ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΙΚΩΝ ΣΧΕΔΙΑΣΜΩΝ, ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΧΡΙΣΤΙΝΑΣ Α. ΠΑΡΠΟΥΛΑ

Διπλωματούχου Μαθηματικού Πανεπιστημίου Ιωαννίνων

ΕΠΙΒΛΕΠΩΝ:

Χ. ΚΟΥΚΟΥΒΙΝΟΣ

Καθηγητής Ε.Μ.Π.

ΑΘΗΝΑ, Φεβρουάριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΙΚΩΝ ΣΧΕΔΙΑΣΜΩΝ, ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΧΡΙΣΤΙΝΑΣ Α. ΠΑΡΠΟΥΛΑ

Διπλωματούχου Μαθηματικού Πανεπιστημίου Ιωαννίνων

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

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

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

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

ΑΘΗΝΑ, Φεβρουάριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Στατιστική Ανάλυση Παραγοντικών Σχεδιασμών, Γενικευμένα Γραμμικά Μοντέλα και Κριτήρια Επιλογής Μοντέλων

Διδακτορική Διατριβή
Χριστίνα Α. Παρπούλα

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

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

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

Ιωάννης Σπηλιώτης
Αναπληρωτής Καθηγητής Ε.Μ.Π. (Μέλος της Τριμελούς Επιτροπής)

Χαράλαμπος Ευαγγελάρας
Επίκουρος Καθηγητής Πανεπιστημίου Πειραιώς

Αλέξανδρος Καραγρηγορίου
Καθηγητής Πανεπιστημίου Αιγαίου

Χρυσηίς Καρώνη
Καθηγήτρια Ε.Μ.Π.

Μάρκος Κούτρας
Καθηγητής Πανεπιστημίου Πειραιώς

ΠΕΡΙΛΗΨΗ

Η διδακτορική αυτή διατριβή αποτελείται από τρία μέρη και συνολικά δεκαπέντε κεφάλαια, πέντε για το πρώτο μέρος, πέντε για το δεύτερο μέρος και πέντε για το τρίτο μέρος. Στο πρώτο μέρος της διατριβής "Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Μέτρα της Θεωρίας Πληροφοριών", μελετώνται διάφορες μέθοδοι στατιστικής ανάλυσης των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων. Οι διάφορες μέθοδοι ανάλυσης που παρουσιάζονται βασίζονται πάνω σε ποικίλα μέτρα της Θεωρίας Πληροφοριών. Στο δεύτερο μέρος της διατριβής "Επιλογή Μεταβλητών και Επιλογή Μοντέλου", μελετώνται διάφορες μέθοδοι επιλογής των καταλληλότερων μεταβλητών καθώς επίσης και κριτήρια πληροφορίας για την επιλογή του βέλτιστου μοντέλου. Επιπλέον, μελετώνται μέθοδοι επιλογής βέλτιστου σχεδιασμού μέσα από μία βάση πραγματικών δεδομένων. Στο τρίτο και τελευταίο μέρος της διατριβής "Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις", μελετώνται μέθοδοι κατασκευής και ανάλυσης υπερκορεσμένων σχεδιασμών που συμπεριλαμβάνουν αλληλεπιδράσεις. Επιπλέον, εισάγεται μία νέα κλάση σχεδιασμών που ονομάζονται μερικώς ισορροπημένοι υπερκορεσμένοι σχεδιασμοί και μελετώνται οι ιδιότητές τους και κάτω φράγματα για ευρέως γνωστά κριτήρια βελτιστοποίησης.

ABSTRACT

The present PhD thesis is divided into three parts consisting of fifteen chapters, five chapters for each part respectively. In the first part of the thesis "Statistical Analysis of Supersaturated Designs Via Information Theory Measures", several statistical analysis methods of main effects supersaturated designs are being studied, under generalized linear models. The presented analysis methods are based upon various information theory measures. In the second part of the thesis "Variable Selection and Model Selection", several best subset methods as well as information based criteria for optimal model selection are being studied. Moreover, optimal design selection methods in databases are being studied. In the third and last part of the thesis "Statistical Analysis of Supersaturated Designs Involving Interactions", construction and analysis methods of supersaturated designs involving interactions are being studied. Moreover, a new class of designs called as partially balanced supersaturated designs is being introduced, and their properties as well as the lower bounds of well-known optimality criteria are being studied.

