορισμένο αριθμό ενεργών μεταβλητών στον πίνακα σχεδιασμού. Οι συντελεστές των μη ενεργών μεταβλητών, στο πραγματικό μοντέλο, τέθηκαν ίσες με μηδέν. Η τιμή της παραμέτρου α για τη SCAD, επιλέχθηκε ίση με 3.7, σύμφωνα με την εισήγηση στη δημοσίευση όπου εμφανίζεται [53]. Επίσης, ορίσαμε το επίπεδο σημαντικότητας να είναι 0.1 για την F-enter (a_E) και F-remove (a_R), στη διαδικασία της κατά βήματα επιλογής μεταβλητών, την οποία και χρησιμοποιήσαμε για την επιλογή μιας αρχικής τιμής για τις ποινικοποιημένες μεθόδους. Ωστόσο, η επιλογή αυτή θα μπορούσε να επηρεάσει την απόδοση των μεθόδων, καθώς και τα αποτελέσματα. Για αυτό το λόγο, διερευνήσαμε και την περίπτωση κατά την οποία το επίπεδο σημαντικότητας είναι ρυθμισμένο στην τιμή 0.05 για την F-enter και F-remove αλλά και την περίπτωση όπου η τιμή για την F-enter είναι 0.05 ενώ για την F-remove είναι 0.1. Αναφέρουμε επίσης ότι όπως η νέα μέθοδος $\eta_\nu(\lambda)$, έτσι και η GCV χρησιμοποιεί κάποιες αρχικές τιμές για την παράμετρο λ στην αντίστοιχη ρουτίνα, συνεπώς οι ίδιες τιμές δόθηκαν στις δύο μεθόδους.

Στο σημείο αυτό, να τονίσουμε ότι παράμετρος ν στην (3.11) μπορεί να λάβει κάθε πραγματική τιμή. Οι Brezinski et al. [23] διαπίστωσαν ότι για την επίλυση προβλημάτων με την Tikhonov κανονικοποίηση, μια τιμή $\nu \geq 2$ οδηγεί σε καλύτερα αποτελέσματα. Αυτό είναι πιθανόν να οφείλεται στην ανάγκη να δοθεί επαρκές βάρος στο υπόλοιπο $\|\mathbf{r}_\lambda\|$ της $\eta_\nu(\lambda)$. Βάσει των προσομοιώσεων που κάναμε και έπειτα από πολλές δοκιμές, καταλήξαμε στην επιλογή $\nu \geq 3$. Ως εκ τούτου, χρησιμοποιήσαμε την τιμή $\nu = 4$ σε όλα τα τελικά πειράματα προσομοίωσης.

3.4.2 Κριτήρια Σύγκρισης και Αξιολόγησης

Προκειμένου να αξιολογήσουμε την προτεινόμενη μέθοδο αλλά και να τη συγκρίνουμε με τη γενικευμένη διασταυρωμένη επικύρωση, υπολογίσαμε τα ποσοστά σφάλματος Τύπου I και Τύπου II. Λόγω της χρήσης αρκετών παραγοντικών σχεδιασμών στις προσομοιώσεις, αυτό που μας ενδιέφερε να ελέγξουμε είναι το κόστος του να δηλώσουμε έναν ανενεργό παράγοντα ως ενεργό (σφάλμα Τύπου I), καθώς και το κόστος του να δηλώσουμε έναν ενεργό παράγοντα ως ανενεργό (σφάλμα Τύπου II).

3.4.3 Αποτελέσματα Προσομοιώσεων

Στους παρακάτω Πίνακες 3.1-3.21, παρουσιάζονται τα αποτελέσματα των πειραμάτων προσομοίωσης. Ειδικότερα, στην πρώτη στήλη των εν λόγω πινάκων παρουσιάζεται το είδος του χρησιμοποιούμενου σχεδιασμού ή συνόλου δεδομένων. Στη δεύτερη στήλη, αναφέρεται ο αριθμός των πραγματικά ενεργών παραγόντων (q) που χρησιμοποιήθηκε στα προσομοιωμένα μοντέλα. Στις επόμενες στήλες παρουσιάζουμε τα ποσοστά σφάλματος Τύπου I (Type I) και Τύπου II (Type II) για τις συγκρινόμενες μεθόδους. Για κάθε είδος σχεδιασμού και προσομοιωμένου συνόλου δεδομένων και για κάθε αριθμό ενεργών παραγόντων, δημιουργήσαμε τυχαία 1000 γραμμικά μοντέλα (αναφορικά με τους ενεργούς παράγοντες που περιλήφθησαν στο μοντέλο και τους συντελεστές τους) τα οποία και χρησιμοποιήσαμε για την αξιολόγηση των μεθόδων.

Πίνακας 3.1: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = \alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^5FD	1	0.12	0.00	0.00	0.00	0.11	0.00	0.03	0.00	0.11	0.00	0.01	0.00
	2	0.11	0.00	0.01	0.00	0.10	0.00	0.02	0.00	0.10	0.00	0.01	0.00
	3	0.10	0.00	0.00	0.00	0.11	0.00	0.02	0.00	0.11	0.00	0.01	0.00
	4	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.01	0.00

Πίνακας 3.2: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν κλασματικό παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = \alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^{8-2}FFD	1	0.11	0.00	0.00	0.00	0.10	0.00	0.02	0.00	0.11	0.00	0.01	0.00
	2	0.10	0.00	0.00	0.00	0.11	0.00	0.01	0.00	0.11	0.00	0.01	0.00
	3	0.11	0.00	0.00	0.00	0.11	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	4	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	5	0.10	0.00	0.00	0.00	0.11	0.00	0.01	0.00	0.10	0.00	0.00	0.00
	6	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	7	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.12	0.00	0.01	0.00

Πίνακας 3.3: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν κλασματικό παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = \alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^{10-3}FFD	1	0.10	0.00	0.00	0.00	0.11	0.00	0.01	0.00	0.10	0.00	0.00	0.00
	2	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.11	0.00	0.00	0.00
	3	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00
	4	0.10	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00
	5	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	6	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	7	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00
	8	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	9	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.09	0.00	0.01	0.00

Πίνακας 3.4: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν ορθογώνιο σχεδιασμό τριών επιπέδων με ισχύ 2, για $\alpha_E = \alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
OA(18,7,3,2)	1	0.13	0.00	0.00	0.00	0.13	0.00	0.07	0.00	0.13	0.00	0.04	0.00
	2	0.12	0.00	0.01	0.03	0.13	0.00	0.06	0.01	0.13	0.00	0.04	0.01
	3	0.12	0.00	0.01	0.03	0.13	0.00	0.07	0.01	0.12	0.00	0.04	0.00
	4	0.12	0.00	0.01	0.02	0.12	0.00	0.07	0.00	0.12	0.00	0.03	0.01
	5	0.10	0.00	0.01	0.02	0.11	0.00	0.09	0.00	0.12	0.00	0.04	0.01
	6	0.11	0.00	0.01	0.02	0.11	0.00	0.09	0.00	0.12	0.00	0.05	0.00

Πίνακας 3.5: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν ορθογώνιο σχεδιασμό τριών επιπέδων με ισχύ 2, για $\alpha_E = \alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
OA(27,13,3,2)	1	0.14	0.00	0.00	0.00	0.14	0.00	0.06	0.00	0.14	0.00	0.03	0.00
	2	0.14	0.00	0.01	0.01	0.14	0.00	0.05	0.00	0.14	0.00	0.03	0.00
	3	0.13	0.00	0.00	0.01	0.13	0.00	0.04	0.00	0.14	0.00	0.03	0.00
	4	0.13	0.00	0.00	0.01	0.14	0.00	0.05	0.00	0.14	0.00	0.04	0.00
	5	0.13	0.00	0.01	0.01	0.13	0.00	0.06	0.00	0.12	0.00	0.03	0.00
	6	0.13	0.00	0.01	0.01	0.12	0.00	0.06	0.00	0.13	0.00	0.03	0.00
	7	0.13	0.00	0.01	0.01	0.12	0.00	0.07	0.00	0.13	0.00	0.03	0.00
	8	0.11	0.00	0.01	0.01	0.11	0.00	0.08	0.00	0.12	0.00	0.03	0.00
	9	0.11	0.00	0.01	0.01	0.12	0.00	0.08	0.00	0.11	0.00	0.04	0.00
	10	0.11	0.00	0.01	0.01	0.12	0.00	0.09	0.00	0.12	0.00	0.04	0.00
	11	0.11	0.00	0.01	0.00	0.09	0.00	0.08	0.00	0.10	0.00	0.03	0.00
	12	0.11	0.00	0.00	0.00	0.11	0.00	0.10	0.00	0.11	0.00	0.04	0.00

Πίνακας 3.6: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων στο dataset 1, για $\alpha_E = \alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Dataset 1	1	0.10	0.00	0.00	0.00	0.10	0.00	0.02	0.00	0.09	0.00	0.02	0.00
	2	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	3	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	4	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	5	0.11	0.00	0.00	0.00	0.09	0.00	0.02	0.00	0.10	0.00	0.02	0.00
	6	0.10	0.00	0.00	0.00	0.10	0.00	0.02	0.00	0.10	0.00	0.02	0.00
	7	0.10	0.00	0.00	0.00	0.09	0.00	0.02	0.00	0.09	0.00	0.02	0.00
	8	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	9	0.11	0.00	0.00	0.00	0.10	0.00	0.02	0.00	0.11	0.00	0.02	0.00

Πίνακας 3.7: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων στο dataset 2, για $\alpha_E = \alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Dataset 2	1	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.11	0.00	0.02	0.00
	2	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.11	0.00	0.02	0.00
	3	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.11	0.00	0.02	0.00
	4	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	5	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.09	0.00	0.02	0.00
	6	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.09	0.00	0.01	0.00
	7	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.09	0.00	0.01	0.00
	8	0.09	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	9	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.09	0.00	0.01	0.00

Πίνακας 3.8: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = \alpha_R = 0.05$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^5FD	1	0.06	0.00	0.00	0.00	0.05	0.00	0.02	0.00	0.06	0.00	0.01	0.00
	2	0.06	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.06	0.00	0.01	0.00
	3	0.05	0.00	0.01	0.00	0.06	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	4	0.06	0.00	0.01	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.00	0.00

Πίνακας 3.9: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν κλασματικό παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = \alpha_R = 0.05$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^{8-2}FFD	1	0.05	0.00	0.00	0.00	0.05	0.00	0.02	0.00	0.05	0.00	0.00	0.00
	2	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.00	0.00
	3	0.05	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.06	0.00	0.01	0.00
	4	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00
	5	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.06	0.00	0.01	0.00
	6	0.05	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.05	0.00	0.01	0.00
	7	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.06	0.00	0.01	0.00

Πίνακας 3.10: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν κλασματικό παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = \alpha_R = 0.05$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^{10-3}FFD	1	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.00	0.00
	2	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.00	0.00
	3	0.04	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	4	0.06	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.05	0.00	0.00	0.00
	5	0.06	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.05	0.00	0.00	0.00
	6	0.06	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00
	7	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.06	0.00	0.01	0.00
	8	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.05	0.00	0.00	0.00
	9	0.05	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.06	0.00	0.01	0.00

Πίνακας 3.11: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν ορθογώνιο σχεδιασμό τριών επιπέδων με ισχύ 2, για $\alpha_E = \alpha_R = 0.05$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
OA(18,7,3,2)	1	0.07	0.00	0.01	0.03	0.07	0.00	0.04	0.00	0.06	0.01	0.03	0.01
	2	0.06	0.00	0.00	0.03	0.06	0.01	0.03	0.01	0.06	0.00	0.03	0.00
	3	0.07	0.00	0.01	0.02	0.06	0.00	0.04	0.00	0.06	0.00	0.04	0.00
	4	0.05	0.00	0.01	0.02	0.06	0.00	0.06	0.00	0.05	0.01	0.03	0.01
	5	0.06	0.01	0.01	0.02	0.06	0.00	0.06	0.00	0.06	0.01	0.04	0.01
	6	0.05	0.01	0.01	0.02	0.06	0.01	0.06	0.01	0.05	0.01	0.04	0.01

Πίνακας 3.12: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν ορθογώνιο σχεδιασμό τριών επιπέδων με ισχύ 2, για $\alpha_E = \alpha_R = 0.05$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
OA(27,13,3,2)	1	0.07	0.00	0.00	0.02	0.07	0.00	0.04	0.00	0.07	0.00	0.03	0.00
	2	0.07	0.00	0.00	0.01	0.07	0.00	0.03	0.00	0.06	0.00	0.03	0.00
	3	0.07	0.00	0.01	0.01	0.07	0.00	0.03	0.00	0.06	0.00	0.03	0.00
	4	0.06	0.00	0.01	0.01	0.07	0.00	0.03	0.00	0.07	0.00	0.03	0.00
	5	0.06	0.00	0.00	0.01	0.06	0.00	0.04	0.00	0.06	0.00	0.03	0.00
	6	0.06	0.00	0.01	0.01	0.06	0.00	0.04	0.00	0.06	0.00	0.03	0.00
	7	0.06	0.00	0.01	0.01	0.06	0.00	0.05	0.00	0.05	0.00	0.03	0.00
	8	0.06	0.00	0.01	0.00	0.06	0.00	0.05	0.00	0.05	0.00	0.03	0.00
	9	0.06	0.00	0.01	0.00	0.06	0.00	0.06	0.00	0.05	0.00	0.03	0.00
	10	0.06	0.01	0.00	0.01	0.07	0.01	0.06	0.01	0.06	0.00	0.03	0.00
	11	0.05	0.03	0.01	0.03	0.05	0.03	0.05	0.02	0.05	0.03	0.03	0.03
	12	0.05	0.08	0.01	0.08	0.05	0.08	0.04	0.07	0.05	0.07	0.03	0.07

Πίνακας 3.13: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων στο dataset 1, για $\alpha_E = \alpha_R = 0.05$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Dataset 1	1	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	2	0.04	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	3	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	4	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	5	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	6	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	7	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	8	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.02	0.00
	9	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.04	0.00	0.01	0.00

Πίνακας 3.14: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων στο dataset 2, για $\alpha_E = \alpha_R = 0.05$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Dataset 2	1	0.05	0.00	0.00	0.00	0.05	0.00	0.02	0.00	0.05	0.00	0.01	0.00
	2	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	3	0.04	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	4	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	5	0.06	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	6	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	7	0.05	0.00	0.00	0.00	0.05	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	8	0.06	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.05	0.00	0.01	0.00
	9	0.05	0.00	0.00	0.00	0.04	0.00	0.01	0.00	0.06	0.00	0.01	0.00

Πίνακας 3.15: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = 0.05$ και $\alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^5FD	1	0.06	0.00	0.00	0.00	0.06	0.00	0.03	0.00	0.06	0.00	0.01	0.00
	2	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.12	0.00	0.01	0.00
	3	0.11	0.00	0.01	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	4	0.10	0.00	0.01	0.00	0.10	0.00	0.01	0.00	0.12	0.00	0.01	0.00

Πίνακας 3.16: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν κλασματικό παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = 0.05$ και $\alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^{8-2}FFD	1	0.07	0.00	0.00	0.00	0.07	0.00	0.02	0.00	0.07	0.00	0.01	0.00
	2	0.10	0.00	0.00	0.00	0.11	0.00	0.01	0.00	0.11	0.00	0.00	0.00
	3	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.11	0.00	0.00	0.00
	4	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	5	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.11	0.00	0.01	0.00
	6	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.11	0.00	0.01	0.00
	7	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00

Πίνακας 3.17: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν κλασματικό παραγοντικό σχεδιασμό δύο επιπέδων, για $\alpha_E = 0.05$ και $\alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
2^{10-3}FFD	1	0.08	0.00	0.00	0.00	0.08	0.00	0.02	0.00	0.07	0.00	0.00	0.00
	2	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.00	0.00
	3	0.12	0.00	0.00	0.00	0.12	0.00	0.01	0.00	0.11	0.00	0.00	0.00
	4	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	5	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	6	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.09	0.00	0.00	0.00
	7	0.10	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	8	0.11	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00
	9	0.10	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.11	0.00	0.00	0.00

Πίνακας 3.18: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν ορθογώνιο σχεδιασμό τριών επιπέδων με ισχύ 2, για $\alpha_E = 0.05$ και $\alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
OA(18,7,3,2)	1	0.08	0.00	0.01	0.04	0.08	0.00	0.04	0.00	0.08	0.00	0.04	0.00
	2	0.13	0.01	0.01	0.02	0.12	0.01	0.05	0.01	0.12	0.00	0.03	0.00
	3	0.13	0.00	0.01	0.03	0.12	0.00	0.06	0.00	0.12	0.00	0.04	0.00
	4	0.13	0.00	0.01	0.02	0.12	0.00	0.08	0.00	0.11	0.01	0.04	0.01
	5	0.10	0.00	0.01	0.02	0.10	0.00	0.08	0.00	0.11	0.00	0.04	0.00
	6	0.09	0.00	0.01	0.02	0.11	0.00	0.09	0.00	0.10	0.01	0.03	0.01

Πίνακας 3.19: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων σε έναν ορθογώνιο σχεδιασμό τριών επιπέδων με ισχύ 2, για $\alpha_E = 0.05$ και $\alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
OA(27,13,3,2)	1	0.10	0.00	0.00	0.02	0.10	0.00	0.05	0.00	0.10	0.00	0.03	0.00
	2	0.14	0.00	0.00	0.01	0.14	0.00	0.05	0.00	0.14	0.00	0.04	0.00
	3	0.14	0.00	0.01	0.01	0.13	0.00	0.04	0.00	0.13	0.00	0.03	0.00
	4	0.13	0.00	0.01	0.01	0.14	0.00	0.05	0.00	0.13	0.00	0.03	0.00
	5	0.14	0.00	0.01	0.01	0.14	0.00	0.06	0.00	0.14	0.00	0.03	0.00
	6	0.12	0.00	0.01	0.01	0.13	0.00	0.07	0.00	0.13	0.00	0.03	0.00
	7	0.13	0.00	0.01	0.01	0.12	0.00	0.08	0.00	0.12	0.00	0.04	0.00
	8	0.12	0.00	0.01	0.01	0.12	0.00	0.09	0.00	0.12	0.00	0.03	0.00
	9	0.12	0.00	0.01	0.00	0.11	0.00	0.08	0.00	0.11	0.00	0.03	0.00
	10	0.11	0.01	0.01	0.01	0.12	0.01	0.10	0.01	0.11	0.00	0.04	0.00
	11	0.11	0.01	0.01	0.01	0.11	0.02	0.10	0.02	0.10	0.02	0.04	0.02
	12	0.10	0.04	0.00	0.04	0.09	0.03	0.08	0.03	0.10	0.03	0.04	0.03

Πίνακας 3.20: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων στο dataset 1, για $\alpha_E = 0.05$ και $\alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Dataset 1	1	0.07	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.06	0.00	0.01	0.00
	2	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	3	0.11	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.09	0.00	0.02	0.00
	4	0.10	0.00	0.00	0.00	0.09	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	5	0.09	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	6	0.10	0.00	0.00	0.00	0.10	0.00	0.02	0.00	0.10	0.00	0.03	0.00
	7	0.09	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.09	0.00	0.02	0.00
	8	0.10	0.00	0.00	0.00	0.11	0.00	0.02	0.00	0.10	0.00	0.02	0.00
	9	0.10	0.00	0.00	0.00	0.11	0.00	0.02	0.00	0.11	0.00	0.02	0.00

Πίνακας 3.21: Απόδοση των μεθόδων, με χρήση 1000 προσομοιώσεων στο dataset 2, για $\alpha_E = 0.05$ και $\alpha_R = 0.1$

		SCAD(gcv)		SCAD($\eta_\nu(\lambda)$)		LASSO(gcv)		LASSO($\eta_\nu(\lambda)$)		Hard(gcv)		Hard($\eta_\nu(\lambda)$)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Dataset 2	1	0.07	0.00	0.00	0.00	0.07	0.00	0.02	0.00	0.07	0.00	0.01	0.00
	2	0.10	0.00	0.00	0.00	0.11	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	3	0.11	0.00	0.00	0.00	0.11	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	4	0.10	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	5	0.10	0.00	0.00	0.00	0.09	0.00	0.01	0.00	0.11	0.00	0.02	0.00
	6	0.10	0.00	0.00	0.00	0.09	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	7	0.10	0.00	0.00	0.00	0.09	0.00	0.01	0.00	0.10	0.00	0.02	0.00
	8	0.09	0.00	0.00	0.00	0.09	0.00	0.01	0.00	0.10	0.00	0.01	0.00
	9	0.09	0.00	0.00	0.00	0.10	0.00	0.01	0.00	0.09	0.00	0.01	0.00

Στη συνέχεια, παραθέτουμε κάποια γενικά σχόλια, τα οποία προκύπτουν από τα σφάλματα Τύπου I και II των παραπάνω Πινάκων 3.1-3.21. Καταρχήν, η βελτίωση των ποσοστών σφάλματος Τύπου I είναι εμφανής και εξαιρετικά ικανοποιητική κατά την εφαρμογή της νέας μεθόδου για την επιλογή της ρυθμιστικής παραμέτρου. Όσον αφορά τις τιμές των σφαλμάτων Τύπου II, γενικά παραμένουν σε μηδενικά επίπεδα, ανεξάρτητα από τη μέθοδο επιλογής της ρυθμιστικής παραμέτρου που χρησιμοποιούμε. Υπάρχει μόνο μια αμελητέα διαφορά, όταν χρησιμοποιούμε τους δύο OAs.

Μεταξύ των ποινικοποιημένων μεθόδων και με τη χρήση της $\eta_\nu(\lambda)$, παρατηρούμε ότι στην περίπτωση που ο πίνακας σχεδιασμού είναι ένας εκ των πέντε στατιστικών σχεδιασμών, τα σφάλματα Τύπου I είναι μικρότερα για τη μέθοδο SCAD($\eta_\nu(\lambda)$), ακολουθεί η Hard($\eta_\nu(\lambda)$) και τέλος η LASSO($\eta_\nu(\lambda)$). Όσον αφορά τα σφάλματα Τύπου II, είναι μικρότερα όταν χρησιμοποιείται η LASSO($\eta_\nu(\lambda)$) και ακολουθούν η Hard($\eta_\nu(\lambda)$) και η SCAD($\eta_\nu(\lambda)$) με παρόμοια απόδοση. Η τελευταία, έχει μόνο μια ελάχιστη διαφορά στα σφάλματα Τύπου II αναφορικά με τους δύο OAs. Σχετικά με τα δύο προσομοιωμένα Datasets 1 και 2, οι LASSO($\eta_\nu(\lambda)$) και Hard($\eta_\nu(\lambda)$) παρουσιάζουν παρόμοια συμπεριφορά, ενώ η SCAD($\eta_\nu(\lambda)$) οδηγεί σε μηδενικά σφάλματα.

Το σημαντικό πλεονέκτημα της $\eta_\nu(\lambda)$, είναι ότι οι ποινικοποιημένες μέθοδοι παρουσιάζουν μια σταθερή απόδοση, δίνοντας σφάλματα που γενικά παραμένουν σε μηδενικά επίπεδα, ανεξάρτητα από την αύξηση των πραγματικά ενεργών παραγόντων (q). Επιπλέον, η διαφοροποίηση στην τιμή του επιπέδου σημαντικότητας για τις F-enter και F-remove στην κατά βήματα διαδικασία επιλογής μεταβλητών, δεν επηρεάζει σημαντικά τα αποτελέσματα.

3.5 Συμπεράσματα

Σε αυτό το κεφάλαιο, επεκτείναμε τις εκτιμήσεις της νόρμας του σφάλματος των Brezinski et al. [22], [23], κατά την επίλυση συστημάτων γραμμικών εξισώσεων, στην περίπτωση των ποινικοποιημένων ελαχίστων τετραγώνων. Δημιουργήσαμε μέσω αυτών, μια νέα μέθοδο επιλογής της ρυθμιστικής παραμέτρου, την $\eta_\nu(\lambda)$. Η παράμετρος αυτή καθορίζει σημαντικά το μέγεθος της ποινικοποιήσης που θα επιβληθεί στους συντελεστές του μοντέλου, καθώς επίσης επηρεάζει τις τιμές των σφαλμάτων Τύπου I και II των ποινικοποιημένων μεθόδων. Να υπενθυμίσουμε ότι υπάρχει πάντα το κόστος δήλωσης ενός μη σημαντικού παράγοντα ως σημαντικού (σφάλμα Τύπου I), καθώς και το κόστος δήλωσης ενός σημαντικού παράγοντα ως μη σημαντικού (σφάλμα Τύπου II). Συνεπώς, ο στόχος μας ήταν η δημιουργία μιας μεθόδου που να διατηρεί τα σφάλματα αυτά σε μηδενικά επίπεδα. Συγκριτικά με το ευρέως χρησιμοποιούμενο κριτήριο της γενικευμένης διασταυρωμένης επικύρωσης και βάσει των αποτελεσμάτων των προσομοιώσεων που εκτελέσαμε, καταλήγουμε στο συμπέρασμα ότι η $\eta_\nu(\lambda)$ μπορεί να θεωρηθεί ως μια αρκετά αξιόπιστη διαδικασία.

ΚΕΦΑΛΑΙΟ **4**

Ποινικοποιημένα Γενικευμένα Γραμμικά Μοντέλα για Διακριτά Δεδομένα

The best thing about being a statistician,
is that you get to play in everyone's backyard.

—John Tukey (1915–2000)

Στο τέταρτο αυτό κεφάλαιο, μελετάμε την οικογένεια των ποινικοποιημένων γενικευμένων γραμμικών μοντέλων και αναπτύσσουμε μια νέα μέθοδο επιλογής της ρυθμιστικής παραμέτρου. Η διαδικασία που παρουσιάζεται στο προηγούμενο κεφάλαιο, εφαρμόζεται μόνο στην περίπτωση όπου τα δεδομένα απόκρισης ακολουθούν την Κανονική κατανομή. Εδώ βασιζόμαστε στις Kantorovich ανισώσεις και προτείνουμε αρχικά νέες εκτιμήσεις της νόρμας του σφάλματος στα γενικευμένα γραμμικά μοντέλα. Εν συνεχεία, οι εκτιμήσεις αυτές χρησιμοποιούνται κατάλληλα για την εύρεση μιας νέας μεθόδου επιλογής της ρυθμιστικής παραμέτρου, στην περίπτωση των ποινικοποιημένων γενικευμένων γραμμικών μοντέλων για διακριτά δεδομένα.

4.1 Ερευνητικό Πρόβλημα

Η μέθοδος $\eta_\nu(\lambda)$ εύρεσης της ρυθμιστικής παραμέτρου που αναπτύχθηκε στο Κεφάλαιο 3, έδωσε εξαιρετικά αποτελέσματα και συγκεκριμένα μηδενικά σφάλματα. Η χρήση της όμως περιορίζεται στην περίπτωση των ποινικοποιημένων ελαχίστων τετραγώνων. Σκεφτήκαμε λοιπόν να επεκτείνουμε την ιδέα της χρήσης εκτιμήσεων σφάλματος από το πεδίο της Αριθμητικής Ανάλυσης, για την ανάπτυξη μιας πιο γενικής μεθόδου επιλογής της παραμέτρου λ. Σκοπός μας είναι να χρησιμοποιείται στα ποινικοποιημένα γενικευμένα γραμμικά μοντέλα, δίνοντας έμφαση στην περίπτωση των διακριτών δεδομένων και συγκεκριμένα, στην περίπτωση όπου η απόκριση είναι κατανομής Bernoulli και Poisson. Για τον λόγο αυτόν, βασιστήκαμε στις Kantorovich ανισώσεις και στην εργασία του Galantai [77], ο οποίος γενίκευσε την εκτίμηση σφάλματος του Auchmuty [13], στα μη γραμμικά συστήματα. Βασιζόμενοι στα βήματα του Galantai [77], προσαρμόσαμε την εκτίμηση σφάλματος που πρότεινε, στο δικό μας πρόβλημα, δηλαδή στα γενικευμένα γραμμικά μοντέλα.

4.2 Γενικευμένα Γραμμικά Μοντέλα

Η οικογένεια των γενικευμένων γραμμικών μοντέλων [135] είναι μια αξιοσημείωτη σύνθεση και επέκταση των γνωστών μοντέλων παλινδρόμησης, όπως τα γραμμικά μοντέλα. Ενοποιεί διαφορετικές προσεγγίσεις προκειμένου να εξηγηθεί η μεταβλητότητα στα δεδομένα, από την άποψη ενός γραμμικού συνδυασμού των συμμεταβλητών. Ως αποτέλεσμα, οι επιστήμονες μπορούν πλέον να χρησιμοποιούν κατανομές οι οποίες γενικά ανήκουν στη λεγόμενη εκθετική οικογένεια, η οποία περιλαμβάνει την Κανονική, τη Διωνυμική, την Poisson, τη Γεωμετρική, την Αρνητική Διωνυμική, την Εκθετική, τη Γάμμα και την Αντίστροφη Γκαουσιανή κατανομή. Βασικές αναφορές στο πλαίσιο των μοντέλων αυτών είναι τα βιβλία των McCullagh και Nelder [130] και Myers et al. [134].

Έστω ότι τα δεδομένα αποτελούνται από τις τυχαίες μεταβλητές $(y_i, \mathbf{x}_i),\ i = 1, \ldots, n$ όπου $\mathbf{x}_i$ είναι ένα διάνυσμα διάστασης d των προβλεπουσών μεταβλητών. Υποθέτουμε επίσης ότι στο $i-$οστό σημείο, $i = 1, 2, ..., n$, η απόκριση y_i ακολουθεί μια κατανομή που ανήκει στη μονο-παραμετρική εκθετική οικογένεια με παράμετρο p_i και με στήριγμα που δεν εξαρτάται από άγνωστες παραμέτρους. Η συνάρτηση πιθανοφάνειας, στη φυσική της μορφή [15] γράφεται ως

$$Lik(y_1, \ldots, y_n, \eta_1, \ldots, \eta_n) = e^{\sum_{i=1}^n y_i\eta_i - \sum_{i=1}^n d_0(\eta_i) + \sum_{i=1}^n S(y_i)} \tag{4.1}$$

όπου $(\eta_1, \ldots, \eta_n) \in \mathbf{H}$ και όπου $\mathbf{H}$ είναι μια συλλογή όλων των $(\eta_1, \ldots, \eta_n)$ ώστε το $d_0(\eta_i)$ είναι πεπερασμένο, για τις πραγματικές συναρτήσεις d_0 και S.

Έστω ότι η φυσική παράμετρος (natural parameter) η_i γράφεται ως $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$. Κατά τα γνωστά, έχουμε ότι η παράγωγος του λογαρίθμου της πιθανοφάνειας ως προς $\boldsymbol{\beta}$ δίνεται ως

$$\frac{\vartheta \ln Lik}{\vartheta \beta_j} = \sum_{i=1}^n \Big(y_i - d_0{}'(\mathbf{x}_i^T\boldsymbol{\beta})\Big)\, x_{ij}, \quad j = 1, \ldots d. \tag{4.2}$$

Συνεπώς, η εκτιμήτρια μέγιστης πιθανοφάνειας (ΕΜΠ) του $\boldsymbol{\beta}$ είναι η λύση των παρακάτω εξισώσεων, οι οποίες με συμβολισμό πινάκων γράφονται ως

$$X^T(\mathbf{y} - d_0{}'(X, \boldsymbol{\beta})) = X^T(\mathbf{y} - \boldsymbol{\mu}) = 0 \tag{4.3}$$

διότι είναι γνωστό ότι $d_0{}'(\boldsymbol{\eta}) = E_{\boldsymbol{\eta}}(\mathbf{y}) = \boldsymbol{\mu}$, όπου $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ και $d_0{}'(X, \boldsymbol{\beta}) = (d_0{}'(\mathbf{x}_1^T\boldsymbol{\beta}), \ldots, d_0{}'(\mathbf{x}_n^T\boldsymbol{\beta}))^T$. Η εξίσωση (4.3) είναι γνωστή ως 'score equation' στα γενικευμένα γραμμικά μοντέλα και η ΕΜΠ $\hat{\boldsymbol{\beta}}$ είναι η λύση της.

Η αρνητική δεύτερη παράγωγος ως προς $\boldsymbol{\beta}$ του παραπάνω λογαρίθμου της πιθανοφάνειας,

δίνεται ως

$$\sum_{i=1}^{n} x_{ij} d_0''(\mathbf{x}_i^T \boldsymbol{\beta}) x_{ik},\ j = 1, \ldots, d, \quad k = 1, \ldots, d \tag{4.4}$$

ή ισοδύναμα με συμβολισμό πινάκων ως

$$X^T W X \tag{4.5}$$

όπου W ο πίνακας διασποράς - συνδιασποράς της απόκρισης $\mathbf{y}$ από τη στιγμή που είναι γνωστό ότι $d_0''(\boldsymbol{\eta}) = Var_{\boldsymbol{\eta}}(\mathbf{y})$. Συνεπώς, ο W είναι ένας $n \times n$ διαγώνιος πίνακας με στοιχεία $Var(y_i), i = 1, \ldots, n$.

4.2.1 Το Λογιστικό Μοντέλο Παλινδρόμησης

Έστω τώρα ότι στο $i-$οστό σημείο, $i = 1, 2, ..., n$, η απόκριση y_i είναι μια τυχαία μεταβλητή που ακολουθεί την κατανομή Bernoulli ώστε $y_i \epsilon \{0, 1\}$, με $E(y_i) = p_i = p(\mathbf{x}_i)$ και $Var(y_i) = p_i(1-p_i)$. Εδώ η p_i δηλώνει την πιθανότητα επιτυχίας σε μια διαδικασία Bernoulli. Είναι γνωστό ότι η ΕΜΠ του $\boldsymbol{\beta}$, με τη χρήση μιας κανονικής συνάρτησης σύνδεσης, είναι η λύση των 'score' εξισώσεων [130], [134]

$$\sum_{i=1}^{n} x_{ij} \left(y_i - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) = 0,\ j = 1, \ldots, d \tag{4.6}$$

ή ισοδύναμα με συμβολισμό πινάκων

$$X^T(\mathbf{y} - \boldsymbol{\mu}) = 0 \tag{4.7}$$

όπου $\mu_i = E_{\boldsymbol{\eta}}(y_i) = d_0'(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} = p_i$. Επιπροσθέτως, ο πίνακας $X^T W X$ λαμβάνει τη μορφή

$$\sum_{i=1}^{n} x_{ij} \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \left(1 - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) \right) x_{ik},\ j = 1, \ldots, d, \quad k = 1, \ldots, d, \tag{4.8}$$

όπου W ο $n \times n$ διαγώνιος πίνακας με στοιχεία $Var(y_i) = p_i(1 - p_i),\ i = 1, \ldots, n$.

4.2.2 Το Μοντέλο Παλινδρόμησης Poisson

Ας υποθέσουμε ότι οι αποκρίσεις $y_1, \ldots, y_n$ είναι μετρήσεις (counts) που ακολουθούν ανεξάρτητες Poisson κατανομές με $E(y_i) = p_i = \mu_i$ και $Var(y_i) = \mu_i$, καθότι η διασπορά είναι ίση με το μέσο. Για το μοντέλο παλινδρόμησης Poisson και με τη χρήση μιας κανονικής συνάρτησης σύνδεσης, η ΕΜΠ του $\boldsymbol{\beta}$ είναι η λύση των 'score' εξισώσεων [130], [134]

$$\sum_{i=1}^{n} x_{ij} \left(y_i - e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right) = 0,\ j = 1, \ldots, d \tag{4.9}$$

ή ισοδύναμα με συμβολισμό πινάκων

$$X^T(\mathbf{y} - \boldsymbol{\mu}) = 0. \tag{4.10}$$

Επιπλέον, ο πίνακας $X^T W X$ λαμβάνει τη μορφή

$$\sum_{i=1}^{n} x_{ij} \left(e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right) x_{ik},\ j = 1, \ldots, d, \quad k = 1, \ldots, d, \tag{4.11}$$

όπου W ένας $n \times n$ διαγώνιος πίνακας με στοιχεία $\mu_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}},\ i = 1, \ldots, n$.

4.3 Εκτίμηση Σφάλματος στα Γενικευμένα Γραμμικά Μοντέλα

Σε αυτό το σημείο, προτείνουμε νέες εκτιμήσεις της νόρμας του σφάλματος στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, μέσω της χρήσης της Kantorovich ανισότητας (βλ. [84], [92] και [127]) η οποία δίνεται στην ακόλουθη μορφή

$$\|z\|_2^4 \leq (z^T B z)(z^T B^{-1} z) \leq \frac{1}{4}\frac{(w_1 + w_d)^2}{w_1 w_d}\ \|z\|_2^4, \tag{4.12}$$

όπου $B \in \mathbb{R}^{d\times d}$ είναι ένας συμμετρικός, θετικά ορισμένος πίνακας με ιδιοτιμές $w_1 \geq w_2 \geq ... \geq w_d > 0$, $z \in \mathbb{R}^d$ ένα αυθαίρετο διάνυσμα και η νόρμα $\|.\|_2$ είναι η Ευκλίδεια νόρμα για διανύσματα ή η Frobenius νόρμα για πίνακες. Η εκτιμήτρια της νόρμας του σφάλματος του Galantai [77], στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, προσαρμόζεται ως ακολούθως:

Πρόταση 4.1 *Έστω οι μη γραμμικές 'score' εξισώσεις $X^T(\boldsymbol{y}-\boldsymbol{\mu}) = 0$ πολλαπλασιασμένες με -1 και ο πίνακας των αρνητικών δεύτερων παραγώγων της πιθανοφάνειας, $X^T W X$. Έστω επίσης $\boldsymbol{\beta}$ κάποια προσεγγιστική λύση και $\boldsymbol{\beta}^*$ η σωστή λύση. Μια εκτίμηση της νόρμας του σφάλματος δίνεται ως*

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 = c\frac{\|X^T(\boldsymbol{y}-\boldsymbol{\mu})\|_2^2}{\|(X^T W X)^T X^T(\boldsymbol{y}-\boldsymbol{\mu})\|_2} \tag{4.13}$$

με $1 \lessapprox c \lessapprox C_2(F'(\boldsymbol{\beta})) = \frac{1}{2}\frac{\sigma_1^2+\sigma_d^2}{\sigma_1\sigma_d}$ και $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d > 0$ οι ιδιάζουσες τιμές του $F'(\boldsymbol{\beta}) = X^T W X$.

Απόδειξη: Η απόδειξη της πρότασης ακολουθεί τα βήματα του Galantai [77], την οποία όμως παραθέτουμε για λόγους πληρότητας. Θεωρούμε τις μη γραμμικές 'score' εξισώσεις $X^T(\mathbf{y} - \boldsymbol{\mu}) = 0$ πολλαπλασιασμένες με -1. Τις συμβολίζουμε ως $F(\boldsymbol{\beta}) = 0$ με $F : R^d \to R^d$ και έστω $\boldsymbol{\beta}$ μια προσεγγιστική λύση και $\boldsymbol{\beta}^*$ η ακριβής λύση του συστήματος $F(\boldsymbol{\beta}) = 0$. Ο πίνακας $X^T W X$ στην (4.5) είναι επίσης ίσος με τον Ιακωβιανό πίνακα $F'(\boldsymbol{\beta})$. Υποθέτουμε ότι ο $F'(\boldsymbol{\beta}^*)$ είναι αντιστρέψιμος, $F' \in C^1(\overline{S(\boldsymbol{\beta}^*, \delta)})$ και $\|F'(\boldsymbol{\beta_1}) - F'(\boldsymbol{\beta_2})\|_2 \leq M\|\boldsymbol{\beta_1} - \boldsymbol{\beta_2}\|_2$, $\forall \boldsymbol{\beta_1}, \boldsymbol{\beta_2} \in \overline{S(\boldsymbol{\beta}^*, \delta)}$, όπου

$$\overline{S(\boldsymbol{\beta}^*, \delta)} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2 \leq \delta\}, \tag{4.14}$$

για $\delta > 0$. Έστω τώρα ότι το $\boldsymbol{\beta}$ είναι κοντά στο $\boldsymbol{\beta}^*$ και έστω

$$B = F'(\boldsymbol{\beta})F'(\boldsymbol{\beta})^T = (X^T W X)(X^T W X)^T. \tag{4.15}$$

Έστω επίσης ότι $U\Sigma V^T$ είναι η ανάλυση σε ιδιάζουσες τιμές (Singular Value Decomposition) του F', όπου $\Sigma = diag(\sigma_1, ..., \sigma_d)$ έτσι ώστε $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d > 0$. Αν $z \in \mathbb{R}^d$ είναι ένα τυχαίο διάνυσμα, έχουμε ότι

$$z^T B z = \|F'(\boldsymbol{\beta})^T z\|_2^2 \tag{4.16}$$

$$z^T B^{-1} z = \|F'(\boldsymbol{\beta})^{-1} z\|_2^2 \tag{4.17}$$

και $w_i = w_i(B) = \sigma_i^2(F') = \sigma_i^2$, όπου w_i και σ_i είναι οι $i-$οστές ιδιοτιμές και ιδιάζουσες τιμές του B και F', αντίστοιχα. Οπότε τώρα μπορούμε να εφαρμόσουμε την Kantorovich ανισότητα (4.12) η οποία δίνει

$$\|z\|_2^2 \leq \|F'(\boldsymbol{\beta})^T z\|_2 \|F'(\boldsymbol{\beta})^{-1} z\|_2 \leq \frac{1}{2}\frac{\sigma_1^2 + \sigma_d^2}{\sigma_1\sigma_d}\ \|z\|_2^2,\ z \in \mathbb{R}^d. \tag{4.18}$$

Έστω $z = F(\boldsymbol{\beta})$. Από τη Lipschitz συνέχεια, προκύπτει ότι

$$F(\boldsymbol{\beta}) = F'(\boldsymbol{\beta})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + O(\|e\|_2^2) \tag{4.19}$$

και

$$F'(\boldsymbol{\beta})^{-1}F(\boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\beta}^* + O(\|e\|_2^2). \tag{4.20}$$

Συνεπώς,

$$\|F(\boldsymbol{\beta})\|_2^2 \leq \|F'(\boldsymbol{\beta})^T F(\boldsymbol{\beta})\|_2(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + O(\|e\|_2^2)) \leq C_2(F'(\boldsymbol{\beta}))\|F(\boldsymbol{\beta})\|_2^2 \tag{4.21}$$

και

$$\frac{\|F(\boldsymbol{\beta})\|_2^2}{\|F'(\boldsymbol{\beta})^T F(\boldsymbol{\beta})\|_2} \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + O(\|e\|_2^2) \leq C_2(F'(\boldsymbol{\beta}))\frac{\|F(\boldsymbol{\beta})\|_2^2}{\|F'(\boldsymbol{\beta})^T F(\boldsymbol{\beta})\|_2}, \tag{4.22}$$

όπου $C_2(F'(\boldsymbol{\beta}))=\frac{1}{2}\frac{\sigma_1^2+\sigma_d^2}{\sigma_1\sigma_d}$. Άρα τελικά, προκύπτει ότι η προσεγγιστική εκτίμηση της νόρμας του σφάλματος δίνεται ως

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 = c\frac{\|F(\boldsymbol{\beta})\|_2^2}{\|F'(\boldsymbol{\beta})^T F(\boldsymbol{\beta})\|_2} \tag{4.23}$$

ή ισοδύναμα

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 = c\frac{\|X^T(\mathbf{y} - \boldsymbol{\mu})\|_2^2}{\|(X^T W X)^T X^T(\mathbf{y} - \boldsymbol{\mu})\|_2} \tag{4.24}$$

όπου $1 \lessapprox c \lessapprox C_2(F'(\boldsymbol{\beta})) = \frac{1}{2}\frac{\sigma_1^2+\sigma_d^2}{\sigma_1\sigma_d}$. □

Στο σημείο αυτό, χρειάζεται να τονίσουμε ότι για να εφαρμόσουμε την ανισότητα Kantorovich ώστε να αποκτήσουμε την (4.24) στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, πρέπει να ισχύουν οι ακόλουθες υποθέσεις:

Α. Οι παλινδρομητές $\mathbf{x}$ λαμβάνουν πεπερασμένες τιμές, είναι μη εκφυλισμένες (nondegenerate) τυχαίες μεταβλητές και γραμμικά ανεξάρτητες.

Β. Η συνάρτηση $d_0 \in \mathcal{C}^3$ που ορίζεται στη (4.1) έχει ομοιόμορφα φραγμένη τρίτη παράγωγο για $\boldsymbol{\beta} \in \overline{S(\boldsymbol{\beta}^*, \delta)}$.

Οπότε, η συνέχεια του $X^T W X$ ως προς $\boldsymbol{\beta}$ και η Lipschitz συνέχεια, προκύπτουν από τις υποθέσεις **Α** και **Β** για $\boldsymbol{\beta} \in \overline{S(\boldsymbol{\beta}^*, \delta)}$. Επιπλέον, ο πίνακας $X^T W X$ είναι συμμετρικός, θετικά ορισμένος και αντιστρέψιμος στο $\boldsymbol{\beta}^*$ λόγω της μορφής του και των υποθέσεων μας. Συνεπώς, ο πίνακας B είναι συμμετρικός και θετικά ορισμένος. Συνεπώς, οι προϋποθέσεις που απαιτούνται για την εφαρμογή της ανισότητας Kantorovich πληρούνται για το πρόβλημά μας.

4.4 Η Προτεινόμενη Μέθοδος Επιλογής της Ρυθμιστικής Παραμέτρου

Παραθέτουμε τώρα μια εναλλακτική μέθοδο για την επιλογή της ρυθμιστικής παραμέτρου λ, στα ποινικοποιημένα γενικευμένα γραμμικά μοντέλα, μέσω της κατάλληλης χρήσης της εκτιμήτριας της νόρμας του σφάλματος (4.24). Η διαδικασία που προτείνουμε περιγράφεται ρητά ως εξής:

1. Ξεκινάμε με ένα σύνολο από αρχικές τιμές για τη ρυθμιστική παράμετρο λ.
2. Για κάθε λ στο σύνολο αυτό, υπολογίζουμε έναν ποινικοποιημένο εκτιμητή $\hat{\boldsymbol{\beta}}$.

3. Έπειτα, υπολογίζουμε τη νόρμα του σφάλματος βάσει του τύπου (4.24), με τη σταθερά c να λαμβάνει την τιμή του άνω φράγματός της, χρησιμοποιώντας τις κατάλληλες μορφές των $F(\hat{\boldsymbol{\beta}})$ και $F'(\hat{\boldsymbol{\beta}})$, για παράδειγμα, βασιζόμενοι στις εξισώσεις (4.6) και (4.8) για το ποινικοποιημένο μοντέλο Λογιστικής παλινδρόμησης ή στις εξισώσεις (4.9) και (4.11) για το ποινικοποιημένο μοντέλο Poisson παλινδρόμησης.

4. Η τιμή της παραμέτρου λ του συνόλου που ελαχιστοποιεί τη (4.24) για την προαναφερθείσα σταθερά c, είναι η επιθυμητή ρυθμιστική παράμετρος και ο προκύπτων ποινικοποιημένος εκτιμητής $\hat{\boldsymbol{\beta}}$ είναι ο τελικός εκτιμητής του $\boldsymbol{\beta}$.

Παρατήρηση 4.1 *Εύκολα παρατηρούμε ότι η $C_2(F'(\boldsymbol{\beta}))$ είναι κατ΄ ουσίαν μια απλή συνάρτηση του δείκτη κατάστασης $k_2(F'(\boldsymbol{\beta})) = \|F'(\boldsymbol{\beta})\|_2\|F'(\boldsymbol{\beta})^{-1}\|_2$ του πίνακα $F'(\boldsymbol{\beta})$. Συνεπώς, πρέπει να τονίσουμε εδώ ότι, το γεγονός ότι ο πίνακας $F'(\boldsymbol{\beta}) = X^TWX$ είναι μη ιδιάζων, σε συνδυασμό με το ότι η ποσότητα $C_2(F'(\boldsymbol{\beta}))$ δεν ξεπερνάει την τιμή του $k_2(F'(\boldsymbol{\beta}))$ [77], εξασφαλίζουν πως το άνω φράγμα της σταθεράς c είναι πεπερασμένο.*

Παρατήρηση 4.2 *Για την περίπτωση ενός γραμμικού συστήματος $\boldsymbol{X\beta} = \boldsymbol{y}$, ο Galantai [77] διερεύνησε μέσα από προσομοιώσεις, τη συμπεριφορά της σταθεράς c η οποία στην πραγματικότητα εξαρτάται από την διάσταση του πίνακα $\boldsymbol{X}$ ή ισοδύναμα από τη διάσταση d του $\boldsymbol{\beta}$. Το συμπέρασμα ήταν ότι η σταθερά c κατά μέσο όρο, στην εκτιμήτρια του Auchmuty για αυτήν την περίπτωση, αυξάνει αργά ως προς το d και ότι με μεγάλη πιθανότητα*

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \lessapprox \frac{0.5dim(X)\|res(\boldsymbol{\beta})\|_2^2}{\|X^Tres(\boldsymbol{\beta})\|_2}, \tag{4.25}$$

όπου $res(\boldsymbol{\beta}) = \boldsymbol{X\beta} - \boldsymbol{y}$ είναι το υπολειπόμενο σφάλμα για οποιαδήποτε προσεγγιστική λύση $\boldsymbol{\beta}$ του γραμμικού συστήματος.

Ωστόσο, χρειάζεται περαιτέρω έρευνα για τη διερεύνηση της συμπεριφοράς της σταθεράς c στην περίπτωσή μας, ειδικά για μεγάλης διάστασης δεδομένα. Εν τούτοις, παρατίθενται κάποια σχετικά σχόλια στη συνέχεια στην Παρατήρηση 4.3.

4.5 Συγκριτική Μελέτη Προσομοίωσης

Σε αυτήν την ενότητα, παρουσιάζουμε τα πειράματα προσομοίωσης που εκτελέσαμε, ώστε να συγκρίνουμε την προτεινόμενη μέθοδο, την οποία στο εξής τη συμβολίζουμε ως glmErrest, με τη συμβατική τεχνική επιλογής της ρυθμιστικής παραμέτρου, τη γενικευμένη διασταυρωμένη επικύρωση (GCV).

4.5.1 Σχεδιασμός Προσομοιώσεων

Καταρχήν, να σημειώσουμε ότι ως ποινικοποιημένο εκτιμητή του $\boldsymbol{\beta}$ στο βήμα 2 της διαδικασίας που περιγράψαμε παραπάνω, θεωρήσαμε τον εκτιμητή μέγιστης ποινικοποιημένης πιθανοφάνειας των Fan και Li [53] και ακολουθήσαμε το δικό τους σχήμα προσομοιώσεων. Συνεπώς, προσομοιώσαμε 100 σύνολα δεδομένων, αποτελούμενα από n παρατηρήσεις, βάσει των μοντέλων

1. $\mathbf{y} \sim Bernoulli(p(\mathbf{x}^T\boldsymbol{\beta}))$, όπου $p(u) = \frac{1}{1+e^{-u}}$.

2. $\mathbf{y} \sim Poisson(\mu(\mathbf{x}^T\boldsymbol{\beta}))$, όπου $\mu(u) = e^u$.

Οι τιμές που δόθησαν στο n ήταν 100, 200 και 500. Οι πρώτες έξι συνιστώσες του $\mathbf{x}$ προέκυψαν από την τυποποιημένη Κανονική κατανομή με συσχέτιση μεταξύ των x_i και x_j

ίση με $\rho^{|i-j|}$ όπου $\rho = 0.5$. Οι υπόλοιπες δύο συνιστώσες ήταν ανεξάρτητες και ισόνομα κατανεμημένες τυχαίες μεταβλητές από την κατανομή Bernoulli με πιθανότητα επιτυχίας 0.5. Το διάνυσμα με τις πραγματικές παραμέτρους είναι $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ για το μοντέλο Λογιστικής παλινδρόμησης, ενώ για το μοντέλο της Poisson παλινδρόμησης έχουμε ότι $\boldsymbol{\beta} = (1.2, 0.6, 0, 0, 0.8, 0, 0, 0)^T$.

4.5.2 Κριτήρια Σύγκρισης και Αξιολόγησης

Στις προσομοιώσεις που εκτελέσαμε, αξιολογήσαμε και συγκρίναμε την απόδοση των ποινικοποιημένων μεθόδων, με τις ποινές SCAD, LASSO και Hard, με γνώμονα, όπως και στο Κεφάλαιο 2, τα σφάλματα μοντέλου, την πολυπλοκότητα και την ακρίβεια του μοντέλου. Τα σφάλματα μοντέλου των διαδικασιών αυτών συγκρίθηκαν με αυτά των εκτιμητριών μέγιστης πιθανοφάνειας. Συγκεκριμένα, παρουσιάζουμε τη διάμεσο του σχετικού σφάλματος του μοντέλου (Median of Relative Model Error-MRME) των 100 προσομοιωμένων συνόλων δεδομένων, για τις προαναφερθείσες τιμές του n. Σημειώνουμε, ότι στα μοντέλα Λογιστικής και Poisson παλινδρόμησης, το σφάλμα μοντέλου ορίζεται ως [184]

$$ME = E[\mu(\mathbf{x}^T\hat{\boldsymbol{\beta}}) - \mu(\mathbf{x}^T\boldsymbol{\beta}_0)]^2. \tag{4.26}$$

ενώ για το σχετικό σφάλμα έχουμε ότι $RME = ME/ME_{full}$, με ME_{full} το σφάλμα μοντέλου υπολογιζόμενο προσαρμόζοντας στα δεδομένα το πλήρες μοντέλο. Επιπλέον, αναφέρουμε το μέσο αριθμό των σωστά και λανθασμένα αναγνωρισμένων μηδενικών συντελεστών (Aver. no. of 0 coeff., στήλες correct και incorrect αντίστοιχα στους επόμενους πίνακες). Ελέγξαμε επίσης την ακρίβεια του τύπου εύρεσης του τυπικού σφάλματος των Fan και Li [53]. Η διάμεση απόλυτη απόκλιση διαιρούμενη με 0.6745, η οποία συμβολίζεται με SD, των 100 εκτιμηθέντων συντελεστών στις προσομοιώσεις, μπορεί να θεωρηθεί ως το πραγματικό τυπικό σφάλμα. Επιπροσθέτως, η διάμεσος των 100 εκτιμηθέντων τυπικών αποκλίσεων (SDs), την οποία συμβολίζουμε ως SD_m και το διάμεσο απόλυτο σφάλμα απόκλισης των 100 εκτιμηθέντων τυπικών σφαλμάτων, διαιρούμενο με 0.6745, το οποίο συμβολίζεται ως SD_{mad}, καθορίζουν τη γενική συμπεριφορά και απόδοση του τύπου εύρεσης του τυπικού σφάλματος. Να σημειώσουμε, ότι στους πίνακας παρουσιάζονται τα αποτελέσματα μόνο για τους μη μηδενικούς συντελεστές. Επίσης, παραθέτουμε για κάθε μέθοδο τις μέσες τιμές των 100 εκτιμηθέντων συντελεστών των πραγματικά ενεργών μεταβλητών X_1, X_2 και X_5. Στους πίνακες παραθέτουμε και τις μέσες τιμές της παραμέτρου λ που επιλέχθηκε από τις συγκρινόμενες μεθόδους, glmErrest και GCV. Αναφέρουμε για λόγους σύγκρισης και τα αποτελέσματα του Oracle εκτιμητή, ο οποίος αντλείται προσαρμόζοντας το ιδανικό μοντέλο που περιλαμβάνει μόνο τις μεταβλητές X_1, X_2 και X_5. Να παρατηρήσουμε σε αυτό το σημείο ότι, από τη στιγμή που και η GCV χρησιμοποιεί κάποιες αρχικές τιμές της παραμέτρου λ στη ρουτίνα της, οι ίδιες δόθησαν και στη glmErrest.

4.5.3 Αποτελέσματα Προσομοιώσεων

Στους επόμενους Πίνακες 4.1-4.6, δίνονται τα αποτελέσματα από τις προσομοιώσεις που εκτελέσαμε και ακολουθεί σχετική συζήτησή τους.