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

Περιεχόμενα

Ευχαριστίες **ix**

Ερευνητικό Έργο **xi**

Περίληψη **xiii**

Κατάλογος Σχημάτων **xvi**

Κατάλογος Πινάκων **xix**

I Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Μέτρα της Θεωρίας Πληροφοριών **1**

1 Εισαγωγικά Στοιχεία και Βασικές Έννοιες **3**
1.1 Μέτρα της Θεωρίας Πληροφοριών 4
1.1.1 Μέτρα Εντροπίας 4
1.1.2 Μέτρα Πληροφορίας 6
1.2 Σχεδιασμοί Πειραμάτων: Ορισμοί και Ιδιότητες 8
1.2.1 Παραγοντικοί Σχεδιασμοί 8
1.2.2 Κλασματικοί Παραγοντικοί Σχεδιασμοί 10
1.2.3 Υπερκορεσμένοι Σχεδιασμοί 12
1.3 Ιστορική Αναδρομή Υπερκορεσμένων Σχεδιασμών 14
1.3.1 Προηγούμενη Συνεισφορά στην Κατασκευή Υπερκορεσμένων Σχεδιασμών 14
1.3.2 Προηγούμενη Συνεισφορά στην Ανάλυση Υπερκορεσμένων Σχεδιασμών 14
1.4 Ερευνητικά Προβλήματα στην Ανάλυση Υπερκορεσμένων Σχεδιασμών . . . 16
1.4.1 Ύπαρξη Μεροληψίας 16
1.4.2 Επέκταση από το Γραμμικό στο Γενικευμένο Γραμμικό Μοντέλο . . 17

2 Ανάλυση Υπερκορεσμένων Σχεδιασμών με Μέτρα Εντροπίας **19**
2.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 20
2.2 Η Προτεινόμενη Μέθοδος Ανάλυσης 20
2.2.1 Περιγραφή της Προτεινόμενης Μεθόδου 20
2.2.2 Αξιολόγηση της Προτεινόμενης Μεθόδου 21
2.3 Συμπεράσματα 23

3 Ανάλυση Υπερκορεσμένων Σχεδιασμών με το Κριτήριο της Συμμετρικής Αβεβαιότητας **25**
3.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 26
3.2 Η Προτεινόμενη Μέθοδος Ανάλυσης 26
3.2.1 Περιγραφή της Προτεινόμενης Μεθόδου 27
3.2.2 Κριτήρια Αξιολόγησης 28

3.3 Αξιολόγηση της Προτεινόμενης Μεθόδου 29
3.3.1 Εμπειρική Μελέτη 29
3.3.2 Συγκριτική Μελέτη 31
3.4 Συμπεράσματα 33

4 Αλγοριθμική Προσέγγιση του Προβλήματος Ανάλυσης Υπερκορεσμένων Σχεδιασμών 35
4.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 36
4.2 Η Προτεινόμενη Μέθοδος Ανάλυσης 36
4.2.1 Περιγραφή της Προτεινόμενης Μεθόδου 36
4.2.2 Κριτήρια Αξιολόγησης 37
4.3 Αξιολόγηση της Προτεινόμενης Μεθόδου 38
4.3.1 Εμπειρική Μελέτη 39
4.3.2 Συγκριτική Μελέτη 43
4.4 Συμπεράσματα 47

5 Επέκταση του Προβλήματος σε Γενικευμένα Γραμμικά Μοντέλα με Διακριτή Απόκριση 49
5.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 50
5.2 Γενικευμένα Γραμμικά Μοντέλα 50
5.2.1 Το Λογιστικό Μοντέλο Παλινδρόμησης 51
5.2.2 Μοντέλα Παλινδρόμησης Poisson, Αρνητικής Διωνυμικής και Γεωμετρικής Κατανομής 52
5.3 Η Προτεινόμενη Μέθοδος Ανάλυσης 54
5.3.1 Περιγραφή της Προτεινόμενης Μεθόδου 55
5.3.2 Αξιολόγηση της Προτεινόμενης Μεθόδου 56
5.4 Συγκριτική Μελέτη 57
5.4.1 Συγκριτική Μελέτη: Bernoulli Απόκριση 58
5.4.2 Συγκριτική Μελέτη: Count Απόκριση 60
5.4.3 Συγκριτική Μελέτη: Επιλογή Τιμών Κατωφλιού 66
5.5 Συμπεράσματα 72