Πίνακας 4.1: Αποτελέσματα προσομοιώσεων για το ποινικοποιημένο Λογιστικό μοντέλο παλινδρόμησης

Μέθοδος	**GCV**: **MRME**	**GCV**: **Aver. no. of 0 coeff.**		**glmErrest**: **MRME**	**glmErrest**: **Aver. no. of 0 coeff.**	
		correct	incorrect		correct	incorrect
n=100						
SCAD	0.7279	4.0400	0.1300	0.5923	4.4500	0.0900
LASSO	0.7445	4.0000	0.0900	0.6704	4.3200	0.1100
Hard	0.8341	3.5600	0.0400	0.7187	4.1700	0.0200
Oracle	0.3314	5	0	0.3314	5	0
n=200						
SCAD	0.3394	4.5400	0.2000	0.2838	5.0000	0.0500
LASSO	0.7894	3.1300	0	0.6540	3.2800	0
Hard	0.3507	4.6900	0.0200	0.2890	4.9200	0.0200
Oracle	0.2420	5	0	0.2420	5	0
n=500						
SCAD	0.3952	4.6600	0	0.2503	5.0000	0
LASSO	0.8119	3.0200	0.02	0.5898	3.1700	0
Hard	0.4306	4.3800	0	0.3010	4.9000	0
Oracle	0.2668	5	0	0.2668	5	0

Πίνακας 4.2: Τυπικές αποκλίσεις για το ποινικοποιημένο Λογιστικό μοντέλο παλινδρόμησης, με χρήση της glmErrest και της GCV (σε παρένθεση)

	$\hat{\boldsymbol{\beta}}_1$			$\hat{\boldsymbol{\beta}}_2$			$\hat{\boldsymbol{\beta}}_5$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=100									
SCAD	0.8570	0.8021	0.2136	0.6795	0.5688	0.1813	0.7445	0.6018	0.2337
	(0.9962)	(0.8802)	(0.3335)	(0.7220)	(0.6030)	(0.1895)	(0.8253)	(0.6427)	(0.2603)
LASSO	0.3973	0.4090	0.1274	0.3910	0.3329	0.1112	0.4606	0.3300	0.1032
	(0.2658)	(0.4589)	(0.1347)	(0.2543)	(0.3474)	(0.1342)	(0.3108)	(0.3517)	(0.1356)
Hard	0.8607	0.8371	0.2926	0.6509	0.6185	0.2017	0.7884	0.6735	0.2357
	(0.9200)	(0.8961)	(0.2612)	(0.6322)	(0.6290)	(0.1901)	(0.9251)	(0.6900)	(0.2519)
Oracle	0.8447	0.7465	0.2315	0.4601	0.5310	0.1047	0.6334	0.5686	0.1725
n=200									
SCAD	0.5989	0.5126	0.0990	0.3783	0.3665	0.0649	0.3477	0.3708	0.0574
	(0.4474)	(0.5420)	(0.0736)	(0.4885)	(0.3587)	(0.0751)	(0.3132)	(0.3867)	(0.0560)
LASSO	0.6262	0.5122	0.0970	0.3511	0.3680	0.0480	0.4102	0.3856	0.0728
	(0.4817)	(0.5376)	(0.0894)	(0.4497)	(0.3670)	(0.0562)	(0.4297)	(0.4034)	(0.0706)
Hard	0.6023	0.5300	0.1031	0.3479	0.3795	0.0567	0.4235	0.3880	0.0651
	(0.4387)	(0.5468)	(0.0783)	(0.4457)	(0.3699)	(0.0546)	(0.3534)	(0.3974)	(0.0577)
Oracle	0.5751	0.5271	0.0946	0.3263	0.3794	0.0483	0.3341	0.3883	0.0624
n=500									
SCAD	0.2958	0.3246	0.0347	0.2173	0.2319	0.0199	0.2361	0.2452	0.0253
	(0.2778)	(0.3297)	(0.0326)	(0.2443)	(0.2281)	(0.0191)	(0.2628)	(0.2447)	(0.0261)
LASSO	0.3033	0.3206	0.0325	0.2365	0.2309	0.0225	0.2568	0.2457	0.0294
	(0.2698)	(0.3254)	(0.0317)	(0.2577)	(0.2287)	(0.0177)	(0.2635)	(0.2205)	(0.0256)
Hard	0.2954	0.3264	0.0334	0.2099	0.2323	0.0198	0.2633	0.2464	0.0266
	(0.2589)	(0.3312)	(0.0322)	(0.2552)	(0.2283)	(0.0199)	(0.2632)	(0.2450)	(0.0283)
Oracle	0.3415	0.3328	0.0405	0.2350	0.2307	0.0251	0.2565	0.2437	0.0266

Πίνακας 4.3: Μέσες τιμές των μη μηδενικών συντελεστών και της παραμέτρου λ, με χρήση της glmErrest και της GCV (σε παρένθεση) για το ποινικοποιημένο Λογιστικό μοντέλο παλινδρόμησης

Μέθοδος	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$	$\hat{\lambda}$
n=100				
SCAD	3.2655(3.5394)	1.6271(1.8601)	2.2506(2.5467)	0.2212(0.2094)
LASSO	2.7264(2.2912)	1.3041(1.0556)	1.8238(1.4431)	0.0194(0.0138)
Hard	3.3662(3.8886)	1.6013(1.9754)	2.3054(2.6431)	0.2401(0.2295)
Oracle	3.3144	1.6140	2.2040	-
n=200				
SCAD	3.0753(3.2004)	1.4882(1.3401)	1.9761(2.0918)	0.3351(0.3319)
LASSO	3.1151(3.2227)	1.5650(1.5658)	2.0359(2.1330)	0.0015(0.0009)
Hard	3.0962(3.1923)	1.5446(1.5452)	2.0474(2.1276)	0.8735(0.8278)
Oracle	3.0749	1.5353	2.0873	-
n=500				
SCAD	3.0532(3.0963)	1.5415(1.3995)	2.0397(2.0969)	0.1959(0.1964)
LASSO	3.0821(3.1142)	1.5610(1.4165)	2.0834(2.0931)	0.0008(0.0005)
Hard	3.0615(3.0888)	1.5379(1.5513)	2.0364(2.0451)	0.4652(0.4731)
Oracle	3.0641	1.5236	2.0501	-

Πίνακας 4.4: Αποτελέσματα προσομοιώσεων για το ποινικοποιημένο Poisson μοντέλο παλινδρόμησης

Μέθοδος	**GCV**: **MRME**	**GCV**: **Aver. no. of 0 coeff.**		**glmErrest**: **MRME**	**glmErrest**: **Aver. no. of 0 coeff.**	
		correct	incorrect		correct	incorrect
n=100						
SCAD	0.6975	3.6100	0.0800	0.5306	3.8700	0.0200
LASSO	0.8303	3.4500	0.1100	0.6370	3.7200	0.0500
Hard	0.8614	3.5300	0.0500	0.6511	3.7800	0.0500
Oracle	0.3536	5	0	0.3536	5	0
n=200						
SCAD	0.8252	3.5200	0	0.6376	3.8700	0
LASSO	0.8383	3.4700	0	0.6374	3.8600	0
Hard	0.9431	3.5000	0	0.7567	3.8200	0
Oracle	0.3595	5	0	0.3595	5	0
n=500						
SCAD	0.8838	3.5700	0	0.5934	3.8800	0
LASSO	0.8913	3.5200	0	0.6335	3.8400	0
Hard	0.9277	3.5700	0	0.7421	3.8100	0
Oracle	0.3985	5	0	0.3985	5	0

Πίνακας 4.5: Τυπικές αποκλίσεις για το ποινικοποιημένο Poisson μοντέλο παλινδρόμησης, με χρήση της glmErrest και της GCV (σε παρένθεση)

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_5$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=100									
SCAD	0.0566	0.0576	0.0109	0.0658	0.0611	0.0096	0.0627	0.0527	0.0103
	(0.0609)	(0.0548)	(0.0113)	(0.0703)	(0.0604)	(0.0105)	(0.0552)	(0.0532)	(0.0082)
LASSO	0.0563	0.0560	0.0090	0.0649	0.0600	0.0090	0.0668	0.0541	0.0112
	(0.0670)	(0.0551)	(0.0104)	(0.0769)	(0.0614)	(0.0110)	(0.0652)	(0.0562)	(0.0098)
Hard	0.0566	0.0596	0.0105	0.0659	0.0629	0.0108	0.0625	0.0547	0.0117
	(0.0665)	(0.0556)	(0.0107)	(0.0740)	(0.0626)	(0.0110)	(0.0623)	(0.0557)	(0.0096)
Oracle	0.0566	0.0552	0.0094	0.0626	0.0613	0.0099	0.0542	0.0500	0.0087
n=200									
SCAD	0.0357	0.0363	0.0048	0.0460	0.0406	0.0060	0.0291	0.0348	0.0063
	(0.0282)	(0.0353)	(0.0052)	(0.0333)	(0.0399)	(0.0058)	(0.0429)	(0.0349)	(0.0062)
LASSO	0.0368	0.0354	0.0052	0.0451	0.0407	0.0052	0.0325	0.0348	0.0062
	(0.0268)	(0.0357)	(0.0053)	(0.0362)	(0.0403)	(0.0057)	(0.0429)	(0.0367)	(0.0067)
Hard	0.0345	0.0364	0.0050	0.0440	0.0410	0.0068	0.0347	0.0359	0.0066
	(0.0272)	(0.0357)	(0.0053)	(0.0374)	(0.0408)	(0.0063)	(0.0442)	(0.0365)	(0.0066)
Oracle	0.0315	0.0333	0.0061	0.0394	0.0385	0.0052	0.0359	0.0337	0.0041
n=500									
SCAD	0.0238	0.0215	0.0022	0.0252	0.0242	0.0028	0.0237	0.0202	0.0027
	(0.0194)	(0.0215)	(0.0018)	(0.0249)	(0.0243)	(0.0021)	(0.0216)	(0.0204)	(0.0035)
LASSO	0.0233	0.0212	0.0021	0.0249	0.0239	0.0024	0.0235	0.0202	0.0028
	(0.0188)	(0.0215)	(0.0018)	(0.0254)	(0.0246)	(0.0024)	(0.0221)	(0.0210)	(0.0038)
Hard	0.0245	0.0216	0.0023	0.0271	0.0246	0.0027	0.0234	0.0204	0.0030
	(0.0188)	(0.0216)	(0.0017)	(0.0252)	(0.0246)	(0.0023)	(0.0225)	(0.0210)	(0.0039)
Oracle	0.0216	0.0210	0.0027	0.0227	0.0235	0.0026	0.0214	0.0196	0.0022

Πίνακας 4.6: Μέσες τιμές των μη μηδενικών συντελεστών και της παραμέτρου λ, με χρήση της glmErrest και της GCV (σε παρένθεση) για το ποινικοποιημένο Poisson μοντέλο παλινδρόμησης

Μέθοδος	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$	$\hat{\lambda}$
n=100				
SCAD	1.2027(1.1846)	0.5958(0.6116)	0.7956(0.7803)	0.1083(0.0697)
LASSO	1.1968(1.1931)	0.5970(0.6109)	0.7832(0.7859)	0.0547(0.0089)
Hard	1.1960(1.1933)	0.6008(0.6129)	0.7949(0.7904)	0.0826(0.0739)
Oracle	1.2032	0.5914	0.7898	-
n=200				
SCAD	1.2023(1.1937)	0.5989(0.6170)	0.7952(0.7831)	0.0507(0.0393)
LASSO	1.2026(1.1878)	0.5934(0.6257)	0.7859(0.7910)	0.0456(0.0052)
Hard	1.2022(1.1954)	0.5986(0.6166)	0.7939(0.7927)	0.0550(0.0452)
Oracle	1.2009	0.5991	0.8003	-
n=500				
SCAD	1.2003(1.1961)	0.6007(0.5977)	0.8046(0.7980)	0.0340(0.0240)
LASSO	1.2011(1.1956)	0.5982(0.5971)	0.7997(0.7974)	0.0294(0.0031)
Hard	1.2009(1.1987)	0.6012(0.5981)	0.8047(0.7981)	0.0348(0.0264)
Oracle	1.1968	0.6019	0.8020	-

Από τους παραπάνω πίνακες παρατηρούμε ότι με τη χρήση της μεθόδου glmErrest για την επιλογή της ρυθμιστικής παραμέτρου, παρουσιάζεται μια σημαντική βελτίωση στη γενική απόδοση των ποινικοποιημένων μεθόδων, με γνώμονα τα αποτελέσματα του Oracle εκτιμητή. Πιο συγκεκριμένα, το MRME βελτιώνεται για όλες τις μεθόδους, ανεξάρτητα από το μέγεθος του δείγματος ή το μοντέλο που χρησιμοποιείται. Επιπλέον, παρατηρούμε ότι οι ποινικοποιημένες μέθοδοι και ειδικά η SCAD και η Hard, στην περίπτωση της ποινικοποιημένης Λογιστικής παλινδρόμησης, έχουν μεγαλύτερη πιθανότητα μηδενισμού των συντελεστών των μη σημαντικών μεταβλητών, όταν χρησιμοποιείται η glmErrest. Ταυτόχρονα, υπάρχει μικρότερη πιθανότητα μέσω της νέας μεθόδου για λάθος εκτίμηση των σημαντικών συντελεστών ως μηδενικούς. Στην περίπτωση τώρα της ποινικοποιημένης Poisson παλινδρόμησης, η νέα μέθοδος βελτιώνει εμφανώς την απόδοση όλων των ποινικοποιημένων μεθόδων, με καλύτερα αποτελέσματα για τις SCAD και Hard, συγκριτικά με τη LASSO. Ωστόσο, η απόδοση της LASSO αυξάνει σημαντικά με τη χρήση της glmErrest.

Παρατηρούμε επίσης ότι ο τύπος εύρεσης του τυπικού σφάλματος των Fan και Li [53] δίνει αρκετά καλά αποτελέσματα σε πολλές περιπτώσεις, όταν εφαρμόζουμε τη glmErrest, σε σύγκριση με τη GCV. Σχετικά με τις μέσες τιμές των μη μηδενικών συντελεστών, είναι εμφανές ότι σε όλες τις περιπτώσεις, οι εκτιμημένες και οι πραγματικές τιμές βρίσκονται αρκετά κοντά κατά τη χρήση της glmErrest, οδηγώντας φυσικά σε καλύτερη απόδοση των ποινικοποιημένων μεθόδων. Όσον αφορά τέλος τις μέσες τιμές για τις επιλεγμένες παραμέτρους λ, μια γενική παρατήρηση είναι ότι η προτεινόμενη μέθοδος οδηγεί σε υψηλότερες ή ελαφρώς υψηλότερες τιμές, σε σύγκριση με τη GCV.

Παρατήρηση 4.3 *Ο Galantai τονίζει στην εργασία του [77] ότι στην πράξη, στην περίπτωση μη τυχαίων πινάκων, η σταθερά c φαίνεται να λαμβάνει τιμές συνήθως μικρότερες του 10. Για τυχαίους πίνακες όμως, είναι πιθανόν να λαμβάνει τιμές αρκετά μεγαλύτερες, κάτι που σίγουρα εξαρτάται από την τιμή της διάστασης d. Ως εκ τούτου, εκτελέσαμε επιπλέον πειράματα προσομοίωσης, με σκοπό να διερευνηθούν και οι πιθανές τιμές του άνω φράγματος $C_2(F'(\boldsymbol{\beta}))$ της σταθεράς c. Τα αποτελέσματα σχετικά με το MRME και την αναγνωρισιμότητα του μοντέλου, για τα νέα πειράματα προσομοίωσης, ήταν παρόμοια με αυτά που αναλύονται παραπάνω. Θα τονίσουμε όμως ότι, στο Λογιστικό μοντέλο η μέση τιμή της $C_2(F'(\boldsymbol{\beta}))$ ήταν πράγματι σχεδόν πάντα μικρότερη του 10, στην περίπτωση όπου το πλήθος των παραμέτρων παλινδρόμησης ήταν μέχρι και 20 και για n=200 και 500. Αυτό συνέβη και για τις τρεις ποινικοποιημένες μεθόδους. Αντιθέτως, στο Poisson μοντέλο προέκυψαν αρκετά μεγαλύτερες μέσες τιμές για τη $C_2(F'(\boldsymbol{\beta}))$ σε κάποιες περιπτώσεις. Για παράδειγμα, για $d = 20$ και $n = 500$ προέκυψε μία μέση τιμή της $C_2(F'(\boldsymbol{\beta}))$ κοντά στο 275 και για τις τρεις ποινικοποιημένες μεθόδους. Ταυτόχρονα όμως, το ελάχιστο άνω φράγμα που έδωσε η (4.24) ήταν αρκετά χαμηλό, λαμβάνοντας την τιμή 0.24, για όλες τις μεθόδους, ενώ η μέση μεροληψία του $\hat{\boldsymbol{\beta}}$ ήταν ίση με 0.02. Επιπλέον, αναφέρουμε ότι ο Galantai [77] σημειώνει πως η $C_2(F'(\boldsymbol{\beta}))$ καταλήγει προσεγγιστικά ίση με το $k_2(F'(\boldsymbol{\beta}))/2$, αν το $k_2(F'(\boldsymbol{\beta}))$ είναι αρκετά μεγάλο. Η διαπίστωση αυτή ισχύει στα αποτελέσματά μας, όπως στην περίπτωση του ποινικοποιημένου Λογιστικού και Poisson μοντέλου παλινδρόμησης, για $d = 20$ και n=200 ή 500.*

Η προαναφερθείσα συζήτηση, υποδεικνύει ότι η προτεινόμενη μέθοδος λειτουργεί πολύ αποδοτικά στην επιλογή της ρυθμιστικής παραμέτρου λ και ως αποτέλεσμα, οδηγεί σε βελτιωμένη εκτίμηση της ποινικοποιημένης παραμέτρου $\boldsymbol{\beta}$.

4.6 Συμπεράσματα

Σε αυτό το κεφάλαιο, προτείναμε και αναλύσαμε νέες εκτιμήσεις της νόρμας του σφάλματος, στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, οι οποίες βασίζονται σε ήδη

υπάρχοντα θεωρητικά αποτελέσματα από το πεδίο των μη γραμμικών συστημάτων της Αριθμητικής Ανάλυσης. Προτείναμε επίσης μια νέα μέθοδο για την επιλογή της ρυθμιστικής παραμέτρου στα ποινικοποιημένα γενικευμένα γραμμικά μοντέλα για διακριτά δεδομένα, με βάση αυτές τις εκτιμήσεις. Όπως έχουμε αναφέρει και σε προηγούμενο κεφάλαιο, η επιλογή αυτής της παραμέτρου είναι κρίσιμη, καθότι ελέγχει την έκταση της ποινικοποίησης. Η μέθοδος της γενικευμένης διασταυρωμένης επικύρωσης χρησιμοποιείται συνήθως για το σκοπό αυτό. Παρότι αποτελεί μια αρκετά γνωστή και αποτελεσματική διαδικασία, οι προσομοιώσεις που εκτελέσαμε επιβεβαιώνουν την αξιοπιστία της νέας μεθόδου, δεδομένου ότι παράγει βελτιωμένα αποτελέσματα. Επιπλέον, είναι πιο εύκολη στην εφαρμογή της και υπολογιστικά λιγότερο χρονοβόρα. Να σημειώσουμε ότι χρησιμοποιώντας τη νέα μέθοδο, όχι μόνο αντλούμε τη ρυθμιστική παράμετρο, αλλά επίσης έχουμε ταυτόχρονα εκτίμηση των συντελεστών του μοντέλου, ελαχιστοποιώντας το μέγιστο του σφάλματος. Συμπερασματικά, η προτεινόμενη μέθοδος θα μπορούσε να θεωρηθεί μια αρκετά αξιόπιστη εναλλακτική προσέγγιση για την επιλογή της ρυθμιστικής παραμέτρου.

ΚΕΦΑΛΑΙΟ **5**

Ποινικοποιημένα Μοντέλα Ευπάθειας με Ομαδοποιημένα Δεδομένα

An approximate answer to the right problem
is worth a good deal more than
an exact answer to an approximate problem.

—John Tukey (1915–2000)

Η διαδικασία επιλογής της ρυθμιστικής παραμέτρου που αναπτύχθηκε στο προηγούμενο Κεφάλαιο 4, έδωσε πολύ καλά αποτελέσματα στο πλαίσιο των ποινικοποιημένων γενικευμένων γραμμικών μοντέλων. Στο παρόν κεφάλαιο, η μέθοδος επεκτείνεται θεωρώντας μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα. Συνεπώς, βασιζόμενοι και εδώ στις Kantorovich ανισώσεις, θα ξεκινήσουμε προτείνοντας νέες εκτιμήσεις της νόρμας του σφάλματος στη συγκεκριμένη κατηγορία μοντέλων. Έπειτα, οι εκτιμήσεις αυτές θα χρησιμοποιηθούν κατάλληλα για την εύρεση μιας νέας μεθόδου επιλογής της ρυθμιστικής παραμέτρου.

5.1 Ερευνητικό Πρόβλημα

Η αποδοτικότητα των μεθόδων ποινικοποιημένης πιθανοφάνειας, εξαρτάται κατά μεγάλο βαθμό από τη σωστή επιλογή της ρυθμιστικής παραμέτρου. Συνήθως, επιλέγεται μέσω της διασταυρωμένης επικύρωσης (CV) ή της γενικευμένης διασταυρωμένης επικύρωσης (GCV), τεχνικές που εφαρμόστηκαν στις εργασίες των Fan και Li [53], [54]. Λόγω των προβλημάτων υπολογισμού που απορρέουν από διαδικασίες αναδειγματοληψίας (CV) καθώς και από την πολυπλοκότητα ορισμένων κριτηρίων (GCV), σκεφτήκαμε στα δύο προηγούμενα κεφάλαια να χρησιμοποιήσουμε ή να τροποποιήσουμε διάφορες μορφές εκτίμησης του σφάλματος από το πεδίο της Αριθμητικής Ανάλυσης, οι οποίες να εφαρμοστούν κατάλληλα για την επιλογή της ρυθμιστικής παραμέτρου. Ο στόχος είναι να επιλέγεται με αποδοτικό τρόπο η ρυθμιστική παράμετρος, ελαχιστοποιώντας ταυτόχρονα το μέγιστο του σφάλματος εκτίμησης των ποινικοποιημένων συντελεστών παλινδρόμησης.

Λόγω της μη γραμμικότητας του μοντέλου ευπάθειας, θα βασιστούμε και στο παρόν κεφάλαιο στα βήματα του Galantai [77], προσαρμόζοντας την εκτίμηση σφάλματος που πρότεινε, στο δικό μας πρόβλημα. Οπότε, αρχικά θα αναπτύξουμε νέες εκτιμήσεις της νόρμας του σφάλματος, που σχετίζονται με την εκτίμηση των συντελεστών παλινδρόμησης στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα, μέσω της χρήσης των Kantorovich ανισοτήτων. Με βάση αυτές τις εκτιμήσεις, θα προτείνουμε μια νέα μέθοδο επιλογής της ρυθμιστικής παραμέτρου στα ποινικοποιημένα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα, με έμφαση στα μοντέλα ευπάθειας Γάμμα και Αντίστροφης Γκαουσιανής κατανομής. Θα χρησιμοποιήσουμε επίσης τις ειδικές μορφές των πιθανοφανειών των ως άνω μοντέλων, που αναλύθηκαν εκτενώς στο Κεφάλαιο 2. Η μέθοδος θα συγκριθεί, όπως και στα προηγούμενα κεφάλαια, με τη γενικευμένη διασταυρωμένη επικύρωση η οποία εξετάστηκε στο ποινικοποιημένο μοντέλο ευπάθειας Γάμμα των Fan και Li, στην εργασία [54].

5.2 Εκτίμηση Σφάλματος στα Μοντέλα Ευπάθειας με Ομαδοποιημένα Δεδομένα

Ξεκινάμε υπενθυμίζοντας κάποια βασικά θεωρητικά αποτελέσματα του Κεφαλαίου 2, όπου προτείναμε την ακόλουθη γενικευμένη μορφή ποινικοποιημένης λογαριθμο-πιθανοφάνειας:

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}^T\boldsymbol{\beta}+\sum_{l=1}^{N}\ln\mu_l$$

$$+\sum_{i=1}^{n}\ln\left(\int_0^\infty \exp\Big(-u_i\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})\Big)\,u_i^{A_i}\,dF_{u_i}(u_i)\right)-n\sum_{j=1}^{d}p_\lambda(|\beta_j|). \quad (5.1)$$

Δώσαμε επίσης τις μορφές της πρώτης και δεύτερης παραγώγου της (5.1) χωρίς τον όρο ποινής, ως προς $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_d)^T$. Συγκεκριμένα, η παράγωγος ως προς β_{k_1}, $k_1=1,\ldots,d$ είναι

$$\frac{\vartheta l(\boldsymbol{\beta})}{\vartheta\beta_{k_1}}=\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}x_{ijk_1}-\sum_{i=1}^{n}\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})\;x_{ijk_1} \quad (5.2)$$

ενώ για τη δεύτερη παράγωγο $\frac{\vartheta^2 l(\boldsymbol{\beta})}{\vartheta\beta_{k_1}\vartheta\beta_{k_2}}$, $k_1=1,\ldots,d$, $k_2=1,\ldots,d$ έχουμε ότι ισούται με την ποσότητα

$$\sum_{i=1}^{n}\left\{\frac{|L^{(A_i+2)}(x)|}{|L^{(A_i)}(x)|}-\left(\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}\right)^2\right\}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})x_{ijk_1}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})x_{ijk_2}$$

$$-\sum_{i=1}^{n}\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})x_{ijk_1}x_{ijk_2} \tag{5.3}$$

όπου $x=\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\mu_l I(z_l\le z_{ij})$, που ουσιαστικά εξαρτάται από το i.

Παρατήρηση 5.1 *Χρησιμοποιώντας συμβολισμό πινάκων, ο παρατηρούμενος πίνακας πληροφορίας είναι της μορφής*

$$\mathcal{J}(\beta)=-\nabla_\beta\nabla_\beta^T l(\beta)=-\sum_{i=1}^{n}C_iC_i^T+\sum_{i=1}^{n}X_i^TW_iX_i \tag{5.4}$$

όπου X_i, $i=1,\dots,n$ ένας πίνακας διάστασης $J_i\times d$,

$$X_i=\begin{pmatrix} x_{i1k_1} & x_{i1k_2} & \dots & x_{i1k_d}\\ & & \dots & \\ x_{iJ_ik_1} & x_{iJ_ik_2} & \dots & x_{iJ_ik_d}\end{pmatrix},$$

W_i ένας διαγώνιος πίνακας διάστασης $J_i\times J_i$ της μορφής

$$W_i=\begin{pmatrix}\Delta_i\gamma_{i1} & 0 & \dots & 0\\ 0 & \Delta_i\gamma_{i2} & \dots & 0\\ 0 & 0 & \dots & \Delta_i\gamma_{iJ_i}\end{pmatrix}$$

όπου $\Delta_i=\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}>0$ και $\gamma_{ij}=e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\lambda_l I(z_l\le z_{ij})>0$. Το διάνυσμα C_i είναι διάστασης d και ορίζεται ως

$$C_i^T=\left(\sum_{j=1}^{J_i}\sqrt{\Gamma_i}\gamma_{ij}x_{ijk_1},\sum_{j=1}^{J_i}\sqrt{\Gamma_i}\gamma_{ij}x_{ijk_2},\dots,\sum_{j=1}^{J_i}\sqrt{\Gamma_i}\gamma_{ij}x_{ijk_d}\right)^T$$

όπου $\Gamma_i=\frac{|L^{(A_i+2)}(x)|}{|L^{(A_i)}(x)|}-\left(\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}\right)^2$.

Συνεχίζουμε δίνοντας τις ειδικές αναλυτικές μορφές της πρώτης και δεύτερης παραγώγου της (5.1) χωρίς τον όρο ποινής, για τα μοντέλα ευπάθειας Γάμμα και Αντίστροφης Γκαουσιανής κατανομής. Αυτές θα τις χρειαστούμε στη συνέχεια για την εφαρμογή της νέας μεθόδου επιλογής της ρυθμιστικής παραμέτρου κατά τις προσομοιώσεις.

Στο μοντέλο ευπάθειας Γάμμα, έχουμε ότι η πρώτη παράγωγος της (5.1) χωρίς τον όρο ποινής, ως προς β_{k_1}, $k_1=1,\dots,d$ προκύπτει ως

$$\frac{\vartheta l(\boldsymbol{\beta})}{\vartheta\beta_{k_1}}=\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}x_{ijk_1}-\sum_{i=1}^{n}(A_i+\alpha)\frac{\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\lambda_l I(z_l\le z_{ij})\ x_{ijk_1}}{(\alpha+x)} \tag{5.5}$$

ενώ για τη δεύτερη παράγωγο $\frac{\vartheta^2 l(\boldsymbol{\beta})}{\vartheta\beta_{k_1}\vartheta\beta_{k_2}}$, $k_1=1,\dots,d$, $k_2=1,\dots,d$ έχουμε ότι ισούται με την ποσότητα

$$-\sum_{i=1}^{n}(A_i+\alpha)\frac{\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\lambda_l I(z_l\le z_{ij})\ x_{ijk_1}x_{ijk_2}}{(\alpha+x)}$$

$$+\sum_{i=1}^{n}(A_i+\alpha)\frac{\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\lambda_l I(z_l\le z_{ij})\ x_{ijk_1}\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\lambda_l I(z_l\le z_{ij})\ x_{ijk_2}}{(\alpha+x)(\alpha+x)} \tag{5.6}$$

όπου $x=\sum_{j=1}^{J_i}e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}}\sum_{l=1}^{N}\lambda_l I(z_l\le z_{ij})$.

Αναφορικά με το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής, η πρώτη παράγωγος της (5.1) χωρίς τον όρο ποινής, ως προς β_{k_1}, $k_1 = 1, \ldots, d$ προκύπτει ως

$$\frac{\vartheta l(\boldsymbol{\beta})}{\vartheta \beta_{k_1}} = \sum_{i=1}^{n}\sum_{j=1}^{J_i} \delta_{ij} x_{ijk_1} - \sum_{i=1}^{n} \sqrt{\frac{b}{b+x}} \frac{K_{1/2-A_i-1}\sqrt{4(b+x)b}}{K_{1/2-A_i}\sqrt{4(b+x)b}} \sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} \sum_{l=1}^{N} \lambda_l I(z_l \leq z_{ij})\ x_{ijk_1} \tag{5.7}$$

ενώ για τη δεύτερη παράγωγο $\frac{\vartheta^2 l(\boldsymbol{\beta})}{\vartheta\beta_{k_1}\vartheta\beta_{k_2}}$, $k_1 = 1, \ldots, d$, $k_2 = 1, \ldots, d$ έχουμε ότι ισούται με την ποσότητα

$$\begin{aligned}
&\sum_{i=1}^{n} \left\{ \left(\frac{b}{b+x}\right) \frac{K_{1/2-A_i-2}\sqrt{4(b+x)b}}{K_{1/2-A_i}\sqrt{4(b+x)b}} - \left(\sqrt{\frac{b}{b+x}} \frac{K_{1/2-A_i-1}\sqrt{4(b+x)b}}{K_{1/2-A_i}\sqrt{4(b+x)b}}\right)^2 \right\} \times \\
&\quad \times \sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} \sum_{l=1}^{N} \lambda_l I(z_l \leq z_{ij})\ x_{ijk_1} \sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} \sum_{l=1}^{N} \lambda_l I(z_l \leq z_{ij})\ x_{ijk_2} \\
&\quad - \sqrt{\frac{b}{b+x}} \frac{K_{1/2-A_i-1}\sqrt{4(b+x)b}}{K_{1/2-A_i}\sqrt{4(b+x)b}} \sum_{j=1}^{J_i} e^{\mathbf{x}_{ij}^T\boldsymbol{\beta}} \sum_{l=1}^{N} \lambda_l I(z_l \leq z_{ij})\ x_{ijk_1} x_{ijk_2}.
\end{aligned} \tag{5.8}$$

Στο σημείο αυτό, παραθέτουμε την προτεινόμενη εκτιμήτρια σφάλματος στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα.

Πρόταση 5.1 *Έστω $F(\boldsymbol{\beta})$ και $F'(\boldsymbol{\beta})$ η (αρνητική) πρώτη και δεύτερη παράγωγος της λογαριθμισμένης πιθανοφάνειας αντίστοιχα, στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα. Έστω $\boldsymbol{\beta}$ κάποια προσεγγιστική λύση και $\boldsymbol{\beta}^*$ η σωστή λύση. Μια εκτίμηση της νόρμας του σφάλματος δίνεται ως*

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 = c \frac{\|F(\boldsymbol{\beta})\|_2^2}{\|F'(\boldsymbol{\beta})^T F(\boldsymbol{\beta})\|_2}. \tag{5.9}$$

με $1 \lessapprox c \lessapprox C_2(F'(\boldsymbol{\beta})) = \frac{1}{2}\frac{\sigma_1^2+\sigma_d^2}{\sigma_1\sigma_d}$ και $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ οι ιδιάζουσες τιμές του $F'(\boldsymbol{\beta})$.

Η απόδειξη της πρότασης με χρήση της ανισότητας Kantorovich, στην περίπτωση των γενικευμένων γραμμικών μοντέλων για διακριτά δεδομένα, δίνεται στο προηγούμενο Κεφάλαιο 4. Η επέκτασή της στα μοντέλα ευπάθειας δε διαφέρει, θεωρώντας ως $F(\boldsymbol{\beta})$ τις μη γραμμικές 'score' εξισώσεις (5.2) πολλαπλασιασμένες με -1 και ως $F'(\boldsymbol{\beta})$ τον Ιακωβιανό πίνακα που ταυτίζεται με την αρνητική δεύτερη παράγωγο της (5.1), χωρίς τον όρο ποινής. Οπότε παραλείπουμε την απόδειξη και ο ενδιαφερόμενος αναγνώστης παραπέμπεται στην ενότητα 4.3 για τις τεχνικές λεπτομέρειες. Για να εφαρμόσουμε όμως την ανισότητα Kantorovich και να αποκτήσουμε την (5.9) στο πλαίσιο των μοντέλων ευπάθειας, πρέπει να ισχύουν οι ακόλουθες υποθέσεις [77]:

A. Όλες οι ομάδες - clusters είναι μη κενές. Οι παλινδρομητές $\mathbf{x}$ λαμβάνουν πεπερασμένες τιμές, είναι μη εκφυλισμένες (nondegenerate) τυχαίες μεταβλητές και είναι γραμμικά ανεξάρτητες εντός κάθε ομάδας i, $i = 1, \ldots, n$.

B. Για τη βασική σωρευτική συνάρτηση διακινδύνεσης ισχύει ότι $H(t) < \infty$ για $t < \infty$.

C. Ο μετασχηματισμός Laplace της κατανομής της ευπάθειας καθώς και οι παράγωγοί του μέχρι τάξης $A_i + 3$ για όλα τα i υπάρχουν (αλλά είναι πιθανόν να μη δίνονται σε κλειστή μορφή). Υποθέτουμε επίσης ότι τα $\frac{|L^{(A_i+3)}(x)|}{|L^{(A_i)}(x)|}$, $\frac{|L^{(A_i+2)}(x)|}{|L^{(A_i)}(x)|}$ και $\frac{|L^{(A_i+1)}(x)|}{|L^{(A_i)}(x)|}$ είναι ομοιόμορφα φραγμένα για όλα τα i.

Να παρατηρήσουμε ότι η υπόθεση **C** ικανοποιείται για τα ευρέως χρησιμοποιούμενα μοντέλα ευπάθειας Γάμμα και Αντίστροφης Γκαουσιανής κατανομής. Επίσης, η $F'(\boldsymbol{\beta})$ είναι συνεχής ως προς $\boldsymbol{\beta}$. Από τις υποθέσεις **A-C** λαμβάνουμε τη Lipschitz συνέχεια της $F'(\boldsymbol{\beta})$

βάσει του Θεωρήματος Μέσης Τιμής και του γεγονότος ότι η $F^{''}(\boldsymbol{\beta})$ είναι ομοιόμορφα φραγμένη για $\boldsymbol{\beta} \in \overline{S(\boldsymbol{\beta}^*, \delta)}$.

Επιπλέον, ο πίνακας $F^{'}(\boldsymbol{\beta}^*)$ που συμπίπτει με τον πίνακα $J(\boldsymbol{\beta}^*)$ στην (5.4), είναι συμμετρικός και μη αρνητικά ορισμένος ως παρατηρούμενος πίνακας πληροφορίας. Η υπόθεση **A** αποκλείει εκφυλισμένες περιπτώσεις που ίσως είχαν ως αποτέλεσμα έναν παρατηρούμενο πίνακα πληροφορίας για το $\boldsymbol{\beta}$ με μηδενική ορίζουσα και συνεπώς ιδιάζων. Είναι πιθανόν παρ΄ όλα αυτά να συναντήσουμε αυτό το πρόβλημα στην πράξη, ειδικά σε δεδομένα μεγάλης διάστασης. Αν προκύψει κάτι τέτοιο, επηρεάζεται όχι μόνο η μέθοδος επιλογής της ρυθμιστικής παραμέτρου που θα αναπτύξουμε στη συνέχεια, αλλά γενικά η όλη μεθοδολογία ποινικοποιημένης πιθανοφάνειας και με τη σειρά της η μέθοδος της γενικευμένης διασταυρωμένης επικύρωσης, ακόμα και η ύπαρξη της ΕΜΠ του $\boldsymbol{\beta}$. Συνεπώς, υποθέτουμε στο παρόν κεφάλαιο, τυπικές συνθήκες κανονικότητας και μη εκφυλισιμότητας για την παράμετρο $\boldsymbol{\beta}$ (Bickel και Doksum [16], Θεώρημα 6.2.2, Van der Vaart [166], Θεώρημα 5.39) οι οποίες συνεπάγονται ότι ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ του $\boldsymbol{\beta}$ υπάρχει και είναι μοναδικά ορισμένος ως η λύση των 'score' εξισώσεων $F(\boldsymbol{\beta}) = 0$, καθώς επίσης είναι αποτελεσματικός (efficient), με πίνακα διασποράς - συνδιασποράς $I^{-1}(\boldsymbol{\beta}_0)$ όπου $I(\boldsymbol{\beta}_0)$ ο πραγματικός πίνακας πληροφορίας του Fisher για το $\boldsymbol{\beta}$. Αν το μέγεθος του δείγματος είναι μεγάλο, ο νόμος των μεγάλων αριθμών σε συνδυασμό με τη συνέχεια της ιδιότητας της θετικής οριστικότητας (positive definiteness) σε χώρο πινάκων, είναι αρκετά για να εξασφαλίσουν ότι ο παρατηρούμενος πίνακας πληροφορίας είναι θετικά ορισμένος, αν είναι και ο πραγματικός.

Να σημειώσουμε στο σημείο αυτό, ότι η εκτιμήτρια σφάλματος (5.9) που προτείναμε, δεν απαιτεί τον υπολογισμό του αντιστρόφου του παρατηρούμενου πίνακα πληροφορίας, κάτι που συμβαίνει στο κριτήριο της γενικευμένης διασταυρωμένης επικύρωσης. Για τον υπολογισμό του τελευταίου κριτηρίου, χρησιμοποιείται το ίχνος του αντιστρόφου ενός πίνακα, σχετιζόμενου με τον παρατηρούμενο πίνακα πληροφορίας και με έναν όρο ποινής [53], [54]. Συνεπώς, η προτεινόμενη εκτιμήτρια υπερέχει σε υπολογιστικό χρόνο και εφαρμόζεται εύκολα.

5.3 Η Προτεινόμενη Μέθοδος Επιλογής της Ρυθμιστικής Παραμέτρου

Σε αυτήν την ενότητα, προτείνουμε μια νέα μέθοδο για την επιλογή της ρυθμιστικής παραμέτρου λ, στα ποινικοποιηνένα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα, μέσω της κατάλληλης χρήσης της (5.9). Η διαδικασία που προτείνουμε περιγράφεται ρητά ως εξής:

1. Ξεκινάμε με ένα σύνολο από αρχικές τιμές για τη ρυθμιστική παράμετρο λ.

2. Για κάθε λ στο σύνολο αυτό, υπολογίζουμε έναν ποινικοποιημένο εκτιμητή $\hat{\boldsymbol{\beta}}$.

3. Έπειτα, υπολογίζουμε τη νόρμα του σφάλματος βάσει του τύπου (5.9), με τη σταθερά c να λαμβάνει την τιμή του άνω φράγματός της, χρησιμοποιώντας τις κατάλληλες μορφές των $F(\hat{\boldsymbol{\beta}})$ και $F'(\hat{\boldsymbol{\beta}})$, ανάλογα με τη θεωρούμενη κατανομή της ευπάθειας.

4. Η τιμή της παραμέτρου λ του συνόλου, που ελαχιστοποιεί τη (5.9) για την προαναφερθείσα σταθερά c, είναι η επιθυμητή ρυθμιστική παράμετρος και ο προκύπτων ποινικοποιημένος εκτιμητής $\hat{\boldsymbol{\beta}}$ είναι ο τελικός εκτιμητής του $\boldsymbol{\beta}$.

Παρατήρηση 5.2 *Το άνω φράγμα της σταθεράς c στην (5.9), η ποσότητα $C_2(F'(\boldsymbol{\beta}))$ δηλαδή, είναι μια απλή συνάρτηση του δείκτη κατάστασης $k_2(F'(\boldsymbol{\beta})) = \|F'(\boldsymbol{\beta})\|_2 \|F'(\boldsymbol{\beta})^{-1}\|_2$ του πίνακα $F'(\boldsymbol{\beta})$. Οπότε, να τονίσουμε εδώ ότι λόγω των υποθέσεων **A**-**C** και της συζήτησης που ακολούθησε, ο πίνακας $F'(\boldsymbol{\beta})$ θεωρείται μη ιδιάζων. Το γεγονός επίσης ότι η ποσότητα $C_2(F'(\boldsymbol{\beta}))$ είναι μικρότερη ή ίση του $k_2(F'(\boldsymbol{\beta}))$ [77], εξασφαλίζουν πως το άνω φράγμα $C_2(F'(\boldsymbol{\beta}))$ είναι πεπερασμένο. Αν όμως ο πίνακας είναι ιδιάζων ή σχεδόν ιδιάζων, κάτι που*

είναι πιθανό να συμβεί στην πράξη, μπορούμε να βασιστούμε στο δείκτη κατάστασης με χρήση του γενικευμένου αντιστρόφου Moore-Penrose. Η τεχνική αυτή εφαρμόζεται και από τους Fan και Li στις εργασίες [53] και [54], όταν εμφανίζεται το συγκεκριμένο πρόβλημα.

5.4 Συγκριτική Μελέτη Προσομοίωσης

Συνεχίζουμε σε αυτήν την ενότητα, παρουσιάζοντας και αναλύοντας τα πειράματα προσομοίωσης που εκτελέσαμε. Θα αξιολογήσουμε την προτεινόμενη μέθοδο, την οποία στο εξής τη συμβολίζουμε ως frailErrest και θα τη συγκρίνουμε με τη συμβατική τεχνική επιλογής της ρυθμιστικής παραμέτρου, τη γενικευμένη διασταυρωμένη επικύρωση (GCV).

5.4.1 Σχεδιασμός Προσομοιώσεων

Το σχήμα προσομοιώσεων που ακολουθήσαμε βασίζεται σε αυτό του Κεφαλαίου 2. Συνεπώς, προσομοιώσαμε 100 σύνολα δεδομένων αποτελούμενα από n ομάδες και J άτομα σε κάθε ομάδα, βάσει του μοντέλου

$$h(t|\mathbf{x}, u) = u \exp(\mathbf{x}^T \boldsymbol{\beta}), \tag{5.10}$$

όπου $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0.6, 0)$, τα x_i παράγονται από την τυποποιημένη Κανονική κατανομή και η συσχέτιση μεταξύ των x_i και x_j είναι $\rho^{|i-j|}$ με $\rho = 0.5$. Το ποσοστό αποκοπής που επιβάλαμε ήταν της τάξης του 30%. Εξετάσαμε τις περιπτώσεις της ευπάθειας Γάμμα κατανομής με $\alpha = 4$ και της Αντίστροφης Γκαουσιανής κατανομής με $b = 2$, ώστε η διασπορά να είναι ίση με 1/4. Ως ποινικοποιημένο εκτιμητή του $\boldsymbol{\beta}$ στο βήμα 2 της διαδικασίας που περιγράψαμε παραπάνω, θεωρήσαμε τον εκτιμητή μέγιστης ποινικοποιημένης πιθανοφάνειας των Fan και Li [54].

5.4.2 Κριτήρια Σύγκρισης και Αξιολόγησης

Αξιολογήσαμε την απόδοση των ποινικοποιημένων μεθόδων στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα, με τις συναρτήσεις ποινής SCAD, LASSO και Hard, χρησιμοποιώντας ως συγκρινόμενες μεθόδους επιλογής της ρυθμιστικής παραμέτρου, την προτεινόμενη frailErrest και τη GCV. Θα παρουσιάσουμε στη συνέχεια τα αποτελέσματα σχετικά με τη διάμεσο του σχετικού σφάλματος του μοντέλου (Median of Relative Model Error-MRME) των 100 προσομοιωμένων συνόλων δεδομένων, για διάφορους συνδυασμούς των n και J. Επίσης, δίνουμε το μέσο αριθμό των σωστά και λανθασμένα αναγνωρισμένων μηδενικών συντελεστών (Aver. no. of 0 coeff., στήλες correct και incorrect αντίστοιχα στους επόμενους πίνακες). Εξετάσαμε επιπλέον και την ακρίβεια του τύπου εύρεσης του τυπικού σφάλματος των Fan και Li [54], παρουσιάζοντας μόνο τα αποτελέσματα για τους μη μηδενικούς συντελεστές.

Εν συνεχεία, παραθέτουμε για κάθε ποινικοποιημένη μέθοδο τις μέσες τιμές των 100 εκτιμηθέντων συντελεστών των πραγματικά ενεργών μεταβλητών X_1, X_4 και X_7, καθώς επίσης και τις μέσες τιμές των α και b για τις οποίες μεγιστοποιήθηκε η πιθανοφάνεια. Δίνουμε επίσης τις μέσες τιμές της παραμέτρου λ που επιλέχθηκε από τις μεθόδους frailErrest και GCV. Για όλα τα παραπάνω, θεωρήσαμε ως τιμές σύγκρισης τα αποτελέσματα που λάβαμε από τον Oracle εκτιμητή, τα οποία προέκυψαν προσαρμόζοντας το ιδανικό μοντέλο που συμπεριλαμβάνει μόνο τις X_1, X_4 και X_7. Δόθησαν οι ίδιες αρχικές τιμές της παραμέτρου λ στις ρουτίνες της GCV και της frailErrest.

5.4.3 Αποτελέσματα Προσομοιώσεων

Στους παρακάτω Πίνακες 5.1-5.6, δίνονται όλα τα αποτελέσματα των προσομοιώσεων που εκτελέσαμε, τα οποία και σχολιάζουμε στη συνέχεια.

Πίνακας 5.1: Αποτελέσματα προσομοιώσεων για το μοντέλο ευπάθειας Γάμμα κατανομής

Μέθοδος	**GCV**: **MRME**	**GCV**: **Aver. no. of 0 coeff.**		**frailErrest**: **MRME**	**frailErrest**: **Aver. no. of 0 coeff.**	
		correct	incorrect		correct	incorrect
n=50, J=2						
SCAD	0.7637	4.4300	0.0200	0.6302	4.5800	0.0200
LASSO	1.1644	3.6200	0.0200	0.7597	3.7800	0.0200
Hard	0.7430	4.3100	0.1100	0.6921	4.6900	0.0700
Oracle	0.4206	5	0	0.4206	5	0
n=50, J=5						
SCAD	0.6231	4.4600	0	0.5589	4.8800	0
LASSO	0.9445	3.5500	0	0.7511	3.8700	0
Hard	0.7548	4.1700	0.0400	0.6205	4.7300	0
Oracle	0.5156	5	0	0.5156	5	0
n=100, J=2						
SCAD	0.7903	4.4200	0	0.6432	4.9900	0
LASSO	0.9249	3.7400	0	0.7226	3.8600	0
Hard	0.8572	4.4800	0.0700	0.6797	4.7700	0
Oracle	0.6044	5	0	0.6044	5	0
n=100, J=5						
SCAD	0.6341	4.4500	0	0.5309	4.8600	0
LASSO	0.7474	3.5200	0	0.6743	3.6500	0
Hard	0.8115	4.5600	0	0.6392	4.7100	0
Oracle	0.5872	5	0	0.5872	5	0

Πίνακας 5.2: Μέσες τιμές των μη μηδενικών συντελεστών και των α και λ, με χρήση των μεθόδων frailErrest και GCV (σε παρένθεση) για το μοντέλο ευπάθειας Γάμμα κατανομής

Μέθοδος	β_1	β_4	β_7	α	λ
n=50, J=2					
SCAD	0.7205(0.7142)	0.9735(0.9718)	0.4853(0.3693)	4.2559(4.5775)	0.1736(0.1903)
LASSO	0.7517(0.5277)	0.9241(0.6828)	0.5202(0.3333)	4.5680(4.8958)	0.0200(0.0988)
Hard	0.8301(0.7917)	1.0245(1.0085)	0.5882(0.4267)	4.0720(4.6383)	0.4100(0.5376)
Oracle	0.8175	1.0099	0.5905	4.1101	-
n=50, J=5					
SCAD	0.8580(0.8757)	1.1000(1.0909)	0.6407(0.5885)	4.1454(4.2176)	0.1199(0.1679)
LASSO	0.8404(0.7349)	1.0434(0.9192)	0.5909(0.5026)	4.1620(4.1680)	0.0136(0.0424)
Hard	0.8376(0.8860)	1.1031(1.1004)	0.6178(0.6258)	4.1784(4.2544)	0.2765(0.3251)
Oracle	0.8878	1.1014	0.6462	4.1227	-
n=100, J=2					
SCAD	0.8061(0.7932)	1.0036(0.9980)	0.5696(0.5293)	4.2726(4.5430)	0.1275(0.1643)
LASSO	0.8048(0.6366)	1.0046(0.8124)	0.5844(0.4322)	4.4425(4.4605)	0.0038(0.0565)
Hard	0.8122(0.8025)	1.0114(1.0021)	0.5895(0.5338)	4.4517(4.5832)	0.3296(0.3712)
Oracle	0.8105	1.0077	0.5896	4.2897	-
n=100, J=5					
SCAD	0.8458(0.8368)	1.0375(1.0304)	0.6099(0.5703)	4.2338(4.3721)	0.0998(0.1584)
LASSO	0.8389(0.7351)	1.0231(0.9090)	0.6068(0.5153)	4.2939(4.5920)	0.0033(0.0318)
Hard	0.8479(0.8482)	1.0378(1.0391)	0.6161(0.6164)	4.1754(4.1990)	0.1948(0.2304)
Oracle	0.8464	1.0369	0.6151	4.1566	-

Πίνακας 5.3: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Γάμμα κατανομής, με χρήση των μεθόδων frailErrest και GCV

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD(frailErrest)	0.1155	0.1177	0.0191	0.0993	0.1243	0.0166	0.1651	0.1702	0.0454
SCAD(GCV)	0.1062	0.1177	0.0211	0.1115	0.1230	0.0166	0.1703	0.1533	0.0506
LASSO(frailErrest)	0.1049	0.1240	0.0142	0.1228	0.1347	0.0168	0.1282	0.1211	0.0149
LASSO(GCV)	0.0977	0.0902	0.0110	0.1250	0.0973	0.0114	0.1371	0.0799	0.0126
Hard(frailErrest)	0.0918	0.1091	0.0159	0.1143	0.1340	0.0171	0.1186	0.1262	0.0163
Hard(GCV)	0.1057	0.1083	0.0420	0.1200	0.1059	0.0460	0.1324	0.1071	0.0194
Oracle	0.0924	0.1257	0.0130	0.0986	0.1322	0.0161	0.1058	0.1281	0.0138
n=50, J=5									
SCAD(frailErrest)	0.0612	0.0849	0.0085	0.0803	0.0853	0.0072	0.0826	0.0834	0.0064
SCAD(GCV)	0.0615	0.0834	0.0101	0.0781	0.0829	0.0105	0.0802	0.0814	0.0125
LASSO(frailErrest)	0.0689	0.0857	0.0089	0.0831	0.0888	0.0091	0.0790	0.0849	0.0076
LASSO(GCV)	0.0732	0.0744	0.0072	0.0935	0.0766	0.0057	0.0627	0.0705	0.0064
Hard(frailErrest)	0.0702	0.0849	0.0086	0.0883	0.0856	0.0081	0.0611	0.0846	0.0078
Hard(GCV)	0.0585	0.0830	0.0092	0.0814	0.0832	0.0098	0.0422	0.0813	0.0091
Oracle	0.0580	0.0848	0.0084	0.0827	0.0854	0.0070	0.0471	0.0837	0.0064
n=100, J=2									
SCAD(frailErrest)	0.0842	0.0960	0.0071	0.0730	0.0950	0.0070	0.0796	0.0910	0.0085
SCAD(GCV)	0.0805	0.0930	0.0082	0.0720	0.0926	0.0081	0.1081	0.0840	0.0128
LASSO(frailErrest)	0.0722	0.0996	0.0089	0.1105	0.1053	0.0111	0.0845	0.1061	0.0098
LASSO(GCV)	0.0765	0.0981	0.0062	0.0915	0.0809	0.0054	0.0769	0.0725	0.0057
Hard(frailErrest)	0.0836	0.0968	0.0083	0.0812	0.0967	0.0079	0.0784	0.0950	0.0081
Hard(GCV)	0.0902	0.0819	0.0086	0.0752	0.0924	0.0092	0.0833	0.0897	0.0103
Oracle	0.0866	0.0963	0.0072	0.0744	0.0959	0.0072	0.0742	0.0925	0.0079
n=100, J=5									
SCAD(frailErrest)	0.0393	0.0401	0.0043	0.0448	0.0598	0.0037	0.0438	0.0601	0.0037
SCAD(GCV)	0.0325	0.0570	0.0071	0.0499	0.0566	0.0079	0.0554	0.0668	0.0069
LASSO(frailErrest)	0.0519	0.0633	0.0048	0.0534	0.0663	0.0055	0.0519	0.0661	0.0054
LASSO(GCV)	0.0431	0.0544	0.0043	0.0447	0.0554	0.0028	0.0533	0.0524	0.0029
Hard(frailErrest)	0.0392	0.0506	0.0042	0.0441	0.0612	0.0043	0.0491	0.0614	0.0038
Hard(GCV)	0.0376	0.0599	0.0040	0.0447	0.0605	0.0039	0.0464	0.0606	0.0038
Oracle	0.0385	0.0604	0.0039	0.0453	0.0604	0.0039	0.0464	0.0605	0.0038

Πίνακας 5.4: Αποτελέσματα προσομοιώσεων για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής

Μέθοδος	**GCV**: **MRME**	**GCV**: **Aver. no. of 0 coeff.**		**frailErrest**: **MRME**	**frailErrest**: **Aver. no. of 0 coeff.**	
		correct	incorrect		correct	incorrect
n=50, J=2						
SCAD	0.5315	4.2700	0.0800	0.4104	4.6600	0.0400
LASSO	0.5987	3.5800	0.0900	0.5589	4.0500	0.0700
Hard	0.5815	4.2600	0.2200	0.5110	4.5500	0.1200
Oracle	0.3979	5	0	0.3979	5	0
n=50, J=5						
SCAD	0.5978	4.3400	0	0.4332	4.7500	0
LASSO	1.1452	3.5100	0	0.8137	3.6400	0
Hard	0.4778	4.3600	0.0200	0.4395	4.7400	0
Oracle	0.4012	5	0	0.4012	5	0
n=100, J=2						
SCAD	0.7548	4.4600	0	0.6805	4.8000	0
LASSO	1.1272	3.5300	0	0.8299	3.7500	0
Hard	0.9403	4.3700	0.0700	0.8176	4.7200	0
Oracle	0.6360	5	0	0.6360	5	0
n=100, J=5						
SCAD	1.0308	4.4800	0	0.8371	4.7600	0
LASSO	1.8110	3.2600	0	1.1799	3.5300	0
Hard	0.9118	4.4700	0	0.8059	4.7200	0
Oracle	0.8438	5	0	0.8438	5	0

Πίνακας 5.5: Μέσες τιμές των μη μηδενικών συντελεστών και των b και λ, με χρήση των μεθόδων frailErrest και GCV (σε παρένθεση) για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής

Μέθοδος	β_1	β_4	β_7	b	λ
n=50, J=2					
SCAD	0.7620(0.7350)	1.0411(1.0376)	0.4570(0.4552)	2.4996(2.6609)	0.1765(0.1921)
LASSO	0.7652(0.5597)	0.9823(0.7695)	0.5676(0.3784)	2.5939(2.7799)	0.0213(0.0900)
Hard	0.8393(0.7843)	1.0854(1.0684)	0.5934(0.5323)	2.5482(2.4978)	0.4573(0.5378)
Oracle	0.8267	1.0578	0.6176	2.3931	-
n=50, J=5					
SCAD	0.8197(0.8167)	1.0138(1.013)	0.6016(0.5711)	2.5319(2.7111)	0.1222(0.1719)
LASSO	0.7818(0.6668)	0.9696(0.8364)	0.5800(0.4716)	2.6353(2.7961)	0.0123(0.0484)
Hard	0.8260(0.8244)	1.0205(1.0186)	0.6204(0.6105)	2.5936(2.7172)	0.3065(0.3261)
Oracle	0.8205	1.0130	0.6142	2.4889	-
n=100, J=2					
SCAD	0.7953(0.7932)	1.0009(1.0040)	0.5775(0.5537)	2.3541(2.3986)	0.1279(0.1694)
LASSO	0.7518(0.6228)	0.9510(0.8014)	0.5604(0.4355)	2.7373(2.7380)	0.0144(0.0596)
Hard	0.8083(0.7971)	1.0170(1.0029)	0.6021(0.5604)	2.4550(2.4638)	0.3130(0.3850)
Oracle	0.8000	1.0007	0.5934	2.2354	-
n=100, J=5					
SCAD	0.7583(0.7526)	0.9577(0.9546)	0.5633(0.5319)	2.4547(2.4959)	0.0928(0.1493)
LASSO	0.7326(0.6556)	0.9326(0.8417)	0.5428(0.4717)	2.6742(2.6285)	0.0089(0.0354)
Hard	0.7631(0.7594)	0.9667(0.9578)	0.5732(0.5684)	2.4907(2.6524)	0.2062(0.2202)
Oracle	0.7581	0.9563	0.5674	2.2965	-

Πίνακας 5.6: Τυπικές αποκλίσεις για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής, με χρήση των μεθόδων frailErrest και GCV

	$\hat{\beta}_1$			$\hat{\beta}_4$			$\hat{\beta}_7$		
Μέθοδος	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}	SD	SD_m	SD_{mad}
n=50, J=2									
SCAD(frailErrest)	0.1291	0.1203	0.0119	0.1550	0.1394	0.0147	0.1861	0.1990	0.0215
SCAD(GCV)	0.1406	0.1180	0.0161	0.1579	0.1161	0.0164	0.1943	0.1692	0.0515
LASSO(frailErrest)	0.1247	0.1215	0.0129	0.1553	0.1267	0.0143	0.1426	0.1222	0.0138
LASSO(GCV)	0.1253	0.0945	0.0105	0.1475	0.0983	0.0105	0.1457	0.0844	0.0154
Hard(frailErrest)	0.1416	0.1259	0.0148	0.1544	0.1217	0.0169	0.1263	0.1222	0.0212
Hard(GCV)	0.1427	0.0926	0.0730	0.1682	0.0936	0.0535	0.1137	0.1276	0.0136
Oracle	0.1276	0.1278	0.0112	0.1464	0.1232	0.0131	0.1137	0.1276	0.0136
n=50, J=5									
SCAD(frailErrest)	0.0657	0.0805	0.0074	0.0604	0.0827	0.0088	0.0733	0.0811	0.0086
SCAD(GCV)	0.0601	0.0796	0.0086	0.0603	0.0791	0.0120	0.0946	0.0760	0.0117
LASSO(frailErrest)	0.0703	0.0794	0.0062	0.0714	0.0869	0.0086	0.0892	0.0838	0.0084
LASSO(GCV)	0.0574	0.0675	0.0050	0.0627	0.0724	0.0074	0.0739	0.0655	0.0066
Hard(frailErrest)	0.0614	0.0816	0.0076	0.0657	0.0842	0.0090	0.0686	0.0834	0.0077
Hard(GCV)	0.0551	0.0798	0.0080	0.0641	0.0823	0.0103	0.0637	0.0812	0.0090
Oracle	0.0542	0.0811	0.0072	0.0638	0.0837	0.0091	0.0600	0.0829	0.0076
n=100, J=2									
SCAD(frailErrest)	0.0795	0.0903	0.0077	0.0827	0.0925	0.0081	0.0887	0.0910	0.0081
SCAD(GCV)	0.0770	0.0904	0.0083	0.0917	0.0909	0.0083	0.0812	0.0893	0.0113
LASSO(frailErrest)	0.0830	0.0906	0.0084	0.0951	0.0965	0.0086	0.0760	0.0922	0.0076
LASSO(GCV)	0.0807	0.0759	0.0057	0.0813	0.0779	0.0052	0.0617	0.0718	0.0054
Hard(frailErrest)	0.0837	0.0916	0.0075	0.0925	0.0932	0.0088	0.0796	0.0916	0.0077
Hard(GCV)	0.0771	0.0885	0.0097	0.0876	0.0896	0.0095	0.0675	0.0900	0.0095
Oracle	0.0774	0.0915	0.0074	0.0833	0.0924	0.0075	0.0608	0.0927	0.0077
n=100, J=5									
SCAD(frailErrest)	0.0441	0.0588	0.0039	0.0532	0.0588	0.0039	0.0520	0.0587	0.0045
SCAD(GCV)	0.0507	0.0580	0.0055	0.0639	0.0571	0.0062	0.0591	0.0549	0.0059
LASSO(frailErrest)	0.0471	0.0589	0.0040	0.0631	0.0626	0.0048	0.0504	0.0606	0.0042
LASSO(GCV)	0.0442	0.0522	0.0040	0.0535	0.0541	0.0036	0.0500	0.0506	0.0032
Hard(frailErrest)	0.0425	0.0587	0.0041	0.0589	0.0611	0.0043	0.0466	0.0609	0.0043
Hard(GCV)	0.0437	0.0587	0.0040	0.0557	0.0589	0.0039	0.0415	0.0591	0.0053
Oracle	0.0439	0.0587	0.0040	0.0556	0.0587	0.0040	0.0414	0.0590	0.0042

Ένα γενικό σχόλιο που μπορούμε να κάνουμε, παρατηρώντας τους Πίνακες 5.1-5.6, είναι ότι η χρήση της προτεινόμενης μεθόδου επιλογής της ρυθμιστικής παραμέτρου, frailErrest, αυξάνει την απόδοση των ποινικοποιημένων μεθόδων, θεωρώντας ως βέλτιστα τα αποτελέσματα του Oracle εκτιμητή. Συγκεκριμένα, μειώνεται το MRME, ανεξάρτητα από το μέγεθος του δείγματος, το μέγεθος των ομάδων ή την κατανομή της ευπάθειας που χρησιμοποιήσαμε. Όσον αφορά την αναγνωρισιμότητα του μοντέλου, η χρήση των συναρτήσεων ποινής SCAD και Hard, σε συνδυασμό με την frailErrest, οδηγούν σε καλύτερα αποτελέσματα, συγκριτικά με την περίπτωση της GCV, εκτιμώντας σωστά τους συντελεστές των μη σημαντικών μεταβλητών ως μηδενικούς. Ακολουθεί η LASSO, της οποίας η απόδοση βελτιώνεται αρκετά με χρήση της frailErrest.

Αναφορικά με τις μέσες τιμές των μη μηδενικών συντελεστών και των παραμέτρων α ή b, βλέπουμε ότι σε όλες τις περιπτώσεις και με τη χρήση της frailErrest, οι εκτιμηθείσες και οι πραγματικές τιμές βρίσκονται πολύ κοντά. Σχετικά τώρα με τις μέσες τιμές των επιλεγμένων ρυθμιστικών παραμέτρων λ, η νέα μέθοδος οδηγεί γενικά σε μικρότερες τιμές, σε σχέση με τη GCV. Επίσης, ο τύπος εύρεσης του τυπικού σφάλματος των Fan και Li [54] λειτουργεί πολύ αποδοτικά όταν εφαρμόζουμε τη frailErrest.

Παρατήρηση 5.3 *Στο προηγούμενο Κεφάλαιο 4, αναφέραμε σε αντίστοιχη παρατήρηση ότι σύμφωνα με τον Galantai [77], στην πράξη η σταθερά c είναι πιθανόν να λαμβάνει αρκετά μεγάλες τιμές, ξεπερνώντας και τη συνήθη τιμή 10, κάτι που σίγουρα εξαρτάται από την τιμή της διάστασης d. Οπότε εκτελέσαμε επιπλέον πειράματα προσομοίωσης, ώστε να διερευνηθούν οι πιθανές τιμές του άνω φράγματος $C_2(F'(\boldsymbol{\beta}))$. Όσον αφορά το MRME και την αναγνωρισιμότητα του μοντέλου, λάβαμε παρόμοια αποτελέσματα με τα παραπάνω. Οι μέσες τιμές για το $C_2(F'(\boldsymbol{\beta}))$ προέκυψαν τελικά μικρότερες του 10, για πλήθος μεταβλητών έως και 20 και για $n = 100$ και $J = 5$, σε όλες τις ποινικοποιημένες μεθόδους και για τις δύο εξεταζόμενες κατανομές ευπάθειας. Ταυτόχρονα, το ελάχιστο άνω φράγμα που προέκυψε από την (5.9) ήταν αρκετά χαμηλό, λαμβάνοντας για παράδειγμα την τιμή 0.55 για $d = 20$ και στις τρεις ποινικοποιημένες μεθόδους, ενώ η μέση μεροληψία του $\hat{\boldsymbol{\beta}}$ ήταν περίπου ίση με 0.32.*

Παρατήρηση 5.4 *Για να ελέγξουμε αν και κατά πόσο η επιλογή της σταθεράς c, η οποία λαμβάνει τιμές μεταξύ 1 και $C_2(F'(\boldsymbol{\beta}))$, είναι κρίσιμη για την απόδοση της νέας μεθόδου, θεωρήσαμε διαφορετικές τιμές αυτής. Για παράδειγμα, εξετάσαμε τις περιπτώσεις όπου $c = 1$ και $c = (1+C_2(F'(\boldsymbol{\beta})))/2$. Δεν παρατηρήσαμε κάποια σημαντική βελτίωση στα αποτελέσματά μας. Εξετάσαμε επιπλέον και διαφορετικούς τρόπους επιλογής της ρυθμιστικής παραμέτρου λ στο βήμα 4 της προτεινόμενης διαδικασίας, επιλέγοντας για παράδειγμα την παράμετρο λ που μεγιστοποιεί την (5.9) με τη σταθερά c να λαμβάνει την τιμή του άνω φράγματός της. Όπως αναμενόταν, μια τέτοια ενέργεια οδήγησε σε πολύ υψηλές τιμές του MRME. Συνεπώς, επιμείναμε στην τελική μας επιλογή, να επιλέξουμε δηλαδή το λ που ελαχιστοποιεί την (5.9), θέτοντας ως $c = C_2(F'(\boldsymbol{\beta}))$.*

Παρατήρηση 5.5 *Εξετάσαμε επίσης την περίπτωση όπου υπάρχει μεγαλύτερη συσχέτιση μεταξύ των συμμεταβλητών, καθώς η ανεξαρτησία αυτών αποτελεί σημαντική υπόθεση για την εφαρμογή της ανισότητας Kantorovich, ώστε να αξιολογήσουμε την όποια επίδραση θα είχε κάτι τέτοιο στα αποτελέσματά μας. Συγκεκριμένα, εκτελέσαμε επιπλέον πειράματα προσομοίωσης για την περίπτωση όπου η συσχέτιση μεταξύ των x_i και x_j είναι $\rho^{|i-j|}$ με $\rho = 0.8$. Η νέα μέθοδος, έδωσε ξανά καλύτερα αποτελέσματα σε σύγκριση με την GCV, όπως άλλωστε φαίνεται και από τους παρακάτω Πίνακες 5.7 και 5.8.*

Πίνακας 5.7: Αποτελέσματα προσομοιώσεων για το μοντέλο ευπάθειας Γάμμα κατανομής, για $\rho = 0.8$

Μέθοδος	**GCV: MRME**	**GCV: Aver. no. of 0 coeff.**		**frailErrest: MRME**	**frailErrest: Aver. no. of 0 coeff.**	
		correct	incorrect		correct	incorrect
n=50, J=2						
SCAD	0.8700	4.3700	0.2500	0.7655	4.5300	0.1300
LASSO	0.9340	3.6500	0.1400	0.8462	3.8500	0.1300
Hard	0.8285	4.2100	0.3100	0.7714	4.4700	0.1100
Oracle	0.6711	5	0	0.6711	5	0
n=50, J=5						
SCAD	0.7384	4.1000	0.0100	0.6831	4.7500	0.0100
LASSO	0.9749	3.7200	0.0100	0.8276	3.8600	0.0100
Hard	0.8380	4.4900	0.0400	0.7304	4.6400	0.0200
Oracle	0.5916	5	0	0.5916	5	0
n=100, J=2						
SCAD	0.8463	4.1300	0.0400	0.7549	4.4300	0.0400
LASSO	1.1062	3.1200	0.0500	0.8513	3.4000	0.0400
Hard	0.8173	4.3400	0.1600	0.7674	4.5900	0.0800
Oracle	0.6508	5	0	0.6508	5	0
n=100, J=5						
SCAD	0.7029	4.5300	0.0400	0.5828	4.7800	0
LASSO	0.8499	3.4200	0.0500	0.6915	3.7500	0
Hard	0.8374	4.1400	0.1300	0.6935	4.5700	0.0100
Oracle	0.6116	5	0	0.6116	5	0

Πίνακας 5.8: Αποτελέσματα προσομοιώσεων για το μοντέλο ευπάθειας Αντίστροφης Γκαουσιανής κατανομής, για $\rho = 0.8$

Μέθοδος	**GCV: MRME**	**GCV: Aver. no. of 0 coeff.**		**frailErrest: MRME**	**frailErrest: Aver. no. of 0 coeff.**	
		correct	incorrect		correct	incorrect
n=50, J=2						
SCAD	0.6581	4.4700	0.1200	0.5302	4.6100	0.0800
LASSO	0.7638	3.5800	0.1400	0.6810	3.9800	0.1000
Hard	0.6680	4.4600	0.3200	0.5053	4.6800	0.1400
Oracle	0.4755	5	0	0.4755	5	0
n=50, J=5						
SCAD	0.7924	4.3400	0	0.6940	4.5300	0
LASSO	1.0976	3.5100	0	0.9492	3.6200	0
Hard	0.8264	4.4600	0.0200	0.7617	4.7000	0.0100
Oracle	0.5323	5	0	0.5323	5	0
n=100, J=2						
SCAD	0.9080	4.4100	0.0800	0.8661	4.7000	0.0100
LASSO	1.2279	3.1000	0.0100	1.0859	3.4400	0.0100
Hard	0.9804	4.3200	0.1900	0.8720	4.5500	0.0800
Oracle	0.8187	5	0	0.8187	5	0
n=100, J=5						
SCAD	1.0553	4.2500	0	0.8860	4.6100	0
LASSO	1.7017	3.7400	0	1.0932	3.8200	0
Hard	0.8998	4.2800	0.0500	0.8370	4.5500	0
Oracle	0.9254	5	0	0.9254	5	0

5.5 Συμπεράσματα

Σε αυτό το κεφάλαιο, προτείναμε νέες εκτιμήσεις της νόρμας του σφάλματος στα μοντέλα ευπάθειας με ομαδοποιημένα δεδομένα, επεκτείνοντας έτσι τη μεθοδολογία που αναπτύξαμε στα γενικευμένα γραμμικά μοντέλα του Κεφαλαίου 4. Μέσω αυτών, δημουργήσαμε μια νεά μέθοδο επιλογής της ρυθμιστικής παραμέτρου, την frailErrest, στην περίπτωση χρήσης ποινικοποιημένης πιθανοφάνειας. Από τα αποτελέσματα μιας εκτενούς μελέτης προσομοίωσης που διενεργήσαμε, τα οποία ήταν ιδιαίτερα ενθαρρυντικά, συμπεραίνουμε ότι η frailErrest υπερέχει της συχνά χρησιμοποιούμενης γενικευμένης διασταυρωμένης επικύρωσης. Επίσης, η ευκολία εφαρμογής της νέας μεθόδου, την καθιστά αρκετά πιο εύχρηστη, μειώνοντας συγχρόνως και τον υπολογιστικό χρόνο που απαιτείται.

Μέρος **III**

Μέθοδοι Ανάλυσης Παραγοντικών Σχεδιασμών

ΚΕΦΑΛΑΙΟ **6**

Σχεδιασμοί Πειραμάτων: Βασικές Έννοιες και Ορισμοί

By a small sample,
we may judge of the whole piece.

—Miguel de Cervantes (1547–1616)

Στο κεφάλαιο αυτό, παραθέτουμε κάποια από τα βασικά είδη των πειραματικών σχεδιασμών, καθώς και τους σχετικούς ορισμούς και τις βασικές ιδιότητές τους, όπως αυτά αναφέρονται στα βιβλία [99] και [100]. Θα δώσουμε έμφαση στους παραγοντικούς και κλασματικούς παραγοντικούς σχεδιασμούς, στους μη επαναλαμβανόμενους παραγοντικούς σχεδιασμούς, στους υπερκορεσμένους και στους ομοιόμορφους σχεδιασμούς.

6.1 Παραγοντικοί Σχεδιασμοί

Ο στατιστικός κλάδος του σχεδιασμού πειραμάτων (designs of experiments) προσφέρει μια αρκετά μεγάλη ποικιλία διαδικασιών, με τελικό στόχο τον έλεγχο της ποιότητας των προϊόντων που διατίθενται στην αγορά, καθώς και την εξαγωγή στατιστικών συμπερασμάτων στον τομέα της Υγείας, της Γεωργίας και της Βιομηχανίας. Με τη χρήση των διαδικασιών αυτών, ο ερευνητής μελετά και εντοπίζει τους στατιστικά σημαντικούς παράγοντες (factors) με επίδραση στην απόκριση (response).

Ορισμός 6.1 *Η επίδραση ενός παράγοντα ορίζεται να είναι η αλλαγή που γίνεται στην απόκριση, από την αλλαγή στο επίπεδο του παράγοντα. Αυτή συχνά ονομάζεται κύρια επίδραση (main effect), επειδή αναφέρεται στους παράγοντες που είναι πρωταρχικής σημασίας στο πείραμα.*

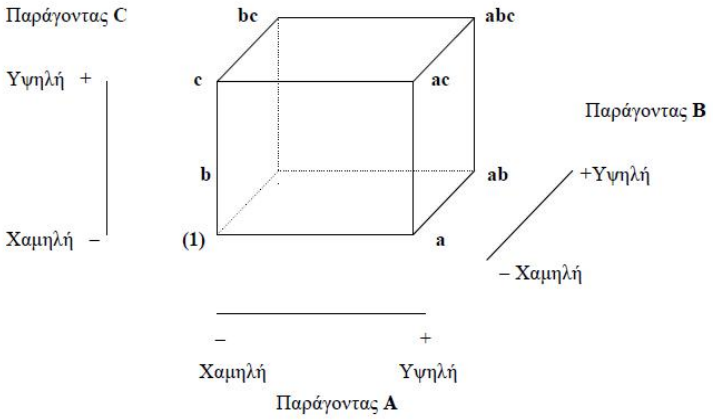
Αρχικά, είναι απαραίτητο να οριστεί με σαφήνεια και πληρότητα το πρόβλημα που χρήζει λύσης. Για την αντιμετώπισή του ορίζεται στη συνέχεια κάποια υπόθεση. Η υπόθεση αυτή πρέπει να ελεγχθεί για την ισχύ της και για τον έλεγχο αυτό σχεδιάζεται ένα κατάλληλο πείραμα. Ο στόχος λοιπόν του πειράματος είναι είτε να επιβεβαιώσει την ως άνω υπόθεση, είτε να απορρίψει την πιθανότητα ορθότητάς της. Μια κατηγορία σχεδιασμών, ιδιαίτερα αποδοτικών για πειράματα αυτού του τύπου, είναι οι παραγοντικοί σχεδιασμοί. Ιστορικά, η πρώτη αναφορά στην έννοια και τη χρησιμότητα του παραγοντικού σχεδιασμού έγινε το 1926 από τον Άγγλο στατιστικό, βιολόγο και γενετιστή R. A. Ficher, ο οποίος μελέτησε προβλήματα, κυρίως του κλάδου της Βιολογίας, αλλά και της Γεωργίας [108]. Έκτοτε, άρχισαν σταδιακά να χρησιμοποιούνται οι παραγοντικοί σχεδιασμοί, αρχικά σε γεωργικά πειράματα και στη συνέχεια σε εφαρμογές της Βιομηχανίας, καθώς και της Ιατρικής.

Ορισμός 6.2 *Με τον όρο παραγοντικό σχεδιασμό (factorial design), εννοούμε ότι σε κάθε πλήρη δοκιμή ή επανάληψη του πειράματος, εξετάζονται όλοι οι δυνατοί συνδυασμοί των επιπέδων (ή σταθμών) των παραγόντων.*

Να τονίσουμε ότι σε μερικά πειράματα μπορεί να βρούμε ότι η διαφορά στην απόκριση μεταξύ των επιπέδων ενός παράγοντα δεν είναι η ίδια σε όλα τα επίπεδα των άλλων παραγόντων. Σε τέτοιες περιπτώσεις, αυτό σημαίνει ότι υπάρχει μια αλληλεπίδραση (interaction) μεταξύ των παραγόντων. Υπάρχουν επίσης αρκετές ειδικές περιπτώσεις του γενικού παραγοντικού σχεδιασμού που είναι πολύ σημαντικές, καθώς χρησιμοποιούνται ευρύτατα σε ερευνητικές εργασίες και επειδή αποτελούν τη βάση άλλων σχεδιασμών με μεγάλη πρακτική αξία. Η πιο σημαντική από αυτές, είναι όταν έχουμε k παράγοντες, καθένας σε δύο μόνο στάθμες. Αυτές οι στάθμες μπορεί να είναι ποσοτικές ή ποιοτικές ή ακόμα και η παρουσία και απουσία ενός παράγοντα. Οι στάθμες αυτές μπορούν αυθαίρετα να ονομασθούν 'χαμηλή' και 'υψηλή' και κωδικοποιούνται ως -1 και +1 αντίστοιχα.

Ορισμός 6.3 *Μία πλήρης επανάληψη ενός σχεδιασμού με k παράγοντες, καθένας σε δύο στάθμες, απαιτεί 2^k παρατηρήσεις και ονομάζεται 2^k παραγοντικός σχεδιασμός.*

Αν για παράδειγμα, έχουμε τρεις παράγοντες A, B και C, τότε ο σχεδιασμός ονομάζεται 2^3 παραγοντικός σχεδιασμός και οι οκτώ συνδυασμοί αγωγών μπορούν να παρουσιασθούν γραφικά ως ένας κύβος, όπως φαίνεται στο παρακάτω Σχήμα 6.1. Οι ακμές του κύβου δείχνουν τη μετακίνηση των παραγόντων από τη χαμηλή στην υψηλή στάθμη και αντίστροφα, ενώ οι κορυφές αντιστοιχούν στους 2^3 διαφορετικούς συνδυασμούς των επιπέδων των παραγόντων.



Σχήμα 6.1: Ο 2^3 παραγοντικός σχεδιασμός

6.2 Κλασματικοί Παραγοντικοί Σχεδιασμοί

Όπως είδαμε στην προηγούμενη ενότητα, ο 2^k παραγοντικός σχεδιασμός απαιτεί 2^k εκτελέσεις. Κατ' αντιστοιχία ο 3^k σχεδιασμός, δηλαδή ο σχεδιασμός όπου οι k παράγοντες δέχονται τιμές σε τρεις στάθμες, απαιτεί ακριβώς 3^k εκτελέσεις. Επομένως για μεγάλο αριθμό παραγόντων, ο αριθμός των εκτελέσεων που απαιτείται είναι τεράστιος. Γίνεται λοιπόν εμφανής η επιτακτική ανάγκη για περιορισμό τους σε προβλήματα με μεγάλο αριθμό παραγόντων.

Η αρχική ιδέα για τον περιορισμό των εκτελέσεων ενός πειράματος, ήταν ο πειραματιστής να επιλέγει με προσοχή τους συνδυασμούς παραγόντων των οποίων οι αλληλεπιδράσεις τον ενδιαφέρουν περισσότερο και να αγνοεί ορισμένες αλληλεπιδράσεις υψηλής τάξης τις οποίες θεωρεί αμελητέες. Με τον τρόπο αυτό εκτελείται μόνο ένα κλάσμα του πλήρους παραγοντικού σχεδιασμού.

Ορισμός 6.4 *Με τον όρο κλασματικό παραγοντικό σχεδιασμό (fractional factorial design), εννοούμε ότι σε κάθε πλήρη δοκιμή ή επανάληψη του πειράματος, περιλαμβάνεται ένα υποσύνολο των δυνατών συνδυασμών των επιπέδων (ή σταθμών) των παραγόντων.*

Για παράδειγμα, στην περίπτωση όπου ο κλασματικός παραγοντικός σχεδιασμός είναι δύο επιπέδων, τότε συμβολίζεται ως 2^{k-p}, με $p < k$, όπου λαμβάνονται υπόψη k παράγοντες και ο αριθμός των εκτελέσεων είναι 2^{k-p}.

Ορισμός 6.5 *Ένας σχεδιασμός είναι αναλυτικής τάξης (resolution) R, αν καμία επίδραση p παραγόντων δεν είναι ταυτόσημη με άλλη επίδραση που περιέχει λιγότερους από $R - p$ παράγοντες.*

Ορισμός 6.6 *Ορίζουσα σχέση (defining relation) ενός κλασματικού παραγοντικού σχεδιασμού είναι το σύνολο όλων των στηλών που είναι ίσες με τη μοναδιαία στήλη I.*

Η αναλυτική τάξη ενός κλασματικού παραγοντικού σχεδιασμού με δύο επίπεδα είναι ίση με το μικρότερο αριθμό γραμμάτων σε οποιαδήποτε λέξη στην ορίζουσα σχέση. Συνήθως, ο στόχος είναι η χρήση κλασματικών παραγοντικών σχεδιασμών που έχουν την υψηλότερη δυνατή αναλυτική τάξη. Η υψηλή αναλυτική τάξη, βάζει λιγότερους περιορισμούς στις υποθέσεις που απαιτούνται όσον αφορά ποιες αλληλεπιδράσεις είναι αμελητέες ούτως ώστε να έχουμε μια αξιόπιστη ερμηνεία των δεδομένων.

Για την επιτυχή χρήση των κλασματικών παραγοντικών σχεδιασμών, βασιζόμαστε στις παρακάτω βασικές ιδέες - κλειδιά.

- Αρχή της σποραδικότητας των επιδράσεων [19] (sparsity of effects principle). Όταν υπάρχουν αρκετοί παράγοντες, το σύστημα ή η διαδικασία είναι πιθανό να οδηγείται αρχικά από μερικές από τις κύριες επιδράσεις και τις αλληλεπιδράσεις χαμηλής τάξης.

- Προβολική ιδιότητα (projective property). Οι κλασματικοί παραγοντικοί σχεδιασμοί μπορούν να προβάλλονται σε ισχυρότερους (μεγαλύτερους) σχεδιασμούς με αντικείμενο τους σημαντικούς παράγοντες.

- Ακολουθιακός πειραματισμός (sequential experimentation). Είναι δυνατόν να συνδυάσουμε τις εκτελέσεις δύο (ή περισσότερων) κλασματικών παραγοντικών σχεδιασμών ούτως ώστε να συγκεντρώσουμε ακολουθιακά ένα μεγαλύτερο σχεδιασμό και να εκτιμηθούν οι επιδράσεις και αλληλεπιδράσεις των παραγόντων που μας ενδιαφέρουν.

Να αναφέρουμε στο σημείο αυτό, ότι κυριότερη χρήση των κλασματικών παραγοντικών σχεδιασμών είναι σε πειράματα κρησαρίσματος (screening experiments). Αφορούν πειράματα στα οποία θεωρούμε ότι μετέχει μεγάλο πλήθος παραγόντων με σκοπό την αναγνώριση εκείνων (αν υπάρχουν) που έχουν μεγάλες επιδράσεις. Τέτοιου είδους πειράματα εκτελούνται συνήθως στα αρχικά στάδια μιας έρευνας, όταν είναι πιθανό ότι πολλοί από τους αρχικά θεωρούμενους παράγοντες, έχουν μικρή ή μηδενική επίδραση στην απόκριση. Οι παράγοντες που αναγνωρίζονται ως σημαντικοί, ερευνώνται τότε περισσότερο λεπτομερειακά σε επακόλουθα πειράματα. Οι σχεδιασμοί κρησαρίσματος γενικά απαιτούν λίγες πειραματικές εκτελέσεις, συνεπώς έχουν πολύ μικρό κόστος. Άρα είναι αρκετά ελκυστικοί επειδή παρέχουν ένα φθηνό και αποτελεσματικό τρόπο για να ξεκινήσει μια διαδικασία ανάλυσης ενός πειράματος και ως εκ τούτου, αποτελούν τη βάση για περαιτέρω αναλύσεις.

6.3 Μη Επαναλαμβανόμενοι Παραγοντικοί Σχεδιασμοί

Πολλές φορές, οι διαθέσιμοι πόροι για την εκτέλεση ενός πειράματος κρησαρίσματος, είναι περιορισμένοι. Γεγονός που έχει ως αποτέλεσμα, να μην επιτρέπεται στον πειραματιστή να εκτελέσει πάνω από μια επανάληψη του πλήρους πειραματικού σχεδιασμού. Αυτό συνεπάγεται κορεσμένο (saturated) σχεδιασμό, καθώς το πλήθος των υπό μελέτη παραγόντων k είναι ίσο με $n-1$, όπου n το πλήθος των εκτελέσεων.

Ορισμός 6.7 *Η μία επανάληψη ενός 2^k παραγοντικού σχεδιασμού, καλείται μη επαναλαμβανόμενος παραγοντικός σχεδιασμός (unreplicated factorial design).*

Αυτοί οι σχεδιασμοί είναι συνήθως κορεσμένοι όπως προαναφέραμε, συνεπώς ο πειραματιστής μπορεί να εκτιμήσει όλες τις κύριες επιδράσεις και αλληλεπιδράσεις, χωρίς όμως να μένουν βαθμοί ελευθερίας για την εκτίμηση του πειραματικού σφάλματος, οπότε η συμβατική τεχνική της ανάλυσης διασποράς για την αναγνώριση των σημαντικών παραγόντων, δε μπορεί να εφαρμοσθεί. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος, θα μπορούσε να στηριχθεί στην υπόθεση ότι μερικές υψηλής τάξης αλληλεπιδράσεις είναι αμελητέες, οπότε και να συνδυάσουμε τα μέσα τετράγωνα αυτών για να εκτιμήσουμε το σφάλμα. Όταν όμως αναλύουμε δεδομένα από μη επαναλαμβανόμενους παραγοντικούς σχεδιασμούς, συχνά εμφανίζονται αλληλεπιδράσεις υψηλής τάξης που είναι πραγματικά ενεργές. Η χρήση του μέσου τετραγώνου για το σφάλμα που παίρνουμε συνενώνοντας αυτές τις αλληλεπιδράσεις, δεν είναι κατάλληλη για αυτές τις περιπτώσεις. Παρ΄ όλα αυτά, έχουν προταθεί αρκετές αποδοτικές μέθοδοι ανάλυσης στη βιβλιογραφία, ορισμένες εκ των οποίων αναφέρουμε στο Κεφάλαιο 8.

6.4 Υπερκορεσμένοι Σχεδιασμοί

Οι υπερκορεσμένοι σχεδιασμοί αποτελούν μια κλάση των κλασματικών παραγοντικών σχεδιασμών και χρησιμοποιούνται κυρίως σε πειράματα κρησαρίσματος, με στόχο των προσ-

διορισμό των ενεργών παραγόντων. Συνεπώς, βασίζονται στην αρχή της σποραδικότητας των επιδράσεων. Αποτελούν ιδιαίτερα σημαντική κατηγορία σχεδιασμών, για αυτό άλλωστε η ανάπτυξη τους είναι ραγδαία σήμερα. Μπορούν να εξοικονομήσουν σημαντικό κόστος, όταν ο αριθμός των παραγόντων είναι μεγάλος και ένας μικρός αριθμός εκτελέσεων είναι επιθυμητός ή διαθέσιμος. Διαχωρίζονται σε τρεις κατηγορίες, στους υπερκορεσμένους σχεδιασμούς δύο επιπέδων (two-level), πολλών επιπέδων (multi-level) και μικτών επιπέδων (mixed-level). Στην παρούσα διδακτορική διατριβή θα μας απασχολήσουν οι υπερκορεσμένοι σχεδιασμοί δύο επιπέδων.

Ορισμός 6.8 *Υπερκορεσμένοι σχεδιασμοί (supersaturated designs) δύο επιπέδων, λέγονται οι κλασματικοί παραγοντικοί σχεδιασμοί των οποίων ο αριθμός των παραγόντων k είναι μεγαλύτερος ή ίσος του αριθμού των πειραματικών εκτελέσεων n, δηλαδή $k \geq n$.*

Η κατασκευή των υπερκορεσμένων σχεδιασμών δεν είναι εύκολη και απαιτούνται αρκετά μαθηματικά και στατιστικά εργαλεία και γνώσεις. Ο πρώτος που εισήγαγε την ιδέα των υπερκορεσμένων σχεδιασμών, ως τυχαίους ισορροπημένους σχεδιασμούς, ήταν ο Satterthwaite [150]. Έκτοτε, αρκετοί ερευνητές έχουν ασχοληθεί με την κατασκευή και την εξέταση των ιδιοτήτων τους. Να τονίσουμε ότι το πρόβλημα της κατασκευής των υπερκορεσμένων σχεδιασμών θεωρείται ότι είναι ένα πρόβλημα δυσκολίας NP. Αυτό σημαίνει ότι δεν υπάρχει αλγόριθμος που όταν του δίνουμε τον αριθμό των παραγόντων και τον αριθμό των πειραματικών εκτελέσεων, να αποκρίνεται σε πολυωνυμικό χρόνο αν υπάρχει υπερκορεσμένος σχεδιασμός.

Αναφορικά τώρα με τη στατιστική ανάλυση των υπερκορεσμένων σχεδιασμών, μέχρι στιγμής δεν έχει βρεθεί κάποια μέθοδος η οποία όταν χρησιμοποιείται να αναγνωρίζονται οι σημαντικοί παράγοντες και να εκτιμώνται με αμελητέο σφάλμα. Το πρόβλημα έγκειται καταρχήν στο ότι ο αριθμός των παραγόντων είναι μεγαλύτερος από τον αριθμό των εκτελέσεων, συνεπώς δε μπορεί να χρησιμοποιηθεί η μέθοδος των ελαχίστων τετραγώνων. Επιπλέον, η μη ορθογωνιότητα του πίνακα σχεδιασμού, συνεπάγεται μια μικρή μεροληψία ανάμεσα σε όλες τις εκτιμημένες επιδράσεις, εφόσον οι ενεργοί παράγοντες μπορούν ακόμη να αναγνωριστούν. Ένας ενεργός παράγοντας μπορεί να αναγνωριστεί αν η επίδραση του είναι αρκετά μεγάλη ώστε να μην επισκιάζεται από το πειραματικό σφάλμα και τη συνδυασμένη επίδραση μη σημαντικών παραγόντων. Θα δούμε στο επόμενο κεφάλαιο ορισμένες από τις εισηγήσεις και τους τρόπους που προτείνονται για την κατασκευή και ανάλυση των υπερκορεσμένων σχεδιασμών.

6.5 Ομοιόμορφοι Σχεδιασμοί

Ας θεωρήσουμε την περίπτωση όπου ο πειραματιστής θέλει να εξετάσει πολλούς παράγοντες, κάθε ένας με μεγάλο αριθμό επιπέδων. Είναι προφανές ότι οι συνδυασμοί των επιπέδων θα είναι επίσης πολλοί. Επομένως, χωρίς τη χρήση μοντέρων πειραματικών σχεδιασμών, δε μπορεί να αποκτηθεί ένα σημαντικό ποσό πληροφορίας που να εκφράζει τη σχέση μεταξύ της μεταβλητής απόκρισης και των συμβαλλόμενων παραγόντων. Έστω ότι υπάρχουν s παράγοντες, με q επίπεδα ο καθένας. Αυτό συνεπάγεται q^s συνδυασμούς επιπέδων και ταυτόχρονα υψηλό πειραματικό κόστος και κατανάλωση χρόνου. Ως εκ τούτου, συχνά χρησιμοποιούνται οι παραγοντικοί σχεδιασμοί δύο και τριών επιπέδων, ή οι κλασματικοί παραγοντικοί σχεδιασμοί δεδομένου ότι οι αλληλεπιδράσεις υψηλής τάξης μπορούν να αγνοηθούν. Μεταξύ της τελευταίας κλάσης σχεδιασμών, οι ορθογώνιοι σχηματισμοί [88] είναι πιθανόν οι πιο δημοφιλείς και αποτελεσματικοί.

Οι περισσότεροι πειραματικοί σχεδιασμοί, υποθέτουν ότι το υπό μελέτη μοντέλο είναι γνωστό με κάποιες άγνωστες παραμέτρους, όπως είναι οι κύριες επιδράσεις και οι αλληλεπιδράσεις, οπότε γίνεται η επιλογή του σχεδιασμού εκείνου που δίνει την καλύτερη-αποδοτικότερη εκτίμηση των παραμέτρων. Ωστόσο, ο ερευνητής σε κάποια πειράματα δε γνωρίζει το μοντέλο που υποβόσκει (underlying model). Ένας σχεδιασμός γεμίσματος του χώρου (space

filling design) είναι η καλύτερη επιλογή σε αυτή την περίπτωση. Μια κατηγορία σχεδιασμών γεμίσματος του χώρου, οι οποίοι αναζητούν πειραματικά σημεία που να είναι ομοιόμορφα διασκορπισμένα στο πεδίο ορισμού είναι η κλάση των ομοιόμορφων σχεδιασμών. Προτάθηκαν από τους Fang [61], Wang και Fang [171] και έχουν ευρέως χρησιμοποιηθεί από το 1980.

Έστω ότι έχουμε s παράγοντες που μας ενδιαφέρουν, σε έναν πειραματικό χώρο C^s. Ο στόχος των ομοιόμορφων σχεδιασμών είναι η επιλογή ενός συνόλου n σημείων, $P_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset C^s$, ούτως ώστε αυτά τα σημεία να είναι ομοιόμορφα διασκορπισμένα στο C^s, βάσει ενός μέτρου ασυμφωνίας (discrepancy measure) D. Το μέτρο αυτό, καθορίζει πόσο απέχει η εμπειρική κατανομή των σημείων του σχεδιασμού, από την Ομοιόμορφη κατανομή [89]. Το κατάλληλο σύνολο θα πρέπει να ελαχιστοποιεί το D, κάτι που, ισοδύναμα, σημαίνει ότι θα μεγιστοποιεί την ομοιομορφία πάνω σε όλα τα πιθανά n σημεία του C^s. Με λίγα λόγια, για δοσμένο μέτρο ασυμφωνίας D, ο ομοιόμορφος σχεδιασμός συμβολιζόμενος ως $U_n(q^s)$, θα έχει τη μικρότερη D τιμή μεταξύ όλων των σχεδιασμών με n εκτελέσεις και s παράγοντες q επιπέδων.

Η αναζήτηση και εύρεση ενός ομοιόμορφου σχεδιασμού $U_n(q^s)$ θεωρείται ως ένα NP δύσκολο πρόβλημα βελτιστοποίησης, ειδικά όταν τα (n, q, s) αυξάνουν. Συνεπώς, απαιτείται μια λογική δομή για αυτόν. Οι λεγόμενοι U-type σχεδιασμοί δίνουν μία τέτοια δομή.

Ορισμός 6.9 *Ένας U-type σχεδιασμός είναι ένας $n \times s$ πίνακας με στοιχεία $\{1, \ldots, q\}$, τέτοια ώστε αυτά τα q στοιχεία να εμφανίζονται τον ίδιο αριθμό φορών σε κάθε στήλη.*

Οι σχεδιασμοί αυτοί περιορίζουν τον αριθμό των εκτελέσεων n να είναι πολλαπλάσιος του αριθμού q των επιπέδων των παραγόντων. Έστω τώρα $\mathcal{U}_n(q^s)$ το σύνολο όλων των U-type σχεδιασμών.

Ορισμός 6.10 *Ένας σχεδιασμός $\boldsymbol{U} \in \mathcal{U}_n(q^s)$ καλείται ομοιόμορφος σχεδιασμός, υπό του μέτρου ασυμφωνίας D, αν*

$$D(\boldsymbol{U}) = \min_{\boldsymbol{V} \in \mathcal{U}_n(q^s)} D(\boldsymbol{V}) \tag{6.1}$$

και συμβολίζεται ως $U_n(q^s)$.

Ο παρακάτω πίνακας παρουσιάζει ένα παράδειγμα ομοιόμορφου σχεδιασμού $U_{14}(14^7)$, με 14 εκτελέσεις και 7 παράγοντες με 14 επίπεδα.

Πίνακας 6.1: Ένας ομοιόμορφος σχεδιασμός $U_{14}(14^7)$

A	B	C	D	E	F	G
5	11	10	1	3	12	8
14	1	8	5	4	5	9
2	6	12	9	5	1	6
13	12	5	12	6	11	5
8	13	14	4	8	4	2
4	14	2	10	10	6	10
9	5	1	8	2	9	1
11	7	3	2	12	2	7
12	10	11	7	14	8	14
1	8	7	6	11	14	3
6	2	9	14	13	7	4
7	9	6	13	1	3	12
3	3	4	3	7	10	13
10	4	13	11	9	13	11

Οι ομοιόμορφοι σχεδιασμοί έχουν δυο μεγάλα πλεονεκτήματα. Καταρχήν, παρέχουν μια καλή αναπαράσταση των πειραματικών σημείων με ένα λογικό αριθμό εκτελέσεων, γεγονός που είναι αρκετά χρήσιμο ειδικά όταν ο αριθμός των επιπέδων είναι μεγάλος. Επιπλέον, είναι

εύρωστοι ως προς την υπόθεση του μοντέλου που υποβόσκει. Λόγω αυτών των πλεονεκτημάτων, τα τελευταία 20 χρόνια οι σχεδιασμοί αυτοί έχουν εφαρμοστεί με επιτυχία σε διάφορους τομείς, όπως η Βιομηχανία, η Μηχανική, η Χημεία και η Φαρμακευτική. Ενδεικτικά αναφέρουμε τις εργασίες των Liang et al. [119] και των Fang και Lin [66].

ΚΕΦΑΛΑΙΟ 7

Μέθοδος Ανάλυσης Υπερκορεσμένων Σχεδιασμών με Τροποποίηση του **PageRank** Αλγορίθμου

First, solve the problem.
Then, write the code.

—*John Johnson (1977)*

Στο έβδομο αυτό κεφάλαιο, ασχολούμαστε με την ανάλυση των υπερκορεσμένων σχεδιασμών δύο επιπέδων. Στόχος μας είναι η δημιουργία μιας κατάλληλης μεθόδου επιλογής μεταβλητών στους υπερκορεσμένους σχεδιασμούς για τις περιπτώσεις όπου η απόκριση είναι κατανομής Bernoulli και Poisson. Η μέθοδος βασίζεται στη τροποποίηση ενός αρκετά γνωστού αλγορίθμου βαθμιδότησης της μηχανής αναζήτησης Google, του PageRank αλγορίθμου, ώστε να μπορούμε να τον χρησιμοποιήσουμε αποτελεσματικά ως μέθοδο κρησαρίσματος. Αυτό επιτυγχάνεται με το συνδυασμό του εν λόγω αλγορίθμου με κατάλληλα μέτρα από τη Θεωρία Πληροφορίας.

7.1 Ερευνητικό Πρόβλημα

Τα τελευταία χρόνια, το ενδιαφέρον των ερευνητών έχει στραφεί αρκετά σε μεθόδους κατασκευής και ανάλυσης των υπερκορεσμένων σχεδιασμών, κυρίως λόγω του μικρού πλήθους εκτελέσεων και του ταυτόχρονα μεγάλου αριθμού παραγόντων που διαθέτουν. Η ιδέα των υπερκορεσμένων σχεδιασμών, ως τυχαίων ισορροπημένων σχεδιασμών, εισήχθη από τον Satterthwaite [150]. Οι Booth και Cox [18] ήταν οι πρώτοι που τους εξέτασαν συστηματικά. Από το χρονικό σημείο αυτό και για επιπλέον 30 περίπου χρόνια, δε συναντάμε κάποια σχετική εργασία στη διεθνή βιβλιογραφία. Ωστόσο, το 1993, οι εργασίες των Lin [120] και Wu [178] πυροδότησαν ξανά το ενδιαφέρον των ερευνητών σε αυτόν τον τομέα. Συγκεκριμένα, ο Lin [120] πρότεινε μια νέα κλάση υπερκορεσμένων σχεδιασμών, βασιζόμενος σε κλάσματα των πινάκων Hadamard. Επίσης, ο Wu [178] επαύξησε τους πίνακες Hadamard, προσθέτοντας στήλες αλληλεπιδράσεων. Έκτοτε, δημοσιεύθηκε ένας μεγάλος αριθμός εργασιών, κυρίως για μεθόδους κατασκευής των υπερκορεσμένων σχεδιασμών. Μεταξύ άλλων, αναφέρουμε τις εργασίες των Lin [121], Nguyen [136], Cheng [31], Li και Wu [118], Tang και Wu [161], Fang et al. [67], Bulutoglu [25], Bulutoglu και Cheng [26], Bulutoglu και Ryan [27], Liu και Dean [123], Xu και Wu [181], Ryan και Bulutoglu [149], Koukouvinos και Mylona [102] και Koukouvinos et al. [101, 104–106].

Σε αντίθεση με την την ευρεία μελέτη των μεθόδων κατασκευής των υπερκορεσμένων σχεδιασμών, η ανάλυσή τους παραμένει ακόμα μια πρόκληση για τους επιστήμονες, παρά το γεγονός ότι έχουν αναπτυχθεί πολλές ενδιαφέρουσες μέθοδοι, τις πιο σημαντικές από τις οποίες παρουσιάζουμε στη συνέχεια. Ένα κοινό χαρακτηριστικό τους είναι ότι στηρίζονται στο γενικό γραμμικό μοντέλο, με τις γνωστές υποθέσεις περί Κανονικής κατανομής της μεταβλητής απόκρισης, ασυσχέτιστων σφαλμάτων και ομοιογένειας της διασποράς. Ξεκινάμε με τον Lin [120] ο οποίος εφάρμοσε τη μέθοδο επιλογής μεταβλητών κατά βήματα, για την επιλογή των σημαντικών παραγόντων, χρησιμοποιώντας 1/2 κλάσματα των Plackett-Burman σχεδιασμών, ενώ ο Wang [170] εφάρμοσε την ίδια ανάλυση στα υπόλοιπα 1/2 κλάσματα. Οι Chipman et al. [35] πρότειναν μια Μπεϋζιανή μέθοδο επιλογής μεταβλητών για την ανάλυση υπερκορεσμένων σχεδιασμών με πολύπλοκη δομή. Επίσης, οι Westfall et al. [174] ανέπτυξαν μια τεχνική ελέγχου του σφάλματος Τύπου Ι στη διαδικασία της προς τα εμπρός επιλογής μεταβλητών. Οι Abraham et al. [3] εφάρμοσαν την κατά βήματα μέθοδο επιλογής μεταβλητών και τη μέθοδο όλων των μοντέλων για να διερευνήσουν τους σημαντικούς παράγοντες. Οι Li και Lin [116] εισήγαγαν μια νέα προσέγγιση του προβλήματος ανάλυσης των υπερκορεσμένων σχεδιασμών, βασισμένη στα ποινικοποιημένα ελάχιστα τετράγωνα. Οι Holcomb et al. [91] πρότειναν μια μέθοδο βασισμένη στις αντιθέσεις, ενώ οι Lu και Wu [124] ανέπτυξαν μια τροποποιημένη κατά βήματα διαδικασία επιλογής μεταβλητών, βασισμένη στη σταδιακή μείωση των διαστάσεων του προβλήματος. Οι Zhang et al. [185] πρότειναν μια μέθοδο που στηρίζεται στα μερικώς ελάχιστα τετράγωνα. Οι Koukouvinos και Stylianou [107] ανέπτυξαν μια τροποποιημένη μέθοδο μεταβλητότητας των αντιθέσεων. Ο Georgiou [79] πρότεινε μια μέθοδο που βασίζεται στο συνδυασμό της ανάλυσης ιδιαζουσών τιμών, της ανάλυσης κυρίων συνιστωσών και της ανάλυσης παλινδρόμησης για την αναγνώριση των ενεργών επιδράσεων. Οι Koukouvinos και Mylona [103] ανέπτυξαν μια μέθοδο κρησαρίσματος κατά ομάδες για την ανάλυση των $E(f_{NOD})$-βέλτιστων υπερκορεσμένων σχεδιασμών μεικτών επιπέδων. Οι Phoa et al. [143] ανέπτυξαν μια μέθοδο επιλογής μεταβλητών με χρήση του Dantzig selector [29] και οι Marley και Woods [128] πρότειναν μια νέα επαναληπτική μέθοδο για την ανάλυση των σχεδιασμών αυτών. Πέραν όμως των παραπάνω, ο ενδιαφερόμενος αναγνώστης παραπέμπεται και στις εργασίες [80] και [82] για μια χρήσιμη ανασκόπηση μεθόδων ανάλυσης αλλά και κατασκευής των υπερκορεσμένων σχεδιασμών.

Ένα επίσης σημαντικό πεδίο έρευνας, μη ανεπτυγμένο στην έως σήμερα διεθνή βιβλιογραφία, αποτελεί η ανάλυση των υπερκορεσμένων σχεδιασμών, όταν όμως τα δεδομένα απόκρισης δεν προέρχονται από την Κανονική κατανομή. Οι παραδοσιακές μέθοδοι επιλογής

μεταβλητών, ακόμα και αυτές που βασίζονται σε μεθόδους συρρίκνωσης, όπως η μεθοδολογία ποινικοποιημένης πιθανοφάνειας των Fan και Li [53], δε μπορούν να εφαρμοστούν άμεσα, λόγω του μικρού αριθμού των διαθέσιμων πειραματικών εκτελέσεων. Το γεγονός αυτό μας ώθησε να αναπτύξουμε μια νέα μέθοδο κρησαρίσματος των σημαντικών μεταβλητών στους υπερκορεσμένους σχεδιασμούς, για διακριτά δεδομένα απόκρισης.

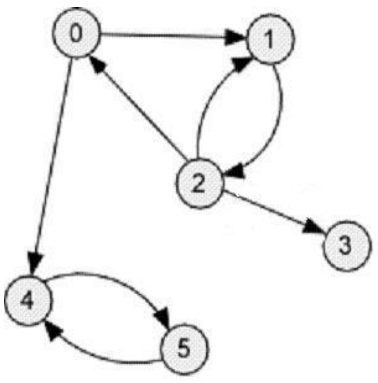
Η μέθοδος που θα προτείνουμε, βασίζεται στο συνδυασμό του γνωστού αλγορίθμου PageRank [110], με ένα κατάλληλο μέτρο πληροφορίας, τη δεσμευμένη αμοιβαία πληροφορία. Ο αλγόριθμος PageRank αποτελεί μια δημοφιλή διαδικασία βαθμιδότησης και κατάταξης των ιστοσελίδων, βάσει της σημαντικότητάς τους, της μηχανής αναζήτησης Google. Υπάρχουν βέβαια και άλλοι παρόμοιοι αλγόριθμοι βαθμιδότησης και κατάταξης, καθένας με τα πλεονεκτήματα και τα μειονεκτήματά του. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται σχετικά στο βιβλίο των Langville και Meyer [111]. Δεν έχουν όμως έως τώρα τροποποιηθεί ή χρησιμοποιηθεί κατάλληλα στους υπερκορεσμένους σχεδιασμούς. Κατά τη δικιά μας γνώση, είναι η πρώτη φορά που τροποποιείται ο συγκεκριμένος αλγόριθμος και χρησιμοποιείται για σκοπούς επιλογής μεταβλητών.

7.2 Ο Αλγόριθμος PageRank

Ο Παγκόσμιος Ιστός (World Wide Web) χαρακτηρίζεται από μια εξαιρετικά ευμετάβλητη δομή. Κάθε μέρα προστίθενται, διαγράφονται ή τροποποιούνται ιστοσελίδες. Ως αποτέλεσμα, ήταν προφανής η ανάγκη για την εύρεση ενός εργαλείου αναζήτησης που να μπορεί να διακρίνει τις ιστοσελίδες υψηλής ποιότητας από αυτές χαμηλής ποιότητας. Η ιδέα αυτή υλοποιήθηκε στον αλγόριθμο PageRank, ο οποίος χρησιμοποιείται στη μηχανή αναζήτησης Google, που προτάθηκε από τους Larry Page και Sergey Brin [24], [139]. Αποτελεί τον κυρίαρχο αλγόριθμο που βοήθησε να καθιερωθεί τόσο η τεχνική ανωτερότητα της Google όσο και η οικονομική επιτυχία της και εξακολουθεί να είναι το κλειδί για την παροχή ακριβών βαθμιδοτήσεων-κατατάξεων των ιστοσελίδων στα αποτελέσματα αναζήτησης.

Ο αλγόριθμος PageRank, είναι ένας αλγόριθμος βαθμιδότησης, βασιζόμενος σε γραφήματα, που χρησιμοποιείται στην κατάταξη των ιστοσελίδων με κριτήριο τη σημαντικότητα αυτών. Αν θεωρήσουμε ότι ο Ιστός είναι ένα μεγάλο γράφημα, τότε οι ιστοσελίδες μπορούν να μοντελοποιηθούν ως οι κορυφές του γραφήματος (nodes) και τα links τους (απευθείας συνδέσεις με άλλες ιστοσελίδες) ως οι κατευθυνόμενες ακμές (edges). Για να επεξηγήσουμε τα βασικά χαρακτηριστικά του αλγορίθμου, παραθέτουμε ένα απλό μεν, χαρακτηριστικό δε, παράδειγμα.

Παράδειγμα 7.1 *Έστω ένα μικρό απομονωμένο μέρος του Παγκόσμιου Ιστού με μόνο 6 URLs,* $P_0, \ldots, P_5$*, όπως φαίνεται στο παρακάτω Σχήμα 7.1.*



Σχήμα 7.1: Ένα μικρό μέρος του Παγκόσμιου Ιστού με 6 URLs

Ας παρατηρήσουμε την ακμή που συνδέει την κορυφή 1 με τη 2. Αυτό στη πράξη, σημαίνει ότι η ιστοσελίδα 1 οδηγεί στη 2. Εκεί βασίζεται η κεντρική ιδέα του PageRank αλγορίθμου: Ανάθεση ενός συνολικού βαθμού σημαντικότητας, που καλείται PageRank τιμή, σε κάθε ιστοσελίδα του Παγκόσμιου Ιστού, βάσει του αριθμού των links που κατευθύνονται προς την ιστοσελίδα αυτή καθώς και της σημαντικότητας των ιστοσελίδων από τις οποίες προέρχονται αυτά τα links. Ως αποτέλεσμα, προκύπτει ένα διάνυσμα αποτελούμενο από τους βαθμούς κάθε ιστοσελίδας. Προφανώς, η ιστοσελίδα με τον μεγαλύτερο βαθμό θα εμφανίζεται πρώτη στη μηχανή αναζήτησης Google. Αυτός ο αναδρομικός ορισμός της σημαντικότητας μπορεί να περιγραφεί μαθηματικά, από τη στάσιμη κατανομή ενός απλού τυχαίου περιπάτου πάνω στο γράφημα, όπου αρχικά ξεκινάμε από μια αυθαίρετη κορυφή και σε κάθε βήμα επιλέγουμε τυχαία μια εξερχόμενη ακμή από την κορυφή που βρισκόμαστε εκείνη τη στιγμή.