II Επιλογή Μεταβλητών και Επιλογή Μοντέλου 73

6 Εισαγωγικά Στοιχεία και Βασικές Έννοιες 75
6.1 Ανάπτυξη Βέλτιστου Μοντέλου 76
6.2 Έλεγχοι και Κριτήρια 76
6.3 Μέθοδοι Επιλογής των Καταλληλότερων Μεταβλητών 77
6.3.1 Μέθοδοι Επιλογής του Καλύτερου Υποσυνόλου Μεταβλητών 77
6.3.2 Μέθοδοι Επιλογής Μοντέλου με Βήματα 78
6.3.3 Ποινικοποιημένες Μέθοδοι Επιλογής Μεταβλητών 79
6.4 Πολυπλοκότητα Μοντέλου 80

7 Στατιστική Πολυπλοκότητα και Περιθώρια Κατανομή σε Πειράματα Bernoulli 83
7.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 84
7.2 Η Παράμετρος Prior σε Ένα Πείραμα Bernoulli 84
7.2.1 Σύνδεση της Παραμέτρου Prior με την Πολυπλοκότητα ενός Μοντέλου 84
7.2.2 Ένα Νέο Μέτρο Στατιστικής Πολυπλοκότητας (EPC) 88
7.2.3 Υπολογισμός της Εκ των Προτέρων Περιθώριας Κατανομής (MPD) 89
7.2.4 Εμπειρική Μελέτη 91

7.3 Πολυπλοκότητα Μοντέλου και Σχεδιασμοί Πειραμάτων 93
7.3.1 Επιλογή Βέλτιστου Μοντέλου στους Υπερκορεσμένους Σχεδιασμούς Μέσω του Κριτηρίου EPC 93
7.3.2 Επιλογή Τιμών Κατωφλιού 93
7.3.3 Αξιολόγηση της Μεθόδου Επιλογής Βέλτιστου Μοντέλου 94
7.4 Συμπεράσματα 97

8 Επιλογή του Καλύτερου Υποσυνόλου Μεταβλητών Μέσω του Κριτηρίου Πληροφορίας ϵBIC 99
8.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 100
8.2 Η Προτεινόμενη Μέθοδος Επιλογής του Καλύτερου Υποσυνόλου Μεταβλητών 101
8.3 Συγκριτική Μελέτη 102
8.3.1 Κριτήρια Αξιολόγησης 102
8.3.2 Μοτίβο Προσομοιώσεων 103
8.3.3 Η Παράμετρος ϵ 104
8.3.4 Αποτελέσματα Μελέτης Προσομοίωσης 105
8.3.5 Ανάλυση Πραγματικών Δεδομένων 108
8.4 Συμπεράσματα 111

9 Επιλογή Μεταβλητών σε Προβλήματα Υψηλής Διάστασης: Μοντελοποίηση Μέσω Υπερκορεσμένων Σχεδιασμών 113
9.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 114
9.2 Περιγραφή Μεθόδων Επιλογής Μεταβλητών 114
9.3 Η Προτεινόμενη Μέθοδος Επιλογής Μεταβλητών 115
9.4 Ανάλυση Πραγματικών Δεδομένων 115
9.4.1 Διαδοχική Επιλογή Μεταβλητών 116
9.4.2 Επιλογή Μεταβλητών Μέσω Υπερκορεσμένων Σχεδιασμών 119
9.4.3 Συγκριτικά Αποτελέσματα 120
9.5 Συμπεράσματα 121

10 Επιλογή Σχεδιασμού σε Βάσεις Δεδομένων 123
10.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 124
10.2 Τεχνικές Εξόρυξης Δεδομένων και Γενετικοί Αλγόριθμοι 125
10.2.1 Εξόρυξη Κανόνων Συσχέτισης 125
10.2.2 Απλός Γενετικός Αλγόριθμος 125
10.2.3 L_1-norm Μηχανές Διανυσματικής Υποστήριξης 126
10.3 Η Προτεινόμενη Μέθοδος Επιλογής Σχεδιασμού 126
10.4 Εφαρμογή και Πειραματικά Αποτελέσματα 127
10.4.1 Περιγραφή της Βάσης Δεδομένων 127
10.4.2 Κριτήρια Αξιολόγησης της Απόδοσης 128
10.4.3 Επιλογή Υπερκορεσμένων Σχεδιασμών 130
10.4.4 Συγκριτικά Αποτελέσματα 131
10.4.5 Περαιτέρω Ανάλυση 133
10.5 Συμπεράσματα 133

III Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις 135

11 Εισαγωγικά Στοιχεία και Βασικές Έννοιες 137
11.1 Ιστορική Αναδρομή 138
11.2 Τεχνικές Κρησαρίσματος με Αλληλεπιδράσεις 138

11.3 Βασικές Αρχές Τεχνικών Κρησαρίσματος 139

12 Ανάλυση Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις Μέσω Ποινικοποιημένων Μηχανών Διανυσματικής Υποστήριξης 141
12.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 142
12.2 Ποινικοποιημένες Μηχανές Διανυσματικής Υποστήριξης 142
12.2.1 Μαθηματική Περιγραφή των Μηχανών Διανυσματικής Υποστήριξης 143
12.2.2 Η L_1-norm SVM 144
12.2.3 Η L_1-norm Huberized SVM 144
12.3 Εμπειρική Μελέτη Προσομοίωσης 145
12.3.1 Κριτήρια Αξιολόγησης της Απόδοσης 146
12.3.2 Κρησάρισμα Υπερκορεσμένων Σχεδιασμών Κύριων Επιδράσεων . . . 146
12.3.3 Κρησάρισμα Υπερκορεσμένων Σχεδιασμών Κύριων Επιδράσεων και Αλληλεπιδράσεων σε Ένα Στάδιο 150
12.3.4 Ιδιότητες Ποινικοποιημένων Μηχανών Διανυσματικής Υποστήριξης . 155
12.4 Ανάλυση Πραγματικών Δεδομένων 155
12.5 Συμπεράσματα 157

13 Κατασκευή Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις 159
13.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 160
13.2 Τροποποίηση του Cronbach Άλφα Συντελεστή 161
13.2.1 Κίνητρο Χρήσης και Τροποποίησης του Cronbach Άλφα Συντελεστή 161
13.2.2 Cronbach Άλφα Συντελεστής 162
13.2.3 Το Τροποποιημένο Κριτήριο 162
13.2.4 Συγκριτική Απόδοση του Τροποποιημένου Κριτηρίου 165
13.3 Η Προτεινόμενη Μέθοδος Βελτιστοποίησης 166
13.3.1 Στατιστικά Μέτρα Συσχέτισης 166
13.3.2 Ο Προτεινόμενος Αλγόριθμος Βελτιστοποίησης (BEDDO) 167
13.4 Κατασκευή Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών 168
13.4.1 Παράδειγμα 1: $\mathrm{SSD}_{(6,10)}$ 169
13.4.2 Παράδειγμα 2: $\mathrm{SSD}_{(8,14)}$ 170
13.4.3 Παράδειγμα 3: $\mathrm{SSD}_{(10,18)}$ 170
13.4.4 Σύγκριση Μη Ισορροπημένων και Ισορροπημένων Υπερκορεσμένων Σχεδιασμών 171
13.5 Συμπεράσματα 173

14 Ανάλυση Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις 175
14.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 176
14.2 Ανάλυση Ισορροπημένων και Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών 176
14.3 Σύγκριση Ισορροπημένων ΥΣ και Μη Ισορροπημένων Bayesian D-Βέλτιστων ΥΣ 178
14.3.1 Ανάλυση Δεδομένων Μέσω Γενικευμένων Γραμμικών Μοντέλων . . 179
14.3.2 Ιδιότητες Βελτιστότητας 180
14.4 Σύγκριση Ισορροπημένων ΥΣ και Μη Ισορροπημένων ΥΣ με Αλληλεπιδράσεις 181
14.4.1 Ανάλυση Δεδομένων Μέσω Γενικευμένων Γραμμικών Μοντέλων . . 181
14.4.2 Ιδιότητες Βελτιστότητας 183
14.5 Συμπεράσματα 184

15 Κάτω Φράγματα του Μέτρου $E(S^2)$ για Μερικώς Ισορροπημένους Σχεδιασμούς Δύο Επιπέδων **187**
15.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα 188
15.2 Το Κριτήριο Βελτιστοποίησης $E(s^2)$ 188
15.2.1 Συμβολισμοί και Προκαταρκτικά 189
15.2.2 Νέο Κάτω Φράγμα για το $E(s^2)$ Κριτήριο Βελτιστοποίησης 191
15.2.3 Νέοι $E(s^2)$-Βέλτιστοι Μερικώς Ισορροπημένοι Σχεδιασμοί 193
15.2.4 Εφαρμογή του Νέου Κάτω Φράγματος σε Υπάρχοντες Σχεδιασμούς 195
15.3 Συμπεράσματα 195

Βιβλιογραφία **197**

Ευχαριστίες

Κατά τη διάρκεια των διδακτορικών μου σπουδών στο Εθνικό Μετσόβιο Πολυτεχνείο (Ε.Μ.Π.), αρκετοί άνθρωποι συνέβαλαν, ο καθένας με τον τρόπο του, στο να αποκτήσω το μορφωτικό επίπεδο που έχω σήμερα και θα ήθελα σε αυτό το σημείο να τους εκφράσω τις ευχαριστίες μου, με την πεποίθηση ότι φάνηκα αντάξια των προσδοκιών τους.

Η εκπόνηση αυτής της διατριβής θα ήταν αδύνατη χωρίς τη συμβολή του Επιβλέποντος κ. Χρήστου Κουκουβίνου, Καθηγητή της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π. Η καθοδήγησή του υπήρξε αρωγός στην ερευνητική μου δραστηριότητα, καθώς μέσω των ιδεών, προτάσεων και στοχευμένων παρατηρήσεών του, που μπορούν να βρεθούν με τη μορφή ερευνητικών προβλημάτων στην αρχή κάθε κεφαλαίου της διατριβής, συνέβαλε καθοριστικά στο να εμβαθύνω και να παρουσιάσω αρτιότερα τα ερευνητικά μου αποτελέσματα. Συγχρόνως, η διαρκής υποστήριξη και αδιάκοπη ενθάρρυνση που μου παρείχε αποτέλεσαν κινητήριο μοχλό για την επιτυχή ολοκλήρωση της διατριβής αυτής. Για αυτόν το λόγο, όπως και για τη δυνατότητα που μου έδωσε να ασχοληθώ ερευνητικά με τις δύο μεγάλες επιστημονικές μου αγάπες, τις επιστήμες των Μαθηματικών και της Στατιστικής, τον ευχαριστώ θερμά.

Τις ευχαριστίες μου θα ήθελα επίσης να εκφράσω στα άλλα δύο μέλη της Τριμελούς Συμβουλευτικής Επιτροπής, την κα. Φιλία Βόντα, Επίκουρη Καθηγήτρια της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π. και τον κ. Ιωάννη Σπηλιώτη, Αναπληρωτή Καθηγητή της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π., των οποίων οι πολύτιμες συμβουλές και υποδείξεις καθώς επίσης και η ηθική τους συμπαράσταση με βοήθησαν τα μέγιστα. Θα ήθελα επίσης να εκφράσω τις ευχαριστίες μου στα υπόλοιπα μέλη της Επταμελούς Εξεταστικής Επιτροπής, την κα. Χρυσηΐς Καρώνη, Καθηγήτρια της Σ.Ε.Μ.Φ.Ε. του Ε.Μ.Π., τον κ. Μάρκο Κούτρα, Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, τον κ. Χαράλαμπο Ευαγγελάρα, Επίκουρο Καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς και τον κ. Αλέξανδρο Καραγρηγορίου, Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Αιγαίου, για το χρόνο που αφιέρωσαν στην ανάγνωση της διατριβής, καθώς και για τις χρήσιμες υποδείξεις τους.

Ευχαριστήρια οφείλω προς τον Τομέα Μαθηματικών της Σ.Ε.Μ.Φ.Ε του Ε.Μ.Π., για τη χορήγηση υποτροφίας εκ του Κληροδοτήματος Χρ. Παπακυριακόπουλου κατά τη διάρκεια εκπόνησης της διδακτορικής μου διατριβής, υποστηρίζοντας οικονομικά την έρευνά μου.