Έστω τώρα ένα διάνυσμα $\mathbf{r}_{(k)}$ που αποτελεί το PageRank διάνυσμα στην $k-$οστή επανάληψη. Τότε, η PageRank τιμή της ιστοσελίδας P_i στην $k+1$ επανάληψη θα είναι

$$\mathbf{r}_{(k+1)}(P_i) = \sum_{P_j \in B_{P_i}} \frac{\mathbf{r}_{(k)}(P_j)}{N_j}. \tag{7.1}$$

Η προαναφερθείσα έκφραση ήταν ο αρχικός επίσημος μαθηματικός τύπος του PageRank αλγορίθμου. Η διαδικασία ξεκινάει με αρχική τιμή $r_0(P_i) = 1/d$ για όλες τις ιστοσελίδες P_i, $i = 1, \ldots, d$ και επαναλαμβάνεται μέχρις ότου οι τιμές - βαθμοί να συγκλίνουν σε κάποιες τελικές σταθερές τιμές. Αυτές θα αποτελούν τα στοιχεία του PageRank διανύσματος. Σημειώνουμε ότι N_j είναι ο αριθμός των εξερχόμενων συνδέσεων (outlinks) από την ιστοσελίδα P_j και B_{P_i} είναι το σύνολο των ιστοσελίδων που οδηγούν στη P_i.

Εναλλακτικά, για να μπορέσουμε να υπολογίσουμε το PageRank διάνυσμα χρησιμοποιώντας πίνακες, πρέπει να ξεκινήσουμε με ένα μαθηματικό μοντέλο που να περιγράφει τη δομή των συνδέσεων στον Παγκόσμιο Ιστό. Επιστρέφοντας λοιπόν στο παράδειγμά μας, μπορούμε αρχικά να κατασκευάσουμε έναν πίνακα υπερσύνδεσης (hyperlink matrix) όπως λέγεται, έστω $\mathbf{M}$, για το προαναφερθέν γράφημα με $(j,i)-$στοιχείο

$$\mathbf{M}_{ji} = \begin{cases} 1/N_j & \text{αν υπάρχει link από την } P_j \text{ προς την } P_i, \\ 0 & \text{αλλιώς} \end{cases}$$

ώστε το PageRank πρόβλημά μας να μπορεί πλέον να θεωρηθεί ως πρόβλημα πινάκων. Συνεπώς, η δομή των συνδέσεων γράφεται τώρα ως

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Ο τετραγωνικός πίνακας $\mathbf{M}$ περιέχει τα ίδια μη μηδενικά στοιχεία όπως και ο πίνακας γειτνίασης του γραφήματος, με τη διαφορά όμως ότι τα στοιχεία αυτά είναι πλέον πιθανότητες μετάβασης. Η μαθηματική διατύπωση της αναδρομικά οριζόμενης PageRank διαδικασίας, γράφεται τώρα ως

$$\mathbf{r}^T_{(k+1)} = \mathbf{r}^T_{(k)}\mathbf{M}, k = 0, 1, \ldots \tag{7.2}$$

Σε αυτό το στάδιο, μια επαναληπτική μέθοδος μπορεί να χρησιμοποιηθεί για τον υπολογισμό του διανύσματος $\mathbf{r}$, ξεκινώντας με ένα αυθαίρετο διάνυσμα που θα περιγράφει τις αρχικές PageRank τιμές r_i για όλες τις ιστοσελίδες P_i. Η Power μέθοδος [109] είναι η συνήθης

επιλογή του αλγορίθμου για την εύρεση του διανύσματος $\mathbf{r}^T_{(k+1)}$.

Όμως, ο υπολογισμός του $\mathbf{r}^T_{(k+1)}$ δεν είναι τόσο άμεσος όσο φαίνεται. Η προαναφερθείσα μοντελοποίηση, υποθέτει καταρχήν ότι κάθε ιστοσελίδα-κορυφή του γραφήματος, έχει τουλάχιστον μια εξερχόμενη ακμή, εξερχόμενη σύνδεση δηλαδή προς άλλες ιστοσελίδες. Αυτό φυσικά δεν ισχύει στην πράξη ακόμα και για ολόκληρο τον Παγκόσμιο Ιστό, καθότι υπάρχουν ιστοσελίδες με out-degree μηδέν. Σημειώνουμε ότι out-degree μιας ιστοσελίδας είναι ο αριθμός των διακριτών εξερχόμενων συνδέσεών της. Ας παρατηρήσουμε για παράδειγμα την τέταρτη γραμμή του πίνακα $\mathbf{M}$. Η γραμμή αυτή αντιστοιχεί στην ιστοσελίδα P_3 και το γεγονός ότι έχει μηδενικά στοιχεία, σημαίνει πολύ απλά ότι δεν υπάρχουν εξερχόμενες συνδέσεις για την P_3. Συνεπώς θα είμαστε περιορισμένοι στην P_3 για πάντα. Για να επιλυθεί αυτό το πρόβλημα, όλες οι γραμμές με μόνο μηδενικά στοιχεία αντικαθιστώνται με την τιμή $1/d$, όπου d η διάσταση του πίνακα. Ως αποτέλεσμα, προκύπτει ο ακόλουθος στοχαστικός κατά γραμμή πίνακας $\mathbf{M}^*$.

$$\mathbf{M}^* = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Παρ΄ όλα αυτά, το γράφημα δε θεωρείται ακόμα ισχυρά συνεκτικό. Πιθανόν να υπάρχουν πολλές μικρότερες ή ακόμη και μερικές μεγαλύτερες ομάδες ιστοσελίδων που αποτελούν χωριστές ισχυρά συνδεδεμένες συνιστώσες, ίσως με συνδέσεις που εισέρχονται σε μια ομάδα, αλλά με καμία εξερχόμενη σύνδεση. Αυτές οι ομάδες μπορούν να λειτουργήσουν ως 'rank sinks', που 'παγιδεύουν' μεγάλες ποσότητες τιμών κατάταξης των ιστοσελίδων. Αυτό θα συμβεί για παράδειγμα, αν θεωρήσουμε το κατώτερο μέρος της δομής στο Σχήμα 7.1, γεγονός που αντιστοιχεί σε αναγώγιμο πίνακα. Για να ξεπεραστεί αυτό το πρόβλημα, ο PageRank αλγόριθμος χρησιμοποιεί ένα συντελεστή απόσβεσης α, $0 < \alpha < 1$, ο οποίος προστίθεται στον τυχαίο περίπατο ως εξής: Σε κάθε βήμα, επιλέγουμε μια εξερχόμενη σύνδεση με πιθανότητα α, και εκτελείται ένα άλμα σε έναν τυχαίο κόμβο στο γράφημα με πιθανότητα $1 - \alpha$. Ως αποτέλεσμα, θα έχουμε ένα νέο πίνακα $\tilde{\mathbf{M}}$, ο οποίος θα είναι θετικός, στοχαστικός και μη αναγώγιμος, χρησιμοποιώντας τον ακόλουθο μετασχηματισμό

$$\tilde{\mathbf{M}} = \alpha\mathbf{M}^* + (1-\alpha)\frac{1}{d}\mathbf{e}\mathbf{e}^T \tag{7.3}$$

όπου $\mathbf{M}^*$ είναι ο στοχαστικός κατά γραμμή πίνακας του πρηγούμενο βήματος και $\mathbf{e}^T$ ένα διάνυσμα γραμμή με μονάδες. Οπότε, ο υπολογισμός του PageRank διανύσματος είναι πλέον εφικτός. Να σημειώσουμε ότι η μηχανή αναζήτησης Google χρησιμοποιεί ένα συντελεστή απόσβεσης $\alpha = 0.85$ ως προεπιλογή. Επιστρέφοντας τώρα στο παράδειγμά μας, ο τελικός πίνακας $\tilde{\mathbf{M}}$ θα είναι

$$\tilde{\mathbf{M}} = \begin{pmatrix} 0.0250 & 0.4500 & 0.0250 & 0.0250 & 0.4500 & 0.0250 \\ 0.0250 & 0.0250 & 0.8750 & 0.0250 & 0.0250 & 0.0250 \\ 0.3083 & 0.3083 & 0.0250 & 0.3083 & 0.0250 & 0.0250 \\ 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 & 0.1667 \\ 0.0250 & 0.0250 & 0.0250 & 0.0250 & 0.0250 & 0.8750 \\ 0.0250 & 0.0250 & 0.0250 & 0.0250 & 0.8750 & 0.0250 \end{pmatrix}$$

Υπενθυμίζουμε σε αυτό το σημείο, ότι βάσει του Θεωρήματος Perron-Frobenius, ένας θετικός, στοχαστικός κατά γραμμή και μη αναγώγιμος πίνακας, είναι εγγυημένο ότι θα έχει ένα θετικό κυρίαρχο ιδιοδιάνυσμα και η αντίστοιχη και μεγαλύτερη ιδιοτιμή του θα είναι ίση με 1. Επιπλέον, η Power μέθοδος που αναφέραμε προηγουμένως, είναι μία από τις

παλαιότερες και πιο απλές επαναληπτικές διαδικασίες για την εύρεση της κυρίαρχης ιδιοτιμής και του ιδιοδιανύσματος ενός πίνακα. Σημειώνουμε επίσης ότι, ο υπολογισμός του διανύσματος PageRank συνίσταται ουσιαστικά στον επαναληπτικό υπολογισμό του ιδιοδιανύσματος που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή του αρχικού πίνακα $\mathbf{M}$ [110]. Ως εκ τούτου, στον αλγόριθμο PageRank με την Power μέθοδο, μετατρέπουμε πρώτα τον πίνακα $\mathbf{M}$ σε σε ένα στοχαστικό κατά γραμμή πίνακα $\mathbf{M}^*$. Στη συνέχεια, ο αλγόριθμος παράγει το νέο πίνακα $\tilde{\mathbf{M}}$, χρησιμοποιώντας την προαναφερθείσα διαδικασία και τέλος, υπολογίζει το διάνυσμα με τις PageRank τιμές των ιστοσελίδων. Επιστρέφοντας πάλι στο παράδειγμά μας, προκύπτει το ακόλουθο διάνυσμα PageRank: $\mathbf{r} = [0.0675, 0.0962, 0.1164, 0.0675, 0.3339, 0.3184]$. Άρα τελικά, βλέπουμε ότι οι ιστοσελίδες P_4 και P_5 έχουν τους μεγαλύτερους βαθμούς κατάταξης, πράγμα που σημαίνει και μεγαλύτερη σημαντικότητα για αυτές. Συγκεκριμένα, από τη στιγμή που η ιστοσελίδα P_4 έχει δύο εισερχόμενες συνδέσεις αυτομάτως λαμβάνει μεγάλο βαθμό κατάταξης. Η ιστοσελίδα P_5 έχει επίσης μεγάλο βαθμό παρότι της αντιστοιχεί μόνο μια εισερχόμενη σύνδεση από την P_4. Όμως, αυτή η συγκεκριμένη σύνδεση που προέρχεται από ιστοσελίδα μεγάλης σημαντικότητας είναι αυτή που καθορίζει την P_5 ως μια από τις υψηλότερες βαθμολογικά ιστοσελίδες, μολονότι για παράδειγμα η P_1 έχει περισσότερες εισερχόμενες συνδέσεις. Τονίζουμε ότι αυτό ακριβώς είναι το επιθυμητό αποτέλεσμα που πρέπει να έχει ο αλγόριθμος PageRank.

7.3 Αναζήτηση των Ενεργών Επιδράσεων σε Υπερκορεσμένους Σχεδιασμούς με Διακριτά Δεδομένα

Σε αυτήν την ενότητα, ξεκινάμε με μια συζήτηση αναφορικά με το πλαίσιο των γενικευμένων γραμμικών μοντέλων στο οποίο δουλέψαμε, συνεχίζουμε με τα βασικά εργαλεία της Θεωρίας Πληροφορίας που θα χρησιμοποιηθούν σε συνδυασμό με τον αλγόριθμο PageRank και καταλήγουμε με την προτεινόμενη μέθοδο.

7.3.1 Τροποποίηση του PageRank Αλγορίθμου με Χρήση Μέτρων Πληροφορίας

Θα μας απασχολήσει η κατηγορία των υπερκορεσμένων σχεδιασμών δύο επιπέδων με d παράγοντες και n πειραματικές εκτελέσεις, με στόχο μια κατάλληλη μετατροπή του PageRank αλγορίθμου ώστε να μπορούμε να τον χρησιμοποιήσουμε ως μέθοδο κρησαρίσματος. Έστω X ο $n \times d$ πίνακας σχεδιασμού. Επικεντρωνόμαστε στην περίπτωση όπου έχουμε διακριτή απόκριση $\mathbf{y}$ και συγκεκριμένα, απόκριση κατανομής Bernoulli και Poisson.

Αρχικά, έστω ότι στο $c-$οστό πειραματικό σημείο, $c = 1, \ldots, n$ η απόκριση y_c είναι μια τυχαία μεταβλητή κατανομής Bernoulli, όπου $\mu_c = E(y_c) = P_c = P(\mathbf{x}_c)$. Εδώ, P_c είναι η πιθανότητα επιτυχίας σε μια διαδικασία Bernoulli, $\mathbf{x}_c$ είναι το $d-$διάστατο διάνυσμα προβλεπουσών μεταβλητών και $Var(y_c) = P_c(1 - P_c)$ η διασπορά της απόκρισης. Συνεπώς, η διασπορά είναι συνάρτηση της μέσης απόκρισης. Το Λογιστικό μοντέλο παλινδρόμησης για τη μέση απόκριση $P(\mathbf{x}_c)$ ορίζεται ως

$$P(\mathbf{x}_c) = \frac{1}{1 + e^{-\mathbf{x}_c^T \boldsymbol{\beta}}}, \tag{7.4}$$

όπου ο όρος $\mathbf{x}_c^T \boldsymbol{\beta}$ είναι η γραμμική προβλέπουσα και $\boldsymbol{\beta}$ ένα διάνυσμα διάστασης d των συντελεστών παλινδρόμησης.

Ως δεύτερη περίπτωση, θεωρούμε αυτήν όπου οι αποκρίσεις y_c, $c = 1, \ldots, n$ αποτελούν μετρήσεις (counts) που ακολουθούν ανεξάρτητες Poisson κατανομές, με $y_c \sim Poisson(\lambda_c)$ και λ_c η παράμετρος της κατανομής Poisson, η οποία ισούται με τη μέση απόκριση $\mu_c = E(y_c)$ καθώς και με τη διασπορά $Var(y_c)$. Θεωρώντας τη συμβατική περίπτωση χρήσης της log

συνάρτησης σύνδεσης, αποκτάμε ένα πολλαπλασιαστικό μοντέλο για το μέσο:

$$\mu_c = \lambda_c = e^{\mathbf{x}_c^T \beta}, \tag{7.5}$$

όπου $\mathbf{x}_c$ το $d-$διάστατο διάνυσμα προβλεπουσών μεταβλητών.

Πριν από την παρουσίαση της νέας μεθόδου, παραθέτουμε μια σύντομη περιγραφή και βασικούς ορισμούς των μέτρων πληροφορίας που θα χρειαστούμε. Η Θεωρία Πληροφορίας παρέχει διαισθητικά εργαλεία για την ποσοτικοποίηση της αβεβαιότητας τυχαίων μεταβλητών, ή πόση πληροφορία διαμοιράζεται από μερικές εξ αυτών. Η πιο θεμελιώδης έννοια στη Θεωρία Πληροφορίας είναι η εντροπία (entropy), που θεσπίστηκε από τον Shannon [156].

Ορισμός 7.1 *Έστω U διακριτή τυχαία μεταβλητή με συνάρτηση μάζας πιθανότητας $p(u) = P(U = u)$, $u \in \mathcal{U}$. Η εντροπία κατά Shannon της μεταβλητής U δίνεται από τη σχέση*

$$H(U) = -\sum_{u \in \mathcal{U}} p(u) \log_2(p(u)). \tag{7.6}$$

Ως βάση του λογαρίθμου επιλέγεται συνήθως το 2, οπότε η μονάδα μέτρησης της πληροφορίας είναι το bit. Εξίσου σημαντικά μέτρα είναι η από κοινού εντροπία (joint entropy) και η δεσμευμένη εντροπία (conditional entropy).

Ορισμός 7.2 *Έστω U, V διακριτές τυχαίες μεταβλητές, $u \in \mathcal{U}$ και $v \in \mathcal{V}$, με από κοινού συνάρτηση μάζας πιθανότητας $p(u, v)$. Η από κοινού εντροπία κατά Shannon δίνεται ως*

$$H(U, V) = -\sum_{v \in \mathcal{V}} \sum_{u \in \mathcal{U}} p(u, v) \log_2(p(u, v)). \tag{7.7}$$

Ορισμός 7.3 *Έστω U, V διακριτές τυχαίες μεταβλητές, $u \in \mathcal{U}$ και $v \in \mathcal{V}$, με δεσμευμένη συνάρτηση μάζας πιθανότητας $p(u|v)$. Η δεσμευμένη εντροπία κατά Shannon δίνεται ως*

$$H(U|V) = -\sum_{v \in \mathcal{V}} p(v) \sum_{u \in \mathcal{U}} p(u|v) \log_2(p(u|v)). \tag{7.8}$$

Συνεχίζουμε με το μέτρο της αμοιβαίας πληροφορίας (mutual information-MI) δύο τυχαίων μεταβλητών, το οποίο μετράει την αμοιβαία εξάρτηση των μεταβλητών. Εναλλακτικά, καλείται και κέρδος πληροφορίας (information gain) [146].

Ορισμός 7.4 *Έστω U, V διακριτές τυχαίες μεταβλητές, $u \in \mathcal{U}$ και $v \in \mathcal{V}$, με από κοινού συνάρτηση μάζας πιθανότητας $p(u, v)$ και περιθώριες συναρτήσεις μάζας πιθανότητας $p(u)$ και $p(v)$. Η αμοιβαία πληροφορία δίνεται ως*

$$I(U, V) = \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V}} p(u, v) \log_2 \frac{p(u, v)}{p(u)p(v)}. \tag{7.9}$$

Αφού δώσαμε τους παραπάνω βασικούς ορισμούς από τη Θεωρία Πληροφορίας, προχωράμε στο μέτρο πληροφορίας στο οποίο βασιστήκαμε και συνδυάσαμε με τον αλγόριθμο PageRank. Πρόκειται για το μέτρο της δεσμευμένης αμοιβαίας πληροφορίας (conditional mutual information-CMI). Αν U, V και W είναι διακριτές τυχαίες μεταβλητές, το μέτρο αυτό μπορεί να θεωρηθεί ως η διαφορά μεταξύ της μέσης υπολειπόμενης αβεβαιότητας της μεταβλητής U όταν η W είναι γνωστή και της ίδιας αβεβαιότητας όταν οι W και V είναι και οι δύο γνωστές.

Ορισμός 7.5 *Έστω U, V και W διακριτές τυχαίες μεταβλητές, $u \in \mathcal{U}$, $v \in \mathcal{V}$ και $w \in \mathcal{W}$. Η δεσμευμένη αμοιβαία πληροφορία των μεταβλητών U και V δεδομένης της W, δίνεται ως*

$$I(U, V|W) = H(U|W) - H(U|V, W). \tag{7.10}$$

Για μια διεξοδική ανάλυση των μέτρων της Θεωρίας Πληροφορίας, συνίσταται ιδιαίτερα το βιβλίο των Cover και Thomas [40].

Τόσο το μέτρο της αμοιβαίας πληροφορίας, όσο και της δεσμευμένης αμοιβαίας πληροφορίας μπορούν να χρησιμοποιηθούν για σκοπούς επιλογής μεταβλητών, ώστε να επιλέγονται οι παράγοντες που περιέχουν όσο το δυνατόν περισσότερη πληροφορία. Σχετικές εργασίες έχουν γίνει απο τους Fleuret [74] και Novovičová et al. [138]. Αναφορικά με το δικό μας ερευνητικό πλαίσιο αυτής της ενότητας και από τη σκοπιά του μέτρου αμοιβαίας πληροφορίας, το πρόβλημα επιλογής των σημαντικών παραγόντων μπορεί να διατυπωθεί ως εξής: Δοθέντος ενός αρχικού συνόλου X με d παράγοντες, εύρεση του υποσυνόλου $S \subset X$ με $d' < d$ που να πετυχαίνει την μεγαλύτερη δυνατή τιμή του $I(S, \mathbf{y})$. Σχετικά τώρα με το μέτρο της δεσμευμένης αμοιβαίας πληροφορίας, αποτελεί έναν τρόπο ποσοτικοποίησης της κοινής πληροφορίας μεταξύ του παράγοντα $\mathbf{x}_i$ του X και της απόκρισης $\mathbf{y}$, δοθέντος ενός άλλου παράγοντα $\mathbf{x}_j$. Στην περίπτωση όπου οι $\mathbf{x}_i$ και $\mathbf{y}$ είναι ανεξάρτητες μεταβλητές δεδομένης της $\mathbf{x}_j$, έχουμε ότι $\mathrm{I}(\mathbf{x}_i, \mathbf{y}|\mathbf{x}_j) = 0$, το οποίο σημαίνει ότι ο παράγοντας $\mathbf{x}_i$ δεν παρέχει καμμία πληροφορία σε σχέση με την απόκριση $\mathbf{y}$ όταν ο $\mathbf{x}_j$ είναι γνωστός.

Για να χρησιμοποιήσουμε τον αλγόριθμο PageRank έτσι ώστε να προσδιοριστούν οι παράγοντες που έχουν μη μηδενικές επιδράσεις στους υπερκορεσμένους σχεδιασμούς, θα πρέπει να τον τροποποιήσουμε σε αλγόριθμο με επίβλεψη (supervised algorithm)-άρα θα πρέπει να λαμβάνει υπόψη και την απόκριση $\mathbf{y}$. Οπότε, θεωρούμε ένα μη κατευθυνόμενο γράφημα $G = (R, C)$, όπου R είναι το σύνολο των κόμβων, και αντιπροσωπεύει το σύνολο των παραγόντων του σχεδιασμού και C το σύνολο των ακμών. Κάθε ζεύγος κόμβων συνδέεται με μια (μη κατευθυνόμενη) ακμή. Στη συνέχεια, δημιουργούμε έναν κατάλληλο πίνακα εισόδου κατασκευάζοντας έναν $d \times d$ πίνακα $\mathbf{M}$, με διαφορετικό τρόπο σύνδεσης των κόμβων-παραγόντων, αφού όπως προαναφέραμε λαμβάνουμε υπόψη και την απόκριση.

Συγκεκριμένα έστω ότι ξεκινάμε από έναν αυθαίρετα επιλεγμένο παράγοντα $\mathbf{x}_j$. Το στοιχείο M_{ji} δε θα αποτελεί την πιθανότητα μετάβασης από τον παράγοντα $\mathbf{x}_j$ στον $\mathbf{x}_i$ (το οποίο από τη σκοπιά του αλγορίθμου PageRank σημαίνει ότι υπάρχει μια εξερχόμενη σύνδεση από την ιστοσελίδα $\mathbf{x}_j$ στη $\mathbf{x}_i$) αλλά τη δεσμευμένη αμοιβαία πληροφορία του παράγοντα $\mathbf{x}_i$ και της απόκρισης $\mathbf{y}$ δοθέντος του παράγοντα $\mathbf{x}_j$. Συνεπώς για $i = 1, \ldots, d$, $j = 1, \ldots, d$ και $i \neq j$ υπολογίζουμε τις τιμές $\mathrm{I}(\mathbf{x}_i, \mathbf{y}|\mathbf{x}_j)$. Έπειτα, ο πίνακας που προκύπτει αφού μετασχηματιστεί σε στοχαστικό κατά γραμμή, δίνεται ως είσοδο στον αλγόριθμο PageRank. Υπολογίζεται ο τελικός πίνακας $\tilde{\mathbf{M}}$ και έπειτα από την υλοποίηση της Power μεθόδου, το τελικό PageRank διάνυσμα θα περιλαμβάνει τους βαθμούς των παραγόντων, εκ των οποίων όσοι έχουν τιμές πάνω από ένα προκαθορισμένο όριο, θεωρούνται σημαντικοί. Τονίζουμε ότι, στις προσομοιώσεις που κάναμε, εξετάστηκαν πολλά και διαφορετικά όρια, όπως η μέση τιμή, ο γεωμετρικός μέσος και η διάμεσος των PageRank τιμών. Βάσει των αποτελεσμάτων που προέκυψαν, επιλέχθηκε ο γεωμετρικός μέσος ως το κατάλληλο όριο.

7.3.2 Η Προτεινόμενη Μέθοδος

Η διαδικασία επιλογής μεταβλητών που προτείνουμε, εφαρμόζεται στους υπερκορεσμένους σχεδιασμούς κάτω από την υπόθεση ενός γενικευμένου γραμμικού μοντέλου, του οποίου τα δεδομένα απόκρισης προέρχονται από κατανομή Bernoulli ή Poisson, και περιγράφεται ρητά ως εξής:

1. Έστω ένας $n \times d$ υπερκορεσμένος σχεδιασμός $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d]$, όπου $\mathbf{x}_l$, $l = 1, 2, \ldots, d$, η $l-$οστή στήλη του. Έστω επίσης ένα $n \times 1$ διάνυσμα απόκρισης $\mathbf{y}$, κατανομής Bernoulli ή Poisson.

2. Για κάθε $i = 1, \ldots, d$, $j = 1, \ldots, d$ με $i \neq j$, υπολογίζουμε τις τιμές $I(\mathbf{x}_i, \mathbf{y}|\mathbf{x}_j)$ βάσει της (7.10).

3. Προκύπτει ο $d \times d$ πίνακας $\mathbf{M}$ όπου κάθε στοιχείο του $\mathbf{M}_{ji}$, στη γραμμή j και τη στήλη i αντιστοιχεί στην παραπάνω αμοιβαία πληροφορία μεταξύ του παράγοντα $\mathbf{x}_i$ και της απόκρισης $\mathbf{y}$, δεδομένου του παράγοντα $\mathbf{x}_j$.

4. Ο πίνακας $\mathbf{M}$ μετατρέπεται σε στοχαστικό κατά γραμμή και δίνεται ως είσοδο στον αλγόριθμο PageRank, οπότε και υλοποιείται η Power μέθοδος.

5. Προκύπτει ο θετικός, στοχαστικός και μη αναγώγιμος πίνακας $\tilde{\mathbf{M}}$ έπειτα από τη χρήση του παράγοντα απόσβεσης $\alpha = 0.85$.

6. Λαμβάνουμε το τελικό PageRank διάνυσμα που θα περιέχει τις τιμές - βαθμούς για κάθε παράγοντα, βάσει των οποίων επιλέγονται οι τελικά σημαντικοί, με χρήση συγκεκριμένου ορίου.

7. Επιλέγονται οι παράγοντες με τιμές μεγαλύτερες του γεωμετρικού μέσου των τιμών του PageRank διανύσματος.

7.4 Πειράματα Προσομοίωσης

Σε αυτήν την ενότητα παρουσιάζουμε μια λεπτομερή μελέτη προσομοίωσης που πραγματοποιήσαμε, προκειμένου να αξιολογηθεί η απόδοση της προτεινόμενης μεθόδου. Για το σκοπό αυτό, εκτελέσαμε προσομοιώσεις για ένα ευρύ φάσμα μοντέλων και υπερκορεσμένων σχεδιασμών. Για την περίπτωση των δίτιμων δεδομένων, εξετάσαμε επίσης τον Conditional Mutual Information Maximization (CMIM) αλγόριθμο που πρότεινε ο Fleuret [74] μαζί με τον minimal-redundancy-maximal-relevance feature selection (mRMR) αλγόριθμο των Peng et al. [142]. Η επιλογή των συγκεκριμένων αλγορίθμων, έγινε καθότι και οι δύο, όπως και η προτεινόμενη μέθοδος, βασίζονται σε μέτρα της Θεωρίας Πληροφορίας. Συγκεκριμένα, ο CMIM αποτελεί μια τεχνική επιλογής μεταβλητών βασισμένη στο μέτρο της δεσμευμένης αμοιβαίας πληροφορίας, η οποία επιλέγει επαναληπτικά τις μεταβλητές που μεγιστοποιούν την αμοιβαία πληροφορία σε σχέση με την απόκριση, δεδομένου οποιουδήποτε παράγοντα έχει ήδη επιλεχθεί. Ας σημειώσουμε εδώ ότι η συγκεκριμένη διαδικασία μπορεί να χρησιμοποιηθεί σε προβλήματα ταξινόμησης δύο μόνο κλάσεων, άρα προϋποθέτει δίτιμη απόκριση. Από την άλλη πλευρά, ο αλγόριθμος mRMR μεγιστοποιεί την αμοιβαία πληροφορία μεταξύ των επιλεγμένων μεταβλητών και της απόκρισης (relevance) και ελαχιστοποιεί την αμοιβαία πληροφορία μεταξύ των επιλεγμένων μεταβλητών (redundancy). Σημειώνουμε ότι και σε αυτόν τον αλγόριθμο, το εξαγόμενο αποτέλεσμα πρέπει να είναι μια μεταβλητή με στόχο την ταξινόμηση σε δύο κλάσεις, οπότε τον χρησιμοποιήσαμε μόνο στην περίπτωση της δίτιμης απόκρισης. Αναφορικά τώρα με την περίπτωση όπου τα δεδομένα απόκρισης προέρχονται από την κατανομή Poisson και βάσει της δικιάς μας γνώσης, υπάρχει έλλειψη στη βιβλιογραφία κατάλληλων μεθόδων ανάλυσης υπερκορεσμένων σχεδιασμών. Συνεπώς, δεν ήταν δυνατή η πραγματοποίηση συγκρίσεων με άλλες μεθόδους της υπάρχουσας βιβλιογραφίας για Poisson δεδομένα.

7.4.1 Σχεδιασμός Προσομοιώσεων

Στη μελέτη προσομοίωσης, θεωρήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς, που κατασκευάστηκαν από τους Koukouvinos et al. [105]. Οι συγκεκριμένοι σχεδιασμοί, έχουν n εκτελέσεις και $d = q(n-1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Συγκεκριμένα, χρησιμοποιήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, d) τιμές: (6, 10), (8, 14), (10, 18), (12, 22), (14, 26), (16, 30), (18, 34), (20, 38) και (22, 42). Επιπλέον, θεωρήσαμε και τους s-block-ορθογώνιους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με n εκτελέσεις και $d = s(n-1)$ παράγοντες, που κατασκευάστηκαν από τους

Tang και Wu [161]. Συγκεκριμένα, συμπεριλάβαμε το σχεδιασμό με $d = 22$ παράγοντες και $n = 12$ εκτελέσεις. Τέλος, χρησιμοποιήσαμε και τον υπερκορεσμένο σχεδιασμό του Lin [120] με $n = 6$ εκτελέσεις και $d = 10$ παράγοντες.

Οι πραγματικά ενεργές επιδράσεις επιλέχθηκαν τυχαία από το σύνολο των $1, \ldots, d$ πιθανά ενεργών παραγόντων. Οι συντελεστές των μη ενεργών μεταβλητών, στο πραγματικό μοντέλο, τέθηκαν ίσες με μηδέν. Θεωρήσαμε μοντέλα κύριων επιδράσεων μόνο. Τονίζουμε ότι, οι συνθήκες στην πράξη είναι συνήθως διαφορετικές από εκείνες των προσομοιώσεων και φυσικά ο πειραματιστής δε γνωρίζει εκ των προτέρων, πόσοι και ποιοί παράγοντες είναι ενεργοί. Συνεπώς, για μεγαλύτερη αξιοπιστία των αποτελεσμάτων μας και για να εξετάσουμε το πόσο ευαίσθητα είναι στην επιλογή και στον αριθμό των ενεργών παραγόντων, αλλάξαμε τη διάταξη των ενεργών μεταβλητών, χρησιμοποιήσαμε διαφορετικές τιμές για τους συντελεστές β και διαφορετικό πλήθος ενεργών παραγόντων, σε κάθε υπερκορεσμένο σχεδιασμό. Ως αποτέλεσμα, εξετάσαμε ένα ευρύ φάσμα διαφορετικών μοντέλων στα πειράματά μας.

7.4.2 Κριτήρια Αξιολόγησης της Απόδοσης

Όπως συμβαίνει σε αρκετά προβλήματα απόφασης, τα σφάλματα πρέπει να εξισορροπούνται σε σχέση με το κόστος. Στους σχεδιασμούς κρησαρίσματος, υπάρχει η πιθανότητα να δηλώσουμε μία ανενεργή επίδραση ως ενεργή, κάτι που είναι γνωστό ως σφάλμα Τύπου I, καθώς και η πιθανότητα να δηλώσουμε μία ενεργή επίδραση ως ανενεργή, δηλαδή το λεγόμενο σφάλμα Τύπου II. Τα σφάλματα Τύπου II δημιουργούν αρκετά προβλήματα, όπως τονίζει ο Lin [121]. Επιπλέον, τα σφάλματα Τύπου I τα οποία απαντώνται συχνά όταν ισχύει η αρχή της σποραδικότητας των επιδράσεων [121], μπορούν να οδηγήσουν σε περιττό υπολογιστικό κόστος στα επερχόμενα πειράματα. Συνεπώς, βασιστήκαμε σε αυτά τα δύο κριτήρια για την αξιολόγηση της απόδοσης και της αποτελεσματικότητας της προτεινόμενης μεθόδου επιλογής μεταβλητών.

7.4.3 Αποτελέσματα Προσομοιώσεων: Bernoulli Απόκριση

Σε αυτό το σημείο, παρουσιάζουμε τα αποτελέσματα για δίτιμη απόκριση. Στον Πίνακα 7.1, περιγράφονται τα εξεταζόμενα μοντέλα. Στην πρώτη στήλη, αναφέρεται ο αριθμός που αντιστοιχεί σε κάθε μοντέλο. Στις υπόλοιπες τρεις στήλες, αναγράφεται ο εκάστοτε υπερκορεσμένος σχεδιασμός που εξετάζεται, μαζί με τον αριθμό των παραγόντων και των εκτελέσεων του σχεδιασμού, d και n αντίστοιχα. Στην τελευταία στήλη καταγράφεται το διάνυσμα που περιέχει τις τυχαία επιλεγμένες τιμές των συντελεστών β. Για κάθε ένα από τα απεικονιζόμενα μοντέλα, παρήχθησαν τυχαία 1000 διανύσματα απόκρισης $\mathbf{y} \sim Bernoulli(P(\mathbf{x}^T\boldsymbol{\beta}))$ όπου $P(u) = \frac{1}{1+e^{-u}}$. Τα λαμβανόμενα αποτελέσματα, έπειτα από την εφαρμογή της προτεινόμενης PageRank μεθόδου καθώς και των αλγορίθμων CMIM και mRMR, συνοψίζονται στον Πίνακα 7.2. Συγκεκριμένα, στην πρώτη στήλη αναγράφεται ο αριθμός που αντιστοιχεί στο μοντέλο που χρησιμοποιήθηκε. Οι στήλες που ονομάζονται 'Type I' και 'Type II' αναφέρονται σε κάθε περίπτωση, στις μέσες τιμές των σφαλμάτων Τύπου I και II, των 1000 προσομοιώσεων.

Πίνακας 7.1: Τα μοντέλα που χρησιμοποιήθηκαν στις προσομοιώσεις, για απόκριση κατανομής Bernoulli

Μοντέλο	SSD	d	n	β
1	Lin [120]	10	6	[0,0,0,0,0,0,0,7,-1,0]^T
2	Lin [120]	10	6	[0,5,0,0,-8,1,0,0,0,0]^T
3	Lin [120]	10	6	[0,-3,0,0,0,0,9,0,0,0]^T
4	Tang και Wu [161]	22	12	[2,-1,1,0,0,0,0,0,0,0,0,0,0,0,-1,1,0,2,0,0,0,0]^T
5	Tang και Wu [161]	22	12	[0,0,6,0,4,8,0,0,0,0,1,9,8,5,0,0,0,0,7,9,3,0]^T
6	Tang και Wu [161]	22	12	[0,-2,-2,1,2,0,0,4,0,0,0,4,0,0,0,0,0,0,-3,0,2,2]^T
7	Tang και Wu [161]	22	12	[0,0,0,0,20,-16,0,0,0,0,0,0,0,0,0,0,20,0,0,-7,0,0]^T
8	Koukouvinos et al. [105]	10	6	[-4,0,0,0,7,0,0,0,8,0]^T
9	Koukouvinos et al. [105]	10	6	[-7,0,0,0,0,-7,0,0,2,0]^T
10	Koukouvinos et al. [105]	10	6	[-9,0,0,3,10,0,0,-9,0,0]^T
11	Koukouvinos et al. [105]	14	8	[-4,0,0,0,0,0,0,0,-9,0,4,-5,0,0]^T
12	Koukouvinos et al. [105]	14	8	[0,0,1,0,0,0,0,3,0,0,0,0,0,0]^T
13	Koukouvinos et al. [105]	14	8	[-6,0,0,0,0,-2,0,0,0,0,-15,0,0,0]^T
14	Koukouvinos et al. [105]	14	8	[5,5,0,0,-5,0,0,3,0,-2,0,0,2,0]^T
15	Koukouvinos et al. [105]	18	10	[-4,2,0,0,0,-3,0,0,7,0,0,17,0,5,0,0,21,0]^T
16	Koukouvinos et al. [105]	18	10	[0,0,0,0,0,-3,0,-8,0,0,0,0,2,0,0,-9,0,0]^T
17	Koukouvinos et al. [105]	18	10	[0,8,0,0,0,0,7,0,0,4,0,-5,0,4,7,0,0,0]^T
18	Koukouvinos et al. [105]	18	10	[0,0,0,0,0,0,2,0,6,10,0,0,0,0,0,0,11,20]^T
19	Koukouvinos et al. [105]	22	12	[0,0,4,6,0,0,0,8,0,6,0,-2,0,0,4,0,9,0,8,0,0,9]^T
20	Koukouvinos et al. [105]	22	12	[0,0,0,0,0,0,0,0,-6,0,0,5,0,0,0,0,0,0,0,0,0,-6,0,0]^T
21	Koukouvinos et al. [105]	22	12	[0,-13,18,0,0,17,0,0,0,0,-8,0,7,0,0,0,0,0,4,0,0,0]^T
22	Koukouvinos et al. [105]	22	12	[0,0,0,0,7,0,0,0,3,0,-6,9,0,0,-3,0,0,0,0,19,13,0]^T
23	Koukouvinos et al. [105]	26	14	[0,4,0,0,0,16,0,-17,0,0,0,19,0,15,17,0,0,0,0,-17,0,3,8,0,3,-7]^T
24	Koukouvinos et al. [105]	26	14	[-2,-1,0,0,2,-3,0,0,0,0,0,2,0,-4,0,0,0,-6,0,0,-4,0,0,0,-5,0]^T
25	Koukouvinos et al. [105]	26	14	[0,2,0,0,0,0,0,0,0,0,0,2,0,2,0,0,0,4,0,0,2,0,0,0,0,0]^T
26	Koukouvinos et al. [105]	26	14	[0,0,0,0,0,0,0,0,0,0,0,0,0,0,-4,2,-2,0,2,2,0,0,2,-2,0,4]^T
27	Koukouvinos et al. [105]	30	16	[0,-4,0,0,0,0,2,0,0,0,-5,0,-2,0,-1,0,0,-7,0,0,-3,0,0,0,0,-7,0,0,0,0]^T
28	Koukouvinos et al. [105]	30	16	[0,5,0,-7,0,0,0,0,0,-4,0,0,0,0,5,-4,-9,4,0,0,6,-7,0,0,3,0,0,0,0,0]^T
29	Koukouvinos et al. [105]	30	16	[0,0,-4,0,2,-15,-17,0,0,0,19,0,0,5,4,-12,0,0,0,0,0,0,9,15,-10,0,0,0,-4,0]^T
30	Koukouvinos et al. [105]	30	16	[0,14,0,0,0,0,0,0,0,-8,-18,0,0,0,0,-13,0,0,-19,18,0,0,0,0,13,0,0,0,2,0]^T
31	Koukouvinos et al. [105]	34	18	[0,0,5,0,0,-4,0,-3,-4,3,0,0,0,0,0,0,0,-2,-2,0,0,-6,0,0,4,0,-5,0,0,0,0,0,0,-2,0]^T
32	Koukouvinos et al. [105]	34	18	[0,3,0,0,0,0,0,0,0,0,0,-18,0,0,0,0,7,0,0,0,16,0,0,0,0,0,0,-1,-1,0,0,0,0,0]^T
33	Koukouvinos et al. [105]	34	18	[5,0,4,0,0,0,0,0,0,-6,0,0,0,0,0,0,0,0,0,0,-12,10,0,0,-14,0,0,0,0,0,0,0,-15,0,0]^T
34	Koukouvinos et al. [105]	34	18	[-15,0,-19,0,0,0,0,0,-18,0,0,-10,3,0,0,0,0,0,0,0,0,0,0,0,0,-8,0,0,-18,0,0,0,0,0]^T
35	Koukouvinos et al. [105]	38	20	[0,0,0,0,0,0,0,0,0,0,0,3,-10,0,0,0,0,0,0,0,0,5,6,0,0,0,0,0,0,0,0,0,0,0,9,0,0]^T
36	Koukouvinos et al. [105]	38	20	[0,18,0,0,11,0,9,0,0,0,0,0,0,0,18,0,17,0,0,0,0,0,0,0,0,13,0,0,0,0,0,0,0,0,0,0,0,7,0]^T
37	Koukouvinos et al. [105]	38	20	[0,0,0,0,-3,0,0,0,5,0,5,5,0,0,0,0,-2,0,0,0,0,0,5,0,10,0,5,0,0,7,0,0,0,0,-3,-4,0,-3]^T
38	Koukouvinos et al. [105]	38	20	[0,0,0,0,0,0,0,0,0,2,0,6,0,4,0,0,0,0,0,0,0,9,0,10,0,0,0,0,0,0,0,0,0,-1,0,6,0,-3,0,0]^T
39	Koukouvinos et al. [105]	42	22	[0,0,0,0,-3,0,0,-2,0,0,0,0,0,1,-1,0,0,0,0,0,-2,0,0,0,5,0,-3,0,0,0,0,0,-3,1,0,0,0,0,0,0,-4,1,-1,3,0,0]^T
40	Koukouvinos et al. [105]	42	22	[0,0,0,0,-2,0,0,0,0,1,0,0,0,0,0,0,-4,0,0,0,0,0,0,0,0,0,0,0,0,-5,0,0,0,0,0,0,3,0,0,0,0,0,0]^T
41	Koukouvinos et al. [105]	42	22	[0,0,11,0,0,0,0,0,-13,0,0,0,13,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,6,0,0,0,0,0,0,0,0,0,0]^T
42	Koukouvinos et al. [105]	42	22	[0,0,0,0,0,-15,0,0,2,0,0,0,6,-2,0,0,0,0,0,0,0,-15,0,0,5,0,0,12,0,-10,0,0,0,0,0,0,0,0,0,0,14,0,7]^T

Πίνακας 7.2: Συγκριτική απόδοση των μεθόδων για τα μοντέλα 1-42

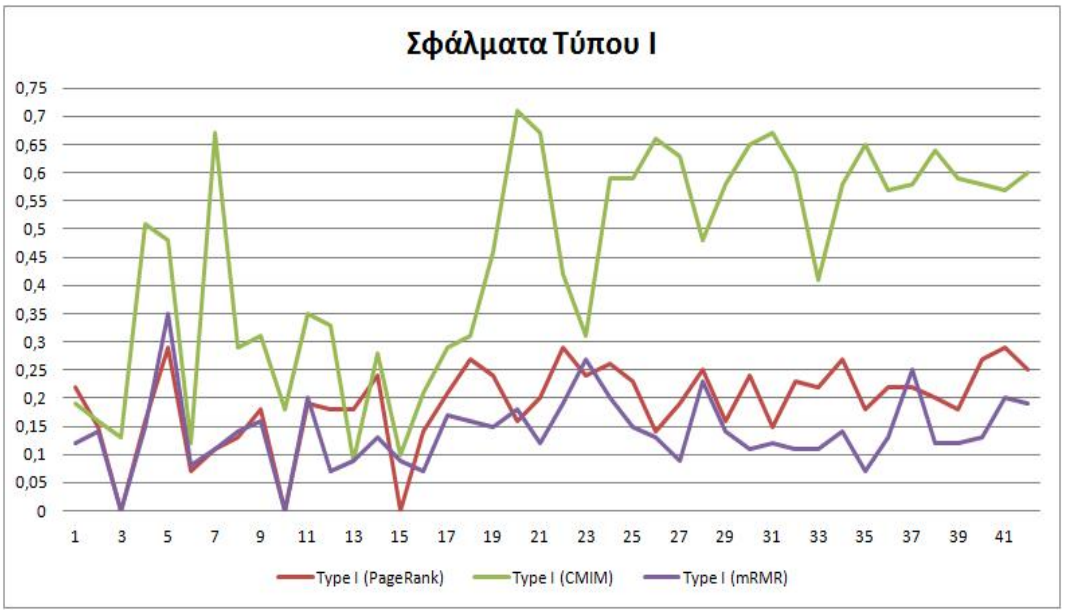
Μοντέλο	Type I (PageRank)	Type I (CMIM)	Type I (mRMR)	Type II (PageRank)	Type II (CMIM)	Type II (mRMR)
1	0.22	0.19	0.12	0.00	0.50	0.00
2	0.15	0.16	0.14	0.01	0.54	0.33
3	0.00	0.13	0.00	0.00	0.50	0.00
4	0.16	0.51	0.15	0.13	0.08	0.33
5	0.29	0.48	0.35	0.15	0.11	0.42
6	0.07	0.12	0.08	0.12	0.15	0.12
7	0.11	0.67	0.11	0.00	0.05	0.50
8	0.13	0.29	0.14	0.03	0.07	0.33
9	0.18	0.31	0.16	0.11	0.11	0.37
10	0.00	0.18	0.00	0.00	0.03	0.01
11	0.19	0.35	0.20	0.03	0.05	0.50
12	0.18	0.33	0.07	0.04	0.36	0.40
13	0.18	0.09	0.09	0.00	0.56	0.33
14	0.24	0.28	0.13	0.01	0.00	0.17
15	0.00	0.10	0.09	0.14	0.14	0.14
16	0.14	0.21	0.07	0.01	0.02	0.26
17	0.21	0.29	0.17	0.01	0.01	0.33
18	0.27	0.31	0.16	0.00	0.00	0.40
19	0.24	0.46	0.15	0.16	0.10	0.25
20	0.16	0.71	0.18	0.00	0.00	0.78
21	0.20	0.67	0.12	0.12	0.09	0.28
22	0.29	0.42	0.19	0.04	0.06	0.41
23	0.24	0.31	0.27	0.19	0.21	0.37
24	0.26	0.59	0.20	0.07	0.02	0.39
25	0.23	0.59	0.15	0.01	0.01	0.50
26	0.14	0.66	0.13	0.11	0.04	0.28
27	0.19	0.63	0.09	0.19	0.07	0.24
28	0.25	0.48	0.23	0.12	0.10	0.46
29	0.16	0.58	0.14	0.16	0.07	0.19
30	0.24	0.65	0.11	0.12	0.06	0.27
31	0.15	0.67	0.12	0.18	0.11	0.24
32	0.23	0.60	0.11	0.17	0.05	0.49
33	0.22	0.41	0.11	0.14	0.15	0.43
34	0.27	0.58	0.14	0.09	0.05	0.40
35	0.18	0.65	0.07	0.00	0.02	0.20
36	0.22	0.57	0.13	0.14	0.05	0.56
37	0.22	0.58	0.25	0.19	0.10	0.46
38	0.20	0.64	0.12	0.17	0.11	0.30
39	0.18	0.59	0.12	0.20	0.08	0.26
40	0.27	0.58	0.13	0.06	0.02	0.50
41	0.29	0.57	0.20	0.00	0.06	0.25
42	0.25	0.60	0.19	0.19	0.05	0.29

Από τον Πίνακα 7.2, εξάγουμε τα ακόλουθα συμπεράσματα: Καταρχήν, παρατηρούμε ότι η προτεινόμενη μέθοδος επιτυγχάνει χαμηλές τιμές και για τα δύο ποσοστά σφάλματος στις περισσότερες περιπτώσεις, ένα γεγονός το οποίο είναι σαφώς απαραίτητο για μια διαδικασία κρησαρίσματος. Ειδικότερα, τα σφάλματα Τύπου I είναι σε χαμηλό επίπεδο για πολλά από τα θεωρούμενα μοντέλα, έχοντας μια μέση τιμή ίση με 0.19. Επιπλέον, τα σφάλματα Τύπου II παραμένουν επίσης σε πολύ χαμηλό επίπεδο, έχοντας μια μέση τιμή ίση με 0.08. Τονίζουμε ότι σε πολλές περιπτώσεις, το σφάλμα Τύπου II έχει τιμές κοντά ή ίσες με μηδέν, με αποτέλεσμα μια πολύ ικανοποιητική ισχύ της μεθόδου (1-Type II). Επιπλέον, για την πλειονότητα των μοντέλων (38/42), τα σφάλματα Τύπου II είναι μικρότερα ή ίσα με τα σφάλματα Τύπου I. Συνεπώς, η νέα μέθοδος έχει την τάση να δηλώνει περισσότερες ανενεργές επιδράσεις ως ενεργές και αρκετά λιγότερες ενεργές επιδράσεις ως ανενεργές. Ως εκ τούτου, η προτεινόμενη μέθοδος είναι πράγματι συντηρητική με αυτή την έννοια.

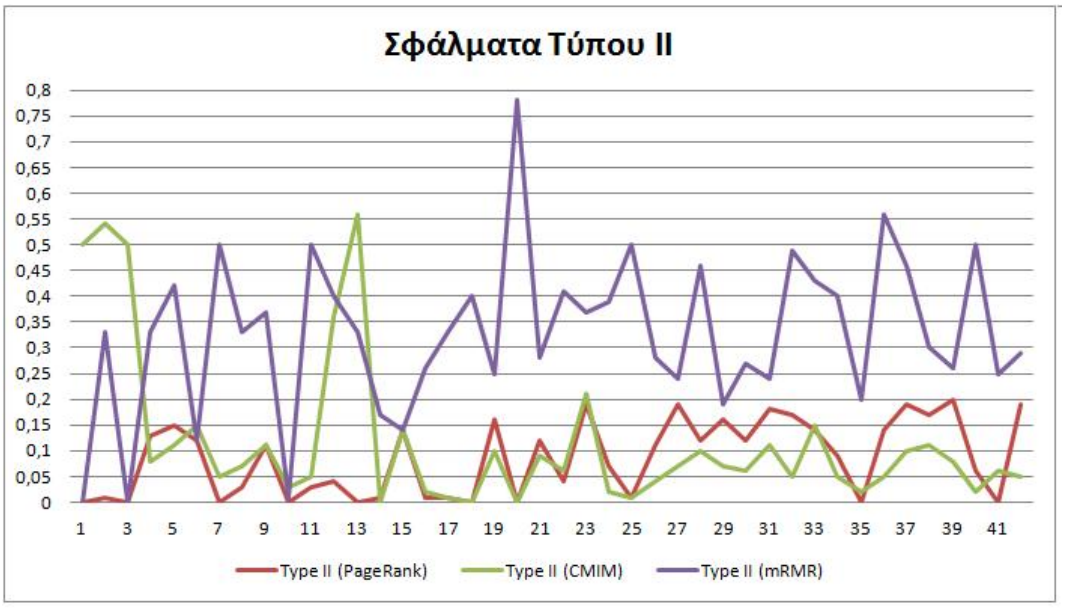
Αναφορικά με τον αλγόριθμο CMIM, παράγει πολύ υψηλά σφάλματα Τύπου I έχοντας μια μέση τιμή ίση με 0.45, με αποτέλεσμα να επιλέγει πολλούς από τους ανενεργούς παράγοντες. Το γεγονός αυτό οδηγεί σε χαμηλά σφάλματα Τύπου II, μέσης τιμής 0.12, κάτι που όμως θεωρείται λογικό. Στα πειράματα κρησαρίσματος και λόγω της αρχής της σποραδικότητας των επιδράσεων [19], ο αριθμός των ενεργών παραγόντων που μπορεί επιτυχώς να αναγνωριστεί, σπανίως υπερβαίνει το 1/2 ή ακόμα και το 1/3 του συνολικού αριθμού των παραγόντων. Συνεπώς, είναι πολύ πιθανό για μια εξεταζόμενη μέθοδο να παράγει πολύ υψηλά σφάλματα Τύπου I και αντιστοίχως χαμηλά σφάλματα Τύπου II. Φυσικά αυτό συνεπάγεται μια ασταθή συμπεριφορά της μεθόδου. Επιπλέον, να σημειώσουμε ότι υπάρχουν και κάποια μοντέλα στα οποία ο αλγόριθμος CMIM οδήγησε σε αρκετά υψηλότερα σφάλματα Τύπου II.

Σχετικά τώρα με τον αλγόριθμο mRMR, επιτυγχάνει ελαφρώς χαμηλότερα σφάλματα Τύπου I για κάποια μοντέλα, με αποτέλεσμα να έχει μια μέση τιμή ίση με 0.14, ενώ παρουσιάζει μια παρόμοια συμπεριφορά για τις υπόλοιπες περιπτώσεις. Παρ΄ όλα αυτά, αυτό συμβαίνει εις βάρος των σφαλμάτων Τύπου II, τα οποία είναι πολύ υψηλά για σχεδόν όλες τις περιπτώσεις, σε σύγκριση με την προτεινόμενη μέθοδο. Η μέση τιμή των σφαλμάτων Τύπου II για αυτόν τον αλγόριθμο ισούται με 0.33.

Συνοψίζοντας, η PageRank μέθοδος, ναι μεν είναι ελαφρώς συντηρητική, αλλά επιτυγχάνει μια γενικά σταθερή απόδοση και οδηγεί σε πολύ χαμηλές τιμές των σφαλμάτων Τύπου II, γεγονός ιδιαίτερα σημαντικό σε διαδικασίες επιλογής μεταβλητών στους υπερκορεσμένους σχεδιασμούς. Οι συγκρινόμενοι αλγόριθμοι, ήταν γενικά ασταθείς, οδηγώντας σε σημαντικά υψηλότερες τιμές σφαλμάτων Τύπου I (CMIM) ή Τύπου II (mRMR), όπως άλλωστε φαίνεται και από τα παρακάτω σχήματα.



Σχήμα 7.2: Σύγκριση σφαλμάτων Τύπου I των μεθόδων PageRank, CMIM και mRMR



Σχήμα 7.3: Σύγκριση σφαλμάτων Τύπου II των μεθόδων PageRank, CMIM και mRMR

7.4.4 Αποτελέσματα Προσομοιώσεων: Poisson Απόκριση

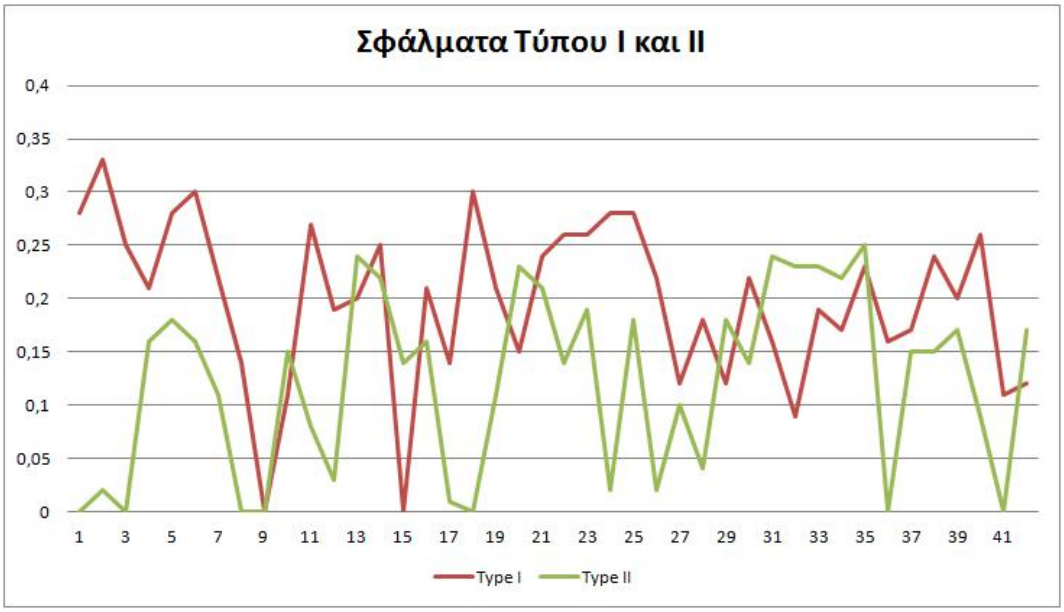
Παρουσιάζουμε στο σημείο αυτό τα αποτελέσματα σχετικά με τα Poisson δεδομένα. Όπως και στην προηγούμενη υποενότητα, τα εξεταζόμενα μοντέλα δίνονται αρχικά στον Πίνακα 7.3. Για κάθε ένα από αυτά, παρήχθησαν τυχαία 1000 διανύσματα απόκρισης $\mathbf{y} \sim Poisson(\lambda(\mathbf{x}^T\boldsymbol{\beta}))$ όπου $\lambda(u) = e^u$. Τα αποτελέσματα που προέκυψαν, έπειτα από την εφαρμογή της PageRank μεθόδου και αναφορικά με τα σφάλματα Τύπου I και II, παρουσιάζονται στον Πίνακα 7.4.