Επίσης, θα ήθελα να ευχαριστήσω θερμά τον Επιβλέποντα Καθηγητή μου κ.Χρήστο Κουκουβίνο για τη δυνατότητα που μου έδωσε να συνεργαστώ με τον Καθηγητή του Πανεπιστημίου McMaster του Καναδά κ. Narayanaswamy Balakrishnan και τον Καθηγητή του Πανεπιστημίου Visva Bharati της Ινδίας κ. Kashinath Chatterjee κατά τη διάρκεια εκπόνησης της διατριβής μου. Είναι μεγάλη μου χαρά και τιμή που έχω συνεργαστεί μαζί τους και τους ευχαριστώ θερμά για τις πολύτιμες υποδείξεις τους και την εποικοδομητική συνεργασία μας σε θέματα Πειραματικών Σχεδιασμών.

Εκφράζω επίσης τις ευχαριστίες μου στους υπόλοιπους συνεργάτες, συναδέλφους και φίλους, τον Επισκέπτη Καθηγητή του Vienna University of Technology της Βιέννης και Κύριο Ερευνητή στο Ινστιτούτο SBA Research της Βιέννης κ. Δημήτριο Σίμο για τη συμπαράσταση, συνεχή ενθάρρυνση και την άψογη συνεργασία μας καθόλη τη διάρκεια εκπόνησης αυτής της διατριβής, τον υποψήφιο διδάκτορα Ε.Μ.Π. κ. Εμμανουήλ Ανδρουλάκη για την υποστήριξη και τις εποικοδομητικές συζητήσεις που είχαμε στα χρόνια των διδακτορικών μου σπουδών. Θερμές ευχαριστίες επίσης εκφράζω στην αδελφή μου Νίκη, στον Γιάννη και σε όλους τους φίλους μου που με στήριξαν όλα αυτά τα χρόνια.

Τέλος, θα ήθελα να ευχαριστήσω μέσα από την καρδιά μου τους γονείς μου, Αθανάσιο και Σοφία, για την αμέριστη ηθική συμπαράσταση, υλική υποστήριξη και καθημερινή ενθάρρυνσή τους, σε όλη τη διάρκεια των διδακτορικών μου σπουδών και ως ένα ελάχιστο δείγμα ευγνωμοσύνης η παρούσα διατριβή αφιερώνεται σε αυτούς.

Ερευνητικό Έργο

Κατά τη διάρκεια εκπόνησης αυτής της διδακτορικής διατριβής προέκυψαν οι παρακάτω δημοσιευμένες ή προς δημοσίευση επιστημονικές εργασίες:

- Analyzing supersaturated designs with entropic measures, (με Χ. Κουκουβίνο, Ε. Μάσσου και Κ. Μυλωνά), *Journal of Statistical Planning and Inference*, 141 (2011), 1307-1312.

- Analyzing supersaturated designs by means of an information based criterion, (με Χ. Κουκουβίνο), *Communications in Statistics-Simulation and Computation*, 41 (2012), 44-57.

- An information theoretical algorithm for analyzing supersaturated designs for a binary response, (με Χ. Κουκουβίνο και N. Balakrishnan), *Metrika*, 76 (2013), 1-18.

- Analyzing supersaturated designs for discrete responses via generalized linear models, (με Χ. Κουκουβίνο και N. Balakrishnan), *Statistical Papers*, (2014), DOI: 10.1007/s00362-013-0569-z.

- Variable selection and computation of the prior probability of a model via ROC curves methodology, (με Χ. Κουκουβίνο), *Journal of Data Science*, 10 (2012), 653-672.

- Analysis of a supersaturated design using entropy prior complexity for binary responses via generalized linear models, (με Χ. Κουκουβίνο και N. Balakrishnan), *Statistical Methodology*, 79 (2012) 478-485.

- Genetic algorithm and data mining techniques for design selection in databases, (με Χ. Κουκουβίνο και Δ. Σίμο), *Proceedings of the 2013 IEEE International Conference on Availability, Reliability and Security*, (2013), Regensburg, pp.743-746, DOI: 10.1109/ARES.2013.98.

- A new variable selection approach inspired by supersaturated designs given a large-dimensional dataset, (με Κ. Δρόσου, Χ. Κουκουβίνο και Κ. Μυλωνά), *Journal of Data Science*, 12 (2014), 35-52.

- On the computation of entropy prior complexity and marginal prior distribution for the Bernoulli model, (με Χ. Κουκουβίνο και N. Balakrishnan), *Journal of Statistical Theory and Practice*, 9 (2015), 59-72.

- A penalized wrapper method for screening main effects and interactions in supersaturated designs, (με Χ. Κουκουβίνο), *Quality and Reliablity Engineering International*, (2014), DOI: 10.1002/qre.1679.

- Computer-aided unbalanced supersaturated designs involving interactions, (με K. Chatterjee και Χ. Κουκουβίνο), έχει υποβληθεί για δημοσίευση.

- On the analysis of unbalanced two-level supersaturated designs via generalized linear models, (με K. Chatterjee και X. Κουκουβίνο), έχει υποβληθεί για δημοσίευση.
- A lower bound to $E(s^2)$-optimality measure for partially balanced two-level designs, (με K. Chatterjee και X. Κουκουβίνο), έχει υποβληθεί για δημοσίευση.

Περίληψη

Η Αρχή είναι το ήμισυ του παντός.

—*Πυθαγόρας (περ. 570 π.Χ.–490 π.Χ)*

Στη διδακτορική αυτή διατριβή, μελετώνται μέθοδοι στατιστικής ανάλυσης μίας κλάσης των παραγοντικών σχεδιασμών που ονομάζονται υπερκορεσμένοι σχεδιασμοί, με ή χωρίς αλληλεπιδράσεις, κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων. Επίσης, μελετώνται μέθοδοι επιλογής των καταλληλότερων μεταβλητών καθώς επίσης και κριτήρια πληροφορίας για την επιλογή του βέλτιστου μοντέλου.

Η διατριβή αποτελείται από τρία μέρη και συνολικά δεκαπέντε κεφάλαια, πέντε για το πρώτο μέρος, πέντε για το δεύτερο μέρος και πέντε για το τρίτο μέρος. Στο πρώτο μέρος της διατριβής "Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Μέτρα της Θεωρίας Πληροφοριών", μελετώνται διάφορες μέθοδοι στατιστικής ανάλυσης των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων. Οι διάφορες μέθοδοι ανάλυσης που παρουσιάζονται βασίζονται πάνω σε ποικίλα μέτρα της Θεωρίας Πληροφοριών. Στο δεύτερο μέρος της διατριβής "Επιλογή Μεταβλητών και Επιλογή Μοντέλου", μελετώνται διάφορες μέθοδοι επιλογής των καταλληλότερων μεταβλητών καθώς επίσης και κριτήρια πληροφορίας για την επιλογή του βέλτιστου μοντέλου. Επιπλέον, μελετώνται μέθοδοι επιλογής βέλτιστου σχεδιασμού μέσα από μία βάση πραγματικών δεδομένων. Στο τρίτο και τελευταίο μέρος της διατριβής "Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις", μελετώνται μέθοδοι κατασκευής και ανάλυσης υπερκορεσμένων σχεδιασμών που συμπεριλαμβάνουν αλληλεπιδράσεις. Επιπλέον, εισάγεται μία νέα κλάση σχεδιασμών που ονομάζονται μερικώς ισορροπημένοι υπερκορεσμένοι σχεδιασμοί και μελετώνται οι ιδιότητές τους και κάτω φράγματα για ευρέως γνωστά κριτήρια βελτιστοποίησης.

Η παρουσίαση των επιμέρους θεμάτων και αποτελεσμάτων της διατριβής αυτής οργανώνεται ως εξής:

Στο Κεφάλαιο 1 παρουσιάζονται οι βασικοί ορισμοί, έννοιες, ιδιότητες, διάφορα στατιστικά εργαλεία και μέτρα από τη Θεωρία Πληροροφοριών, τα οποία ενσωματώνονται στις προτεινόμενες μεθόδους ανάλυσης των υπερκορεσμένων σχεδιασμών. Επιπλέον, παρουσιάζεται μια σύντομη ιστορική αναδρομή σχετικά με το τι έχει προταθεί μέχρι σήμερα στη διεθνή βιβλιογραφία για την κατασκευή και ανάλυση των υπερκορεσμένων σχεδιασμών. Τέλος, αναφέρονται τα ερευνητικά προβλήματα τα οποία καλούμαστε να αντιμετωπίσουμε στην παρούσα διατριβή σχετικά με την ανάλυση των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων.