Πίνακας 7.3: Τα μοντέλα που χρησιμοποιήθηκαν στις προσομοιώσεις, για απόκριση κατανομής Poisson

Μοντέλο	SSD	d	n	β
1	Lin [120]	10	6	$[0,9,0,0,0,-6,0,0,0,7]^T$
2	Lin [120]	10	6	$[0,0,-3,0,0,0,1,0,-3,-3]^T$
3	Lin [120]	10	6	$[0,0,0,0,0,10,0,0,0,17]^T$
4	Tang και Wu [161]	22	12	$[-1,0,0,5,0,7,0,0,0,0,0,0,0,0,0,0,-4,0,0,0,-1,0,0]^T$
5	Tang και Wu [161]	22	12	$[0,0,2,3,0,0,13,1,2,0,5,0,0,1,8,0,0,0,0,0,-1,0]^T$
6	Tang και Wu [161]	22	12	$[0,0,6,0,4,8,0,0,0,0,1,9,8,5,0,0,0,0,7,9,3,0]^T$
7	Tang και Wu [161]	22	12	$[2,-1,1,0,0,0,0,0,0,0,0,0,0,0,-1,1,0,2,0,0,0,0]^T$
8	Koukouvinos et al. [105]	10	6	$[-4,0,0,0,7,0,0,0,8,0]^T$
9	Koukouvinos et al. [105]	10	6	$[-9,0,0,3,10,0,0,-9,0,0]^T$
10	Koukouvinos et al. [105]	10	6	$[-7,0,0,0,0,-7,0,0,2,0]^T$
11	Koukouvinos et al. [105]	14	8	$[1,0,0,2,0,0,-4,0,0,7,0,0,11,0]^T$
12	Koukouvinos et al. [105]	14	8	$[-4,0,0,0,0,0,0,0,-9,0,4,-5,0,0]^T$
13	Koukouvinos et al. [105]	14	8	$[0,0,0,0,0,-15,12,0,0,-3,0,0,0,0]^T$
14	Koukouvinos et al. [105]	14	8	$[0,4,0,0,0,0,0,0,0,0,4,-1,0,0]^T$
15	Koukouvinos et al. [105]	18	10	$[-4,2,0,0,0,-3,0,0,7,0,0,17,0,5,0,0,21,0]^T$
16	Koukouvinos et al. [105]	18	10	$[0,0,-5,0,-3,4,0,0,7,0,0,1,2,0,0,-7,0,3]^T$
17	Koukouvinos et al. [105]	18	10	$[0,0,0,0,0,-3,0,-8,0,0,0,0,2,0,0,-9,0,0]^T$
18	Koukouvinos et al. [105]	18	10	$[0,0,0,0,0,0,0,2,0,6,10,0,0,0,0,0,0,0,11,20]^T$
19	Koukouvinos et al. [105]	22	12	$[0,-13,18,0,0,17,0,0,0,0,-8,0,7,0,0,0,0,0,4,0,0,0]^T$
20	Koukouvinos et al. [105]	22	12	$[1,2,0,0,0,16,0,0,0,0,7,0,0,0,-1,0,0,5,0,18,0,0]^T$
21	Koukouvinos et al. [105]	22	12	$[1,0,0,-13,0,0,4,-3,0,0,-6,0,0,-7,0,-24,0,-5,0,-21,0,0]^T$
22	Koukouvinos et al. [105]	22	12	$[-1,0,0,3,0,0,0,0,0,0,0,0,0,1,2,0,0,0,0,9,0,0]^T$
23	Koukouvinos et al. [105]	26	14	$[0,4,0,0,0,16,0,-17,0,0,0,19,0,15,17,0,0,0,0,-17,0,3,8,0,3,-7]^T$
24	Koukouvinos et al. [105]	26	14	$[-2,-1,0,0,2,-3,0,0,0,0,0,2,0,-4,0,0,0,-6,0,0,-4,0,0,0,-5,0]^T$
25	Koukouvinos et al. [105]	26	14	$[1,0,1,0,2,-4,0,0,5,0,0,-1,0,0,2,1,0,5,3,1,0,0,-2,0,1,0]^T$
26	Koukouvinos et al. [105]	26	14	$[0,2,0,0,0,0,0,0,0,0,0,2,0,2,0,0,0,4,0,0,2,0,0,0,0,0]^T$
27	Koukouvinos et al. [105]	30	16	$[0,5,0,-7,0,0,0,0,0,-4,0,0,0,0,5,-4,-9,4,0,0,6,-7,0,0,3,0,0,0,0,0]^T$
28	Koukouvinos et al. [105]	30	16	$[0,0,0,0,0,0,7,0,0,-1,0,2,0,0,0,0,0,0,0,0,0,0,0,7,0,0,8,0,0,0,0]^T$
29	Koukouvinos et al. [105]	30	16	$[0,0,-4,0,2,-15,-17,0,0,0,19,0,0,5,4,-12,0,0,0,0,0,0,9,15,-10,0,0,0,-4,0]^T$
30	Koukouvinos et al. [105]	30	16	$[0,14,0,0,0,0,0,0,0,-8,-18,0,0,0,0,-13,0,0,-19,18,0,0,0,0,13,0,0,0,2,0]^T$
31	Koukouvinos et al. [105]	34	18	$[0,0,-6,-9,0,10,0,0,6,0,0,0,5,0,0,7,5,0,0,0,0,0,0,0,0,5,0,0,0,0,1,0,0,0]^T$
32	Koukouvinos et al. [105]	34	18	$[0,0,5,0,0,-4,0,-3,-4,3,0,0,0,0,0,0,0,-2,-2,0,0,-6,0,0,4,0,-5,0,0,0,0,0,0,-2,0]^T$
33	Koukouvinos et al. [105]	34	18	$[3,0,0,0,0,0,0,0,0,0,0,0,-9,0,0,0,7,0,0,0,0,0,0,-9,0,0,0,0,0,0,0,0,-8,0,0]^T$
34	Koukouvinos et al. [105]	34	18	$[0,3,0,0,0,0,0,0,0,0,0,0,-18,0,0,0,0,7,0,0,0,16,0,0,0,0,0,0,0,-1,-1,0,0,0,0,0]^T$
35	Koukouvinos et al. [105]	38	20	$[1,2,0,0,4,0,0,0,7,0,0,11,0,0,14,0,0,0,21,0,0,0,16,0,0,0,0,17,0,0,0,21,0,0,0,17,0,0]^T$
36	Koukouvinos et al. [105]	38	20	$[0,0,0,0,0,0,0,0,0,0,0,0,3,-10,0,0,0,0,0,0,0,0,5,6,0,0,0,0,0,0,0,0,0,0,0,0,9,0,0]^T$
37	Koukouvinos et al. [105]	38	20	$[0,0,0,0,0,0,0,0,0,2,0,6,0,4,0,0,0,0,0,0,0,9,0,10,0,0,0,0,0,0,0,0,0,-1,0,6,0,-3,0,0]^T$
38	Koukouvinos et al. [105]	38	20	$[0,18,0,0,11,0,9,0,0,0,0,0,0,0,18,0,17,0,0,0,0,0,0,0,0,0,13,0,0,0,0,0,0,0,0,0,0,0,0,7,0]^T$
39	Koukouvinos et al. [105]	42	22	$[0,0,0,0,-3,0,0,-2,0,0,0,0,0,1,-1,0,0,0,0,-2,0,0,0,5,0,-3,0,0,0,0,-3,1,0,0,0,0,0,-4,1,-1,3,0,0]^T$
40	Koukouvinos et al. [105]	42	22	$[0,0,0,0,-2,0,0,0,0,1,0,0,0,0,0,0,-4,0,0,0,0,0,0,0,0,0,0,0,0,-5,0,0,0,0,0,0,0,3,0,0,0,0,0,0]^T$
41	Koukouvinos et al. [105]	42	22	$[0,0,11,0,0,0,0,0,-13,0,0,0,13,6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,6,0,0,0,0,0,0,0,0,0,0]^T$
42	Koukouvinos et al. [105]	42	22	$[0,0,0,0,0,-15,0,0,2,0,0,0,6,-2,0,0,0,0,0,0,0,-15,0,0,5,0,0,12,0,-10,0,0,0,0,0,0,0,0,0,0,14,0,7]^T$

Πίνακας 7.4: Απόδοση της προτεινόμενης μεθόδου για τα μοντέλα 1-42

Μοντέλο	Type I	Type II	**Model**	Type I	Type II
1	0.28	0.00	**22**	0.26	0.14
2	0.33	0.02	**23**	0.26	0.19
3	0.25	0.00	**24**	0.28	0.02
4	0.21	0.16	**25**	0.28	0.18
5	0.28	0.18	**26**	0.22	0.02
6	0.30	0.16	**27**	0.12	0.10
7	0.22	0.11	**28**	0.18	0.04
8	0.14	0.00	**29**	0.12	0.18
9	0.00	0.00	**30**	0.22	0.14
10	0.11	0.15	**31**	0.16	0.24
11	0.27	0.08	**32**	0.09	0.23
12	0.19	0.03	**33**	0.19	0.23
13	0.20	0.24	**34**	0.17	0.22
14	0.25	0.22	**35**	0.23	0.25
15	0.00	0.14	**36**	0.16	0.00
16	0.21	0.16	**37**	0.17	0.15
17	0.14	0.01	**38**	0.24	0.15
18	0.30	0.00	**39**	0.20	0.17
19	0.21	0.11	**40**	0.26	0.09
20	0.15	0.23	**41**	0.11	0.00
21	0.24	0.21	**42**	0.12	0.17



Σχήμα 7.4: Σφάλματα Τύπου I και II της μεθόδου PageRank για απόκριση κατανομής Poisson

Από τον παραπάνω πίνακα καθώς και το σχετικό Σχήμα 7.4 που παρατίθεται, παρατηρούμε καταρχήν ότι, αναφορικά με την προτεινόμενη μέθοδο, και τα δύο είδη σφαλμάτων παραμένουν σε χαμηλά επίπεδα, γεγονός που ισχύει σε πολλά από τα θεωρούμενα μοντέλα. Οι μέσες τιμές των σφαλμάτων Τύπου I και II είναι ίσες με 0.20 και 0.12 αντίστοιχα. Όπως και στην περίπτωση Bernoulli απόκρισης, τα σφάλματα Τύπου II είναι μικρότερα από αυτά Τύπου I, για την πλειοψηφία των περιπτώσεων (30/42). Συνεπώς, η απόδοση της μεθόδου κρίνεται αρκετά ικανοποιητική, ακόμα και για δεδομένα από την κατανομή Poisson.

7.5 Συμπεράσματα

Η ανάλυση των υπερκορεσμένων σχεδιασμών βρίσκεται ακόμα σε πρώιμο ερευνητικό στάδιο, ιδίως για την περίπτωση όπου χρησιμοποιούνται γενικευμένα γραμμικά μοντέλα. Σε αυτό το κεφάλαιο, εισαγάγαμε μια νέα μέθοδο η οποία μπορεί να χρησιμοποιηθεί για την επιλογή των σημαντικών παραγόντων, όταν η μεταβλητή απόκρισης είναι διακριτή και συγκεκριμένα για απόκριση κατανομής Bernoulli ή Poisson. Καινοτομία της μεθόδου αποτελεί η τροποποίηση του αλγορίθμου PageRank, ο οποίος ακόμα και σήμερα είναι ένας ιδιαίτερα αποδοτικός αλγόριθμος βαθμιδότησης των ιστοσελίδων στη μηχανή αναζήτησης Google, συνδυάζοντάς τον με ένα κατάλληλο μέτρο από τη Θεωρία Πληροφορίας.

Έπειτα από μια εκτενή μελέτη προσομοίωσης που εκτελέσαμε, τα αποτελέσματα προέκυψαν ιδιαίτερα ενθαρρυντικά. Συγκεκριμένα, η νέα μέθοδος επιτυγχάνει το στόχο της διατήρησης των σφαλμάτων Τύπου I και Τύπου II σε αρκετά χαμηλά επίπεδα. Επιπλέον, υπερέχει δύο εναλλακτικών αλγορίθμων βαθμιδότησης και επιλογής μεταβλητών, τους CMIM και mRMR, οι οποίοι και μελετήθηκαν στις προσομοιώσεις. Να τονίσουμε επίσης ότι δώσαμε έμφαση στη συγκεκριμένη κατηγορία σχεδιασμών, λόγω του ότι δεν έχει ακόμα εξεταστεί συστηματικά η χρήση τους για σκοπούς κρησαρίσματος, υπό τη σκέπη ενός γενικευμένου γραμμικού μοντέλου και διακριτών δεδομένων απόκρισης. Παρ΄ όλα αυτά, η προτεινόμενη μέθοδος είναι εφαρμόσιμη σε οποιοδήποτε διακριτό σύνολο δεδομένων ή πειραματικό σχεδιασμό.

ΚΕΦΑΛΑΙΟ **8**

Μέθοδος Ανάλυσης Μη Επαναλαμβανόμενων Παραγοντικών Σχεδιασμών με Τροποποίηση του **PageRank** Αλγορίθμου

The measure of greatness in a scientific idea
is the extent to which it stimulates thought
and opens up new lines of research.

—Paul Dirac (1902–1984)

Η μέθοδος κρησαρίσματος μέσω του PageRank αλγορίθμου, που αναπτύχθηκε στο προηγούμενο κεφάλαιο, μπορεί να εφαρμοστεί και σε άλλους πειραματικούς σχεδιασμούς, πέραν των υπερκορεσμένων. Στο παρόν κεφάλαιο, δίνουμε έμφαση στους μη επαναλαμβανόμενους παραγοντικούς σχεδιασμούς δύο επιπέδων και επεκτείνουμε τη μέθοδό μας για αυτήν την κλάση σχεδιασμών. Μας ενδιαφέρει η περίπτωση όπου η απόκριση είναι κατανομής Bernoulli, συνεπώς υποθέτουμε γενικευμένα γραμμικά μοντέλα. Παρουσιάζεται έλλειψη κατάλληλων μεθόδων κρησαρίσματος στη βιβλιογραφία, για δεδομένα που δεν προέρχονται από την Κανονική κατανομή και αποτελεί ένα φτωχά ανεπτυγμένο πεδίο έρευνας, ακόμα και για τη συγκεκριμένη κλάση σχεδιασμών. Για το σκοπό αυτό, δημιουργούμε μια νέα μέθοδο επιλογής μεταβλητών, ακολουθώντας την ίδια διαδικασία τροποποίησης του αλγορίθμου PageRank.

8.1 Ερευνητικό Πρόβλημα

Υποθέτουμε ότι μας ενδιαφέρει η διεξαγωγή ενός πειράματος που θα βασιστεί σε έναν 2^k παραγοντικό σχεδιασμό. Όταν ειναι περιορισμένοι οι διαθέσιμοι πόροι για την διεξαγωγή του πειράματος, αναγκαζόμαστε να έχουμε τη δυνατότητα μιας μόνο εκτέλεσης του 2^k σχεδιασμού. Συνεπώς, οδηγούμαστε στην περίπτωση ενός μη επαναλαμβανόμενου παραγοντικού σχεδιασμού. Λόγω του ότι οι σχεδιασμοί αυτοί δε μας δίνουν μια ανεξάρτητη εκτίμηση της διασποράς του σφάλματος, η συμβατική μέθοδος της ανάλυσης διασποράς για την έρευση των σημαντικών παραγόντων, δε μπορεί να εφαρμοσθεί.

Η πρώτη μέθοδος για την ανάλυση των μη επαναλαμβανόμενων παραγοντικών σχεδιασμών και την εύρεση των σημαντικών επιδράσεων, ήταν το διάγραμμα πιθανότητας των αντιθέσεων, η οποία προτάθηκε από τον Daniel [44]. Η μέθοδος αυτή συνίσταται στην απεικόνιση των επιδράσεων σε ένα κανονικό διάγραμμα πιθανοτήτων. Στο διάγραμμα αυτό, οι αδρανείς επιδράσεις τείνουν να βρίσκονται πάνω σε μια ευθεία γραμμή, ενώ οι ενεργές απεικονίζονται μακριά από αυτή. Παρ΄ όλα αυτά, η υποκειμενική φύση της εν λόγω μεθόδου, αποτέλεσε κίνητρο για πολλούς ερευνητές να παρέχουν περισσότερο αντικειμενικές διαδικασίες. Για μια επισκόπηση και σύγκριση πολλών μεθόδων, παραπέμπουμε τον ενδιαφερόμενο αναγνώστη στους Hamada και Balakrishnan [86]. Ας αναφέρουμε όμως ορισμένες από τις πιο σημαντικές εργασίες, οι οποίες είναι των Box και Meyer [19], Lenth [115], Dong [46], Chen και Kunert [30], Aboukalam [2], Miller [133], Voss και Wang [169], Angelopoulos και Koukouvinos [7] και Angelopoulos et al. [8, 9].

Όλες οι παραπάνω μέθοδοι για την ανάλυση των μη επαναλαμβανόμενων παραγοντικών σχεδιασμών, αναφέρονται στην περίπτωση απόκρισης Κανονικής κατανομής. Συνεπώς, είναι εμφανής η έλλειψη στη βιβλιογραφία αναφορικά με μεθόδους για μη κανονικά δεδομένα απόκρισης. Ο μεγάλος αριθμός πιθανά ενεργών επιδράσεων (κυρίων και αλληλεπιδράσεων) απαιτεί την ύπαρξη μιας αποτελεσματικής διαδικασίας, ώστε να μπορούν να χρησιμοποιηθούν αυτοί οι σχεδιασμοί για σκοπούς κρησαρίσματος. Αυτό μας ώθησε να εξετάσουμε την απόδοση του PageRank αλγορίθμου που αναπτύξαμε στο Κεφάλαιο 7, για απόκριση κατανομής Bernoulli, σε αυτήν την κατηγορία σχεδιασμών. Να σημειώσουμε, ότι κατά τη δικιά μας γνώση, είναι η πρώτη φορά που ένας τέτοιος αλγόριθμος βαθμιδότησης τροποποιείται και χρησιμοποιείται κατάλληλα για σκοπούς επιλογής μεταβλητών, στη συγκεκριμένη κλάση σχεδιασμών.

8.2 Αναζήτηση των Ενεργών Επιδράσεων σε Μη Επαναλαμβανόμενους Παραγοντικούς Σχεδιασμούς με Απόκριση Κατανομής Bernoulli

Όπως και το προηγούμενο Κεφάλαιο 7, θα χρησιμοποιήσουμε τον PageRank αλγόριθμο σε συνδυασμό με το μέτρο της δεσμευμένης αμοιβαίας πληροφορίας ώστε να τον χρησιμοποιήσουμε ως μέθοδο κρησαρίσματος στους μη επαναλαμβανόμενους παραγοντικούς σχεδιασμούς δύο επιπέδων. Μας ενδιαφέρει η περίπτωση όπου η απόκριση είναι δίτιμη, δηλαδή λαμβάνει τις τιμές 1 και 0, που αντιστοιχούν σε 'επιτυχία' και 'αποτυχία' αντίστοιχα. Συνεπώς, θεωρούμε ένα γενικευμένο γραμμικό μοντέλο. Φυσικά η διαδικασία τροποποίησης του αλγορίθμου είναι ίδια με αυτήν του προηγουμένου κεφαλαίου.

Έστω ένας μη επαναλαμβανόμενος παραγοντικός σχεδιασμός με $d = n - 1$ παράγοντες και n πειραματικές εκτελέσεις, συμπεριλαμβανομένων όλων των κύριων επιδράσεων και αλληλεπιδράσεων. Θεωρούμε ότι στο $c-$οστό πειραματικό σημείο, $c = 1, \ldots, n$ η απόκριση y_c είναι μια τυχαία μεταβλητή κατανομής Bernoulli, με $\mu_c = E(y_c) = P_c = P(\mathbf{x}_c)$, P_c η πιθανότητα επιτυχίας σε μια διαδικασία Bernoulli, $\mathbf{x}_c$ το $d-$διάστατο διάνυσμα προβλεπουσών μεταβλητών και $Var(y_c) = P_c(1 - P_c)$ η διασπορά της απόκρισης. Το Λογιστικό μοντέλο παλινδρόμησης

για τη μέση απόκριση $P(\mathbf{x}_c)$ ορίζεται ως

$$P(\mathbf{x}_c) = \frac{1}{1 + e^{-\mathbf{x}_c^T \boldsymbol{\beta}}}, \tag{8.1}$$

όπου ο όρος $\mathbf{x}_c^T \boldsymbol{\beta}$ είναι η γραμμική προβλέπουσα και $\boldsymbol{\beta}$ ένα διάνυσμα διάστασης d των συντελεστών παλινδρόμησης.

8.2.1 Η Προτεινόμενη Μέθοδος

Η διαδικασία κρησαρίσματος που προτείνουμε, εφαρμόζεται στους μη επαναλαμβανόμενους παραγοντικούς σχεδιασμούς, υπό την υπόθεση ενός γενικευμένου γραμμικού μοντέλου, του οποίου τα δεδομένα απόκρισης προέρχονται από την κατανομή Bernoulli και περιγράφεται ρητά ως εξής:

1. Δοθέντος ενός $n \times d$ μη επαναλαμβανόμενου (πλήρους) παραγοντικού σχεδιασμού $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d]$, όπου $\mathbf{x}_l$, $l = 1, 2, \ldots, d$, η $l-$οστή στήλη του καθώς επίσης και ενός $n \times 1$ διανύσματος απόκρισης $\mathbf{y}$, κατανομής Bernoulli, υπολογίζουμε τη δεσμευμένη αμοιβαία πληροφορία $I(\mathbf{x}_i, \mathbf{y}|\mathbf{x}_j)$ για κάθε $i = 1, \ldots, d$, $j = 1, \ldots, d$ με $i \neq j$,

2. Κατασκευάζεται κατ΄ αυτόν τον τρόπο ο $d \times d$ πίνακας $\mathbf{M}$ όπου κάθε στοιχείο του M_{ji}, στη γραμμή j και τη στήλη i αντιστοιχεί στην παραπάνω αμοιβαία πληροφορία μεταξύ του παράγοντα $\mathbf{x}_i$ και της απόκρισης $\mathbf{y}$, δεδομένου του παράγοντα $\mathbf{x}_j$.

3. Ο πίνακας $\mathbf{M}$ μετατρέπεται σε στοχαστικό κατά γραμμή και δίνεται ως είσοδο στον αλγόριθμο PageRank, οπότε και υλοποιείται η Power μέθοδος.

4. Προκύπτει ο θετικός, στοχαστικός και μη αναγώγιμος πίνακας $\tilde{\mathbf{M}}$ έπειτα από τη χρήση του παράγοντα απόσβεσης $\alpha = 0.85$.

5. Λαμβάνουμε το τελικό PageRank διάνυσμα, που θα περιέχει τις τιμές - βαθμούς για κάθε παράγοντα.

6. Αναγνωρίζονται και επιλέγονται ως σημαντικοί οι παράγοντες με τιμές μεγαλύτερες του γεωμετρικού μέσου των τιμών του PageRank διανύσματος.

8.3 Πειράματα Προσομοίωσης

Για την αξιολόγηση της προτεινόμενης μεθόδου, εκτελέσαμε προσομοιώσεις για ένα ευρύ φάσμα μοντέλων. Για συγκριτικούς σκοπούς εξετάσαμε επίσης τον Conditional Mutual Information Maximization (CMIM) αλγόριθμο που πρότεινε ο Fleuret [74] μαζί με τον minimal-redundancy-maximal-relevance feature selection (mRMR) αλγόριθμο των Peng et al. [142]. Ως κριτήρια αξιολόγησης της απόδοσης, θεωρήσαμε και εδώ τα σφάλματα Τύπου I και II.

8.3.1 Σχεδιασμός Προσομοιώσεων

Χρησιμοποιήσαμε δύο μη επαναλαμβανόμενους παραγοντικούς σχεδιασμούς στις προσομοιώσεις. Έναν 2^4 και έναν 2^5 πλήρη παραγοντικό σχεδιασμό. Ο λόγος για την επιλογή των συγκεκριμένων σχεδιασμών, είναι ότι χρησιμοποιούνται ευρέως από τους ερευνητές για την αξιολόγηση μεθόδων ανάλυσης παραγοντικών σχεδιασμών. Για τους υπό εξέταση σχεδιασμούς, οι πραγματικά ενεργές επιδράσεις επιλέχθηκαν βάσει δύο σεναρίων. Επίσης, για κάθε σχεδιασμό και για κάθε αριθμό ενεργών παραγόντων, παρήχθησαν τυχαία 1000 διανύσματα απόκρισης $\mathbf{y} \sim Bernoulli(P(\mathbf{x}^T \boldsymbol{\beta}))$ όπου $P(u) = \frac{1}{1+e^{-u}}$.

Σενάριο Ι: Δημιουργήσαμε λογιστικά μοντέλα με συντελεστές τυχαία επιλεγμένους από -10 έως 10. Όταν ένας παραγόμενος συντελεστής ήταν 'σχεδόν μηδέν' τότε έγινε αντικατάστασή του από το 50% του μέγιστου συντελεστή. Επιπλέον, οι πραγματικά ενεργές επιδράσεις επιλέχθηκαν τυχαία, σύμφωνα με την Ομοιόμορφη κατανομή, από το σύνολο των $1, \ldots, d$ δυνητικά ενεργών παραγόντων, αναφορικά με τον ήδη καθορισμένο αριθμό ενεργών μεταβλητών στον πίνακα σχεδιασμού. Οι συντελεστές των μη ενεργών μεταβλητών, στο πραγματικό μοντέλο, τέθηκαν ίσες με μηδέν. Να σημειώσουμε, ότι το πλήθος των πραγματικά ενεργών μεταβλητών, δεν ξεπέρασε την τιμή $d/2$, βάσει της αρχής της σποραδικότητας των επιδράσεων [19].

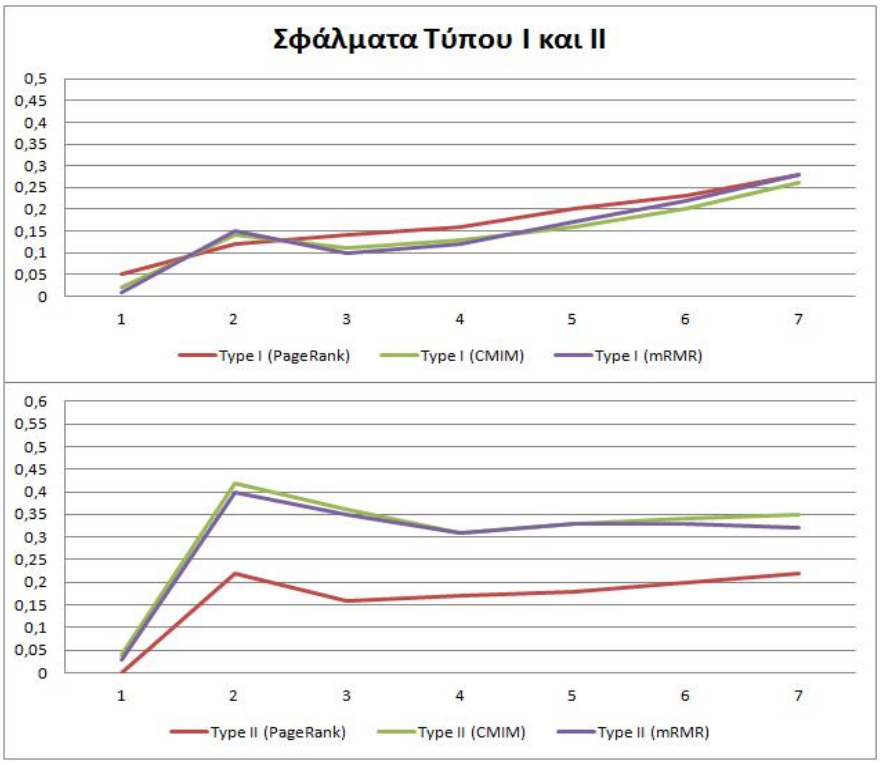
Σενάριο ΙΙ: Δημιουργήσαμε λογιστικά μοντέλα με συντελεστές που έλαβαν προκαθορισμένες τιμές. Για να εξετάσουμε το πόσο ευαίσθητα είναι τα αποτελέσματα στην επιλογή και στον αριθμό των ενεργών παραγόντων, αλλάξαμε τη διάταξη των ενεργών μεταβλητών, χρησιμοποιήσαμε διαφορετικές τιμές για τους συντελεστές β και διαφορετικό πλήθος ενεργών παραγόντων, σε κάθε σχεδιασμό. Συνεπώς, εξετάσαμε ένα ευρύ φάσμα διαφορετικών μοντέλων.

8.3.2 Αποτελέσματα Προσομοιώσεων

Στους παρακάτω πίνακες και στα σχετικά σχήματα, παρουσιάζονται τα αποτελέσματα από την εφαρμογή της προτεινόμενης μεθόδου, καθώς και των μεθόδων CMIM και mRMR. Οι πρώτοι δύο Πίνακες 8.1 και 8.2, καθώς και τα αντίστοιχα Σχήματα 8.1 και 8.2, αναφέρονται στο Σενάριο Ι. Στην πρώτη στήλη των πινάκων, παρουσιάζεται ο αριθμός των πραγματικά ενεργών επιδράσεων (q) των προσομοιωμένων μοντέλων. Οι υπόλοιπες στήλες, παρουσιάζουν τα αποτελέσματα αναφορικά με τα σφάλματα Τύπου Ι και Τύπου ΙΙ, που προέκυψαν από τις 1000 προσομοιώσεις. Στη συνέχεια, οι Πίνακες 8.3-8.6 και τα σχετικά Σχήματα 8.3 και 8.4, αφορούν το Σενάριο ΙΙ. Συγκεκριμένα, οι Πίνακες 8.3 και 8.5 παρουσιάζουν τα εξεταζόμενα μοντέλα. Η πρώτη στήλη δίνει τον αύξοντα αριθμό που αντιστοιχεί στο εκάστοτε μοντέλο με προκαθορισμένες τιμές των συντελεστών, οι οποίες δίνονται στη δεύτερη στήλη. Τα ληφθέντα αποτελέσματα από την υλοποίηση των εξεταζόμενων μεθόδων, παρατίθενται στους Πίνακες 8.4 και 8.6. Στην πρώτη στήλη τους, δίνεται ο αύξων αριθμός του εκάστοτε μοντέλου και οι υπόλοιπες στήλες παρουσιάζουν τα αποτελέσματα για τα σφάλματα Τύπου Ι και Τύπου ΙΙ.

Πίνακας 8.1: 2^4 πλήρης παραγοντικός σχεδιασμός: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων για το Σενάριο Ι

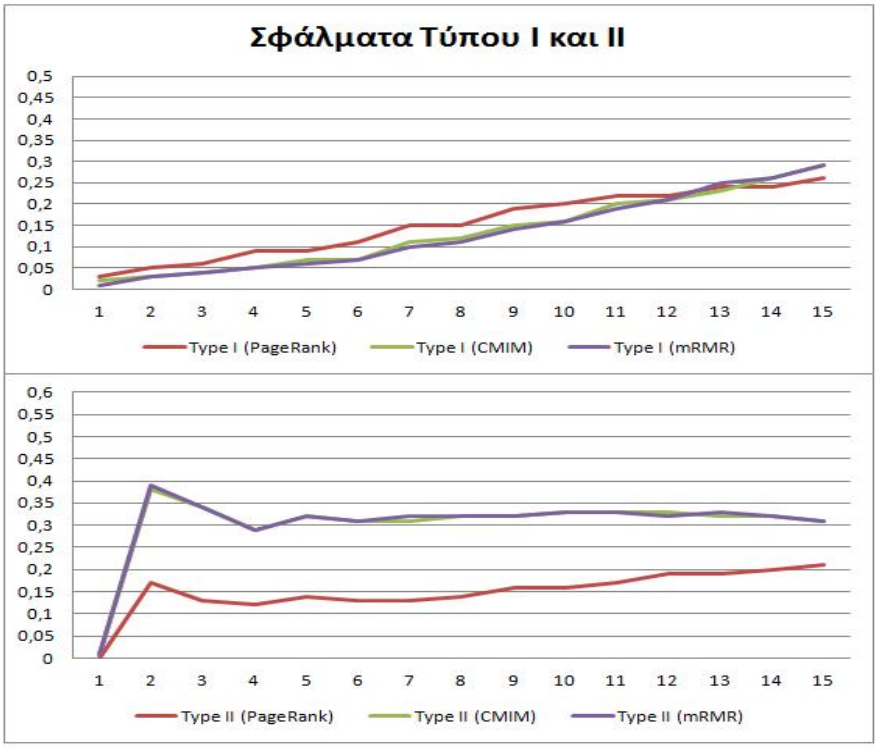
q	Type I (PageRank)	Type I (CMIM)	Type I (mRMR)	Type II (PageRank)	Type II (CMIM)	Type II (mRMR)
1	0.05	0.02	0.01	0	0.04	0.03
2	0.12	0.14	0.15	0.22	0.42	0.40
3	0.14	0.11	0.10	0.16	0.36	0.35
4	0.16	0.13	0.12	0.17	0.31	0.31
5	0.20	0.16	0.17	0.18	0.33	0.33
6	0.23	0.20	0.22	0.20	0.34	0.33
7	0.28	0.26	0.28	0.22	0.35	0.32



Σχήμα 8.1: Σύγκριση σφαλμάτων Τύπου I και Τύπου II για το Σενάριο I και τον 2^4 πλήρη παραγοντικό σχεδιασμό

Πίνακας 8.2: 2^5 πλήρης παραγοντικός σχεδιασμός: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων για το Σενάριο I

q	Type I (PageRank)	Type I (CMIM)	Type I (mRMR)	Type II (PageRank)	Type II (CMIM)	Type II (mRMR)
1	0.03	0.02	0.01	0	0.01	0.01
2	0.05	0.03	0.03	0.17	0.38	0.39
3	0.06	0.04	0.04	0.13	0.34	0.34
4	0.09	0.05	0.05	0.12	0.29	0.29
5	0.09	0.07	0.06	0.14	0.32	0.32
6	0.11	0.07	0.07	0.13	0.31	0.31
7	0.15	0.11	0.10	0.13	0.31	0.32
8	0.15	0.12	0.11	0.14	0.32	0.32
9	0.19	0.15	0.14	0.16	0.32	0.32
10	0.20	0.16	0.16	0.16	0.33	0.33
11	0.22	0.20	0.19	0.17	0.33	0.33
12	0.22	0.21	0.21	0.19	0.33	0.32
13	0.24	0.23	0.25	0.19	0.32	0.33
14	0.24	0.26	0.26	0.20	0.32	0.32
15	0.26	0.29	0.29	0.21	0.31	0.31



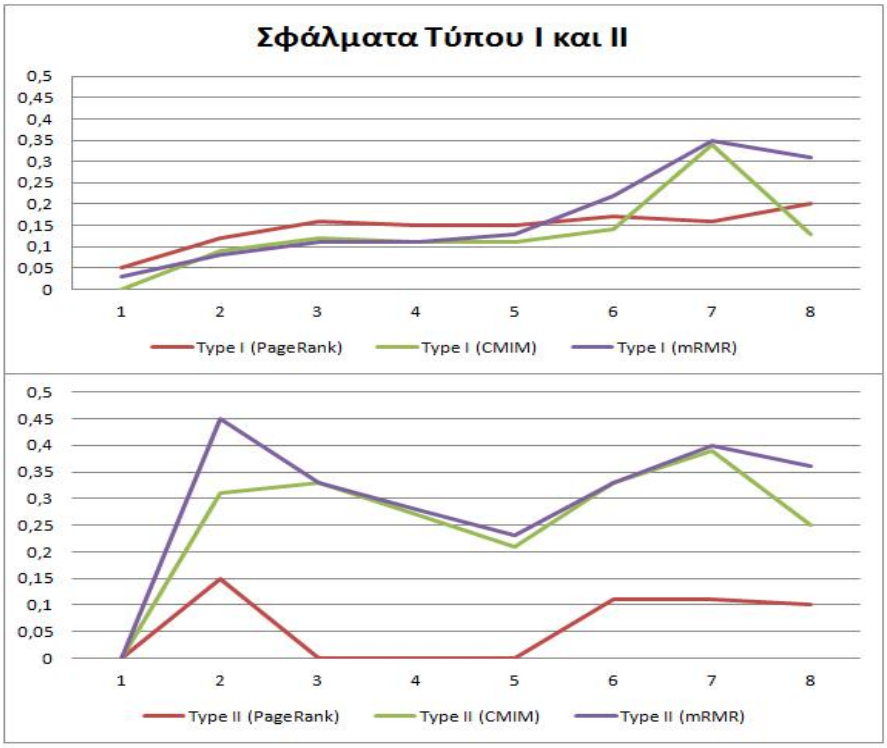
Σχήμα 8.2: Σύγκριση σφαλμάτων Τύπου I και Τύπου II για το Σενάριο I και τον 2^5 πλήρη παραγοντικό σχεδιασμό

Πίνακας 8.3: Θεωρούμενα μοντέλα προσομοιώσης, για τον 2^4 πλήρη παραγοντικό σχεδιασμό και για το Σενάριο II

Μοντέλο	Προκαθορισμένες τιμές συντελεστών
1	$[0,0,0,0,3,0,0,0,0,0,0,0,0,0,0,0]^T$
2	$[0,0,0,0,0,0,0,0,0,0,0,2,0,0,3]^T$
3	$[0,0,-7,0,0,0,0,-8,0,0,0,0,0,0,0,-6]^T$
4	$[0,0,-9,0,4,0,0,0,0,-2,0,0,0,0,10]^T$
5	$[6,0,0,0,0,7,0,0,-5,-5,0,-7,0,0,0]^T$
6	$[0,7,0,-2,0,5,2,0,4,0,0,0,-8,0,0]^T$
7	$[0,0,-9,2,0,0,0,4,5,8,0,0,-5,-7,0]^T$
8	$[0,0,0,0,2,4,8,0,0,-10,4,0,-5,3,0]^T$

Πίνακας 8.4: 2^4 πλήρης παραγοντικός σχεδιασμός: Απόδοση των μεθόδων για τα μοντέλα 1-8, με χρήση 1000 προσομοιώσεων για το Σενάριο II

Μοντέλο	Type I (PageRank)	Type I (CMIM)	Type I (mRMR)	Type II (PageRank)	Type II (CMIM)	Type II (mRMR)
1	0.05	0	0.03	0	0	0
2	0.12	0.09	0.08	0.15	0.31	0.45
3	0.16	0.12	0.11	0	0.33	0.33
4	0.15	0.11	0.11	0	0.27	0.28
5	0.15	0.11	0.13	0	0.21	0.23
6	0.17	0.14	0.22	0.11	0.33	0.33
7	0.16	0.34	0.35	0.11	0.39	0.40
8	0.20	0.13	0.31	0.10	0.25	0.36



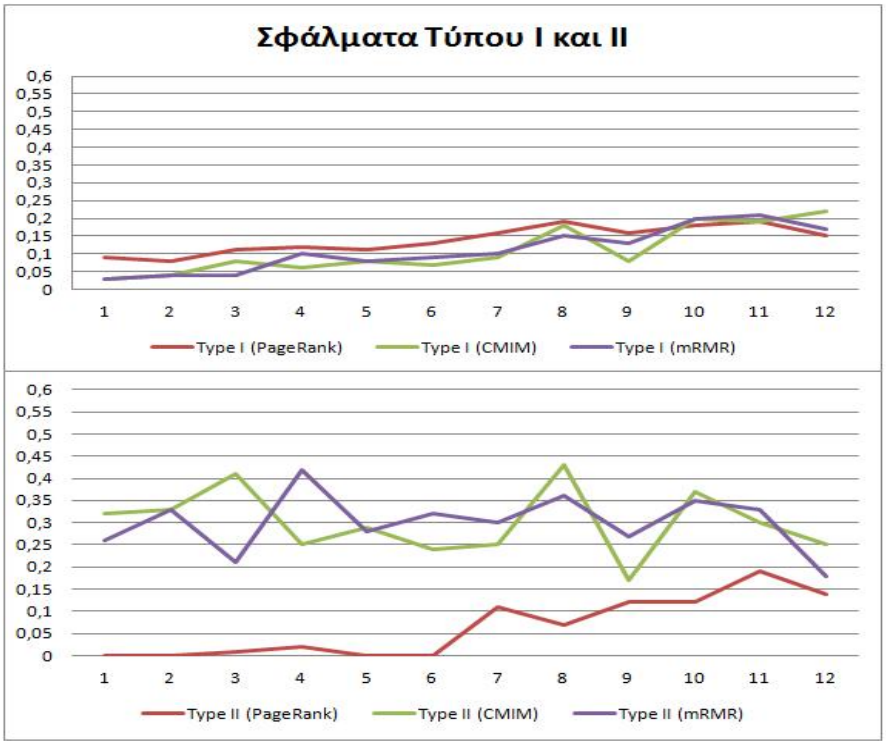
Σχήμα 8.3: Σύγκριση σφαλμάτων Τύπου I και Τύπου II για το Σενάριο II και τον 2^4 πλήρη παραγοντικό σχεδιασμό

Πίνακας 8.5: Θεωρούμενα μοντέλα προσομοιώσης, για τον 2^5 πλήρη παραγοντικό σχεδιασμό και για το Σενάριο II

Μοντέλο	Προκαθορισμένες τιμές συντελεστών
1	$[0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,-4,0,0,0,0,0,0,0,0,0,0,0,0,6,0]^T$
2	$[20,0,-17,0,0,0,0,0,0,0,0,0,12,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]^T$
3	$[7,0,5,0,0,3,0,0,0,0,0,0,0,-5,0,0,0,0,0,0,0,0,0,0,0,0,-7,0,0,0]^T$
4	$[0,0,0,17,0,0,0,0,0,0,0,8,0,0,0,-7,0,12,0,0,0,0,0,„0,3,0,0,0,0,-8,0]^T$
5	$[0,0,-9,-5,0,-9,0,0,0,0,0,0,0,0,0,0,-2,0,0,5,0,0,0,0,„4,0,0,0,0,0,0,8]^T$
6	$[0,0,0,0,0,0,5,0,0,7,0,0,0,0,7,0,0,0,5,0,0,0,0,0,0,0,0,5,9,9,0,0]^T$
7	$[0,2,4,0,0,0,0,0,0,0,0,0,2,0,3,0,0,-2,0,2,0,0,0,„0,0,0,3,2,0,0,0]^T$
8	$[5,0,4,5,0,0,0,0,0,0,0,0,0,0,0,0,0,9,5,0,4,5,0,0,0,0,0,0,0,0,0,9,6]^T$
9	$[0,0,2,-4,-3,0,0,0,0,0,0,-4,3,0,0,0,0,0,-4,0,0,0,0,0,0,0,0,0,0,4,3,0,2,-1]^T$
10	$[0,-5,0,0,0,0,0,-9,0,-7,0,0,0,0,-4,0,0,0,0,-5,-7,0,0,0,0,-2,-9,0,0,0,0,0,-3,-8,0,-5]^T$
11	$[0,7,9,9,0,0,0,0,0,0,17,0,0,0,10,7,19,0,0,0,0,0,10,0,0,0,14,13,0,0,0,0,3,-10,0]^T$
12	$[0,3,0,-2,0,0,0,0,0,1,0,0,-4,-3,2,0,0,0,0,-5,1,4,0,-3,0,2,0,3,0,2,2,4,0]^T$

Πίνακας 8.6: 2^5 πλήρης παραγοντικός σχεδιασμός: Απόδοση των μεθόδων για τα μοντέλα 1-12, με χρήση 1000 προσομοιώσεων για το Σενάριο II

Μοντέλο	Type I (PageRank)	Type I (CMIM)	Type I (mRMR)	Type II (PageRank)	Type II (CMIM)	Type II (mRMR)
1	0.09	0.03	0.03	0	0.32	0.26
2	0.08	0.04	0.04	0	0.33	0.33
3	0.11	0.08	0.04	0.01	0.41	0.21
4	0.12	0.06	0.10	0.02	0.25	0.42
5	0.11	0.08	0.08	0	0.29	0.28
6	0.13	0.07	0.09	0	0.24	0.32
7	0.16	0.09	0.10	0.11	0.25	0.30
8	0.19	0.18	0.15	0.07	0.43	0.36
9	0.16	0.08	0.13	0.12	0.17	0.27
10	0.18	0.20	0.20	0.12	0.37	0.35
11	0.19	0.19	0.21	0.19	0.30	0.33
12	0.15	0.22	0.17	0.14	0.25	0.18



Σχήμα 8.4: Σύγκριση σφαλμάτων Τύπου I και Τύπου II για το Σενάριο II και τον 2^5 πλήρη παραγοντικό σχεδιασμό

Από τους παραπάνω πίνακες και τα σχετικά σχήματα, καταλήγουμε καταρχήν στο συμπέρασμα ότι η προτεινόμενη μέθοδος τείνει να δηλώνει σε υψηλότερο ποσοστό τις μη ενεργές επιδράσεις ως ενεργές και σε πολύ χαμηλότερο ποσοστό τις ενεργές επιδράσεις ως ανενεργές. Συνεπώς, μπορεί να θεωρηθεί συντηρητική υπό αυτήν την έννοια.

Αναφορικά με το Σενάριο Ι, η προτεινόμενη μέθοδος δίνει παρόμοια σφάλματα Τύπου Ι με τους αλγορίθμους CMIM και mRMR, ακόμα και όταν αυξάνει το πλήθος των πραγματικά ενεργών επιδράσεων, τα οποία διατηρούνται σε χαμηλά επίπεδα. Σε ότι αφορά τα σφάλματα Τύπου ΙΙ, η νέα μέθοδος σαφώς υπερέχει των άλλων αλγορίθμων, δίνοντας σημαντικά χαμηλότερες τιμές σφαλμάτων σε όλες τις περιπτώσεις.

Αναφορικά με το Σενάριο ΙΙ, για τον 2^4 πλήρη παραγοντικό σχεδιασμό, οι μέσες τιμές των σφαλμάτων Τύπου Ι και ΙΙ για τη νέα μέθοδο είναι 0.14 και 0.06 αντίστοιχα. Συνεπώς, διατηρούνται σε πολύ χαμηλά επίπεδα. Για τον αλγόριθμο CMIM έχουμε τις μέσες τιμές 0.13 και 0.26 αντίστοιχα, ενώ για τον mRMR, οι μέσες τιμές των σφαλμάτων Τύπου Ι και ΙΙ είναι 0.17 και 0.30. Όσον αφορά τον 2^5 πλήρη παραγοντικό σχεδιασμό, οι μέσες τιμές των σφαλμάτων Τύπου Ι, είναι 0.14, 0.11 και 0.11 για την προτεινόμενη μέθοδο, τον αλγόριθμο CMIM και τον mRMR αντίστοιχα, ενώ για τα σφάλματα Τύπου ΙΙ έχουμε τις τιμές 0.06, 0.30 και 0.30.

Συνεπώς, η προτεινόμενη διαδικασία κρησαρίσματος, σε σύγκριση με τους αλγορίθμους CMIM και mRMR, δίνει ελαφρώς υψηλότερα σφάλματα Τύπου Ι σε ορισμένες μόνο περιπτώσεις, παρουσιάζει όμως γενικά παρόμοια απόδοση. Ωστόσο, τα σφάλματα Τύπου ΙΙ, είναι αρκετά υψηλότερα τόσο για τον CMIM όσο και τον mRMR αλγόριθμο, σε όλες τις περιπτώσεις και τα εξεταζόμενα σενάρια, οδηγώντας σε μια ιδιαίτερα ασταθή απόδοση. Συνεπώς, η νέα μέθοδος διαθέτει πολύ ικανοποιητική ισχύ, γεγονός που είναι αδιαμφισβήτητα απαραίτητο για μια αποτελεσματική διαδικασία κρησαρίσματος. Παρουσιάζει επίσης μια γενικά σταθερή απόδοση δεδομένου ότι προέκυψαν ταυτόχρονα χαμηλές τιμές και για τα δύο είδη σφαλμάτων.

8.4 Συμπεράσματα

Η δημιουργία αποδοτικών διαδικασιών κρησαρίσματος στους μη επαναλαμβανόμενους σχεδιασμούς, είναι ένα σημαντικό ερευνητικό πρόβλημα, ειδικά υπό τη σκέπη ενός γενικευμένου γραμμικού μοντέλου και διακριτών δεδομένων απόκρισης. Για το σκοπό αυτό, επεκτείναμε τη μέθοδο επιλογής μεταβλητών που παρουσιάσαμε στο Κεφάλαιο 7, σε αυτήν την κατηγορία σχεδιασμών, με τη μεταβλητή απόκρισης να προέρχεται από την κατανομή Bernoulli. Η καινοτομία της νέας μεθόδου, έγκειται στην τροποποίηση και χρήση του γνωστού PageRank αλγορίθμου βαθμιδότησης, ο οποίος συνδυάζεται με ένα κατάλληλο μέτρο από τη Θεωρία Πληροφορίας. Η εμπειρική απόδοση της μεθόδου, βάσει μιας εκτενούς συγκριτικής μελέτης προσομοίωσης, υποδεικνύει ότι η νέα μέθοδος μπορεί να θεωρηθεί αρκετά αξιόπιστη, δίνοντας χαμηλά σφάλματα Τύπου Ι και ΙΙ και παρουσιάζοντας μια σταθερή συμπεριφορά. Οι δύο εναλλακτικοί αλγόριθμοι που χρησιμοποιήθηκαν στις προσομοιώσεις, CMIM και mRMR, δεν έδωσαν καλά αποτελέσματα. Παρουσίασαν ιδιαίτερα υψηλές τιμές σφαλμάτων Τύπου ΙΙ, παραλείποντας κατ΄ αυτόν τον τρόπο σημαντικούς παράγοντες, ενώ η απόδοσή τους χαρακτηρίζεται από μια γενική αστάθεια.

ΚΕΦΑΛΑΙΟ **9**

Μέθοδος Ανάλυσης Ομοιόμορφων Σχεδιασμών Βασισμένη στα Ποινικοποιημένα Ελάχιστα Τετράγωνα

Science is what we understand
well enough to explain to a computer.
Art is everything else we do.

—Donald Knuth (1996))

Ως μια σημαντική κλάση σχεδιασμών γεμίσματος του χώρου (space filling designs), οι ομοιόμορφοι σχεδιασμοί (uniform designs) αναζητούν τα σημεία τους να είναι ομοιόμορφα διασκορπισμένα στο πεδίο ορισμού του πειράματος, βάσει ενός μέτρου ασυμφωνίας (discrepancy measure). Από το 1980 που εμφανίστηκαν, έχουν εφαρμοστεί επιτυχώς σε πολλά βιομηχανικά και επιστημονικά πειράματα. Ένα αξιοσημείωτο πλεονέκτημά τους, είναι η ικανότητά τους να διερευνούν ένα μεγάλο αριθμό παραγόντων πολλών επιπέδων, σε συνδυασμό με ένα αρκετά οικονομικό σύνολο πειραματικών εκτελέσεων. Αποτέλεσμα αυτής της ιδιότητάς τους, είναι ότι οι ομοιόμορφοι σχεδιασμοί μπορούν να χρησιμοποιηθούν καταλλήλως ως σχεδιασμοί κρησαρίσματος που προορίζονται για την εξαγωγή των τελικά σημαντικών παραγόντων από ένα μεγάλο πλήθος δυνητικά σημαντικών. Για το σκοπό αυτό, προτείνουμε σε αυτό το κεφάλαιο μια νέα διαδικασία κρησαρίσματος στους ομοιόμορφους σχεδιασμούς, που βασίζεται στα ποινικοποιημένα ελάχιστα τετράγωνα και στη χρήση της μεθόδου $\eta_\nu(\lambda)$ του Κεφαλαίου 3.

9.1 Ερευνητικό Πρόβλημα

Όπως αναφέραμε και σε προηγούμενο κεφάλαιο, οι ομοιόμορφοι σχεδιασμοί παρέχουν μια πολύ καλή αναπαράσταση των πειραματικών σημείων με ένα λογικό αριθμό εκτελέσεων. Επιπλέον, είναι εύρωστοι ως προς την υπόθεση του μοντέλου που υποβόσκει.

Όσον αφορά τις μεθόδους κατασκευής τους, αυτές κατατάσσονται σε δύο μεγάλες κατηγορίες: Σε μεθόδους συνδυαστικής άλγεβρας και σε αλγορίθμους βελτιστοποίησης. Για μια λεπτομερή ανασκόπηση και συζήτηση πολλών μεθόδων κατασκευής ομοιόμορφων ή σχεδόν ομοιόμορφων σχεδιασμών, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο βιβλίο των Fang et al. [64] και τις εκεί αναφορές. Επίσης, σημαντικές εργασίες αποτελούν, μεταξύ άλλων, αυτές των Fang και Wang [73], Hickernell [89], Fang και Mukerjee [72], Fang και Lin [65], Fang et al. [68], Fang και Ma [70], Fang [62], Hickernell και Liu [90], Fang [63] και Fang et al. [69].

Σε αντίθεση με την ευρεία μελέτη των μεθόδων κατασκευής των ομοιόμορφων σχεδιασμών, οι μέθοδοι ανάλυσής τους είναι ακόμα σε πρώιμο ερευνητικό στάδιο. Καθώς πολλοί παράγοντες αποτελούν την κλάση των πιθανά ενεργών, λόγω της μορφής των σχεδιασμών αυτών, είναι εύλογη η ανάγκη εύρεσης μιας αποτελεσματικής μεθόδου επιλογής μεταβλητών. Χάρη στην ικανότητα των ομοιόμορφων σχεδιασμών να διερευνούν πολλούς παράγοντες με μεγάλο πλήθος επιπέδων, ενώ συγχρόνως οι πειραματικές εκτελέσεις είναι αρκετά λίγες, οι σχεδιασμοί αυτοί μπορούν κάλλιστα να χρησιμοποιηθούν ως σχεδιασμοί κρησαρίσματος.

Συνεπώς, ο σκοπός του παρόντος κεφαλαίου, είναι η δημιουργία μιας διαδικασίας εύρεσης των σημαντικών παραγόντων στους ομοιόμορφους σχεδιασμούς, μέσω της χρήσης της μεθοδολογίας ποινικοποιημένων ελαχίστων τετραγώνων των Fan και Li [53]. Η διαδικασία που προτείνουμε, συνδυάζεται με τη μέθοδο επιλογής της ρυθμιστικής παραμέτρου λ, που προτάθηκε στο Κεφάλαιο 3.

9.2 Η Προτεινόμενη Μέθοδος

Έστω ότι η σχέση μεταξύ των ανεξάρτητων μεταβλητών και της μεταβλητής απόκρισης προσεγγίζεται από το παρακάτω μοντέλο παλινδρόμησης

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{9.1}$$

όπου $\mathbf{y}$ είναι ένα $n \times 1$ διάνυσμα απόκρισης, $\mathbf{X}$ είναι ένας $n \times d$ πίνακας σχεδιασμού και $\boldsymbol{\varepsilon}= (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ είναι το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων με $\varepsilon_i \sim N(0, \sigma^2)$ $\forall$ $i = 1, 2, \ldots, n$. Εισάγουμε επίσης και ένα $n-$διάστατο διάνυσμα με μονάδες που αντιστοιχεί στο γενικό μέσο, συνεπώς η διάσταση του πίνακα σχεδιασμού γίνεται $n \times (d + 1)$.

Στο σημείο αυτό, να υπενθυμίσουμε ότι στην εργασία των Fan και Li [53], η μορφή των ποινικοποιημένων ελαχίστων τετραγώνων δίνεται ως

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{9.2}$$

όπου $p_\lambda(.)$ είναι μια συνάρτηση ποινής και λ είναι η άγνωστη ρυθμιστική παράμετρος. Με μια κατάλληλη αρχική τιμή $\boldsymbol{\beta}^{(0)}$, η λύση βρίσκεται υπολογίζοντας επαναληπτικά την ακόλουθη ποσότητα

$$\boldsymbol{\beta}^{(1)} = [\mathbf{X}^T\mathbf{X} + n\sum\nolimits_\lambda(\boldsymbol{\beta}^{(0)})]^{-1}\mathbf{X}^T\mathbf{y}, \tag{9.3}$$

όπου $\sum_\lambda(\boldsymbol{\beta}^{(0)}) = \text{diag}\{p'_\lambda(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \ldots, p'_\lambda(|\beta_{d^*}^{(0)}|)/|\beta_{d^*}^{(0)}|\}$. Επίσης, στο Κεφάλαιο 3, προτείναμε τη χρήση του κριτηρίου $\eta_\nu(\lambda)$ για την επιλογή της παραμέτρου λ, το οποίο ορίζεται

ως

$$\eta_\nu(\lambda) = \|\mathbf{r}_\lambda\|^{\nu-1} \cdot \|n\sum_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}\|^{5-2\nu} \cdot \|\mathbf{X}n\sum_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}^{(1)}\|^{\nu-3}, \nu \in \mathbb{R}, \tag{9.4}$$

όπου $\mathbf{r}_\lambda = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(1)}$. Στην περίπτωση όπου οι προκύπτουσες ποινές είναι όλες μηδέν, αντί της (9.4), ελαχιστοποιούμε την αρχική της έκφραση

$$\eta_\nu(\lambda) = \|\mathbf{r}_\lambda\|^{\nu-1} \cdot \|\mathbf{X}^T\mathbf{r}_\lambda\|^{5-2\nu} \cdot \|\mathbf{X}\mathbf{X}^T\mathbf{r}_\lambda\|^{\nu-3}, \nu \in \mathbb{R}. \tag{9.5}$$

Η νέα διαδικασία κρησαρίσματος που προτείνουμε στην ενότητα αυτή, περιγράφεται ρητά ως εξής

1. Για την αποφυγή ασταθών υπολογισμών, κανονικοποιούμε όλες τις x_i μεταβλητές, $i = 1, \ldots, d$.

2. Ξεκινάμε με ένα στάδιο προεπεξεργασίας, μέσω της εφαρμογής της μεθόδου επιλογής μεταβλητών κατά βήματα, στο πλήρες μοντέλο, χρησιμοποιώντας μεγάλη τιμή του επιπέδου σημαντικότητας α για τις F-enter και F-remove. Το προκύπτων $\boldsymbol{\beta}$ θα χρησιμοποιηθεί και ως αρχική τιμή $\boldsymbol{\beta}^{(0)}$ κατά τη εφαρμογή των ποινικοποιημένων μεθόδων στη συνέχεια.

3. Δίνουμε ένα σύνολο από αρχικές τιμές για την παράμετρο λ.

4. Βάσει του επιλεγμένου μοντέλου από το βήμα 2, βρίσκουμε έναν ποινικοποιημένο εκτιμητή $\boldsymbol{\beta}^{(1)}$ για κάθε λ στο σύνολο του βήματος 3, και εν συνεχεία υπολογίζουμε την (9.4).

5. Επιλέγεται η τιμή του λ που ελαχιστοποιεί την (9.4) και ο προκύπτων εκτιμητής είναι ο τελικός εκτιμητής του $\boldsymbol{\beta}$.

Παρατήρηση 9.1 *Όπως σημειώνεται στην εργασία [53], μια καλή αρχική τιμή $\boldsymbol{\beta}^{(0)}$ που να είναι κοντά στην πραγματική τιμή του $\boldsymbol{\beta}$ είναι απαραίτητη για τον επαναληπτικό υπολογισμό της λύσης (9.3). Ο εκτιμητής ελαχίστων τετραγώνων μπορεί να χρησιμεύσει ως μία αρχική τιμή, ειδικά όταν ο πίνακας σχεδιασμού είναι πλήρους βαθμού. Ωστόσο, λόγω του μεγάλου αριθμού των δυνητικά ενεργών παραγόντων στους ομοιόμορφους σχεδιασμούς, είναι προτιμότερο να χρησιμοποιηθεί η μέθοδος της κατά βήματα επιλογής μεταβλητών, έτσι ώστε να οδηγηθούμε σε μια πρώτη μείωσή τους. Επιπλέον, ο στόχος είναι να συμπεριληφθούν όλοι οι σημαντικοί παράγοντες σε αυτό το βήμα, επομένως πρέπει να χρησιμοποιηθεί μία μεγάλη τιμή του επιπέδου σημαντικότητας α. Προφανώς, κάποιοι μη σημαντικοί παράγοντες μπορεί να παραμείνουν στο μοντέλο κατά αυτό το στάδιο. Βάσει των προσομοιώσεων που πραγματοποιήσαμε, παρατηρήσαμε ότι η χρήση της μεθόδου επιλογής μεταβλητών κατά βήματα, οδήγησε σε καλύτερα αποτελέσματα, σε σύγκριση με τη χρήση της μέθόδου ελαχίστων τετραγώνων. Όσον αφορά την τιμή του α, εξετάσαμε πολλές υποψήφιες τιμές στο διάστημα* [0.05, 0.20] *και τα καλύτερα αποτελέσματα αποκτήθηκαν με χρήση του $\alpha = 0.20$ για τις F-enter και F-remove.*

9.3 Πειράματα Προσομοίωσης

Μια διεξοδική μελέτη προσομοίωσης πραγματοποιήθηκε προκειμένου να αξιολογηθεί η απόδοση της προτεινόμενης μεθόδου. Για λόγους σύγκρισης, χρησιμοποιήσαμε επίσης τη συμβατική τεχνική της γενικευμένης διασταυρωμένης επικύρωσης για την επιλογή της ρυθμιστικής παραμέτρου λ στο βήμα 4 της διαδικασίας (βλ. επίσης [53]).

9.3.1 Σχεδιασμός Προσομοιώσεων

Στις προσομοιώσεις συμπεριλάβαμε ομοιόμορφους σχεδιασμούς με n εκτελέσεις και d παράγοντες με $q = n$ επίπεδα, τους οποίους συμβολίζουμε ως $U_n(n^d)$ και μπορούν να βρεθούν στην ιστοσελίδα http://sites.stat.psu.edu/∼rli/DMCE/UniformDesign/Un_nˆs.html. Συγκεκριμένα, οι σχεδιασμοί αυτοί είναι οι εξής: $U_{14}(14^{10})$, $U_{19}(19^{14})$, $U_{22}(22^{16})$, $U_{23}(23^{19})$, $U_{26}(26^{21})$, $U_{30}(30^{24})$ και $U_{30}(30^{26})$. Οι προαναφερθέντες σχεδιασμοί είναι κατασκευασμένοι βάσει του μέτρου της centered L_2-ασυμφωνίας και με χρήση του Threshold Accepting αλγορίθμου (βλ. για παράδειγμα τις εργασίες [71], [176] και [177]). Δημιουργήσαμε γραμμικά μοντέλα κύριων επιδράσεων με συντελεστές που παίρνουν τιμές τυχαία επιλεγμένες από -5 έως 5. Όταν ένας παραγόμενος συντελεστής ήταν 'σχεδόν μηδέν' τότε έγινε αντικατάστασή του από το 50% του μέγιστου συντελεστή. Ένα τυχαίο σφάλμα $\varepsilon_i \sim N(0, 1)$ για όλα τα $i = 1, 2, \ldots, n$ προστέθηκε σε κάθε αντίστοιχη παρατήρηση y_i. Μόνο μοντέλα κυρίων επιδράσων θεωρήθηκαν στις προσομοιώσεις. Επιπλέον, οι πραγματικά ενεργές επιδράσεις επιλέχθηκαν τυχαία, σύμφωνα με την Ομοιόμορφη κατανομή, από το σύνολο των $1, \ldots, d$ πιθανά ενεργών παραγόντων, αναφορικά με τον ήδη καθορισμένο αριθμό ενεργών μεταβλητών στον πίνακα σχεδιασμού. Οι συντελεστές των μη ενεργών μεταβλητών, στο πραγματικό μοντέλο, τέθηκαν ίσες με μηδέν.

Όπως αναφέραμε και παραπάνω, καθορίσαμε το επίπεδο σημαντικότητας α να είναι 0.20 για τις F-enter και F-remove στη διαδικασία της κατά βήματα επιλογής μεταβλητών, με σκοπό την εύρεση μιας αρχικής τιμής στις μεθόδους των ποινικοποιημένων ελαχίστων τετραγώνων. Οι συναρτήσεις ποινής που εφαρμόστηκαν, είναι η SCAD, L_1 (LASSO) και Hard. Επιπλέον, η τιμή $\nu = 4$ χρησιμοποιήθηκε κατά τον υπολογισμό της $\eta_\nu(\lambda)$ σε όλες τις προσομοιώσεις. Τα κριτήρια βάσει των οποίων αξιολογήσαμε τη προτεινόμενη μέθοδο, ήταν τα ποσοστά σφάλματος Τύπου I (Type I) και Τύπου II (Type II). Στους σχεδιασμούς κρησαρίσματος, υπάρχει πάντα το κόστος της δήλωσης ενός ανενεργού παράγοντα ως ενεργού (σφάλμα Τύπου I) και το κόστος δήλωσης μιας ενεργής επίδρασης ως ανενεργή (σφάλμα Τυπου II). Συνεπώς, οι τιμές τους πρέπει να διατηρούνται όσο τον δυνατόν χαμηλότερες.

9.3.2 Αποτελέσματα Προσομοιώσεων

Στους παρακάτω Πίνακες 9.1-9.7, παρουσιάζουμε τα αποτελέσματα που προέκυψαν από τις προσομοιώσεις. Συγκεκριμένα, η πρώτη στήλη απεικονίζει το συγκεκριμένο είδος ομοιόμορφου σχεδιασμού που χρησιμοποιήθηκε και η δεύτερη το πλήθος των πραγματικά ενεργών παραγόντων (q) στα προσομοιωμένα μοντέλα. Στις επόμενες στήλες αναφέρονται τα σφάλματα Τύπου I και Τύπου II (Type I και Type II αντίστοιχα) για την προτεινόμενη διαδικασία κρησαρίσματος, αναφορικά με τις συναρτήσεις ποινής SCAD, LASSO, Hard και με τη χρήση της $\eta_\nu(\lambda)$ ως μέθοδο επιλογής της ρυθμιστικής παραμέτρου. Σε κάθε περίπτωση, παρουσιάζουμε επίσης και τα αποτελέσματα από τη χρήση της γενικευμένης διασταυρωμένης επικύρωσης (gcv). Για κάθε σχεδιασμό και για κάθε αριθμό ενεργών παραγόντων, αναπτύξαμε τυχαία 1000 γραμμικά μοντέλα, ώστε να αξιολογήσουμε τη νέα μέθοδο.

Πίνακας 9.1: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων σε έναν $U_{14}(14^{10})$

		SCAD($\eta_\nu(\lambda)$)		SCAD(gcv)		LASSO($\eta_\nu(\lambda)$)		LASSO(gcv)		Hard($\eta_\nu(\lambda)$)		Hard(gcv)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
$U_{14}(14^{10})$	1	0.14	0	0.36	0	0.29	0	0.36	0	0.11	0	0.33	0
	2	0.13	0	0.34	0	0.28	0	0.34	0	0.10	0	0.32	0
	3	0.11	0	0.33	0	0.28	0	0.32	0	0.09	0	0.30	0
	4	0.11	0	0.31	0	0.28	0	0.32	0	0.09	0	0.28	0
	5	0.05	0	0.28	0	0.28	0	0.30	0	0.09	0	0.28	0
	6	0.08	0	0.27	0	0.25	0	0.28	0	0.09	0	0.27	0
	7	0.08	0	0.26	0	0.22	0	0.26	0	0.08	0	0.25	0
	8	0.09	0	0.25	0	0.20	0	0.22	0	0.08	0	0.25	0
	9	0.06	0.01	0.20	0.01	0.19	0.01	0.22	0.01	0.08	0.01	0.21	0.01

Πίνακας 9.2: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων σε έναν $U_{19}(19^{14})$

		SCAD($\eta_\nu(\lambda)$)		SCAD(gcv)		LASSO($\eta_\nu(\lambda)$)		LASSO(gcv)		Hard($\eta_\nu(\lambda)$)		Hard(gcv)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
$U_{19}(19^{14})$	1	0.12	0	0.35	0	0.28	0	0.37	0	0.10	0	0.33	0
	2	0.11	0	0.34	0	0.25	0	0.35	0	0.10	0	0.32	0
	3	0.11	0	0.34	0	0.25	0	0.34	0	0.10	0	0.32	0
	4	0.10	0	0.34	0	0.25	0	0.34	0	0.09	0	0.31	0
	5	0.09	0	0.34	0	0.24	0	0.32	0	0.09	0	0.31	0
	6	0.08	0	0.31	0	0.24	0	0.31	0	0.08	0	0.30	0
	7	0.07	0	0.29	0	0.24	0	0.31	0	0.08	0	0.28	0
	8	0.06	0	0.27	0	0.24	0	0.30	0	0.08	0	0.28	0
	9	0.05	0	0.26	0	0.24	0	0.28	0	0.08	0	0.26	0
	10	0.04	0	0.26	0	0.23	0	0.26	0	0.07	0	0.25	0
	11	0.04	0	0.25	0	0.20	0	0.24	0	0.07	0	0.25	0
	12	0.04	0	0.23	0	0.19	0	0.22	0	0.05	0	0.21	0
	13	0.03	0	0.20	0	0.16	0	0.20	0	0.05	0	0.21	0

Πίνακας 9.3: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων σε έναν $U_{22}(22^{16})$

		SCAD($\eta_\nu(\lambda)$)		SCAD(gcv)		LASSO($\eta_\nu(\lambda)$)		LASSO(gcv)		Hard($\eta_\nu(\lambda)$)		Hard(gcv)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
$U_{22}(22^{16})$	1	0.10	0	0.35	0	0.27	0	0.36	0	0.09	0	0.33	0
	2	0.09	0	0.34	0	0.26	0	0.35	0	0.09	0	0.33	0
	3	0.08	0	0.34	0	0.25	0	0.35	0	0.08	0	0.32	0
	4	0.06	0	0.32	0	0.25	0	0.32	0	0.07	0	0.31	0
	5	0.07	0	0.33	0	0.25	0	0.32	0	0.07	0	0.29	0
	6	0.07	0	0.32	0	0.25	0	0.31	0	0.07	0	0.28	0
	7	0.06	0	0.30	0	0.24	0	0.30	0	0.07	0	0.28	0
	8	0.05	0	0.30	0	0.24	0	0.30	0	0.07	0	0.28	0
	9	0.05	0	0.29	0	0.24	0	0.29	0	0.07	0	0.27	0
	10	0.04	0	0.27	0	0.24	0	0.28	0	0.06	0	0.25	0
	11	0.04	0	0.26	0	0.24	0	0.27	0	0.06	0	0.24	0
	12	0.03	0	0.26	0	0.22	0	0.26	0	0.04	0	0.23	0
	13	0.03	0	0.24	0	0.20	0	0.24	0	0.04	0	0.23	0
	14	0.02	0	0.21	0	0.19	0	0.22	0	0.04	0	0.22	0
	15	0.01	0	0.18	0	0.18	0	0.21	0	0.04	0	0.18	0

Πίνακας 9.4: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων σε έναν $U_{23}(23^{19})$

		SCAD($\eta_\nu(\lambda)$)		SCAD(gcv)		LASSO($\eta_\nu(\lambda)$)		LASSO(gcv)		Hard($\eta_\nu(\lambda)$)		Hard(gcv)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
$U_{23}(23^{19})$	1	0.20	0	0.42	0	0.33	0	0.42	0	0.16	0	0.40	0
	2	0.20	0	0.42	0	0.33	0	0.42	0	0.15	0	0.39	0
	3	0.20	0	0.43	0	0.33	0	0.41	0	0.15	0	0.39	0
	4	0.18	0	0.41	0	0.33	0	0.40	0	0.15	0	0.39	0
	5	0.18	0	0.41	0	0.32	0	0.38	0	0.14	0	0.37	0
	6	0.17	0	0.39	0	0.32	0	0.38	0	0.14	0	0.37	0
	7	0.17	0	0.39	0	0.31	0	0.38	0	0.14	0	0.36	0
	8	0.14	0	0.37	0	0.31	0	0.38	0	0.13	0	0.35	0
	9	0.13	0	0.37	0	0.31	0	0.37	0	0.13	0	0.34	0
	10	0.13	0	0.35	0	0.31	0	0.36	0	0.12	0	0.33	0
	11	0.13	0	0.33	0	0.31	0	0.35	0	0.11	0	0.33	0
	12	0.11	0	0.32	0	0.31	0	0.35	0	0.10	0	0.31	0
	13	0.10	0	0.30	0	0.27	0	0.32	0	0.09	0	0.31	0
	14	0.10	0	0.30	0	0.25	0	0.30	0	0.09	0	0.29	0
	15	0.08	0	0.29	0	0.24	0	0.29	0	0.08	0	0.26	0
	16	0.07	0	0.25	0	0.21	0	0.26	0	0.08	0	0.25	0
	17	0.08	0.03	0.23	0.05	0.19	0.01	0.24	0.01	0.07	0.03	0.22	0.03
	18	0.06	0.09	0.18	0.09	0.18	0.07	0.21	0.07	0.07	0.08	0.19	0.08

Πίνακας 9.5: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων σε έναν $U_{26}(26^{21})$

		SCAD($\eta_\nu(\lambda)$)		SCAD(gcv)		LASSO($\eta_\nu(\lambda)$)		LASSO(gcv)		Hard($\eta_\nu(\lambda)$)		Hard(gcv)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
$U_{26}(26^{21})$	1	0.17	0	0.41	0	0.28	0	0.40	0	0.14	0	0.38	0
	2	0.17	0	0.39	0	0.27	0	0.40	0	0.13	0	0.38	0
	3	0.17	0	0.39	0	0.27	0	0.39	0	0.12	0	0.37	0
	4	0.16	0	0.39	0	0.27	0	0.38	0	0.12	0	0.36	0
	5	0.15	0	0.38	0	0.27	0	0.37	0	0.11	0	0.35	0
	6	0.15	0	0.38	0	0.26	0	0.37	0	0.11	0	0.35	0
	7	0.11	0	0.36	0	0.26	0	0.36	0	0.11	0	0.34	0
	8	0.11	0	0.36	0	0.26	0	0.36	0	0.10	0	0.33	0
	9	0.11	0	0.36	0	0.26	0	0.36	0	0.10	0	0.33	0
	10	0.09	0	0.34	0	0.25	0	0.36	0	0.09	0	0.33	0
	11	0.09	0	0.33	0	0.25	0	0.34	0	0.09	0	0.32	0
	12	0.08	0	0.32	0	0.25	0	0.34	0	0.09	0	0.30	0
	13	0.08	0	0.32	0	0.25	0	0.32	0	0.09	0	0.30	0
	14	0.08	0	0.31	0	0.23	0	0.30	0	0.08	0	0.29	0
	15	0.07	0	0.31	0	0.22	0	0.29	0	0.08	0	0.28	0
	16	0.05	0	0.27	0	0.20	0	0.27	0	0.07	0	0.27	0
	17	0.05	0	0.26	0	0.20	0	0.26	0	0.07	0	0.26	0
	18	0.05	0	0.25	0	0.18	0	0.25	0	0.05	0	0.24	0
	19	0.03	0.01	0.21	0.03	0.15	0.01	0.23	0.02	0.04	0.01	0.21	0.01
	20	0.03	0.04	0.18	0.04	0.14	0.04	0.19	0.04	0.05	0.04	0.18	0.05

Πίνακας 9.6: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων σε έναν $U_{30}(30^{24})$

		SCAD($\eta_\nu(\lambda)$)		SCAD(gcv)		LASSO($\eta_\nu(\lambda)$)		LASSO(gcv)		Hard($\eta_\nu(\lambda)$)		Hard(gcv)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
$U_{30}(30^{24})$	1	0.16	0	0.39	0	0.29	0	0.44	0	0.11	0	0.36	0
	2	0.16	0	0.39	0	0.29	0	0.43	0	0.11	0	0.36	0
	3	0.14	0	0.38	0	0.29	0	0.42	0	0.11	0	0.36	0
	4	0.14	0	0.38	0	0.29	0	0.40	0	0.11	0	0.36	0
	5	0.13	0	0.37	0	0.28	0	0.38	0	0.11	0	0.36	0
	6	0.12	0	0.37	0	0.28	0	0.37	0	0.10	0	0.34	0
	7	0.10	0	0.35	0	0.28	0	0.36	0	0.10	0	0.34	0
	8	0.10	0	0.35	0	0.27	0	0.36	0	0.09	0	0.33	0
	9	0.10	0	0.35	0	0.27	0	0.35	0	0.09	0	0.33	0
	10	0.08	0	0.34	0	0.27	0	0.35	0	0.09	0	0.33	0
	11	0.07	0	0.34	0	0.27	0	0.34	0	0.08	0	0.32	0
	12	0.07	0	0.32	0	0.26	0	0.34	0	0.08	0	0.31	0
	13	0.07	0	0.32	0	0.26	0	0.32	0	0.08	0	0.30	0
	14	0.06	0	0.31	0	0.26	0	0.32	0	0.07	0	0.28	0
	15	0.06	0	0.31	0	0.25	0	0.31	0	0.07	0	0.28	0
	16	0.05	0	0.30	0	0.25	0	0.30	0	0.06	0	0.27	0
	17	0.05	0	0.28	0	0.25	0	0.30	0	0.06	0	0.27	0
	18	0.04	0	0.27	0	0.25	0	0.29	0	0.06	0	0.25	0
	19	0.04	0	0.27	0	0.24	0	0.39	0	0.05	0	0.24	0
	20	0.04	0	0.25	0	0.23	0	0.28	0	0.05	0	0.23	0
	21	0.03	0	0.24	0	0.21	0	0.26	0	0.05	0	0.23	0
	22	0.02	0	0.21	0	0.19	0.01	0.22	0.01	0.05	0.01	0.23	0.01
	23	0.01	0.04	0.21	0.04	0.16	0.03	0.20	0.03	0.04	0.05	0.18	0.03

Πίνακας 9.7: Απόδοση των μεθόδων για τυχαίους συντελεστές, με χρήση 1000 προσομοιώσεων σε έναν $U_{30}(30^{26})$

		SCAD($\eta_\nu(\lambda)$)		SCAD(gcv)		LASSO($\eta_\nu(\lambda)$)		LASSO(gcv)		Hard($\eta_\nu(\lambda)$)		Hard(gcv)	
Σχεδιασμός	**q**	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
$U_{30}(30^{26})$	1	0.26	0	0.45	0	0.31	0	0.47	0	0.19	0	0.42	0
	2	0.25	0	0.45	0	0.31	0	0.46	0	0.19	0	0.42	0
	3	0.24	0	0.44	0	0.31	0	0.43	0	0.19	0	0.42	0
	4	0.23	0	0.44	0	0.31	0	0.42	0	0.19	0	0.42	0
	5	0.22	0	0.43	0	0.30	0	0.40	0	0.18	0	0.41	0
	6	0.20	0	0.42	0	0.30	0	0.40	0	0.18	0	0.41	0
	7	0.20	0	0.42	0	0.28	0	0.40	0	0.17	0	0.40	0
	8	0.19	0	0.42	0	0.28	0	0.40	0	0.17	0	0.39	0
	9	0.19	0	0.42	0	0.28	0	0.40	0	0.16	0	0.38	0
	10	0.18	0	0.40	0	0.28	0	0.40	0	0.16	0	0.38	0
	11	0.16	0	0.39	0	0.27	0	0.39	0	0.15	0	0.37	0
	12	0.16	0	0.38	0	0.27	0	0.38	0	0.14	0	0.36	0
	13	0.14	0	0.38	0	0.27	0	0.38	0	0.14	0	0.36	0
	14	0.14	0	0.37	0	0.26	0	0.36	0	0.13	0	0.35	0
	15	0.13	0	0.36	0	0.25	0	0.34	0	0.13	0	0.35	0
	16	0.13	0	0.36	0	0.25	0	0.34	0	0.13	0	0.34	0
	17	0.12	0	0.35	0	0.24	0	0.32	0	0.12	0	0.32	0
	18	0.10	0	0.33	0	0.24	0	0.32	0	0.12	0	0.31	0
	19	0.09	0	0.30	0	0.23	0	0.30	0	0.12	0	0.31	0
	20	0.09	0	0.29	0	0.23	0	0.30	0	0.11	0	0.30	0
	21	0.08	0	0.27	0	0.21	0	0.28	0	0.10	0.01	0.28	0.01
	22	0.07	0.01	0.27	0.01	0.20	0.02	0.27	0.02	0.10	0.02	0.26	0.02
	23	0.06	0.07	0.26	0.09	0.15	0.04	0.22	0.08	0.08	0.08	0.22	0.06
	24	0.05	0.08	0.17	0.12	0.15	0.07	0.18	0.10	0.04	0.12	0.15	0.12
	25	0.04	0.11	0.10	0.19	0.13	0.19	0.17	0.22	0.04	0.18	0.10	0.24

Βάσει των αποτελεσμάτων μας, καταλήγουμε στα εξής: Καταρχήν, μια γενική παρατήρηση είναι ότι σε αρκετές περιπτώσεις, τα σφάλματα Τύπου ΙΙ παραμένουν σε μηδενικά επίπεδα όταν χρησιμοποιείται η μέθοδος επιλογής της ρυθμιστικής παραμέτρου $\eta_\nu(\lambda)$, σε συνδυασμό με την προτεινόμενη διαδικασία κρησαρίσματος. Το γεγονός αυτό συνεπάγεται αρκετά μεγάλη ισχύ. Επιπλέον, η βελτίωση των σφαλμάτων Τύπου Ι είναι εμφανής, ακόμα και όταν αυξάνει ο αριθμός των πραγματικά ενεργών παραγόντων. Μεταξύ των ποινικοποιημένων μεθόδων, η SCAD και η Hard υπερέχουν της LASSO. Η τελευταία, παρουσιάζει σχετικά αυξημένα σφάλματα Τύπου Ι, ωστόσο όμως, η χρήση της $\eta_\nu(\lambda)$ αποφέρει μια σημαντική μείωσή τους. Συνοψίζοντας, οι μέθοδοι SCAD$(\eta_\nu(\lambda))$ και Hard$(\eta_\nu(\lambda))$ παρείχαν τα καλύτερα αποτελέσματα και προτείνεται η χρήση τους ως αποδοτικές διαδικασίες κρησαρίσματος στους ομοιόμορφους σχεδιασμούς.

9.4 Συμπεράσματα

Οι ομοιόμορφοι σχεδιασμοί είναι πολύ αποτελεσματικοί στην επιλογή των ενεργών επιδράσεων, από ένα μεγάλο σύνολο δυνητικά υποψηφίων. Παρακινούμενοι από την ικανότητά τους να διερευνούν πολλούς παράγοντες με μεγάλο πλήθος επιπέδων, ενώ ταυτόχρονα οι πειραματικές εκτελέσεις είναι λίγες, εξετάσαμε σε αυτό το κεφάλαιο τη χρησιμότητά τους για σκοπούς κρησαρίσματος. Προτείναμε μια νεά μέθοδο επιλογής μεταβλητών μέσω της εφαρμογής των ποινικοποιημένων ελαχίστων τετραγώνων. Η απόδοση της μεθόδου, η οποία αξιολογήθηκε με πειράματα προσομοίωσης, ήταν ιδιαίτερα ικανοποιητική. Συγκεκριμένα, τα σφάλματα Τύπου Ι προέκυψαν σε πολύ χαμηλά επίπεδα, το οποίο σημαίνει μείωση του κόστους επιπλέον πειραμάτων, από τη στιγμή που οι μη σημαντικοί παράγοντες αποκλείονται. Επιπλέον, τα μηδενικά ή πολύ χαμηλά σφάλματα Τύπου ΙΙ, συνεπάγονται υψηλή ισχύ της μεθόδου, συνεπώς επιλέγονται επιτυχώς οι ενεργές επιδράσεις. Ως εκ τούτου, η προτεινόμενη διαδικασία μπορεί να θεωρηθεί αρκετά αξιόπιστη, ενώ συγχρόνως είναι αρκετά εύκολη στην υλοποίηση.

Μέρος IV

Ορθογώνιοι Σχηματισμοί Τριών Επιπέδων

ΚΕΦΑΛΑΙΟ **10**

Βασικές Έννοιες και Ορισμοί

To find out what happens when you change something,
it is necessary to change it.

—Box, Hunter and Hunter (1978)

Στο κεφάλαιο αυτό, παραθέτουμε ορισμένα βασικά στοιχεία των ορθογώνιων σχηματισμών, καθώς και τους σχετικούς ορισμούς, βασιζόμενοι στο βιβλίο των Hedayat et al. [88]. Αναλύονται επίσης οι συνδυαστικά ισόμορφοι και γεωμετρικά ισόμορφοι ορθογώνιοι σχηματισμοί, οι οποίοι και θα μας απασχολήσουν στη συνέχεια.

10.1 Ορθογώνιοι Σχηματισμοί

Κατά τη διεξαγωγή ενός πειράματος, ο στόχος είναι η διερεύνηση της σχέσης μεταξύ της μεταβλητής απόκρισης και ορισμένων επεξηγηματικών μεταβλητών. Συνήθως τα βήματα που ακολουθούνται είναι η διατύπωση του προβλήματος, η επιλογή της μεταβλητής απόκρισης, η επιλογή των παραγόντων που θα μπορούσαν να επηρεάζουν την απόκριση, η επιλογή του κατάλληλου πειραματικού σχεδιασμού και η εκτέλεση και ανάλυση του πειράματος. Όταν ο πειραματιστής υποθέτει ότι η απόκριση συνδέεται με τους παράγοντες με μια γραμμική σχέση, τότε μπορεί να επιλεγεί ένας παραγοντικός σχεδιασμός δύο επιπέδων για τη διερεύνηση αυτής της σχέσης. Σε πολλές περιπτώσεις όμως, ειδικά σε προβλήματα Μηχανικής, η γραμμικότητα δεν ισχύει. Οπότε, για να ανιχνεύσουμε την απομάκρυνση από τη γραμμική σχέση που συνδέει τους παράγοντες με την απόκριση, ο σχεδιασμός ενισχύεται με κεντρικά σημεία. Η πρόσθεση κεντρικών σημείων σε έναν παραγοντικό σχεδιασμών δύο επιπέδων, μας επιτρέπει να ελέγξουμε την έλλειψη προσαρμογής του μοντέλου πρώτης τάξης, χωρίς όμως να μας βοηθά να καθορίσουμε τη σχέση μεταξύ των καθαρά τετραγωνικών όρων που μπορεί να υφίστανται και της μεταβλητής απόκρισης. Αυτή η στρατηγική ελέγχου είναι αρκετά δημοφιλής, καθώς προκύπτουν πιο οικονομικοί σχεδιασμοί, όσον αφορά το μέγεθός τους.

Όταν όμως έχει ανιχνευθεί η απόκλιση από τη γραμμικότητα, πρέπει να εκτελεστεί ένα επακόλουθο πείραμα, ώστε να εκτιμηθούν οι επιδράσεις των καθαρά τετραγωνικών όρων. Αυτό έχει πολλές φορές ως αποτέλεσμα ένα αρκετά πιο δαπανηρό πείραμα, ακόμα και σε σύγκριση με την περίπτωση όπου θα χρησιμοποιούσαμε κατευθείαν έναν παραγοντικό σχεδιασμό τριών επιπέδων. Είναι προφανές λοιπόν ότι σε παρόμοιες περιπτώσεις, όπου δηλαδή είναι γνωστή ή έστω υπάρχουν υποψίες για την ύπαρξη τετραγωνικής σχέσης μεταξύ της απόκρισης και των παραγόντων, προτείνεται η χρήση ενός σχεδιασμού τριών επιπέδων.

Αναφορικά με την εκτίμηση των επιδράσεων, μια εξαιρετική επιλογή είναι η χρήση ορθογώνιων σχεδιασμών, καθότι μας επιτρέπουν να έχουμε ασυσχέτιστες εκτιμήσεις των παραμέτρων του μοντέλου. Οι παραγοντικοί σχεδιασμοί τριών επιπέδων είναι παραδείγματα τέτοιων σχεδιασμών. Όμως, οι πλήρεις 3^k σχεδιασμοί απαιτούν ένα μεγάλο αριθμό πειραματικών εκτελέσεων, άρα και πόρων. Μια αρκετά πιο οικονομική επιλογή σχεδιασμών που διατηρούν την ιδιότητα της ορθογωνιότητας, είναι οι ορθογώνιοι σχηματισμοί (orthogonal arrays). Οι σχηματισμοί αυτοί, σε τρία επίπεδα, είναι καθαρά τετραγωνικοί σχεδιασμοί, μέσω των οποίων μπορούμε να διερευνήσουμε αποτελεσματικά τη σχέση μεταξύ της απόκρισης και των επεξηγηματικών μεταβλητών, χωρίς να χρειάζεται η πρόσθεση κεντρικών σημείων. Στον παρακάτω Πίνακα 10.1, παρουσιάζουμε ένα παράδειγμα ενός ορθογώνιου σχηματισμού τριών επιπέδων, με 18 εκτελέσεις και 4 παράγοντες. Ο σχηματισμός αυτός μπορεί να χρησιμοποιηθεί για να εξετασθεί η σχέση της απόκρισης και των 4 παραγόντων, χρησιμοποιώντας μόνο 18 εκτελέσεις, σε αντίθεση με τον πλήρη 3^4 σχεδιασμό ο οποίος θα απαιτούσε 81 εκτελέσεις.

Πίνακας 10.1: Ένας ορθογώνιος σχηματισμός με 4 παράγοντες και 18 εκτελέσεις

Εκτέλεση	A	B	C	D	Εκτέλεση	A	B	C	D
1	1	1	1	1	10	-1	1	-1	-1
2	1	-1	0	-1	11	-1	-1	1	0
3	1	0	0	-1	12	-1	0	1	0
4	1	1	-1	0	13	0	1	0	0
5	1	-1	-1	0	14	0	-1	-1	1
6	1	0	1	1	15	0	0	0	0
7	-1	1	0	1	16	0	1	1	-1
8	-1	-1	0	1	17	0	-1	1	-1
9	-1	0	-1	-1	18	0	0	-1	1

Ορισμός 10.1 *Ένας ορθογώνιος σχηματισμός, συμβολίζεται ως $OA(n, q, l, t)$ και είναι ένας $n \times q$ πίνακας με στοιχεία επιλεγμένα από ένα σύνολο με l διακεκριμένα σύμβολα, διατεταγμένα έτσι ώστε για κάθε επιλογή t στηλών του σχηματισμού, καθένα από τα l^t διαφορετικά διανύσματα γραμμών, να εμφανίζεται το ίδιο συχνά.*

Στις στατιστικές εφαρμογές, ο πρώτος που εισήγαγε τη χρήση τους ήταν ο Rao [147]. Κάθε στήλη αντιστοιχεί σε έναν παράγοντα, τα σύμβολα αντιπροσωπεύουν τα επίπεδα των παραγόντων και κάθε γραμμή αντιστοιχεί σε ένα συνδυασμό των επιπέδων των παραγόντων (πειραματική εκτέλεση). Οπότε, συμβολίζουμε ως n το πλήθος των εκτελέσεων, q τον αριθμό των παραγόντων, l τον αριθμό των επιπέδων κάθε παράγοντα και t την ισχύ (strength) του σχηματισμού. Συνεπώς, κάθε $OA(n, q, l, t)$ ορίζει έναν παραγοντικό σχεδιασμό n εκτελέσεων με q παράγοντες από l επίπεδα. Για περαιτέρω λεπτομέρειες στους ορθογώνιους σχηματισμούς, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο εξαιρετικό βιβλίο των Hedayat et al. [88].

10.2 Συνδυαστικά Ισόμορφοι και Γεωμετρικά Ισόμορφοι Ορθογώνιοι Σχηματισμοί

Κατά την αξιολόγηση δύο ορθογώνιων σχηματισμών, το πρώτο πράγμα που μας ενδιαφέρει είναι αν οι σχηματισμοί αυτοί είναι στοιχειωδώς ίδιοι ή ισοδύναμοι. Οι ισοδύναμοι σχηματισμοί καλούνται ισόμορφοι. Αν τώρα δύο σχηματισμοί είναι ισοδύναμοι, τότε κατέχουν τις ίδιες ιδιότητες και καθένας αποτελεί μια εξίσου καλή επιλογή. Αν μπορεί να προσδιοριστεί μια κλάση ισοδύναμων σχηματισμών, τότε οι ιδιότητες της κλάσης μπορούν να περιγραφούν και να χρησιμοποιηθούν για το χαρακτηρισμό των επιμέρους σχηματισμών αυτής. Μελετώντας συνολικά κλάσεις σχηματισμών αντί μεμονωμένους σχηματισμούς, μειώνεται και το ποσό της δουλειάς. Συνεπώς, ο προσδιορισμός της ισοδυναμίας ενός ορθογώνιου σχηματισμού είναι ένα σημαντικό πρόβλημα και ο ορισμός αυτής εξαρτάται από το είδος των παραγόντων που μελετώνται.

Ορισμός 10.2 *Δύο ορθογώνιοι σχηματισμοί με ποιοτικούς παράγοντες λέγονται συνδυαστικά ισόμορφοι (combinatorial isomorphic) αν ο ένας μπορεί να παραχθεί από τον άλλο με μια σειρά από μεταθέσεις γραμμών ή/και στηλών ή/και συμβόλων σε κάθε στήλη.*

Μη ισόμορφοι ορθογώνιοι σχηματισμοί με τρία επίπεδα βρέθηκαν στις εργασίες [49], [50] και [51]. Να τονίσουμε στο σημείο αυτό, ότι για ποιοτικούς παράγοντες, δεν υπάρχει ουσιαστική διάταξη των επιπέδων των παραγόντων και η αντιστοίχιση των συμβόλων σε επίπεδα για έναν παράγοντα είναι τελείως αυθαίρετη, χωρίς να υπάρχει κάποια ερμηνεία. Για παράδειγμα, σε έναν παράγοντα χρώματος μπορούμε να θεωρήσουμε ότι 0=κόκκινο χρώμα, 1=μπλέ και 2=μαύρο. Για τον παράγοντα αυτόν, μπορεί να εφαρμοστεί οποιαδήποτε μετάθεση των συμβόλων 0, 1 και 2. Δε θα αλλάξει κάτι αναφορικά με την ερμηνεία των επιδράσεων αν τα σύμβολα έχουν ανατεθεί έτσι ώστε 0=μπλέ χρώμα, 1=κόκκινο και 2=μαύρο.

Οι ποιοτικοί παράγοντες, χρησιμοποιούνται αρκετά κατά την κατασκευή και μελέτη των ιδιοτήτων των παραγοντικών σχεδιασμών γενικότερα. Παρ΄ όλα αυτά, υπάρχουν περιπτώσεις κατά τις οποίες εμπεριέχονται ποσοτικοί παράγοντες, όπως για παράδειγμα στη μεθοδολογία αποκριτικών επιφανειών. Όταν λοιπόν έχουμε έναν ορθογώνιο σχηματισμό με ποσοτικούς παράγοντες, η μετάθεση των επιπέδων σε έναν ή περισσότερους παράγοντες στον πίνακα σχεδιασμού μπορεί να οδηγήσει σε διαφορετικές γεωμετρικές δομές οπότε και σε διαφορετικές ιδιότητες του σχηματισμού [34].

Στην περίπτωση ποσοτικών παραγόντων, υφίσταται μια πραγματική διάταξη των επιπέδων. Για παράδειγμα, έστω ένας παράγοντας θερμοκρασίας με επίπεδα 40^oC (υψηλό), 20^oC (μεσαίο) και 0^oC (χαμηλό). Βάσει της διάταξης, στα επίπεδα αυτά μπορούν να αντιστοιχηθούν

τα σύμβολα 2 (υψηλό), 1 (μεσσαίο) και 0 (χαμηλό). Η ανάθεση των συμβόλων αυτών μπορεί να αντιστραφεί, χωρίς να αλλάξει η ουσιαστική δομή της διάταξης. Όμως, αν αναθέσουμε το σύμβολο 1 στους 40^oC, το 0 στους 20^oC και το 2 στους 0^oC, το αποτέλεσμα προφανώς είναι τελείως διαφορετικό.

Όπως προαναφέραμε λοιπόν, η διάταξη των συμβόλων μπορεί να αντιστραφεί, αρκεί φυσικά να διατηρείται η ουσιαστική διάταξη των επιπέδων. Αυτή η απαίτηση περιορίζει τις επιτρεπόμενες εναλλαγές συμβόλων σε μια στήλη του σχηματισμού. Γεγονός που επίσης περιορίζει και τις πιθανές μεταθέσεις που θα οδηγούσαν σε ισόμορφους σχηματισμούς με ποσοτικούς παράγοντες.

Ορισμός 10.3 *Δύο ορθογώνιοι σχηματισμοί με ποσοτικούς παράγοντες λέγονται γεωμετρικά ισόμορφοι (geometrically isomorphic) αν ο ένας μπορεί να παραχθεί από τον άλλο με μια σειρά από μεταθέσεις γραμμών ή/και στηλών ή/και αντιστροφές της σειράς των επιπέδων σε έναν ή περισσότερους παράγοντες (αρκεί να διατηρείται η ουσιαστική διάταξη των επιπέδων).*

Για σχεδιασμούς δύο επιπέδων, ο συνδυαστικός και ο γεωμετρικός ισομορφισμός είναι ισοδύναμοι, καθότι υφίσταται η μετάθεση δύο μόνο συμβόλων, κάτι που διατηρεί φυσικά τη διάταξη των επιπέδων των παραγόντων. Για σχεδιασμούς όμως με περισσότερα επίπεδα, όπως οι ορθογώνιοι σχηματισμοί τριών επιπέδων, αυτοί που είναι συνδυαστικά ισόμορφοι μπορεί να μην είναι και γεωμετρικά ισόμορφοι. Ενώ αν είναι γεωμετρικά ισόμορφοι, θα είναι και συνδυαστικά. Συνεπώς, ο αριθμός των γεωμετρικά μη ισόμορφων σχεδιασμών είναι μεγαλύτερος από τον αριθμό των συνδυαστικά μη ισόμορφων. Για περισσότερες λεπτομέρειες αναφορικά με την έννοια των συνδυαστικά και γεωμετρικά ισόμορφων σχεδιασμών, παραπέμπουμε τον αναγνώστη στις εργασίες των Cheng και Ye [34] και Tsai et al. [163–165].

ΚΕΦΑΛΑΙΟ **11**

Μελέτη Ορθογώνιων Σχηματισμών Τριών Επιπέδων με Συσχετισμένες Παρατηρήσεις

Efficiency = statistical efficiency x usage.

—*John Tukey (1915–2000)*

Όταν ο πειραματιστής υποπτεύεται ότι θα μπορούσε να υπάρχει μια τετραγωνική σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών, ενδείκνυται η χρήση ενός σχεδιασμού τριών επιπέδων (σχεδιασμός δευτέρας τάξης). Μια αρκετά οικονομική επιλογή είναι οι ορθογώνιοι σχηματισμοί, οι οποίοι συχνά χρησιμοποιούνται ως σχεδιασμοί αποκριτικών επιφανειών δευτέρας τάξης. Γενικά, υποθέτουμε ότι τα δεδομένα αποτελούν ανεξάρτητες παρατηρήσεις. Παρ΄ όλα αυτά, υπάρχουν περιπτώσεις που δεν ικανοποιείται αυτή η υπόθεση. Σε αυτό το κεφάλαιο, στόχος μας είναι να συγκρίνουμε τους ορθογώνιους σχηματισμούς τριών επιπέδων με 18, 27 και 36 εκτελέσεις, υπό την παρουσία διαφορετικών μορφών συσχέτισης στις παρατηρήσεις. Απώτερος σκοπός είναι η επιλογή των βέλτιστων σχηματισμών οι οποίοι να μπορούν να χρησιμοποιηθούν αποτελεσματικά ως σχεδιασμοί αποκριτικών επιφανειών.

11.1 Ερευνητικό Πρόβλημα

Οι ορθογώνιοι σχηματισμοί με παραμέτρους $(n, q, 3, t)$ μπορούν να χρησιμοποιηθούν ως σχεδιασμοί αποκριτικών επιφανειών. Στα βιομηχανικά πειράματα, γίνεται συχνή χρήση των αποκριτικών επιφανειών για τη μελέτη της εξάρτησης μεταξύ της μεταβλητής απόκρισης y και ενός συνόλου από q ελεγχόμενες επεξηγηματικές μεταβλητές. Έστω $\mathbf{D}$ ο $n \times q$ πίνακας σχεδιασμού. Η $u-$οστή γραμμή του $\mathbf{D}$ αποτελεί την πειραματική εκτέλεση $(x_{u1}, x_{u2}, \ldots, x_{uq})$, $u = 1, 2, \ldots, n$. Το θεωρούμενο μοντέλο δευτέρας τάξης είναι

$$y_u = \beta_0 + \sum_{i=1}^{q} \beta_i x_{iu} + \sum_{j=1}^{q} \beta_{ii} x_{iu}^2 + \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \beta_{ij} x_{iu} x_{ju} + \epsilon_u, \tag{11.1}$$

όπου το y_u αντιπροσωπεύει την απόκριση στην πειραματική εκτέλεση u. Όλα τα y_u θεωρούνται ασυσχέτιστες παρατηρήσεις με σταθερή διασπορά σ^2, x_{iu} είναι το επίπεδο του $i-$οστού παράγοντα σε αυτή την πειραματική εκτέλεση, $\beta_0, \beta_1, ..., \beta_{qq}$ αποτελούν τις $s = q(q-1)/2 + 2q + 1$ άγνωστες παραμέτρους του μοντέλου και ϵ_u είναι το τυχαίο σφάλμα στην $u-$οστή εκτέλεση. Το μοντέλο (11.1) γράφεται με συμβολισμό πινάκων ως:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{11.2}$$

όπου $\mathbf{y}=[y_1, y_2, ..., y_n]^T$, $\mathbf{X}$ είναι ο $n \times s$ πίνακας μοντέλου, $\boldsymbol{\beta}$ είναι το $s \times 1$ διάνυσμα των αγνώστων παραμέτρων και $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, ..., \epsilon_n]^T$. Η εκτιμήτρια ελαχίστων τετραγώνων του $\boldsymbol{\beta}$ είναι $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ με $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ και με πίνακα διασποράς - συνδιασποράς $Var_1(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = V_1$. Στο εξής, θα χρησιμοποιήσουμε το συμβολισμό $A_1 = \sigma^{-2}V_1$.

Η θεωρία των βέλτιστων σχεδιασμών του Pukelsheim [145], συνήθως υποθέτει ανεξάρτητες παρατηρήσεις. Αρκετές φορές όμως, υπάρχουν περιπτώσεις όπου η υπόθεση αυτή δεν ισχύει, όπως για παράδειγμα στη Βιομηχανία, όπου μπορεί να υπάρχει χρονική εξάρτηση, καθώς και στον τομέα της Γεωργίας, όπου η χωρική εξάρτηση είναι πιθανή. Οι Jenkins και Chanmugam [96] ήταν από τους πρώτους που μελέτησαν πειράματα με έναν παράγοντα δύο επιπέδων, με χρήση μοντέλου αυτοπαλινδρόμησης πρώτης τάξης, AR(1) και θετική συσχέτιση ($\rho > 0$). Οδηγήθηκαν στο συμπέρασμα ότι η αποδοτικότερη σειρά εκτελέσεων προέκυψε μέσω της αλλαγής του επιπέδου του παράγοντα μεταξύ κάθε ζευγαριού εκτελέσεων. Οι Cheng και Steinberg [32] επέκτειναν το πρόβλημα στους 2^k παραγοντικούς σχεδιασμούς, υποθέτοντας αμελητέες αλληλεπιδράσεις. Βέλτιστοι σχεδιασμοί σε συνδυασμό με το μοντέλο AR(1) μελετήθηκαν και από τους Saunders et al. [151], Martin et al. [129] και Elliott et al. [48]. Επιπλέον, ο Constantine [38] θεώρησε το μοντέλο κινητού μέσου πρώτης τάξης, MA(1), ως μορφή εξάρτησης. Πιο πρόσφατα, βέλτιστοι σχεδιασμοί με συσχετισμένες παρατηρήσεις και ενδοκατηγορική (intraclass) μορφή συσχέτισης, μελετήθηκαν από τους Sethuraman et al. [155] και τους Sethuraman και Raghavarao [154]. Επιπλέον, οι Garroi et al. [78] πρότειναν τη χρήση ενός αλγορίθμου αναζήτησης σε συνδυασμό με την απόσταση Hamming για την εύρεση βέλτιστης σειράς εκτελέσεων σε κεντρικούς σύνθετους σχεδιασμούς (central composite designs) υπό την παρουσία AR(1) μορφής συσχέτισης.

Σε αυτό το κεφάλαιο, ο στόχος μας είναι να εξετάσουμε και να συγκρίνουμε τους γεωμετρικά μη ισόμορφους ορθογώνιους σχηματισμούς τριών επιπέδων ώστε να ανιχνεύσουμε τυχόν διαφορές στη χρήση και στην αποδοτικότητά τους ως σχεδιασμοί αποκριτικών επιφανειών. Θα συγκρίνουμε τους διαθέσιμους ορθογώνιους σχηματισμούς με 18, 27 και 36 εκτελέσεις και με 3 και 4 παράγοντες, καθώς και τους ορθογώνιους σχηματισμούς με 27 εκτελέσεις και 5 παράγοντες. Θα θεωρήσουμε επίσης συσχετισμένες παρατηρήσεις. Συγκεκριμένα, θα προτείνουμε τρία κριτήρια τα οποία και θα χρησιμοποιήσουμε ως κριτήρια σύγκρισης, βασιζόμενοι στην εργασία των Ghosh και Shen [81], καθώς επίσης θα θεωρήσουμε τρεις διαφορετικές μορφές συσχέτισης, με θετικές και αρνητικές τιμές αυτών.

11.2 Κριτήρια Σύγκρισης Ορθογώνιων Σχηματισμών

Στα πειράματα, συχνά υποθέτουμε το συμβατικό γραμμικό μοντέλο

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, Var(\mathbf{y}) = \sigma^2\mathbf{I}, \tag{11.3}$$

με ασυσχέτιστες παρατηρήσεις. Στην πραγματικότητα όμως, μπορεί να υπάρχει μια μορφή συσχέτισης και το πραγματικό μοντέλο είναι

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, Var(\mathbf{y}) = \sigma^2\mathbf{P} \tag{11.4}$$

όπου $\mathbf{P}$ ο $n \times n$ πίνακας συσχέτισης, ο οποίος θεωρείται θετικά ορισμένος. Αν γνωρίζουμε εκ των προτέρων την παρουσία συσχέτισης στις παρατηρήσεις, τότε θεωρούμε εξαρχής το μοντέλο (11.4) αντί του (11.3). Εδώ θα θεωρήσουμε την περίπτωση όπου δε διαθέτουμε αυτήν την πληροφορία. Συνεπώς, θα υποθέσουμε το μοντέλο (11.3) και θα χρησιμοποιήσουμε τη συνήθη εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ για το $\boldsymbol{\beta}$. Τονίζουμε ότι $\widehat{\boldsymbol{\beta}}$ είναι μια αμερόληπτη εκτιμήτρια του $\boldsymbol{\beta}$ υπό την ισχύ είτε του μοντέλου (11.3) είτε του (11.4). Επίσης, η διασπορά $Var(\widehat{\boldsymbol{\beta}})$ τροποποιείται ως $Var_2(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = V_2$ υπό το μοντέλο (11.4). Στο εξής, θα χρησιμοποιήσουμε το συμβολισμό $A_2 = \sigma^{-2}V_2$. Η αλλαγή τώρα στη διασπορά $Var(\widehat{\boldsymbol{\beta}})$ καθορίζεται από την εγγύτητα των V_1 και V_2, η οποία με τη σειρά της μετριέται με χρήση του ίχνους (Trace-Tr), της ορίζουσας (Determinant-Det), και της μέγιστης χαρακτηριστικής ρίζας (Maximum Characteristic Root-Ch$_{max}$). Ακολουθώντας τους Ghosh και Shen [81], ορίζουμε τρία νέα κριτήρια σύγκρισης των ορθογώνιων σχηματισμών, συμβολιζόμενα ως CVF (change of variance functions), ως εξής:

$$\begin{aligned} CVF_1 &= \frac{TrV_2}{TrV_1} = \frac{TrA_2}{TrA_1}, \\ CVF_2 &= \frac{DetV_2}{DetV_1} = \frac{DetA_2}{DetA_1}, \\ CVF_3 &= \frac{Ch_{max}V_2}{Ch_{max}V_1} = \frac{Ch_{max}A_2}{Ch_{max}A_1}. \end{aligned} \tag{11.5}$$

Η εγγύτητα των V_1 και V_2, καθορίζεται από την εγγύτητα των CVF_i, $i = 1, 2, 3$, στο 1. Συνεπώς, κατά τη σύγκριση δύο σχεδιασμών D_1 και D_2, πρέπει να καθορίσουμε ποιος οδηγεί σε μεγαλύτερη εγγύτητα μεταξύ των V_1 και V_2. Άρα, θεωρούμε ότι ο D_1 είναι καλύτερος από τον D_2 αν ο D_1 δίνει τιμές των $|CVF_i|$, $i = 1, 2, 3$ πιο κοντά στο 1.

Θα βασιστούμε σε τρεις διαφορετικές μορφές συσχέτισης. Αρχικά, θεωρούμε την περίπτωση μιας ανταλλάξιμης διαδικασίας ή διαδικασίας ισοσυσχέτισης (exchangeable or equicorrelation process), όπου ο πίνακας συσχέτισης είναι της μορφής

$$\mathbf{P} = (1 - \rho)\mathbf{I} + \rho\mathbf{J}, \tag{11.6}$$

όπου $\mathbf{I}$ ο $n \times n$ ταυτοτικός πίνακας και $\mathbf{J}$ ο $n \times n$ πίνακας με στοιχεία μονάδες, σύμφωνα με τους Ghosh και Shen [81]. Βάσει των αποτελεσμάτων μας (βλ. Ενότητα 11.3), παρουσιάζουμε στο παρακάτω Θεώρημα, μια ιδιότητα του CVF_2 κριτηρίου.

Θεώρημα 11.1 *Έστω ο πίνακας ισοσυσχέτισης της μορφής (11.6). Τότε ισχύει ότι*

$$CVF_2 = Det[(1 - \rho)\boldsymbol{I} + \rho\ \boldsymbol{M}], \tag{11.7}$$

όπου $\boldsymbol{M} = \begin{pmatrix} n & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{O} \end{pmatrix}$ *και* $\boldsymbol{I}$ *είναι ο* $s \times s$ *ταυτοτικός πίνακας. Συνεπώς, το κριτήριο* CVF_2 *δεν εξαρτάται από τον εκάστοτε πίνακα μοντέλου, αλλά μόνο από τις τιμές των παραμέτρων*

n, s και ρ.

Απόδειξη: Μπορούμε εύκολα να παρατηρήσουμε ότι

$$(\mathbf{X}^T\mathbf{X})\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{O} \end{pmatrix}(\mathbf{X}^T\mathbf{X}) = \mathbf{X}^T\mathbf{J}\mathbf{X} \tag{11.8}$$

από όπου προκύπτει

$$\mathbf{X}^T\mathbf{J}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{X}^T\mathbf{X}\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{O} \end{pmatrix}. \tag{11.9}$$

Έχουμε επίσης ότι

$$\mathbf{X}^T\mathbf{P}\mathbf{X} = \mathbf{X}^T[(1-\rho)\mathbf{I} + \rho\mathbf{J}]\mathbf{X} = (1-\rho)\mathbf{X}^T\mathbf{X} + \rho\mathbf{X}^T\mathbf{J}\mathbf{X}. \tag{11.10}$$

Συνεπώς, το κριτήριο CVF_2 μετατρέπεται ως

$$CVF_2 = \frac{DetA_2}{DetA_1} = \frac{Det[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]}{Det[(\mathbf{X}^T\mathbf{X})^{-1}]} =$$

$$= Det[\mathbf{X}^T\mathbf{P}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] =$$

$$= Det[(1-\rho)\mathbf{I} + \rho\mathbf{X}^T\mathbf{J}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}]. \tag{11.11}$$

Χρησιμοποιώντας τώρα τη σχέση (11.9), προκύπτει

$$CVF_2 = Det\left[(1-\rho)\mathbf{I} + \rho\mathbf{X}^T\mathbf{X}\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{O} \end{pmatrix}\right] =$$

$$= Det\left[(1-\rho)\mathbf{I} + \rho\begin{pmatrix} n & \mathbf{0}^T \\ \mathbf{0} & \mathbf{O} \end{pmatrix}\right] \tag{11.12}$$

και η απόδειξη ολοκληρώθηκε. □

Η δεύτερη περίπτωση που θεωρήσαμε, αφορά τη διαδικασία κινητού μέσου 1ης τάξης, MA(1), στην οποία ο πίνακας συσχέτισης $\mathbf{P}$ ειναι ένας τριδιαγώνιος πίνακας, με μονάδες στη διαγώνιο και μη διαγώνια στοιχεία ίσα με ρ [186]. Αυτή η μορφή, οδηγεί σε συσχέτιση ρ μεταξύ παρακείμενων παρατηρήσεων και σε μηδενική συσχέτιση σε μη παρακείμενες παρατηρήσεις. Τονίζουμε ότι στη MA(1) διαδικασία, ισχύει ότι $|\rho| \leq 1/2$. Η τρίτη περίπτωση, σχετίζεται με τη διαδικασία αυτοπαλινδρόμησης 1ης τάξης, AR(1), στην οποία ο πίνακας συσχέτισης $\mathbf{P}$ έχει ως (i,j) στοιχείο, $\rho^{|i-j|}$ [157].

Πέραν της γνώσης των CVF τιμών για κάθε σχηματισμό, είναι επιθυμητή και η χρήση ενός κριτηρίου αποδοτικότητας. Ένα αρκετά γνωστό κριτήριο που χρησιμοποιείται συχνά, είναι το κριτήριο της $D-$αποδοτικότητας, το οποίο και θα εφαρμόσουμε στη συνέχεια.

Ορισμός 11.1 *Έστω $\boldsymbol{X}$ ο πίνακας μοντέλου (ο πίνακας που περιέχει s το πλήθος επιδράσεις που εξετάζουμε), και $\boldsymbol{X}'$ ο κανονικοποιημένος πίνακας. Η $D-$αποδοτικότητα του πίνακα σχεδιασμού $\boldsymbol{X}$ ορίζεται ως*

$$D_{eff} = |\boldsymbol{X}'^T\boldsymbol{X}'|^{1/s}. \tag{11.13}$$

Επειδή ο $\mathbf{X}'$ είναι κανονικοποιημένος, η $D-$αποδοτικότητα επιτυγχάνει το μέγιστο 1 όταν οι στήλες του $\mathbf{X}'$ είναι ορθογώνιες μεταξύ τους. Για περισσότερες λεπτομέρειες, παραπέμπουμε στους Wang και Wu [173].

11.3 Αποτελέσματα

Σε αυτήν την ενότητα, παρουσιάζουμε τα αποτελέσματα από τη σύγκριση των διαθέσιμων γεωμετρικά μη ισομορφων ορθογώνιων σχηματισμών (OAs) με 18, 27 και 36 εκτελέσεις, και με 3 και 4 παράγοντες, καθώς και των σχηματισμών με 27 εκτελέσεις και 5 παράγοντες. Θεωρούμε τις περιπτώσεις όπου υπάρχουν αρκετές εκτελέσεις για να είναι εκτιμήσιμο το μοντέλο με κύριες επιδράσεις (γραμμικές και τετραγωνικές) και γραμμικές αλληλεπιδράσεις δύο παραγόντων. Κάθε σχεδιασμός συμβολίζεται στο εξής ως $n.q.id$, όπου n ο αριθμός των εκτελέσεων, q ο αριθμός των παραγόντων και id ο αύξων αριθμός του σχεδιασμού. Για κάθε περίπτωση εξετάζουμε τις τρεις προαναφερθείσες μορφές συσχέτισης με αρνητικές και θετικές τιμές του ρ, ξεκινώντας από $\rho = -0.5$ εως $\rho = 0.5$. Στη συνέχεια, παραθέτουμε τα αποτελέσματα για τους καλύτερους σχηματισμούς ως προς κάθε CVF κριτήριο και για κάθε διαφορετική τιμή συσχέτισης. Παρουσιάζουμε επίσης και τις αντίστοιχες τιμές του κριτηρίου D_{eff}.

11.3.1 Περίπτωση 1: Πίνακας Συσχέτισης P=$(1-\rho)$I+ρ J

Καταρχήν, βάσει του Θεωρήματος που αποδείξαμε παραπάνω, οι τιμές του κριτηρίου CVF_2 είναι προφανώς οι ίδιες σε κάθε κλάση ορθογώνιων σχηματισμών για τη συγκεκριμένη μορφή συσχέτισης, συνεπώς οι σχηματισμοί δε διακρίνονται ως προς την αποδοτικότητά τους με αυτό το κριτήριο. Οπότε απλά θα παρουσιάσουμε αυτές τις τιμές.

Ξεκινάμε τις συγκρίσεις χρησιμοποιώντας τους γεωμετρικά μη ισόμορφους ορθογώνιους σχηματισμούς με 3 παράγονες και 18 εκτελέσεις. Οι σχηματισμοί που έδωσαν τιμές των CVF κοντά στο 1, παρουσιάζονται στον Πίνακα 11.1.

Πίνακας 11.1: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 18 εκτελέσεις και 3 παράγοντες για την Περίπτωση 1

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	103	1.229730	0.9803	-	288.325000	-	1	1.126610	0.9552
-0.4	103	1.183780	0.9803	-	119.834000	-	1	1.089250	0.9552
-0.3	103	1.137840	0.9803	-	43.478400	-	1	1.056300	0.9552
-0.2	103	1.091890	0.9803	-	12.383500	-	1	1.029170	0.9552
-0.1	103	1.045950	0.9803	-	1.650560	-	1	1.009670	0.9552
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	103	0.954054	0.9803	-	1.046040	-	80	0.996938	0.9667
0.2	103	0.908108	0.9803	-	0.590558	-	80	1.007800	0.9667
0.3	103	0.862162	0.9803	-	0.246157	-	123	0.980216	0.8346
0.4	103	0.816216	0.9803	-	0.078606	-	123	1.006270	0.8346
0.5	103	0.770270	0.9803	-	0.018555	-	123	1.047280	0.8346

Σύμφωνα με τις CVF_1 τιμές, ο σχηματισμός 18.3.103 προτείνεται για θετικές και αρνητικές τιμές συσχέτισης. Ωστόσο, το κριτήριο CVF_3 εισηγείται τη χρήση του σχηματισμού 18.3.1 για αρνητικές συσχετίσεις, ενώ οι σχηματισμοί 18.3.80 και 18.3.123 αποτελούν την καλύτερη επιλογή για θετικές συσχετίσεις.

Συνεχίζουμε τις συγκρίσεις με τους ορθογώνιους σχηματισμούς με 4 παράγοντες σε 18 εκτελέσεις. Οι τιμές των CVF κριτηρίων για τους καλύτερους σχηματισμούς παρουσιάζονται στον Πίνακα 11.2. Σύμφωνα με τα αποτελέσματά μας, συνίσταται ιδιαίτερα ο σχηματισμός 18.4.561, ανεξάρτητα από την τιμή της συσχέτισης.

Πίνακας 11.2: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 18 εκτελέσεις και 4 παράγοντες για την Περίπτωση 1

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	561	1.362650	0.8274	-	2189.470000	-	561	1.163650	0.8274
-0.4	561	1.290120	0.8274	-	644.496000	-	561	1.123400	0.8274
-0.3	561	1.217590	0.8274	-	161.432000	-	561	1.086090	0.8274
-0.2	561	1.145060	0.8274	-	30.814000	-	561	1.052460	0.8274
-0.1	561	1.072530	0.8274	-	2.658250	-	561	1.023410	0.8274
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	561	0.927471	0.8274	-	0.617673	-	561	0.983426	0.8274
0.2	561	0.854942	0.8274	-	0.193514	-	561	0.983426	0.8274
0.3	561	0.782413	0.8274	-	0.041372	-	561	0.974861	0.8274
0.4	561	0.709884	0.8274	-	0.006112	-	561	0.975267	0.8274
0.5	561	0.637355	0.8274	-	0.000580	-	561	1.004500	0.8274

Στη συνέχεια παρουσιάζουμε τα αποτελέσματα των σχηματισμών με 27 εκτελέσεις και 3 παράγοντες. Οι καλύτεροι σχηματισμοί δίνονται στον Πίνακα 11.3.

Πίνακας 11.3: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 3 παράγοντες για την Περίπτωση 1

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	302	1.074800	1.0000	-	461.320000	-	175	0.997643	0.9467
-0.4	302	1.059840	1.0000	-	194.214000	-	347	0.999837	0.9767
-0.3	302	1.044880	1.0000	-	72.110600	-	126	1.001200	0.9614
-0.2	302	1.029920	1.0000	-	21.671100	-	12	1.000050	0.9573
-0.1	302	1.014960	1.0000	-	3.772720	-	261	0.998055	0.9529
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	302	0.985039	1.0000	-	1.394710	-	3	0.997022	0.9118
0.2	302	0.970079	1.0000	-	0.832150	-	44	0.993466	0.8235
0.3	302	0.955118	1.0000	-	0.355112	-	173	0.998226	0.9104
0.4	302	0.940157	1.0000	-	0.114886	-	78	0.966635	0.8177
0.5	302	0.925197	1.0000	-	0.027344	-	78	1.022700	0.8177

Παρατηρούμε ότι ο σχηματισμός 27.3.302 οδηγεί στην καλύτερη τιμή του κριτηρίου CVF_1 για $\rho < 0$ και $\rho > 0$, ενώ συγχρόνως επιτυγχάνει τη μέγιστη τιμή της D_{eff}. Συνεπώς, συνίσταται η χρήση του, βάσει του κριτηρίου CVF_1. Αναφορικά με το κριτήριο CVF_3, προτείνεται η χρήση διαφορετικών σχηματισμών για κάθε περίπτωση.

Λόγω του μεγάλου πλήθους βέλτιστων γεωμετρικά μη ισόμορφων ορθογώνιων σχηματισμών με 27 εκτελέσεις και με 4 και 5 παράγοντες, θα παρουσιάσουμε μόνο ένα μικρό μέρος από αυτούς, στους Πίνακες 11.4 και 11.5 αντίστοιχα.

Πίνακας 11.4: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 4 παράγοντες για την Περίπτωση 1

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	63428	1.215190	0.9871	-	3503.150000	-	59860	0.999909	0.9211
-0.4	63428	1.172150	0.9871	-	1044.530000	-	11708	0.999947	0.8910
-0.3	63428	1.129110	0.9871	-	267.742000	-	21214	0.999970	0.8945
-0.2	63428	1.086080	0.9871	-	53.924600	-	49248	0.999996	0.9522
-0.1	63428	1.043040	0.9871	-	6.076000	-	15954	0.999999	0.8742
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	63478	0.956962	0.9871	-	0.823565	-	53338	0.999983	0.8454
0.2	63478	0.913924	0.9871	-	0.272679	-	45736	0.999974	0.9166
0.3	63478	0.870886	0.9871	-	0.059683	-	53648	0.999972	0.8813
0.4	63478	0.827848	0.9871	-	0.008933	-	32340	0.999997	0.8385
0.5	63478	0.784810	0.9871	-	0.000854	-	57745	0.999967	0.8519

Πίνακας 11.5: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 5 παράγοντες για την Περίπτωση 1

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	579768	1.434400	0.8022	-	39903.100000	-	844656,857172	1.500000	0.8468,0.8441
-0.4	579768	1.347520	0.8022	-	7864.820000	-	844656,857172, 572555,859213	1.400000	0.8468,0.8441, 0.8426,0.8387
-0.3	579768	1.260640	0.8022	-	1292.340000	-	844656,857172, 572555,859213	1.300000	0.8468,0.8441, 0.8426,0.8387
-0.2	579768	1.173760	0.8022	-	161.018000	-	844656,857172 572555,859213	1.200000	0.8468,0.8441, 0.8426,0.8387
-0.1	579768	1.086880	0.8022	-	10.764000	-	844656,857172, 572555,617141	1.100000	0.8468,0.8441, 0.8426,0.8498
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	844469	0.994140	0.8993	-	0.437670	-	796760	1.030100	0.8535
0.2	844469	0.998281	0.8993	-	0.071481	-	579836	1.077000	0.9180
0.3	844469	0.982421	0.8993	-	0.007022	-	579836	1.139900	0.9180
0.4	844469	0.976562	0.8993	-	0.000417	-	579836	1.215300	0.9180
0.5	844469, 579836	0.970702	0.8993, 0.9180	-	0.000013	-	579836	1.301000	0.9180

Συνεχίζουμε με τον Πίνακα 11.6, όπου παρουσιάζουμε τα αποτελέσματα για τους σχηματισμούς με 36 εκτελέσεις και 3 παράγοντες.

Πίνακας 11.6: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 36 εκτελέσεις και 3 παράγοντες για την Περίπτωση 1

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	220	1.008850	0.9373	-	634.315000	-	220	1.010310	0.9373
-0.4	220	1.007080	0.9373	-	268.594000	-	220	0.974878	0.9373
-0.3	220	1.005310	0.9373	-	100.743000	-	257	0.991657	0.8992
-0.2	220	1.003540	0.9373	-	30.958700	-	257	0.970157	0.8992
-0.1	220	1.001770	0.9373	-	5.894870	-	132	1.001590	0.9281
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	220	0.998229	0.9373	-	1.743990	-	46	0.993387	0.8863
0.2	220	0.996458	0.9373	-	1.073740	-	46	1.012370	0.8863
0.3	220	0.994687	0.9373	-	0.464066	-	46	1.064200	0.8863
0.4	220	0.992916	0.9373	-	0.151165	-	46	1.135390	0.8863
0.5	220	0.991145	0.9373	-	0.036133	-	46	1.231230	0.8863

Παρατηρούμε στον Πίνακα 11.6 ότι ο σχηματισμός 36.3.220 είναι η βέλτιστη επιλογή, σύμφωνα με το κριτήριο CVF_1. Επιτυγχάνει επίσης μια υψηλή τιμή της D_{eff}. Σχετικά τώρα με το κριτήριο CVF_3, προτείνεται η χρήση του σχηματισμού 36.3.46 για θετικές τιμές συσχέτισης, ενώ για αρνητικές τιμές του ρ, συστήνονται διαφορετικοί σχηματισμοί.

Ακολουθεί η κλάση των σχηματισμών με 36 εκτελέσεις και 4 παράγοντες, δίνοντας τα αποτελέσματα στον Πίνακα 11.7. Σύμφωνα με τα αποτελέσματα που λάβαμε, συμπεραίνουμε ότι ο σχηματισμός 36.4.1163 είναι η καλύτερη επιλογή, αν αναμένουμε ότι η συσχέτιση θα είναι είτε θετική είτε αρνητική, βάσει του κριτηρίου CVF_1. Αντιθέτως, για αρνητικές τιμές συσχέτισης και βάσει του κριτηρίου CVF_3, ο σχηματισμός 36.4.557 δίνει τα καλύτερα αποτελέσματα, ενώ για $\rho > 0$, προτείνεται ο σχηματισμός 36.4.1252.

Πίνακας 11.7: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 36 εκτελέσεις και 4 παράγοντες για την Περίπτωση 1

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	1163	1.130310	0.9744	-	4816.830000	-	557	0.910525	0.9537
-0.4	1163	1.104250	0.9744	-	1444.560000	-	557	0.896266	0.9537
-0.3	1163	1.071800	0.9744	-	374.051000	-	557	0.892590	0.9537
-0.2	1163	1.052120	0.9744	-	77.035100	-	557	0.904383	0.9537
-0.1	1163	1.026060	0.9744	-	9.493750	-	557	0.938043	0.9537
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	1163	0.973939	0.9744	-	1.029460	-	1252	0.965807	0.8637
0.2	1163	0.947877	0.9744	-	0.351844	-	1252	0.948147	0.8637
0.3	1163	0.921816	0.9744	-	0.077996	-	1252	0.951198	0.8637
0.4	1163	0.895755	0.9744	-	0.011755	-	1252	0.976335	0.8637
0.5	1163	0.869694	0.9744	-	0.001129	-	1252	1.021230	0.8637

11.3.2 Περίπτωση 2: MA(1) Μορφή Συσχέτισης

Σε αυτήν την ενότητα, αναφέρουμε τα αποτελέσματα σχετικά με τη MA(1) μορφή συσχέτισης, ξεκινώντας με τους γεωμετρικά μη ισόμορφους ορθογώνιους σχηματισμούς με 18 εκτελέσεις και 3 παράγοντες. Οι καλύτεροι σχηματισμοί αναφορικά με τις τιμές των CVF κριτηρίων, δίνονται στον Πίνακα 11.8.

Πίνακας 11.8: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 18 εκτελέσεις και 3 παράγοντες για την Περίπτωση 2

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	68	0.999569	0.9347	58	0.006852	0.9347	9	1.000370	0.9552
-0.4	68	0.999655	0.9347	58	0.122567	0.9347	42	1.000200	0.9552
-0.3	68	0.999742	0.9347	58	0.366900	0.9347	86	1.015470	0.9347
-0.2	68	0.999828	0.9347	58	0.659465	0.9347	86	1.005730	0.9347
-0.1	68	0.999914	0.9347	58	0.896241	0.9347	86	1.000370	0.9347
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	68	1.000090	0.9347	105	1.013910	0.8458	31	0.993048	0.9281
0.2	68	1.000170	0.9347	50	1.006170	0.8939	31	0.990301	0.9281
0.3	68	1.000260	0.9347	31	1.002490	0.9281	31	0.991989	0.9281
0.4	68	1.000340	0.9347	95	0.990306	0.8346	31	0.998241	0.9281
0.5	68	1.000430	0.9347	52	0.959064	0.9667	31	0.999569	0.9281

Παρατηρούμε ότι κυριαρχεί ο σχηματισμός 18.3.68 σύμφωνα με το CVF_1 κριτήριο. Κατά το CVF_2 κριτήριο, προτείνεται η χρήση του σχηματισμού 18.3.58 για $\rho < 0$, ενώ για $\rho > 0$, προτείνονται διαφορετικοί σχηματισμοί κατά περίπτωση. Αναφορικά τώρα με το τρίτο κριτήριο, ο σχηματισμός 18.3.31 αποτελεί την καλύτερη επιλογή για $\rho > 0$. Ωστόσο, για $\rho = -0.1$, -0.2 ή -0.3, ο σχηματισμός 18.3.86 είναι η καλύτερη επιλογή, ενώ διαφορετικοί σχηματισμοί προτείνονται για τις υπόλοιπες τιμές συσχέτισης.

Συνεχίζουμε τις συγκρίσεις με τους ορθογώνιους σχηματισμούς με 18 εκτελέσεις και 4 παράγοντες. Τα αποτελέσματα δίνονται στον Πίνακα 11.9. Παρατηρούμε ότι ο σχηματισμός 18.4.844 είναι η καλύτερη επιλογή, τόσο για αρνητικές όσο και για θετικές συσχετίσεις, βάσει των CVF_1 τιμών του. Παρ΄ όλα αυτά, δίνει χαμηλή τιμή της D_{eff}, συνεπώς συνίσταται προσοχή στη χρήση του. Για αρνητικές τιμές συσχέτισης και βάσει του CVF_2 κριτηρίου, συστήνεται ο σχηματισμός 18.4.635. Για θετικές τιμές συσχέτισης, το ίδιο κριτήριο προτείνει τη χρήση διαφορετικών σχηματισμών. Τέλος, το τρίτο κριτήριο και για $\rho < 0$, συνιστά το σχηματισμό 18.4.539, ενώ για θετικές συσχετίσεις, συνιστά το σχηματισμό 18.4.473.

Πίνακας 11.9: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 18 εκτελέσεις και 4 παράγοντες για την Περίπτωση 2

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	844	0.999581	0.5017	635	0.001754	0.8274	539	1.004560	0.7757
-0.4	844	0.999665	0.5017	635	0.080923	0.8274	539	0.998198	0.7757
-0.3	844	0.999749	0.5017	635	0.315555	0.8274	539	0.994612	0.7757
-0.2	844	0.999832	0.5017	635	0.643261	0.8274	539	0.993763	0.7757
-0.1	844	0.999916	0.5017	635	0.914679	0.8274	539	0.995588	0.7757
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	844	1.000080	0.5017	132	1.001970	0.7243	473	0.995175	0.7387
0.2	844	1.000170	0.5017	278	1.00909	0.7088	473	0.993106	0.7387
0.3	844	1.000250	0.5017	361	0.792373	0.6982	473	0.993648	0.7387
0.4	844	1.000340	0.5017	320	0.432700	0.6982	473	0.996625	0.7387
0.5	844	1.000420	0.5017	320	0.151273	0.6982	473	1.001840	0.7387

Στη συνέχεια, στον Πίνακα 11.10 παρουσιάζουμε τα αποτελέσματα των ορθογώνιων σχηματισμών με 27 εκτελέσεις και 3 παράγοντες.

Πίνακας 11.10: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 3 παράγοντες για την Περίπτωση 2

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	379	0.830892	0.9376	340	0.000218	1.0000	41	1.000900	0.8681
-0.4	379	0.864714	0.9376	340	0.014748	1.0000	41	0.986422	0.8681
-0.3	379	0.898535	0.9376	340	0.086696	1.0000	41	0.977834	0.8681
-0.2	379	0.932357	0.9376	340	0.266748	1.0000	41	0.976413	0.8681
-0.1	379	0.966178	0.9376	340	0.580310	1.0000	41	0.983446	0.8681
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	379	1.033820	0.9376	340	1.438830	1.0000	173	0.991210	0.9104
0.2	379	1.067640	0.9376	340	1.770980	1.0000	173	0.988874	0.9104
0.3	379	1.101460	0.9376	340	1.876180	1.0000	173	0.993452	0.9104
0.4	379	1.135290	0.9376	340	1.692910	1.0000	173	1.005090	0.9104
0.5	379	1.169110	0.9376	340	1.258440	1.0000	173	1.023580	0.9104

Από τον Πίνακα 11.10, συμπεραίνουμε ότι για αρνητικές τιμές συσχέτισης, οι σχηματισμοί 27.3.379 και 27.3.340 είναι οι καλύτερες επιλογές βάσει των κριτηρίων CVF_1 και CVF_2, αντίστοιχα. Παρατηρούμε επίσης ότι ο σχηματισμός 27.3.340 επιτυγχάνει τη μέγιστη τιμή της D_{eff}, συνεπώς συνίσταται ιδιαίτερα η χρήση του. Αναφορικά τώρα με το CVF_3 κριτήριο, προτείνονται οι σχηματισμοί 27.3.41 για $\rho < 0$ και 27.3.173 για $\rho > 0$.

Συνεχίζουμε με τα αποτελέσματα για 27 εκτελέσεις και 4 παράγοντες, στον Πίνακα 11.11.

Πίνακας 11.11: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 4 παράγοντες για την Περίπτωση 2

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	16354	0.999927	0.8714	64475	0.000240	0.8750	63288	0.999961	0.9519
-0.4	16354	0.999941	0.8714	64475	0.024571	0.8750	25019	0.999983	0.9026
-0.3	16354	0.999956	0.8714	64475	0.147553	0.8750	28655	0.999992	0.8108
-0.2	16354	0.999971	0.8714	64475	0.408889	0.8750	61776	0.999985	0.8968
-0.1	16354	0.999985	0.8714	64475	0.742528	0.8750	47732	0.999998	0.8208
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	30746	0.999994	0.7919	63878	1.038610	0.8974	45748	0.999999	0.9167
0.2	30746	0.999988	0.7919	15262	0.999438	0.8547	56404	0.999997	0.8746
0.3	30746	0.999982	0.7919	60197	0.999852	0.8606	39610	0.999983	0.8706
0.4	30746	0.999976	0.7919	46118	0.999975	0.9634	9551	0.999965	0.8974
0.5	30746	0.999970	0.7919	33454	0.999903	0.8553	22573	0.999981	0.8206

Το CVF_1 κριτήριο συνιστά τη χρήση του σχηματισμού 27.4.16354 για αρνητικές τιμές

συσχέτισης, ενώ $\rho > 0$, ο σχηματισμός 27.4.30746 αποτελεί τη βέλτιστη επιλογή, δίνοντας όμως μια μέτρια τιμή της D_{eff}. Σύμφωνα με το CVF_2 κριτήριο, επιλέγεται ο σχηματισμός 27.4.64475 για $\rho < 0$, ενώ για θετικές τιμές συσχέτισης, προτείνονται διαφορετικοί σχηματισμοί. Τέλος, παρατηρούμε μια μεγάλη διαφοροποίηση στους καλύτερους σχηματισμούς βάσει του CVF_3 κριτηρίου.

Ακολουθεί ο Πίνακας 11.12, όπου δίνονται οι καλύτερες επιλογές σχηματισμών με 27 εκτελέσεις και 5 παράγοντες. Λόγω του μεγάλου πλήθους αυτών, παραθέτουμε μόνο ένα μικρό μέρος τους.

Πίνακας 11.12: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 5 παράγοντες για την Περίπτωση 2

$\boldsymbol{\rho}$	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	308606	1.315910	0.8064	799426	0.000137	0.8207	675776	1.765300	0.8121
-0.4	635870	1.149190	0.8025	675776	0.005912	0.8121	579836	1.005200	0.9180
-0.3	308606	1.189550	0.8064	635870	0.216692	0.8025	675776	1.457040	0.8121
-0.2	308606	1.126360	0.8064	635870	0.585836	0.8025	675776	1.303536	0.8121
-0.1	308606	1.063180	0.8064	294862	0.934435	0.8086	675776	1.150860	0.8121
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	505762	1.047030	0.8127	368309	1.483700	0.8088	596055	1.145080	0.8088
0.2	505762	1.094060	0.8127	368309	1.583650	0.8088	596055	1.291050	0.8088
0.3	505762	1.141090	0.8127	368309	1.151290	0.8088	596055	1.437530	0.8088
0.4	505762	1.188110	0.8127	668663	1.378313	0.8426	596055	1.584340	0.8088
0.5	505762	1.235140	0.8127	668663	1.159864	0.8426	596055	1.731370	0.8088

Ακολουθούν οι καλύτεροι σχηματισμοί με 36 εκτελέσεις και 3 παράγοντες. Τα αποτελέσματα δίνονται στον Πίνακα 11.13.

Πίνακας 11.13: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 36 εκτελέσεις και 3 παράγοντες για την Περίπτωση 2

$\boldsymbol{\rho}$	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	242	1.001230	0.9681	24	0.101900	0.9892	184	1.030950	0.9952
-0.4	242	1.000980	0.9681	24	0.783061	0.9892	184	0.999090	0.9952
-0.3	242	1.000740	0.9681	132	1.009980	0.9281	239	0.999883	0.9490
-0.2	242	1.000490	0.9681	168	1.028440	0.9552	201	1.000040	0.9681
-0.1	242	1.000250	0.9681	70	1.004660	0.9267	226	1.000170	0.9681
0.0	-	1.00000	-	-	1.000000	-	-	1.000000	-
0.1	242	0.999754	0.9681	173	1.005300	0.9803	73	0.979562	0.8863
0.2	242	0.999508	0.9681	256	1.00193	0.8992	73	0.961050	0.8863
0.3	242	0.999263	0.9681	251	0.922398	0.8918	73	0.944581	0.8863
0.4	242	0.999017	0.9681	251	0.675925	0.8918	73	0.930254	0.8863
0.5	242	0.998771	0.9681	251	0.409852	0.8918	86	0.922874	0.8891

Προτείνεται η χρήση του σχηματισμού 36.3.242 βάσει του CVF_1 κριτηρίου. Αναφορικά όμως με τα κριτήρια CVF_2 και CVF_3, ο πειραματιστής μπορεί να επιλέξει μεταξύ διαφορετικών σχηματισμών, αναλόγως της αναμενόμενης συσχέτισης.

Η τελευταία κλάση σχηματισμών αφορά αυτούς με 36 εκτελέσεις και 4 παράγοντες. Τα αποτελέσματα δίνονται στον Πίνακα 11.14. Βάσει των αποτελεσμάτων μας, προτείνεται η χρήση του σχηματισμού 36.4.1274 τόσο για αρνητικές όσο και για θετικές τιμές συσχέτισης, σύμφωνα με το CVF_1 κριτήριο. Τα υπόλοιπα δύο κριτήρια, συνιστούν διαφορετικούς σχηματισμούς κατά περίπτωση, οπότε ο πειραματιστής μπορεί να βασιστεί στον Πίνακα 11.14 και να επιλέξει τους κατάλληλους.

Πίνακας 11.14: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 36 εκτελέσεις και 4 παράγοντες για την Περίπτωση 2

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	1274	0.999960	0.9012	84	0.139208	0.8620	1774	1.000560	0.9191
-0.4	1274	0.999960	0.9012	139	0.848827	0.9559	2425	1.000380	0.8279
-0.3	1274	0.999970	0.9012	276	1.008480	0.9702	1366	1.000140	0.9426
-0.2	1274	0.999980	0.9012	3	1.002200	0.8620	1564	1.000410	0.8872
-0.1	1274	0.999990	0.9012	803	1.000090	0.9339	809	1.000090	0.9265
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	1274	1.000010	0.9012	2292	1.003460	0.8915	852	1.000390	0.8274
0.2	1274	1.000020	0.9012	2342	0.987685	.9159	1976	1.000050	0.9195
0.3	1274	1.000030	0.9012	2299	0.881513	0.8409	2298	1.000070	0.8409
0.4	1274	1.000040	0.9012	2299	0.563872	0.8409	1089	0.999316	0.8985
0.5	1274	1.000050	0.9012	2299	0.261548	0.8409	248	0.999070	0.8713

11.3.3 Περίπτωση 3: AR(1) Μορφή Συσχέτισης

Στην ενότητα αυτή θα εξετάσουμε την τρίτη μορφή συσχέτισης, AR(1). Ξεκινάμε με τον Πίνακα 11.15, παραθέτοντας τα αποτελέσματα για τους σχηματισμούς με 18 εκτελέσεις και 3 παράγοντες.

Πίνακας 11.15: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 18 εκτελέσεις και 3 παράγοντες για την Περίπτωση 3

ρ	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	9	1.007320	0.9552	107	0.977476	0.8458	21	0.983442	0.9281
-0.4	9	0.978570	0.9552	93	1.004980	0.8346	9	1.000930	0.9552
-0.3	23	1.008190	0.9552	128	1.019640	0.9065	78	1.005370	0.9667
-0.2	23	0.987587	0.9552	84	1.029930	0.9347	3	1.004660	0.9552
-0.1	23	0.984981	0.9552	127	1.000000	0.9065	46	1.001450	0.8939
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	127	0.991032	0.9065	58	1.064480	0.9347	33	1.004530	0.9281
0.2	127	1.002880	0.9065	58	1.230400	0.9347	27	1.013470	0.9281
0.3	127	1.042350	0.9065	58	1.558700	0.9347	84	1.000790	0.9347
0.4	101	1.117950	0.8458	58	2.187280	0.9347	127	1.08823	0.9065
0.5	101	1.237090	0.8458	58	3.440880	0.9347	127	1.27013	0.9065

Παρατηρούμε ότι, για θετικές τιμές του ρ, ο σχηματισμός 18.3.58 αποτελεί την καλύτερη επιλογή, σύμφωνα με το κριτήριο CVF_2. Λόγω τώρα της μεγάλης διαφοροποίησης μεταξύ των καλύτερων σχηματισμών, κατά τα υπόλοιπα δύο κριτήρια, ειδικά για το CVF_3, ο πειραματιστής μπορεί να βασιστεί στον Πίνακα 11.15 και να επιλέξει τους κατάλληλους.

Παραθέτουμε τώρα τα αποτελέσματα των σχηματισμών με 18 εκτελέσεις και 4 παράγοντες, στον Πίνακα 11.16. Βάσει του Πίνακα 11.16, προτείνονται διαφορετικοί σχηματισμοί κατά περίπτωση. Επίσης, να τονίσουμε ότι αρκετοί εξ αυτών ναι μεν είναι βέλτιστοι σύμφωνα με τα θεωρούμενα κριτήρια, αλλά έχουν χαμηλές τιμές της D_{eff}, οπότε συνίσταται προσοχή στην επιλογή των κατάλληλων από τον πειραματιστή.

Πίνακας 11.16: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 18 εκτελέσεις και 4 παράγοντες για την Περίπτωση 3

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	708	0.994036	0.4970	468	1.001300	0.7374	888	0.997968	0.5017
-0.4	516	1.015060	0.7395	627	0.994643	0.8212	437	1.000440	0.8534
-0.3	552	1.001190	0.8212	512	1.006000	0.7395	834	0.997552	0.5017
-0.2	838	1.000760	0.5017	716	1.001200	0.5451	239	0.999375	0.7536
-0.1	715	1.000090	0.5979	492	1.001310	0.7930	533	1.000210	0.7395
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	582	1.000210	0.8274	763	0.966713	0.8440	118	0.999385	0.7467
0.2	330	0.999365	0.6982	763	1.004130	0.8440	352	1.000040	0.7088
0.3	479	1.001430	0.7374	664	1.118640	0.7527	702	0.999049	0.7667
0.4	236	1.001880	0.6952	664	1.320710	0.7527	528	1.002320	0.7387
0.5	98	1.000090	0.7369	379	1.665950	0.7896	376	1.000270	0.7451

Συνεχίζουμε με τους καλύτερους σχηματισμούς με 27 εκτελέσεις και 3 παράγοντες. Τα αποτελέσματα δίνονται στον Πίνακα 11.17.

Πίνακας 11.17: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 3 παράγοντες για την Περίπτωση 3

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	142	1.000410	0.9169	377	0.436949	0.9923	43	1.000860	0.9096
-0.4	42	1.002780	0.8681	377	0.372963	0.9923	302	0.998776	1.0000
-0.3	353	0.996802	0.9921	377	0.384394	0.9923	64	0.996345	0.8235
-0.2	385	0.987117	0.9376	377	0.463299	0.9923	170	0.996008	0.8795
-0.1	385	0.979771	0.9376	377	0.639749	0.9923	13	1.000080	0.9118
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	379	1.048240	0.9376	334	1.629330	1.0000	173	0.996923	0.9104
0.2	379	1.128950	0.9376	334	2.887260	1.0000	173	1.016310	0.9104
0.3	9	1.238900	0.9118	334	5.606580	1.0000	173	1.071970	0.9104
0.4	9	1.392140	0.9118	334	12.049500	1.0000	44	1.187280	0.8235
0.5	9	1.609480	0.9118	334	29.066300	1.0000	173	1.420330	0.9104

Σύμφωνα με τον Πίνακα 11.17, το κριτήριο CVF_2 προτείνει τη χρήση των σχηματισμών 27.3.377 και 27.3.334 για αρνητικές και θετικές τιμές συσχέτισης, αντίστοιχα. Οι σχηματισμοί αυτοί, επιτυγχάνουν επίσης υψηλές τιμές της D_{eff}. Αναφορικά με τα υπόλοιπα κριτήρια, παρατηρείται και εδώ υψηλή διαφοροποίηση μεταξύ των βέλτιστων σχηματισμών.

Ακολουθούν τα αποτελέσματα με τους σχηματισμούς με 27 εκτελέσεις και 4 παράγοντες, τα οποία παρατίθενται στον Πίνακα 11.18.

Πίνακας 11.18: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 4 παράγοντες για την Περίπτωση 3

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	55359	0.999403	0.9349	5242	0.999710	0.9026	40661	0.999942	0.8569
-0.4	62651	0.999991	0.9223	36142	0.996801	0.8570	52118	0.999994	0.8868
-0.3	12736	0.999998	0.8046	63929	0.992469	0.9007	60857	0.999916	0.8921
-0.2	23438	0.999999	0.8370	63878	0.975335	0.8974	49126	0.999960	0.8515
-0.1	54220	0.999999	0.8262	63878	0.914909	0.8974	56771	0.999991	0.9051
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	50362	0.999985	0.8068	63110	1.249020	0.9487	48979	0.999998	0.8905
0.2	52828	0.999940	0.8269	63110	1.725610	0.9487	13402	1.000040	0.7581
0.3	11866	0.999608	0.8141	63110	2.648940	0.9487	11883	1.000060	0.7684
0.4	52857	0.996977	0.8269	63110	4.560370	0.9487	51269	1.000030	0.7184
0.5	46506	0.999441	0.7341	63110	8.934180	0.9487	6697	1.000080	0.8363

Εξακολουθεί να υφίσταται μεγάλη διαφοροποίηση μεταξύ των καλύτερων σχηματισμών, σύμφωνα με τον Πίνακα 11.18. Εξαίρεση αποτελεί ο σχηματισμός 27.4.63110, οποίος αποτελεί τη βέλτιστη επιλογή για θετικές τιμές συσχέτισης, βάσει του κριτηρίου CVF_2.

Συνεχίζουμε με τους καλύτερους σχηματισμούς 27 εκτελέσεων και με 5 παράγοντες, στον Πίνακα 11.19. Θα παρουσιάσουμε πάλι ένα μικρό μέρος αυτών.

Πίνακας 11.19: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 27 εκτελέσεις και 5 παράγοντες για την Περίπτωση 3

$\boldsymbol{\rho}$	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	296143	1.931270	0.8029	791748	5.744750	0.8257	771432	3.171470	0.8061
-0.4	513312	1.567500	0.8131	791748	1.346940	0.8257	771432	2.308810	0.8061
-0.3	308606	1.342710	0.8064	791748	2.183470	0.8257	771432	1.218530	0.8061
-0.2	308606	1.188030	0.8064	791748	1.532020	0.8257	771432	1.109630	0.8061
-0.1	308606	1.077450	0.8064	791748	1.037520	0.8257	675776	1.173130	0.8121
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	508592	1.059450	0.8016	855631	1.810760	0.8331	596055	1.164120	0.8088
0.2	508592	1.146720	0.8016	855631	3.473520	0.8331	596055	1.376660	0.8088
0.3	508592	1.272540	0.8016	855631	1.177390	0.8331	596055	1.656390	0.8088
0.4	508592	1.449680	0.8016	855631	14.58520	0.8331	596055	2.032940	0.8088
0.5	508592	1.700150	0.8016	855631	31.62430	0.8331	491655	2.599780	0.8599

Ακολουθεί ο Πίνακας 11.20 με τα αποτελέσματα των βέλτιστων σχηματισμών με 36 εκτελέσεις και 3 παράγοντες. Λόγω πάλι της μεγάλης διαφοροποίησης που παρατηρείται, ο πειραματιστής αφήνεται να επιλέξει τους κατάλληλους, βάσει των CVF κριτηρίων.

Πίνακας 11.20: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 36 εκτελέσεις και 3 παράγοντες για την Περίπτωση 3

$\boldsymbol{\rho}$	**OAs**	$\boldsymbol{CVF_1}$	D_{eff}	**OAs**	$\boldsymbol{CVF_2}$	D_{eff}	**OAs**	$\boldsymbol{CVF_3}$	D_{eff}
-0.5	84	1.229900	0.8891	64	2.246740	0.9267	232	1.288030	0.8992
-0.4	84	1.156680	0.8891	242	1.537970	0.9681	232	1.145040	0.8992
-0.3	250	1.094130	0.8918	242	1.088780	0.9681	230	1.038310	0.9373
-0.2	218	1.036760	0.9681	254	0.992205	0.9373	224	0.996103	0.9373
-0.1	218	1.004750	0.9681	203	1.007200	0.9681	175	0.999960	0.9952
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	84	0.990009	0.8891	156	0.962143	0.9552	73	0.983790	0.8863
0.2	84	0.996058	0.8891	156	1.004400	0.9552	73	0.976458	0.8863
0.3	156	0.998716	0.9552	84	0.999543	0.8891	82	0.999382	0.8891
0.4	171	1.003980	0.9552	141	1.003420	0.9552	73	0.984388	0.8863
0.5	6	1.003900	0.9892	52	0.994254	0.8863	73	1.003770	0.8863

Η τελευταία περίπτωση αφορά τους σχηματισμούς με 36 εκτελέσεις και 4 παράγοντες. Τα αποτελέσματα δίνονται στον Πίνακα 11.21.

Πίνακας 11.21: Οι καλύτεροι ορθογώνιοι σχηματισμοί με 36 εκτελέσεις και 4 παράγοντες για την Περίπτωση 3

ρ	**OAs**	CVF_1	D_{eff}	**OAs**	CVF_2	D_{eff}	**OAs**	CVF_3	D_{eff}
-0.5	858	1.095570	0.8212	908	1.974720	0.8274	698	1.007340	0.8703
-0.4	858	1.031900	0.8212	1254	1.422460	0.8817	1796	1.002350	0.8908
-0.3	858	0.995321	0.8212	2298	1.077980	0.8409	2105	1.002290	0.8366
-0.2	858	0.999021	0.8212	2445	1.001950	0.8401	1698	1.000070	0.9026
-0.1	858	0.999015	0.8212	1917	1.000600	0.9114	2056	1.000080	0.8885
0.0	-	1.000000	-	-	1.000000	-	-	1.000000	-
0.1	613	1.000010	0.9603	1262	1.000110	0.8637	759	1.000150	0.9521
0.2	1122	1.00030	0.9556	638	0.999436	0.8927	2086	1.000730	0.8486
0.3	249	1.000330	0.9316	831	1.003040	0.8212	5	0.999771	0.8954
0.4	322	1.000840	0.8898	103	1.002500	0.9614	274	0.997540	0.9536
0.5	48	1.002760	0.8890	138	0.972818	0.9559	133	1.000190	0.9650

Σύμφωνα με τα αποτελέσματά μας, προτείνεται ο σχηματισμός 36.4.858 για $\rho < 0$, σύμφωνα με το CVF_1 κριτήριο. Για τις υπόλοιπες περιπτώσεις, τα θεωρούμενα κριτήρια συνιστούν διαφορετικούς σχηματισμούς κατά περίπτωση.

11.4 Συμπεράσματα

Ο στόχος αυτού του κεφαλαίου, ήταν η αξιολόγηση ορθογώνιων σχηματισμών τριών επιπέδων, στην περίπτωση συσχετισμένων παρατηρήσεων. Στην πράξη, υπάρχουν αρκετές περιπτώσεις όπου αυτού του είδους παρατηρήσεις είναι αναπόφευκτες. Ως αποτέλεσμα, οι πειραματιστές οφείλουν να είναι ιδιαίτερα προσεκτικοί στην επιλογή ορθογώνιων σχηματισμών, ειδικά όταν ο σκοπός είναι η χρήση τους ως σχεδιασμούς αποκριτικών επιφανειών δευτέρας τάξης. Παρόλο που έχουν γίνει σχετικές μελέτες για παραγοντικούς σχεδιασμούς δύο επιπέδων, ήταν επιθυμητή η εξέταση της περίπτωσης όπου ορθογώνιοι σχηματισμοί τριών επιπέδων θα χρησιμοποιηθούν για τέτοιο σκοπό. Οπότε, στο κεφάλαιο αυτό, προτείναμε και χρησιμοποιήσαμε τρία διαφορετικά κριτήρια και διενεργήσαμε μια εκτενής συγκριτική μελέτη ώστε να αντλήσουμε τους καλύτερους σχηματισμούς από τη μεγάλη κλάση των γεωμετρικά μη ισόμορφων ορθογώνιων σχηματισμών. Αξίζει να αναφέρουμε ότι σε πολλές περιπτώσεις, παρατηρήθηκε μια μεγάλη διαφοροποίηση μεταξύ των καλύτερων σχηματισμών. Επιπλέον, τονίζουμε ότι οι πειραματιστές πρέπει να λαμβάνουν υπόψην και τις αντίστοιχες τιμές της $D-$αποδοτικότητας, ώστε να έχουν συγχρόνως και ένα μέτρο εκτιμητικής ικανότητας για κάθε σχηματισμό.

ΚΕΦΑΛΑΙΟ **12**

Μελέτη Ορθογώνιων Σχηματισμών Τριών Επιπέδων και Αξιολόγηση της Ικανότητάς τους στη Διάκριση Μοντέλων

All models are wrong,
but some are useful.

—George E.P. Box (1919–2013)

Τα τελευταία χρόνια, έχει αυξηθεί το ενδιαφέρον των ερευνητών αναφορικά με το πεδίο της διάκρισης μοντέλων. Λόγω του προβλήματος των ταυτόσημων μοντέλων, τα οποία είναι και μη διακρίσιμα, έχουν προταθεί αρκετά κριτήρια που έχουν ως στόχο την αξιολόγηση της διακριτικής ικανότητας ενός σχεδιασμού. Σε αυτό το κεφάλαιο, επιλέγουμε τρία από αυτά τα κριτήρια, σε συνδυασμό με ένα νέο που προτείνουμε, το οποίο αφενός μεν τα συνδυάζει και αφετέρου αποσκοπεί στην αξιολόγηση της συνολικής διακριτικής ικανότητας ενός σχεδιασμού. Όλα αυτά εφαρμόζονται για να αξιολογηθεί η κλάση των γεωμετρικά μη ισόμορφων, εύρωστων ως προς τα μοντέλα, ορθογώνιων σχηματισμών τριών επιπέδων με 27 εκτελέσεις και έως 10 παράγοντες.

12.1 Ερευνητικό Πρόβλημα

Ένας από τους κύριους στόχους των πειραματικών σχεδιασμών, είναι ο εντοπισμός του συνόλου των ενεργών παραγόντων μεταξύ ενός πολύ μεγαλύτερου συνόλου υποψηφίων, που ενδεχομένως να έχουν αντίκτυπο στην απόκριση του πειράματος. Λόγω του γεγονότος ότι το πραγματικό μοντέλο δεν είναι εκ των προτέρων γνωστό, ο πειραματιστής πρέπει να επιλέξει σχεδιασμούς που να είναι κατάλληλοι να εκτιμήσουν μοντέλα από ένα σύνολο ανταγωνιστικών μοντέλων. Οι συμβατικοί σχεδιασμοί, όπως οι κλασματικοί παραγοντικοί σχεδιασμοί ελάχιστης απόκλισης (minimum aberration fractional factorial designs) των Wu και Hamada [179], ή οι μη-κανονικοί ορθογώνιοι σχεδιασμοί (non-regular orthogonal designs), συνήθως στη μορφή των Plackett-Burman σχεδιασμών, αποδεικνύονται αναποτελεσματικοί σύμφωνα με την εργασία των Li και Nachtsheim [117].

Το γενικό πλαίσιο των μεθόδων εκτίμησης μοντέλων, περιλαμβάνει τρία κύρια στοιχεία, που δηλώνονται ως $\{\mathcal{F}, \mathcal{C}, \mathcal{D}\}$, όπου $\mathcal{F}$ είναι το σύνολο των πιθανών μοντέλων, γνωστό ως χώρος μοντέλων (model space), $\mathcal{C}$ είναι το προτεινόμενο κριτήριο και $\mathcal{D}$ είναι το σύνολο των υποψηφίων σχεδιασμών, γνωστό ως χώρος σχεδιασμών (design space). Θεωρώντας ότι $\mathcal{F}$=$\{\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_g}\}$ είναι ο χώρος των g πιθανών μοντέλων, όπου το συναρτησιακό $\mathbf{f}$ υποδεικνύει ποιες κύριες επιδράσεις και αλληλεπιδράσεις υπάρχουν στο μοντέλο και δοθέντος ενός παραγοντικού σχεδιασμού, ενδιαφερόμαστε για δύο κλάσεων κριτήρια: το ποσοστό των μοντέλων στο σύνολο $\mathcal{F}$ που είναι εκτιμήσιμα και τη μέση αποδοτικότητα εκτίμησης όλων των μοντέλων του συνόλου $\mathcal{F}$ [97].

Πριν ασχοληθούμε με τις πιο γνωστές εργασίες σε αυτό το πεδίο, ας περιγράψουμε πρώτα την έννοια των πλήρως ταυτόσημων μοντέλων (fully aliased models). Δοθέντος ενός σχεδιασμού, έστω d, θεωρούμε τα μοντέλα $\mathbf{f_1}$ και $\mathbf{f_2}$ με πίνακες μοντέλου $\mathbf{X_1}$ και $\mathbf{X_2}$ αντίστοιχα. Στις περιπτώσεις όπου ο πίνακας προβολής (hat matrix) του $\mathbf{X_1}$ είναι ο ίσος με αυτόν του $\mathbf{X_2}$, οι προβλέψεις τους $\widehat{\mathbf{y_1}} = \mathbf{H_1 y}$ και $\widehat{\mathbf{y_2}} = \mathbf{H_2 y}$ θα είναι οι ίδιες για το σχεδιασμό d, ανεξάρτητα από τις τιμές της απόκρισης $\mathbf{y}$. Υπενθυμίζουμε ότι ο πίνακας προβολής του $\mathbf{X_1}$ ορίζεται ως $\mathbf{H_1} = \mathbf{X_1}(\mathbf{X_1^T X_1})^{-1}\mathbf{X_1^T}$ ενώ του $\mathbf{X_2}$ ως $\mathbf{H_2} = \mathbf{X_2}(\mathbf{X_2^T X_2})^{-1}\mathbf{X_2^T}$. Συνεπώς, τα δύο εξεταζόμενα μοντέλα δε μπορούν να διακριθούν στο σχεδιασμό και καλούνται πλήρως ταυτόσημα, στο σχεδιασμό d. Για να γίνουμε περισσότερο κατανοητοί αναφορικά με το θέμα των ταυτόσημων μοντέλων, θα παραθέσουμε το παρακάτω παράδειγμα.

Παράδειγμα 12.1 *Έστω ο 2^{7-1} κλασματικός παραγοντικός σχεδιασμός, με ορίζουσα σχέση I=ABCF. Έστω επίσης το μοντέλο που περιλαμβάνει όλες τις κύριες επιδράσεις A, B, C, D, E, F, G και τις αλληλεπιδράσεις δύο παραγόντων AB και AC, δηλαδή το μοντέλο (A+B+C+D+E+F+G+AB+AC), καθώς επίσης και ένα δεύτερο μοντέλο που περιλαμβάνει και αυτό όλες τις κύριες επιδράσεις, μαζί με τις αλληλεπιδράσεις AB και BF, δηλαδή το (A+B+C+D+E+F+G+AB+BF). Αυτά τα δύο μοντέλα είναι μεν εκτιμήσιμα, αλλά πλήρως ταυτόσημα, συνεπώς δεν είναι διακρίσιμα.*

Από τις παλαιότερες εργασίες στη βιβλιογραφία ήταν αυτή του Läuter [112], ο οποίος πρότεινε τη μεγιστοποίηση της μέσης τιμής του λογαρίθμου της ορίζουσας, λαμβανόμενη στο σύνολο $\mathcal{F}$. Μια εργασία η οποία επίσης αξίζει να αναφερθεί, είναι αυτή του Srivastava [159], στην οποία εισάγεται ο ορισμός της 'ισχύος επίλυσης (resolving power)' για τη μέτρηση της ικανότητας ενός σχεδιασμού στη διάκριση μοντέλων. Σύμφωνα με τον Srivastava [159], οι παραγοντικές επιδράσεις μπορούν να κατηγοριοποιηθούν σε τρεις κλάσεις ως εξής:

1. Οι επιδράσεις που θεωρούνται σίγουρα αμελητέες.
2. Οι επιδράσεις που συμπεριλαμβάνονται υποχρεωτικά στο μοντέλο.
3. Οι υπόλοιπες επιδράσεις, οι περισσότερες εκ των οποίων είναι αμελητέες ενώ κάποιες ίσως όχι.

Ο Srivastava (1975), εισήγαγε επίσης την έννοια των σχεδιασμών αναζήτησης (search designs) οι οποίοι είναι ικανοί να εκτιμήσουν όλες τις επιδράσεις της κατηγορίας (2) και να αναζητήσουν τις μη αμελητέες επιδράσεις της κατηγορίας (3). Ο χώρος μοντέλων που πρότεινε, ορίζεται ως

$$\mathcal{F}=\{\text{τα μοντέλα που περιλαμβάνουν όλες τις επιδράσεις της κατηγορίας (2) + έως και } q \text{ επιδράσεις της κατηγορίας (3)}\}.$$

Ο Srivastava [159] παρείχε μια λύση για το γραμμικό μοντέλο αναζήτησης, υποθέτοντας ότι ο θόρυβος είναι αμελητέος. Παρ' όλα αυτά, όταν η διασπορά του σφάλματος δεν είναι μικρή, οι σχεδιασμοί που ικανοποιούν τις συνθήκες που αναφέρονται στην εργασία [159] μπορούν να παρουσιάσουν αρκετά ανόμοια συμπεριφορά στη διάκριση μοντέλων. Οπότε, η ανάγκη ποσοτικής αξιολόγησης της ταυτοσημότητας μοντέλων, οδήγησε πολλούς συγγραφείς να προτείνουν διάφορα κριτήρια διάκρισής τους.

Οι Atkinson και Fedorov στις εργασίες τους [11] και [12], συνέστησαν τη χρήση μη-Μπεϋζιανών κριτηρίων, όπως το κριτήριο της T-βελτιστοποίησης. Οι Meyer et al. [132] πρότειναν ένα Μπεϋζιανό κριτήριο, βασιζόμενο στην Kullback-Leibler πληροφορία, ενώ οι Bingham και Chipman [17] πρότειναν ένα κριτήριο που βασίστηκε στην απόσταση Hellinger μεταξύ των προβλεπτικών συναρτήσεων πυκνότητας. Το 1982, οι Cook και Nachtsheim [39] ανέπτυξαν ένα νέο κριτήριο το οποίο γενικεύει τη γραμμική βελτιστοποίηση, στην περίπτωση όπου δε χρειάζεται να είναι γνωστή η ακριβής μορφή του μοντέλου παλινδρόμησης.

Από τις πιο σημαντικές εργασίες στον τομέα αυτό, μπορεί να θεωρηθεί αυτή του Sun [160], στην οποία ο χώρος μοντέλου $\mathcal{F}$ ορίζεται ως

$$\mathcal{F}=\{\text{τα μοντέλα που που περιλαμβάνουν όλες τις κύριες επιδράσεις + έως και } q \text{ αλληλεπιδράσεις δύο παραγόντων}\}.$$

Ο Sun [160] εισήγαγε ένα μέτρο της εκτιμητικής ικανότητας ενός σχεδιασμού εντός μιας κλάσης μοντέλων, γνωστό ως κριτήριο της ικανότητας εκτίμησης (estimation capacity-EC). Το μέτρο αυτό ορίζεται ως ο λόγος του αριθμού των εκτιμήσιμων μοντέλων, προς το συνολικό αριθμό των πιθανών μοντέλων. Επίσης, οι Cheng et al. [33] όρισαν το κριτήριο EC στο πλαίσιο των κλασματικών παραγοντικών σχεδιασμών ελάχιστης απόκλισης.

Θεωρώντας το τυπικό D-κριτήριο ως μέτρο αποδοτικότητας, ο Sun [160] πρότεινε επίσης το κριτήριο της πληροφοριακής ικανότητας (information capacity-IC), το οποίο υπολογίζει τη μέση D-αποδοτικότητα ως προς όλα τα εκτιμήσιμα μοντέλα. Αργότερα, οι Li και Nachtsheim [117] ανέπτυξαν μια νέα κλάση σχεδιασμών, γνωστοί ως παραγοντικοί σχεδιασμοί εύρωστοι ως προς τα μοντέλα (model-robust factorial designs), οι οποίοι μεγιστοποιούν τα EC και IC κριτήρια. Πιο συγκεκριμένα, αναγνώρισαν αρχικά τους σχεδιασμούς που μεγιστοποιούσαν το κριτήριο EC και από αυτούς τους EC-βέλτιστους σχεδιασμούς, επέλεξαν όσους μεγιστοποιούσαν το IC.

Μεταγενέστερα, οι Jones et al. [97] πρότειναν τρία μη-Μπεϋζιανά κριτήρια για τη μελέτη της ταυτοσημότητας μεταξύ δύο γραμμικών μοντέλων. Χρησιμοποιώντας αυτά τα μέτρα και σε συνδυασμό με τα EC και IC κριτήρια, αναγνώρισαν τους βέλτιστους, εύρωστους ως προς τα μοντέλα, ορθογώνιους σχεδιασμούς 18 εκτελέσεων και τους αξιολόγησαν ως προς την ικανότητά τους στη διάκριση μοντέλων.

Στο παρόν κεφάλαιο, θα ξεκινήσουμε με μια σύντομη περιγραφή των κριτηρίων των Jones et al. [97]. Ωστόσο, λόγω του ότι δε συμφωνούν αυτά τα κριτήρια σε αρκετές περιπτώσεις, θα προτείνουμε ένα νέο μέτρο το οποίο τα συνδυάζει και αξιολογεί τη συνολική ικανότητα διάκρισης μοντέλων ενός σχεδιασμού. Επιπλέον, θα εφαρμόσουμε αυτά τα κριτήρια για να αξιολογήσουμε την κλάση των ορθογώνιων σχηματισμών τριών επιπέδων με 27 εκτελέσεις.

12.2 Κριτήρια Διάκρισης Μοντέλων

Στην ενότητα αυτή, θα περιγράψουμε αρχικά τα μέτρα των Jones et al. [97], τα οποία και θα εφαρμοστούν στη συνέχεια για να αξιολογήσουμε τους εύρωστους ως προς τα μοντέλα ορθογώνιους σχεδιασμούς με 27 εκτελέσεις. Αυτά τα κριτήρια είναι:

- Η γωνία υποχώρων (subspace angle-SA).
- Η μέγιστη διαφορά προβλέψεων (maximum prediction difference-MPD).
- Η αναμενόμενη διαφορά προβλέψεων (expected prediction difference-EPD).

Όπως αναφέραμε και στην προηγούμενη ενότητα, δύο μοντέλα είναι πλήρως ταυτόσημα ως προς έναν σχεδιασμό, αν οι πίνακες προβολής, $\mathbf{H_1}$ και $\mathbf{H_2}$, των αντίστοιχων πινάκων μοντέλων $\mathbf{X_1}$ και $\mathbf{X_2}$, είναι οι ίδιοι. Συνεπώς, δύο μοντέλα είναι πλήρως ταυτόσημα αν οι γραμμικοί χώροι $\mathrm{V}(\mathbf{X_1})$ και $\mathrm{V}(\mathbf{X_2})$, που παράγονται από τις στήλες των $\mathbf{X_1}$ και $\mathbf{X_2}$, είναι οι ίδιοι.

Ένας τρόπος μέτρησης του βαθμού της ταυτοσημότητας μοντέλων, είναι μέσω ενός γεωμετρικού μέτρου της εγγύτητας μεταξύ των δύο διανυσματικών χώρων $\mathrm{V}(\mathbf{X_1})$ και $\mathrm{V}(\mathbf{X_2})$. Ένα τέτοιο μέτρο είναι αυτό της γωνίας υποχώρων (SA), που αποτελεί μια γενίκευση της γωνίας μεταξύ δύο επιπέδων στον τρισδιάστατο Ευκλείδειο χώρο. Το μέτρο SA μπορεί οριστεί ως:

$$a_{12} = \max_{\mathbf{v_1}\in \mathrm{V}(\mathbf{X_1})} (\min_{\mathbf{v_2}\in \mathrm{V}(\mathbf{X_2})} \arccos\langle \mathbf{v_1}, \mathbf{v_2}\rangle). \tag{12.1}$$

Από τη σκοπιά του γραμμικού μοντέλου παλινδρόμησης, μια ερμηνεία του κριτηρίου της γωνίας υποχώρων SA είναι η ακόλουθη. Το SA a_{12} είναι το arcsin της μεγαλύτερης πιθανής τιμής της ποσότητας $\|\mathbf{y}$ - $\widehat{\mathbf{y}}_\mathbf{2}\|$, δηλαδή της L_2-απόστασης μεταξύ μιας κανονικοποιημένης απόκρισης $\mathbf{y}$ και της αντίστοιχης προσαρμογής της στο άλλο μοντέλο. Στην προηγούμενη έκφραση έχουμε υποθέσει ότι $\mathbf{f_1}$ είναι το πραγματικό μοντέλο, $\mathbf{X_1}$ είναι ο πίνακας μοντέλου και $\mathbf{y}$ είναι η παρατηρούμενη απόκριση που αποκτάται στο $\mathbf{V}(\mathbf{X_1})$. Επιπλέον, το $\mathbf{f_2}$ είναι ένα άλλο μοντέλο και $\widehat{\mathbf{y}}_\mathbf{2}$ η αντίστοιχη προσαρμοσμένη τιμή. Από τη στιγμή που η arcsin είναι αυστηρά μονότονη συνάρτηση, ισχύει ότι όσο μεγαλύτερο είναι το SA μεταξύ του $\mathbf{V}(\mathbf{X_1})$ και $\mathbf{V}(\mathbf{X_2})$, τόσο μεγαλύτερη είναι και η απόσταση μεταξύ του $\mathbf{y} \in \mathbf{V}(\mathbf{X_1})$ και του $\widehat{\mathbf{y}}_\mathbf{2} \in \mathbf{V}(\mathbf{X_2})$. Σε περιπτώσεις όπου το a_{12} είναι μηδέν, αυτό σημαίνει ότι τα μοντέλα $\mathbf{f_1}$ και $\mathbf{f_2}$ είναι πλήρως ταυτόσημα. Συνεπώς, το κριτήριο έχει ως στόχο την επιλογή του σχεδιασμού που μεγιστοποιεί τη γωνία υποχώρων.

Το δεύτερο κριτήριο διάκρισης μοντέλων, είναι αυτό της μέγιστης διαφοράς προβλέψεων (MPD), το οποίο βασίζεται στις μέγιστες διαφορές μεταξύ των προβλέψεων δύο μοντέλων, πάνω σε όλες τις κανονικοποιημένες αποκρίσεις. Έστω πάλι $\mathbf{X_1}$ και $\mathbf{X_2}$ δύο πίνακες μοντέλων και $\mathbf{H_1}, \mathbf{H_2}$ οι αντίστοιχοι πίνακες προβολής. Αν $\mathbf{H_1}-\mathbf{H_2} = 0$, τότε $\widehat{\mathbf{y}}_\mathbf{1}$ - $\widehat{\mathbf{y}}_\mathbf{2} = (\mathbf{H_1}-\mathbf{H_2})\mathbf{y} = 0$ για κάθε απόκριση $\mathbf{y}$. Το κριτήριο MPD ορίζεται ως:

$$\max_{\|\mathbf{y}\|=1} \|\widehat{\mathbf{y}}_\mathbf{1} - \widehat{\mathbf{y}}_\mathbf{2}\|^2 = \max_{\|\mathbf{y}\|=1} \mathbf{y^T}(\mathbf{H_1} - \mathbf{H_2})(\mathbf{H_1} - \mathbf{H_2})\mathbf{y} \tag{12.2}$$

Το τελευταίο κριτήριο, είναι αυτό της αναμενόμενης διαφοράς προβλέψεων (EPD). Έστω δύο μοντέλα με αντίστοιχους πίνακες μοντέλων $\mathbf{X_1}$ και $\mathbf{X_2}$ και πίνακες προβολής $\mathbf{H_1}$ και $\mathbf{H_2}$. Σε αντίθεση με το MPD που θεωρεί μόνο τον καλύτερο δυνατό διαχωρισμό μεταξύ δύο μοντέλων, το EPD μετράει τη μέση απόσταση μεταξύ δύο προσαρμοσμένων τιμών πάνω σε όλες τις κανονικοποιημένες αποκρίσεις. Ορίζεται ως:

$$E(\|\widehat{\mathbf{y}}_\mathbf{1} - \widehat{\mathbf{y}}_\mathbf{2}\|^2 \mid \|\mathbf{y}\| = 1) = E(\mathbf{y}^T D\mathbf{y} \mid \|\mathbf{y}\| = 1) = \frac{1}{n} Trace(D), \tag{12.3}$$

όπου $D = (\mathbf{H_1} - \mathbf{H_2})(\mathbf{H_1} - \mathbf{H_2})$. Βάσει του ορισμού του EPD, πρέπει να τονιστεί ότι EPD=0 αν και μόνο αν τα δύο μοντέλα είναι πλήρως ταυτόσημα. Ο στόχος του κριτηρίου είναι η μεγιστοποίηση της παραπάνω έκφρασης. Είναι επίσης σαφές ότι ο υπολογιστικός μόχθος που εμπλέκεται στο EPD είναι αρκετά λιγότερος συγκριτικά με τα άλλα δύο κριτήρια. Για περισσότερες λεπτομέρειες αναφορικά με τα παραπάνω μέτρα και τους τρόπους υπολογισμού τους, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στην εργασία των Jones et al. [97].

Έστω τώρα ότι $\mathcal{F}$=$\{\mathbf{f_1}$,$\mathbf{f_2}$,...,$\mathbf{f_g}\}$ είναι ο χώρος των g υποψηφίων μοντέλων, οπότε τα πιθανά ζευγάρια μοντέλων θα είναι $k = \binom{g}{2}$. Για αυτά τα μοντέλα, έξι κριτήρια διάκρισης μπορούν να εφαρμοστούν, καθένα εκ των οποίων λαμβάνει είτε την ελάχιστη είτε τη μέση τιμή των προαναφερθέντων κριτηρίων πάνω σε όλα τα k ζευγάρια μοντέλων. Δηλώνουμε ως $\mathbf{f_i}$ και $\mathbf{f_j}$ τα ζευγάρια μοντέλων και $\mathbf{X_i}$, $\mathbf{X_j}$ τους αντίστοιχους πίνακες μοντέλων. Στην ενότητα αυτή, θα εστιάσουμε σε τρία από αυτά τα κριτήρια, τα οποία έχουν τις ακόλουθες μορφές:

$$\text{(minimum SA) MSA} = \min \text{SA}\{\text{V}(\mathbf{X_i}), \text{V}(\mathbf{X_j})\} \tag{12.4}$$

$$\text{(minimum MPD) MMPD} = \min \text{MPD}\{\mathbf{f_i}, \mathbf{f_j}\} \tag{12.5}$$

$$\text{(average EPD) AEPD} = \frac{1}{k} \sum \text{EPD}\{\mathbf{f_i}, \mathbf{f_j}\}. \tag{12.6}$$

Για τα παραπάνω κριτήρια, όσο μεγαλύτερη είναι η τιμή τους, τόσο καλύτερη η ικανότητα διάκρισης μοντέλων του σχεδιασμού. Εντούτοις, τα κριτήρια αυτά δε συμφωνούν πάντα μεταξύ τους, αναφορικά με την κατάταξη των υποψηφίων σχεδιασμών. Επιπλέον, είναι πιθανό να έχουμε σχεδιασμούς που να μεγιστοποιούν ένα κριτήριο ενώ ταυτόχρονα να μηδενίζουν τα υπόλοιπα, κάτι που σημαίνει ότι υπάρχουν τουλάχιστον δύο μοντέλα τα οποία δε μπορούν να διακριθούν. Άρα, τα τρία αυτά κριτήρια δεν είναι ισοδύναμα ως προς την τάση τους να αναγνωρίσουν ένα σχεδιασμό ως μη δυνάμενο να διακρίνει μοντέλα. Το γεγονός αυτό ενισχύεται από τα αποτελέσματα που παρατίθενται στην επόμενη ενότητα. Ως εκ τούτου, παρακινηθήκαμε να συνδυάσουμε τα κριτήρια MSA, MMPD και AEPD και να προτείνουμε στο σημείο αυτό, ένα περισσότερο συντηρητικό και αυστηρό μέτρο, το κριτήριο συνολικής ικανότητας διάκρισης (overall discrimination capability-ODC), το οποίο ορίζεται ως

$$ODC = \prod_{i=1}^{p} \frac{\mathcal{C}_i}{\mathcal{C}_{i_{max}}}, \tag{12.7}$$

όπου $\mathcal{C}_i$ η τιμή του i-οστού κριτηρίου που χρησιμοποιείται, $i = 1, \ldots, p$ και $\mathcal{C}_{i_{max}}$ η μέγιστη τιμή του i-οστού κριτηρίου πάνω σε όλους τους σχεδιασμούς της εκάστοτε εξεταζόμενης κλάσης. Είναι σαφές, ότι στόχος του κριτηρίου είναι η επιλογή του σχεδιασμού που μεγιστοποιεί την (12.7). Να σημειώσουμε επίσης ότι, αν τουλάχιστον ένα από τα θεωρούμενα κριτήρια έχει μηδενική τιμή, τότε και το ODC θα είναι μηδέν και ο εξεταζόμενος σχεδιασμός δε θα είναι χρήσιμος. Στην περίπτωση όμως όπου ένας σχεδιασμός μεγιστοποιεί ταυτόχρονα όλα τα θεωρούμενα κριτήρια, τότε θα επιτυγχάνει και τη μέγιστη τιμή του ODC, ίση με 1.

12.3 Αποτελέσματα

Σε αυτήν την ενότητα, θα παρουσιάσουμε τους σχεδιασμούς 27 εκτελέσεων που μεγιστοποιούν τα κριτήρια που συζητήθηκαν προηγουμένως. Στον παρακάτω Πίνακα 12.1 αναφέρεται το πλήθος των μη ισόμορφων ορθογώνιων σχηματισμών 27 εκτελέσεων, όπως αυτοί αποκτήθηκαν στην εργασία των Evangelaras et al. [51].

Ας θεωρήσουμε δύο χώρους μοντέλων $\mathcal{F}_1$ και $\mathcal{F}_2$, οι οποίοι περιλαμβάνουν τα μοντέλα με όλες τις κύριες επιδράσεις, συν μια ή δύο αλληλεπιδράσεις δύο παραγόντων, αντίστοιχα. Το πρώτο βήμα της προτεινόμενης μεθοδολογίας, συνίσταται στην αναγνώριση των σχεδιασμών οι οποίοι μεγιστοποιούν το κριτήριο EC. Για τους χώρους μοντέλων που μας απασχολούν,

Πίνακας 12.1: Πλήθος μη ισόμορφων ορθογώνιων σχηματισμών με $n = 27$ εκτελέσεις και q παράγοντες

q	3	4	5	6	7	8	9	10	11	12	13
$\|N(27,q)\|$	9	711	187188	922548	157829	21688	9793	3766	1252	341	129

στόχος είναι η μέγιστη τιμή του EC να είναι ίση με 1. Αυτό σημαίνει ότι ένας σχεδιασμός με τη συγκεκριμένη ιδιότητα, θα είναι ικανός να εκτιμά όλες τις επιδράσεις που μελετάμε. Επόμενο βήμα αποτελεί η αναγνώριση των σχεδιασμών που αποφέρουν 99% ή περισσότερο της μέσης D-αποδοτικότητας, πάνω σε όλα τα μοντέλα κάθε συνόλου $\mathcal{F}_j$, $j = 1, 2$. Το υποσύνολο των σχεδιασμών αυτών επιλέχθηκε για περαιτέρω ανάλυση, χρησιμοποιώντας τα κριτήρια της προηγούμενης ενότητας. Στο εξής, τους σχεδιασμούς αυτούς θα τους ονομάζουμε ως εύρωστους ως προς τα μοντέλα σχεδιασμούς (model-robust designs). Να σημειωθεί ότι για να αναγνωρίσουμε τους πιο αποδοτικούς σχεδιασμούς, μεταθέσαμε πρώτα τα επίπεδα των μη ισόμορφων ορθογώνιων σχηματισμών. Αυτό έγινε διότι, όπως αναφέρεται και στους Cheng και Ye [34], από τη στιγμή που χρησιμοποιήσαμε σχεδιασμούς τριών επιπέδων, οι ισόμορφοι σχεδιασμοί είναι πιθανό να μη χαρακτηρίζονται από τις ίδιες στατιστικές ιδιότητες εάν έχει επιβληθεί μια σειρά από μεταθέσεις επιπέδων. Συνεπώς, αξιολογήσαμε τους διαθέσιμους γεωμετρικά μη ισόμορφους, εύρωστους ως προς τα μοντέλα, ορθογώνιους σχηματισμούς τριών επιπέδων με 27 εκτελέσεις και έως $q = 10$ παράγοντες. Βασιστήκαμε επίσης στις γραμμικές και τετραγωνικές αντιθέσεις για κάθε παράγοντα, συνεπώς κωδικοποιήσαμε το (χαμηλό, μεσαίο, υψηλό) επίπεδο ως $(-1,\ 0,\ 1)$ και $(1,\ -2,\ 1)$ για τις γραμμικές και τετραγωνικές αντιθέσεις των κυρίων επιδράσεων, αντίστοιχα.

Στον ακόλουθο Πίνακα 12.2 παρουσιάζουμε τις τιμές των κριτηρίων που χρησιμοποιήσαμε για κάθε χώρο μοντέλων. Συγκεκριμένα, σε κάθε περίπτωση, ο Πίνακας 12.2 περιλαμβάνει τον αύξοντα αριθμό του σχεδιασμού (#) ο οποίος μεγιστοποιεί κάποιο κριτήριο μαζί με την αντίστοιχη τιμή, καθώς επίσης το πλήθος των παραγόντων του (q) όπως και το μεσαίο επίπεδο (mid. lev.) που έχουμε θέσει. Για παράδειγμα, στην έκτη γραμμή του Πίνακα 12.2, ο σχεδιασμός με αριθμό 695 και με 4 κύριες επιδράσεις, αποτελεί τη βέλτιστη επιλογή καθώς μεγιστοποιεί τα κριτήρια MMPD, MSA και ODC και για τους δύο χώρους μοντέλων $\mathcal{F}_1$ και $\mathcal{F}_2$, όταν θέσουμε το μεσαίο επίπεδο στο 2. Ωστόσο παρατηρούμε ότι ο σχεδιασμός 698 είναι καλύτερος βάσει του AEPD κριτηρίου.

Πίνακας 12.2: Βέλτιστοι σχεδιασμοί για κάθε χώρο μοντέλων $\mathcal{F}_j$

	$\mathcal{F}_1$				$\mathcal{F}_2$			
q-mid. lev.	#–MMPD	#–MSA	#–AEPD	#–ODC	#–MMPD	#–MSA	#–AEPD	#–ODC
3–0	8–1.0000	8–1.5708	8–0.0741	8–1.0000	8–1.0000	8–1.5708	8–0.0741	8–1.0000
3–1	8–1.0000	8–1.5708	8–0.0741	8–1.0000	8–1.0000	8–1.5708	8–0.0741	8–1.0000
3–2	8–1.0000	8–1.5708	8–0.0741	8–1.0000	8–1.0000	8–1.5708	8–0.0741	8–1.0000
4–0	581–1.0000	581–1.5708	581–0.0741	581–1.0000	260–0.8962	260–1.1112	260–0.0984	260–1.0000
4–1	563–0.9815	563–1.3781	698–0.0731	563–0.9992	698–0.9682	698–1.3181	698–0.1037	698–1.0000
4–2	695–0.9821	695–1.3815	698–0.0731	695–0.9952	695–0.9743	695–1.3438	698–0.1037	695–0.9926
5–0	6139–0.9606	6139–1.2893	146256–0.0728	6139–0.9878	183780–0.9394	183780–1.2207	35040–0.1166	183780–0.9897
5–1	184264–0.9554	184264–1.2711	155992–0.0723	184264–0.9958	17113–0.9144	17113–1.1541	56747–0.1148	17113–0.9971
5–2	56810–0.9505	56810–1.2547	179381–0.0722	56810–0.9922	-	-	-	-
6–0	584135–0.9202	584135–1.1687	634855–0.0717	584135–0.9968	720136–0.8661	720136–1.0473	632115–0.1201	526040–0.9866
6–1	764222–0.8988	764222–1.1170	795546–0.0714	764222–0.9857	803886–0.7764	803886–0.8889	826421–0.1179	803886–0.9779
6–2	736927–0.9055	736927–1.1327	856134–0.0714	736927–0.9940	826042–0.8365	826042–0.9908	808469–0.1193	826042–0.9704
7–0	1089–0.8485	1089–1.0131	1035–0.0700	1089–0.9861	1089–0.7262	1089–0.8128	131382–0.1188	1089–0.9783
7–1	18294–0.8040	18294–0.9340	95–0.0700	18294–0.9768	177–0.6567	177–0.7164	177–0.1172	177–1.0000
7–2	18095–0.8333	18095–0.9851	127–0.0702	18095–0.9867	2125–0.6433	2125–0.6988	136–0.1184	2125–0.9927
8–0	16404–0.7252	16404–0.8113	16597–0.0683	16404–0.9942	11566–0.5650	11566–0.6005	20293–0.1158	11566–0.9939
8–1	12066–0.7040	12066–0.7810	18637–0.0681	12066–0.9885	18279–0.4840	18279–0.5052	18279–0.1143	18279–1.0000
8–2	20294–0.7210	20294–0.8052	21440–0.0682	20294–0.9903	17203–0.5152	17203–0.5412	14221–0.1150	17203–0.9824
9–0	588–0.4979	588–0.5212	829–0.0655	588–0.9983	588–0.3648	588–0.3734	321–0.1079	588–0.9989
9–1	-	-	6–0.0659	-	-	-	24–0.1072	-
9–2	281–0.4231	281–0.4368	6–0.0658	281–0.9950	-	-	74–0.1085	-
10–0	3638–0.4455	3638–0.4618	3748–0.0628	3638–0.9940	-	-	-	-
10–1	3638–0.4109	3638–0.4234	1246–0.0629	3638–0.9845	-	-	-	-
10–2	3648–0.4070	3648–0.4192	334–0.0628	3648–0.9971	-	-	-	-

Παρατηρώντας τον Πίνακα 12.2, καταλήγουμε στα εξής: Καταρχήν, ένα γενικό σχόλιο είναι ότι σε αρκετές περιπτώσεις, τα κριτήρια MMPD, MSA και ODC συμφωνούν μεταξύ τους ενώ το AEPD παρουσιάζει διαφορετικούς σχεδιασμούς ως βέλτιστη επιλογή. Συνεπώς, συνίσταται η χρήση των σχεδιασμών αυτών που μεγιστοποιούν το ODC κριτήριο, για σκοπούς διάκρισης μοντέλων. Ορισμένοι σχεδιασμοί, όπως για παράδειγμα ο σχεδιασμός 8 που προκύπτει και καθολικά καλύτερος για τους δύο χώρους μοντέλων, επιτυγχάνουν την ανώτατη τιμή του ODC.

Για την περίπτωση των 5 παραγόντων, το χώρο μοντέλων $\mathcal{F}_2$ και μεσαίο επίπεδο=2, δεν υπάρχουν σχεδιασμοί με 99% ή παραπάνω της μέσης D-αποδοτικότητας, συνεπώς δεν τους εξετάσαμε.

Αξίζει επίσης να αναφέρουμε τη περίπτωση όπου έχουμε 6 παράγοντες, μεσαίο επίπεδο=0 και χώρο μοντέλων $\mathcal{F}_2$. Παρατηρούμε ότι ο σχεδιασμός 720136 μεγιστοποιεί τα κριτήρια MMPD και MSA, ενώ ο σχεδιασμός 632115 μεγιστοποιεί το AEPD. Παρ΄ όλα αυτά, το ODC κριτήριο κατατάσσει ως καλύτερο το σχεδιασμό 526040. Αυτό συνέβει λόγω του ότι ο συγκεκριμένος σχεδιασμός κατέλαβε τη δεύτερη θέση στα κριτήρια MMPD, MSA και AEPD ενώ ταυτόχρονα ο σχεδιασμός 720136 είχε πολύ χαμηλή τιμή του AEPD. Συνεπώς, προτείνουμε τη χρήση του σχεδιασμού 526040 για σκοπούς διάκρισης μοντέλων, για αυτή την περίπτωση.

Για 9 παράγοντες, μεσαίο επίπεδο=1 καθώς και για τους δύο χώρους μοντέλων, όλοι οι σχεδιασμοί έδωσαν μηδενικές τιμές των κριτηρίων MMPD και MSA, όποτε μηδενίστηκε και το ODC, άρα δε συνίσταται η χρήση τους. Αν όμως ο πειραματιστής επιθυμεί να βασιστεί μόνο στο AEPD, τότε οι σχεδιασμοί 6 και 24 αποτελούν τις καλύτερες επιλογές για τους χώρους μοντέλων $\mathcal{F}_1$ και $\mathcal{F}_2$, αντίστοιχα. Η παρατήρηση αυτή ισχύει και για την περίπτωση όπου το μεσαίο επίπεδο έχει τεθεί στο 2 και το χώρο μοντέλων $\mathcal{F}_2$, όπου ο σχεδιασμός 74 είναι η βέλτιστη επιλογή, βάσει του AEPD κριτηρίου. Επίσης, στους 10 παράγοντες και για το χώρο μοντέλων $\mathcal{F}_2$, οι τιμές όλων των κριτηρίων προέκυψαν αρκετά κοντά στο μηδέν.

12.4 Συμπεράσματα

Από τα πιο γνωστά και περισσότερο χρησιμοποιούμενα κριτήρια στο θέμα της επιλογής μοντέλου, είναι της ελάχιστης απόκλισης και της D-αποδοτικότητας. Είναι πιθανό, ακόμα και όλα τα μοντέλα σε ένα σύνολο να είναι εκτιμήσιμα σύμφωνα με αυτά τα κριτήρια. Όμως, ένα πρακτικό ζήτημα που προκύπτει είναι ο καθορισμός του σωστού μοντέλου, λόγω του προβλήματος των πλήρως ταυτόσημων μοντέλων. Σε τέτοιες καταστάσεις, τα παραδοσιακά κριτήρια μπορεί να οδηγήσουν σε επιλογή ανεπαρκών και αναποτελεσματικών σχεδιασμών για ένα πείραμα.

Ο προσδιορισμός του κατάλληλου σχεδιασμού είναι κρίσιμης σημασίας, κατά τη διεξαγωγή ενός πειράματος, καθότι η σωστή επιλογή των ενεργών επιδράσεων, που έχουν σημαντική επίδραση στην απόκριση, συνεπάγεται λιγότερο κόστος και λιγότερο χρονοβόρα πειράματα. Για παράδειγμα, στην περίπτωση των 6 παραγόντων υπάρχουν 59601 σχεδιασμοί που είχαν τιμή της μέσης D-αποδοτικότητας 99% ή μεγαλύτερη, για το σύνολο $\mathcal{F}_1$. Συνεπώς, είναι εμφανές ότι δυσκολεύει αρκετά η επιλογή του κατάλληλου σχεδιασμού, άρα κρίνεται απαραίτητη η χρήση κατάλληλων κριτηρίων.

Στο κεφάλαιο αυτό, εξετάσαμε και αξιολογήσαμε τους γεωμετρικά μη ισόμορφους εύρωστους ως προς τα μοντέλα ορθογώνιους σχηματισμούς τριών επιπέδων με 27 εκτελέσεις και έως $q = 10$ παράγοντες, χρησιμοποιώντας τους μη ισόμορφους σχηματισμούς των Evangelaras et. al. [51], εφαρμόζοντας μεταθέσεις των επιπέδων τους. Επεκτείναμε έτσι τη δουλειά των Jones et al. [97], οι οποίοι ασχολήθηκαν με τους σχηματισμούς με 18 εκτελέσεις και έως $q = 7$ παράγοντες. Προτείναμε το συνδυασμό των κριτηρίων αποδοτικότητας EC και IC, καθώς και των κριτηρίων διάκρισης μοντέλων MMPD, MSA και AEPD, με στόχο την επιλογή του κατάλληλου σχεδιασμού. Δημιουργήσαμε επίσης και ένα νέο κριτήριο αξιολόγησης της

συνολικής ικανότητας διάκρισης ενός σχεδιασμού, το ODC. Τα αποτελέσματα που πήραμε, συνιστούν τη χρήση του ODC κριτηρίου, καθώς επίσης και την περαιτέρω εστίαση στα MSA και MMPD στους εύρωστους σχεδιασμούς, από τη στιγμή που αναδείχθηκαν πιο κατάλληλα σε σύγκριση με το AEPD.

Βιβλιογραφία

[1] O.O. Aalen, Heterogeneity in survival analysis. *Statistics in Medicine*, 7 (1998), 1121-1137.

[2] M.A.F. Aboukalam, Quick, easy and powerful analysis of unreplicated factorial designs, *Communications in Statistics-Theory and Methods*, 34 (2005), 1169-1175.

[3] B. Abraham, H. Chipman and K. Vijayan, Some risks in the construction and analysis of supersaturated designs, *Technometrics*, 41 (1999), 135-141.

[4] H. Akaike, Information theory and an extension of the maximum likelihood principle, In *Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki (Eds), 267-281, Akademiai Kiado, Budapest, 1973.

[5] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19 (1974), 716-723.

[6] P.K. Andersen, J.P. Klein and M.J. Zhang, Testing for centre effects in multi-centre survival studies: a Monte Carlo comparison of fixed and random effects tests, *Statistics in Medicine*, 18 (1999), 1489-1500.

[7] P. Angelopoulos and C. Koukouvinos, Detecting active effects in unreplicated designs, *Journal of Applied Statistics*, 35 (2008), 277-281.

[8] P. Angelopoulos, H. Evangelaras and C. Koukouvinos, Analyzing unreplicated 2^k factorial designs by examining their projections into $k-1$ factors, *Quality and Reliability Engineering International*, 26 (2010), 223-233.

[9] P. Angelopoulos, C. Koukouvinos and A. Skountzou, Clustering effects in unreplicated factorial experiments, *Communications in Statistics-Simulation and Computation*, 42 (2013), 1998-2007.

[10] A. Antoniadis, Wavelets in statistics: a review (with discussion), *Journal of the Italian Statistical Society*, 6 (1997), 97-144.

[11] A.C. Atkinson and V.V. Fedorov, The design of experiments for discriminating between two rival models, *Biometrika*, 62 (1975a), 57-70.

[12] A.C. Atkinson and V.V. Fedorov, Optimal design: experiments for discriminating between several models, *Biometrika*, 62 (1975b), 289-303.

[13] G. Auchmuty, A posteriori error estimates for linear equations, *Numerische Mathematik*, 61 (1992), 1-6.

[14] P. Barker and R. Henderson, Small sample bias in the gamma frailty model for univariate survival, *Lifetime Data Analysis*, 11 (2005), 265-284.

[15] P.J. Bickel and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day Inc., San Francisco, 1977.

[16] P.J. Bickel and K.A. Doksum, *Mathematical Statistics*, vol. I, 2nd. ed., updated printing, Pearson Prentice Hall, Upper Saddle River, 2007.

[17] D.R. Bingham and H. A. Chipman, Incorporating prior information in optimal design for model selection, *Technometrics*, 49 (2007), 155-163.

[18] K.H.V Booth and D.R. Cox, Some systematic supersaturated designs, *Technometrics*, 4 (1962), 489-495.

[19] G.E.P. Box and R.D. Meyer, An analysis for unreplicated fractional factorials, *Technometrics*, 28 (1986), 11-18.

[20] L. Breiman, Better subset regression using the non-negative garrote, *Technometrics*, 37 (1995), 373-384.

[21] L. Breiman, Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, 24 (1996), 2350-2383.

[22] C. Brezinski, G. Rodriguez and S. Seatzu, Error estimates for linear systems with applications to regularization, *Numerical Algorithms*, 49 (2008), 85-104.

[23] C. Brezinski, G. Rodriguez and S. Seatzu, Error estimates for the regularization of least squares problems, *Numerical Algorithms*, 51 (2009), 61-76.

[24] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30 (1998), 107-117.

[25] D.A. Bulutoglu, Cyclicly constructed $E(s^2)$-optimal supersaturated designs, *Journal of Statistical Planning and Inference,* 137 (2007), 2413-2428.

[26] D. A. Bulutoglu and C.S. Cheng, Construction of $E(s^2)$-optimal supersaturated designs, *The Annals of Statistics*, 32 (2004), 1662-1678.

[27] D.A. Bulutoglu and K.J. Ryan, $E(s^2)$-optimal supersaturated designs with good minmax properties when N is odd, *Journal of Statistical Planning and Inference,* 138 (2008), 1754-1762.

[28] J. Cai, J. Fan, R. Li and H. Zhou, Variable selection for multivariate failure time data, *Biometrika*, 92 (2005), 303-316.

[29] E.J. Candes and T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, 35 (2007), 2313-2351.

[30] Y. Chen and J. Kunert, A new quantitative method for analysing unreplicated factorial designs, *Biometrical Journal*, 46 (2004), 125-140.

[31] C.S. Cheng, $E(s^2)$-optimal supersaturated designs, *Statistica Sinica*, 7 (1997), 929-939.

[32] C.S. Cheng and D.M. Steinberg, Trend robust two-level factorial designs, *Biometrika*, 78 (1991), 325-336.

[33] C.S. Cheng, D.M. Steinberg and D.X. Sun, Minimum aberration and model robustness for two-level fractional factorial designs, *Journal of the Royal Statistical Society, Ser. B*, 61 (1999), 85-93.

[34] S.W. Cheng and K.Q. Ye, Geometric isomorphism and minimum aberration for factorial designs with quantitative factors, *The Annals of Statistics*, 32 (2004), 2168-2185.

[35] H. Chipman, M. Hamada and C.F.J. Wu, A bayesian variable selection approach for analyzing designed experiments with complex aliasing, *Technometrics*, 39 (1997), 372-381.

[36] D.G. Clayton, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, 65 (1978), 141-151.

[37] D. Collett, *Modelling Survival Data in Medical Research*, Chapman and Hall/CRC, London, 2003.

[38] G.M. Constantine, Robust designs for serially correlated observations, *Biometrika*, 76 (1989), 245-251.

[39] R.D. Cook and C.J. Nachtsheim, Model robust, linear-optimal designs, *Technometrics*, 24 (1982), 49-54.

[40] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley Interscience, New York, 1991.

[41] D.R. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society, Ser. B*, 34 (1972), 187-220.

[42] D.R. Cox, Partial likelihood, *Biometrika*, 62 (1975), 269-276.

[43] P. Craven and G. Wahba, Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, 31 (1979), 377-403.

[44] C. Daniel, Use of half-normal plots in interpreting factorial two-level experiments, *Technometrics*, 1 (1959), 311-341.

[45] I. Daubechies, M. Defrise and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communications on Pure and Applied Mathematics*, 57 (2004), 1413-1457.

[46] F. Dong, On the identification of active contrasts in unreplicated fractional factorials, *Statistica Sinica*, 3 (1993), 209-217.

[47] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression (with discussion), *The Annals of Statistics*, 32 (2004), 407-499.

[48] L.J. Elliott, J.A. Eccleston and R. J. Martin, An algorithm for the design of factorial experiments when the data are correlated, *Statistics and Computing*, 9 (1999), 195-201.

[49] H. Evangelaras, C. Koukouvinos, A.M. Dean and C.A. Dingus, Projection properties of certain three level orthogonal arrays, *Metrika*, 62 (2005), 241-257.

[50] H. Evangelaras, C. Koukouvinos and E. Lappas, 18-run nonisomorphic three level orthogonal arrays, *Metrika*, 66 (2007), 31-37.

[51] H. Evangelaras, C. Koukouvinos and E. Lappas, 27-run nonisomorphic three level orthogonal arrays: Identification, evaluation and projection properties, *Utilitas Mathematica*, 84 (2011), 75-87.

[52] J. Fan, Comments on "Wavelets in statistics: A review" by A. Antoniadis, *Journal of the Italian Statistical Society*, 6 (1997), 131-138.

[53] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association*, 96 (2001), 1348-1360.

[54] J. Fan and R. Li, Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics*, 30 (2002), 74-99.

[55] J. Fan and R. Li, New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *Journal of the American Statistical Association*, 99 (2004), 710-723.

[56] J. Fan and R. Li, Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery, *Proceedings of the International Congress of Mathematicians*, M. Sanz- Sole, J. Soria, J.L. Varona, J. Verdera (Eds), 595-622, European Mathematical Society, Zurich, 2006.

[57] J. Fan and J. Lv, A selective overview of variable selection in high dimensional feature space, *Statistica Sinica*, 20 (2010), 101-148.

[58] J. Fan and J. Lv, Nonconcave penalized likelihood with NP-dimensionality, *IEEE Transactions on Information Theory*, 57 (2011), 5467-5484.

[59] J. Fan and H. Peng, On non-concave penalized likelihood with diverging number of parameters, *The Annals of Statistics*, 32 (2004), 928-961.

[60] Y. Fan and C. Tang, Tuning parameter selection in high dimensional penalized likelihood, *Journal of the Royal Statistical Society, Ser. B*, 75 (2013), 531-552.

[61] K.T. Fang, The uniform design: application of number-theoretic methods in experimental design, *Acta Mathematicae Applicatae Sinica*, 3 (1980), 363-372.

[62] K.T. Fang, Theory, method and applications of the uniform design, *International Journal of Reliability, Quality and Safety Engineering*, 9 (2002), 305-315.

[63] K.T. Fang, Theory, method and applications of the uniform experimental design, a historical review, *Application of Statistics and Management*, 23 (2004), 69-80.

[64] K.T. Fang, R. Li and A. Sudjianto, *Design and Modeling for Computer Experiments*, Chapman and Hall, New York, 2006.

[65] K.T. Fang and D.K.J. Lin, Theory and applications of the uniform design, *Journal of Chinese Statistical Association*, 38 (2000), 331-352.

[66] K.T. Fang and D.K.J. Lin, Uniform designs and their applications in industry, In: *Handbook on Statistics 22: Statistics in Industry*, R. Khattree and C.R. Rao (Eds), 131-170, Elsevier, North-Holland, 2003.

[67] K.T. Fang, D.K.J. Lin and C.X. Ma, On the construction of multi-level supersaturated designs, *Journal of Statistical Planning and Inference*, 86 (2000), 239-252.

[68] K.T. Fang, D.K.J. Lin, P. Winker and Y. Zhang, Uniform design: Theory and Applications, *Technometrics*, 42 (2000), 237-248.

[69] K.T. Fang, X. Lu, Y. Tang and J. Yin, Construction of uniform design by using resolvable packings and coverings, *Discrete Mathematics*, 274 (2004), 25-40.

[70] K.T. Fang and C.X. Ma, Applications of uniformity to factorial designs, *Journal of Chinese Statistical Association*, 38 (2000), 441-464.

[71] K.T. Fang, C.X. Ma and P. Winker, Centered L_2-discrepancy of random sampling and Latin hypercube design, and construction of uniform design, *Mathematics of Computation* 71 (2002), 275-296.

[72] K.T. Fang and R. Mukerjee, A connection between uniformity and aberration in regular fractions of two-level factorials, *Biometrika*, 87 (2000), 193-198.

[73] K.T. Fang and Y. Wang, *Number-theoretic methods in statistics*, Chapman and Hall, London, 1994.

[74] F. Fleuret, Fast binary feature selection with conditional mutual information, *Journal of Machine Learning Research*, 5 (2004), 1531-1555.

[75] I.E. Frank and J.H. Friedman, A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, 35 (1993), 109-148.

[76] W.J. Fu, Penalized regression: the bridge versus the LASSO, *Journal of Computational and Graphical Statistics*, 7 (1998), 397-416.

[77] A. Galantai, A study of Auchmuty's error estimate, *Computers and Mathematics with Applications*, 42 (2001), 1093-1102.

[78] J.J. Garroi, P. Goos and K. Sörensen, A variable-neighbourhood search algorithm for finding optimal run orders in the presence of serial correlation, *Journal of Statistical Planning and Inference*, 139 (2009), 30-44.

[79] S.D. Georgiou, Modelling by supersaturated designs, *Computational Statistics and Data Analysis,* 53 (2008), 428-435.

[80] S.D. Georgiou, Supersaturated designs: a review of their construction and analysis, *Journal of Statistical Planning and Inference,* 144 (2014), 92-109.

[81] S. Ghosh and Y. Shen, Comparison of designs in presence of a possible correlation in observations. *Test*, 15 (2006), 485-504.

[82] S.G. Gilmour, Supersaturated designs in factor screening, In: *Screening*, S.M. Lewis and A.M. Dean (Eds), 169-190, Springer, New York, 2006.

[83] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, 1989.

[84] W. Greub and W. Rheinboldt, On a generalization of an inequality of L. V. Kantorovich, *Proceedings of the American Mathematical Society*, 10 (1959), 407-415.

[85] I.D. Ha, M. Noh and Y. Lee, Bias reduction of likelihood estimators in semiparametric frailty models, *Scandinavian Journal of Statistics*, 37 (2010), 307-320.

[86] M. Hamada and N. Balakrishnan, Analysing unreplicated factorial experiments: A review with some new proposals, *Statistica Sinica*, 8 (1998), 1-41.

[87] E.J. Hannan and B.G. Quinn, The determination of the order of an autoregression, *Journal of the Royal Statistical Society, Ser. B*, 41 (1979), 190-195.

[88] A. Hedayat, N.J.A. Sloane and J. Stufken, *Orthogonal Arrays: Theory and Applications*, Springer-Verlag, New York, 1999.

[89] F.J. Hickernell, Goodness-of-fit statistics, discrepancies and robust designs, *Statistics and Probability Letters*, 44 (1999), 73-78.

[90] F.J. Hickernell and M.Q. Liu, Uniform designs limit aliasing, *Biometrika*, 89 (2002), 893-904.

[91] D.R. Holcomb, D.C. Montgomery and W.M. Carlyle, Analysis of supersaturated designs, *Journal of Quality Technology*, 35 (2003), 13-27.

[92] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, United Kingdom, 1985.

[93] P. Hougaard, Life table methods for heterogeneous populations: Distributions describing the heterogeneity, *Biometrika*, 71 (1984), 75-83.

[94] P. Hougaard, Survival models for heterogeneous populations derived from stable distributions, *Biometrika*, 73 (1986), 387-396.

[95] P. Hougaard, *Analysis of multivariate survival data*, Springer, New York, 2000.

[96] G.M. Jenkins and J. Chanmugam, The estimation of slope when the errors are autocorrelated. *Journal of the Royal Statistical Society, Ser. B*, 24 (1962), 199-214.

[97] B.A.J. Jones, W. Li, C.J. Nachtsheim and K.Q. Ye, Model discrimination-another perspective on model-robust designs, *Journal of Statistical Planning and Inference*, 137 (2007), 1576-1583.

[98] M. Kosorok, B. Lee and J. Fine, Robust inference for univariate proportional hazards frailty regression models, *The Annals of Statistics*, 32 (2004), 1448-1491.

[99] C. Koukouvinos, *Linear Models and Designs*, N.T.U.A. Press, Athens, 2010 (in Greek).

[100] C. Koukouvinos, *Statistical Designs*, N.T.U.A. Press, Athens, 2010 (in Greek).

[101] C. Koukouvinos, P. Mantas and K. Mylona, A general construction of E(s2)-optimal large supersaturated designs, Metrika, 68 (2008), 99-110.

[102] C. Koukouvinos and K. Mylona, A general construction of E(s2)-optimal supersaturated designs via supplementary difference sets, *Metrika,* 70 (2009), 257-265.

[103] C. Koukouvinos and K. Mylona, Group screening method for the statistical analysis of $E(f_{NOD})$-optimal mixed-level supersaturated designs, *Statistical Methodology*, 6 (2009), 380-388.

[104] C. Koukouvinos, K. Mylona and D.E. Simos, Exploring k-circulant supersaturated designs via genetic algorithms, *Computational Statistics and Data Analysis,* 51 (2007), 2958-2968.

[105] C. Koukouvinos, K. Mylona and D.E. Simos, $E(s2)$-optimal and minimax-optimal supersaturated designs via multi-objective simulated annealing, *Journal of Statistical Planning and Inference,* 138 (2008), 1639-1646.

[106] C. Koukouvinos, K. Mylona and D.E. Simos, A hybrid SAGA algorithm for the construction of $E(s2)$-optimal cyclic supersaturated designs, *Journal of Statistical Planning and Inference,* 139 (2009), 478-485.

[107] C. Koukouvinos and S. Stylianou, A method for analyzing supersaturated designs, *Communications in Statistics-Simulation and Computation,* 34 (2005), 929-937.

[108] W. Kruskal, The significance of Fisher: A review of R. A. Fisher, *Journal of the American Statistical Association*, 75 (1980), 1019-1030.

[109] A.N. Langville and C.D. Meyer, A survey of eigenvector methods for Web information retrieval, *SIAM Review*, 47 (2005), 135-161.

[110] A.N. Langville and C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, 2006.

[111] A.N. Langville and C.D. Meyer, *Who's Number 1: The Science of Rating and Ranking*, Princeton University Press, Princeton, 2012.

[112] E. Läuter, Experimental planning in a class of models, *Mathematische Operationsforschung und Statistik*, 5 (1974), 673-708.

[113] S. Lee and J.P. Klein, Bivariate models with a random environmental factor, *Industrial Journal of Productivity, Reliability and Quality Control*, 13 (1988), 1-18.

[114] C. Leng, Y. Lin and G. Wahba, A note on the LASSO and related procedures in model selection, *Statistica Sinica*, 16 (2006), 1273-1284.

[115] R.V. Lenth, Quick and easy analysis of unreplicated factorial, *Technometrics*, 31 (1989), 469-473.

[116] R. Li and D.K.J. Lin, Data analysis in supersaturated designs, *Statistics and Probability Letters*, 59 (2002), 135-144.

[117] W. Li and C.J. Nachtsheim, Model-robust factorial designs, *Technometrics*, 42 (2000), 345-352.

[118] W.W. Li and C.F.J. Wu, Columnwise-pairwise algorithms with applications to the construction of supersaturated designs, *Technometrics*, 39 (1997), 171-179.

[119] Y.Z. Liang, K. T. Fang and Q.S. Xu, Uniform design and its applications in chemistry and chemical engineering, *Chemometrics and Intelligent laboratory Systems*, 58 (2001), 43-57.

[120] D.K.J. Lin, A new class of supersaturated designs, *Technometrics*, 3 (1993), 28-31.

[121] D.K.J. Lin, Generating systematic supersaturated designs, *Technometrics*, 37 (1995), 213-225.

[122] D.V. Lindley and N.A. Singpurwalla, Multivariate distributions for the life lengths of components of a system sharing a common environment, *Journal of Applied Probability*, 23 (1986), 418-431.

[123] Y. Liu and A. Dean, k-circulant supersaturated designs, *Technometrics*, 46 (2004), 32-43.

[124] X. Lu and X. Wu, A strategy of searching active factors in supersaturated screening experiments, *Journal of Quality Technology*, 36 (2004), 392-399.

[125] J. Lv and Y. Fan, A unified approach to model selection and sparse recovery using regularized least squares, *The Annals of Statistics*, 37 (2009), 3498-3528.

[126] C.L. Mallows, Some comments on C_p. *Technometrics*, 15 (1973), 661-675.

[127] M. Marcus and M. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Dover, New York, 1992.

[128] C.J. Marley and D.C. Woods, A comparison of design and model selection methods for supersaturated experiments, *Computational Statistics and Data Analysis*, 54 (2010), 3158-3167.

[129] R.J. Martin, J.A. Eccleston and G. Jones, Some results on multi-level factorial designs with dependent observations, *Journal of Statistical Planning and Inference*, 73 (1998), 91-111.

[130] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1989.

[131] C.A. McGilchrist and C.W. Aisbett, Regression with frailty in survival analysis, *Biometrics*, 47 (1991), 461-466.

[132] R.D. Meyer, D.S. Steinberg and G.E.P. Box, Follow-up designs to resolve the confounding in multifactor experiments, *Technometrics*, 7 (1996), 307-323.

[133] A. Miller, The analysis of unreplicated factorial experiments using all possible comparisons, *Technometrics*, 47 (2005), 51-63.

[134] R.H Myers, D.C. Montgomery and G.G. Vining, *Generalized Linear Models. With Applications in Engineering and the Sciences*, John Wiley and Sons, New York, 2002.

[135] J.A. Nelder and R.W.M. Wedderburn, Generalized linear models, *Journal of the Royal Statistical Society, Ser. A*, 135 (1972), 370-384.

[136] N.K. Nguyen, An algorithmic approach to constructing supersaturated designs, *Technometrics*, 38 (1996), 69-73.

[137] R. Nishii, Asymptotic properties of criteria for selection of variables in multiple regression, *The Annals of Statistics*, 12 (1984), 758-765.

[138] J. Novovičová, P. Somol, M. Haindl and P. Pudil, Conditional mutual information based feature selection for classification task, In *Progress in Pattern Recognition, Image Analysis and Applications*, 417-426, Springer, 2007,.

[139] L. Page, S. Brin, R. Motwani and T. Winograd, *The pagerank citation ranking: Bringing order to the Web*, Technical report 422, Computer Science Department, Stanford University (1999).

[140] H. Park, F. Sakaori and S. Konishi, Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria, *Journal of Statistical Computation and Simulation*, 84 (2014), 1596-1607.

[141] E. Parner, Asymptotic theory for the correlated gamma-frailty model, *The Annals of Statistics*, 26 (1998), 183-214.

[142] H. Peng, F. Long and C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (2005), 1226-1238.

[143] F.K.H Phoa, Y.H. Pan and H. Xu, Analysis of supersaturated designs via the Dantzig selector, *Journal of Statistical Planning and Inference*, 139 (2009), 2362-2372.

[144] B.M. Pötscher and H. Leeb, On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding, *Journal of Multivariate Analysis*, 100 (2009), 2065-2082.

[145] F. Pukelsheim, *Optimal Design of Experiments*, Wiley, New York, 1993.

[146] J.R. Quinlan, Induction of decision trees, *Machine Learning*, 1 (1986), 81-106.

[147] C.R. Rao, Factorial experiments derivable from combinatorial arrangements of array, *Journal of the Royal Statistical Society*, 9 (1947), 128-139.

[148] J. Risannen, Modeling by shortest data description, *Automatica*, 14 (1978), 465-471.

[149] K.J. Ryan and D.A. Bulutoglu, $E(s^2)$-optimal supersaturated designs with good minimax properties, *Journal of Statistical Planning and Inference*, 137 (2007), 2250-2262.

[150] F.E. Satterthwaite, Random Balance Experimentation (with Discussions), *Technometrics*, 1 (1959), 111-137.

[151] I.W. Saunders, J.A. Eccleston and R.J. Martin, An algorithm for the design of 2^p factorial experiments on continuous processes, *The Australian Journal of Statistics*, 37 (1995), 353-365.

[152] G. Schwartz, Estimating the dimension of a model, *The Annals of Statistics*, 6 (1978), 461-464.

[153] V. Seshadri, *The Inverse Gaussian Distribution-Statistical Theory and Applications*, Springer, New York, 1993.

[154] V.S. Sethuraman and D. Raghavarao, Balanced 2^n factorial designs when observations are spatially correlated, *Journal of Biopharmaceutical Statistics*, 19 (2009), 332-344.

[155] V.S. Sethuraman, D. Raghavarao and B.K. Sinha, Optimal s^n factorial designs when observations within-blocks are correlated, *Computational Statistics and Data Analysis*, 50 (2006), 2855-2862.

[156] C.E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, 27 (1948), 379-423 and 623-656.

[157] P. Silvapulle and M.L. King, Testing moving average against autoregressive disturbances in the linear-regression model, *Journal of Business and Economic Statistics*, 9 (1991), 329-335.

[158] E.V. Slud and F. Vonta, Consistency of the NPML Estimator in the Right-Censored Transformation Model, *Scandinavian Journal of Statistics*, 31 (2004), 21-41.

[159] J.N. Srivastava, Designs for searching for non-negligible effects, In: *A Survey of Statistical Designs and Linear Models*, J.N. Srivastava (Ed), 507-519, North-Holland, Amsterdam, 1975.

[160] D.X. Sun, *Estimation Capacity and Related Topics in Experimental Designs*, Unpublished Ph.D. Dissertation, Department of Statistics and Actuarial Science, University of Waterloo, Canada, 1993.

[161] B. Tang and C.F.J. Wu, A method for constructing supersaturated designs and its $E(s^2)$ optimality, *Canadian Journal of Statistics*, 25 (1997), 191-201.

[162] R. Tibshirani, Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Ser. B,*, 58 (1996), 267-288.

[163] P.W. Tsai, S.G. Gilmour and R. Mead, Projective three-level main effects designs robust to model uncertainty, *Biometrika*, 87 (2000), 467-475.

[164] P.W. Tsai, S.G. Gilmour and R. Mead, Some new three-level orthogonal main effects plans robust to model uncertainty, *Statistica Sinica*, 14 (2004), 1075-1084.

[165] P.W. Tsai, S.G. Gilmour and R. Mead, Statistical isomorphism of three-level fractional factorial designs. *Utilitas Mathematica*, 70 (2006), 3-9.

[166] A. Van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, 1998.

[167] J.A. Vaupel, K.G. Manton and E. Stallard, The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, 16 (1979), 439-454.

[168] F. Vonta, Efficient estimation in a non-proportional hazards model in survival analysis, *Scandinavian Journal of Statistics*, 23 (1996), 49-61.

[169] D.T. Voss and W. Wang, On adaptive testing in orthogonal saturated designs, *Statistica Sinica*, 16 (2006), 227-234.

[170] P.C. Wang, Comments on Lin(1993), *Technometrics*, 37 (1995), 358-359.

[171] Y. Wang and K.T. Fang, A note on uniform distribution and experimental design, *KeXue TongBao*, 26 (1981), 485-489.

[172] H. Wang, R. Li and C.L. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, 94 (2007), 553-568.

[173] J.C. Wang and C.F.J. Wu, A hidden projection property of Plackett-Burman and related designs, *Statistica Sinica*, 5 (1995), 235-250.

[174] P.H. Westfall, S.S. Young and D.K.J. Lin, Forward selection error control in the analysis of supersaturated designs, *Statistica Sinica*, 8 (1998), 101-117.

[175] A. Wienke, *Frailty Models in Survival Analysis*, Chapman and Hall/CRC Press, Boca Raton, 2011.

[176] P. Winker and K.T. Fang, Application of threshold accepting to the evaluation of the discrepancy of a set of points, *SIAM Journal on Numerical Analysis* 34 (1997), 2038-2042.

[177] P. Winker and K.T. Fang, Optimal U-type design, In: *Monte Carlo and Quasi-Monte Carlo Methods 1996*, H. Niederreiter, P. Zinterhof and P. Hellekalek (Eds), 436-448, Springer, 1998.

[178] C.F.J. Wu, Construction of supersaturated designs through partially aliased interactions, *Biometrika*, 80 (1993), 661-669.

[179] C.F.J. Wu and M. Hamada, *Experiments: Planning, Analysis and Parameter Design Optimization*, Wiley, New York, 2000.

[180] T.T. Wu and K. Lange, Coordinate descent algorithms for LASSO penalized regression, *The Annals of Applied Statistics*, 2 (2008), 224-244.

[181] H. Xu and C.F.J. Wu, Construction of optimal multi-level supersaturated designs, *The Annals of Statistics*, 33 (2005), 2811-2836.

[182] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, 38 (2010), 894-942.

[183] Y. Zhang and R. Li, Iterative conditional maximization algorithm for nonconcave penalized likelihood, In: *Nonparametric statistics and mixture models: A festschrift in honor of Thomas P. Hettmansperger*, D. R. Hunter, D.St.P. Richards and J.L. Rosenberger (Eds), 336-351, World Scientific Publishing Co., Singapore, 2011.

[184] Y. Zhang, R. Li, and C.L. Tsai, Regularization parameter selections via generalized information criterion, *Journal of the American Statistical Association*, 105 (2010), 312-323.

[185] Q.Z. Zhang, R.C. Zhang and M.Q. Liu, A method for screening active effects in supersaturated designs, *Journal of Statistical Planning and Inference*, 137 (2007), 235-248.

[186] J. Zhou, A robust criterion for experimental designs for serially correlated observations, *Technometrics*, 43 (2001), 462-467.

[187] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Ser. B*, 67 (2005), 301-320.

[188] H. Zou and R. Li, One-step sparse estimates in nonconcave penalized likelihood models (with discussion), *The Annals of Statistics*, 36 (2008), 1509-1566.