Στο Κεφάλαιο 2 παρουσιάζεται μία νέα μέθοδος ανάλυσης υπερκορεσμένων σχεδιασμών στην οποία εξετάζονται τα εξής μέτρα εντροπίας: Rényi εντροπία, Tsallis εντροπία και Havrda-Charvát εντροπία. Τα μέτρα αυτά σε συνδυασμό με το κέρδος πληροφορίας χρησιμοποιούνται για την ανάλυση των υπερκορεσμένων σχεδιασμών και την εύρεση των σημαντικών παραγόντων χρησιμοποιώντας το λογιστικό μοντέλο παλινδρόμησης στο οποίο η μεταβλητή απόκρισης είναι δίτιμη.

Στο Κεφάλαιο 3 παρουσιάζεται μία νέα μέθοδος επιλογής μεταβλητών σε έναν υπερκορεσμένο σχεδιασμό χρησιμοποιώντας τον αλγόριθμο εξόρυξης δεδομένων Fast Correlation Based Filter. Η εντροπία Shannon σε συνδυασμό με το κέρδος πληροφορίας και το κριτήριο

της συμμετρικής αβεβαιότητας χρησιμοποιούνται για την ανάλυση των υπερκορεσμένων σχεδιασμών και την εύρεση των ενεργών κύριων επιδράσεων, θεωρώντας το λογιστικό μοντέλο παλινδρόμησης.

Στο Κεφάλαιο 4 αναπτύσσεται ένας συνεκτικός αλγόριθμος επιλογής μεταβλητών για την ανάλυση των υπερκορεσμένων σχεδιασμών, όπου οι επεξηγηματικές μεταβλητές βρίσκονται σε δύο επίπεδα και τα δεδομένα απόκρισης προέρχονται από το λογιστικό μοντέλο παλινδρόμησης.

Στο Κεφάλαιο 5 και τελευταίο κεφάλαιο του πρώτου μέρους, παρουσιάζεται η γενίκευση του προβλήματος ανάλυσης των υπερκορεσμένων σχεδιασμών για τους οποίους οι επεξηγηματικές μεταβλητές βρίσκονται σε δύο επίπεδα και τα δεδομένα απόκρισης είναι διακριτά. Επεκτείνεται η αλγοριθμική προσέγγιση του προβλήματος ανάλυσης υπερκορεσμένων σχεδιασμών του Κεφαλαίου 4 θεωρώντας γενικευμένα γραμμικά μοντέλα των οποίων τα δεδομένα απόκρισης είναι διακριτά, δηλαδή προέρχονται από την Poisson, γεωμετρική και αρνητική διωνυμική κατανομή.

Στο Κεφάλαιο 6 περιγράφεται συνοπτικά το πρόβλημα εξεύρεσης του βέλτιστου μοντέλου ή υποδείγματος που είναι και ο βασικός στόχος μιας στατιστικής μοντελοποίησης σε ένα σύνολο δεδομένων. Επιπλέον, παρουσιάζονται κριτήρια και έλεγχοι που χρησιμοποιούνται για το σκοπό αυτό, και γίνεται λεπτομερής αναφορά στην έννοια της πολυπλοκότητας ενός μοντέλου. Τέλος, παρουσιάζεται μια σύντομη ιστορική αναδρομή σχετικά με το τι έχει προταθεί βιβλιογραφικά μέχρι τώρα για την καλύτερη δυνατή ανάπτυξη ενός μοντέλου.

Στο Κεφάλαιο 7 παρουσιάζεται ένα νέο μέτρο στατιστικής πολυπλοκότητας EPC (Entropy Prior Complexity), μελετώνται οι ιδιότητες του νέου αυτού μέτρου, και γενικά οι διάφορες πτυχές της πολυπλοκότητας ενός μοντέλου. Επιπλέον, ενσωματώνεται επιτυχώς το μέτρο πολυπλοκότητας EPC στο πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών, με στόχο την αναζήτηση και εύρεση του βέλτιστου μοντέλου μέσα από πολλά υποψήφια λογιστικά μοντέλα παλινδρόμησης. Τέλος, μελετώνται η έννοια και ο υπολογισμός της εκ των προτέρων περιθώριας κατανομής MPD (Marginal Prior Distribution) ενός πειράματος Bernoulli.

Στο Κεφάλαιο 8 προτείνεται μία μέθοδος επιλογής του καλύτερου υποσυνόλου μεταβλητών, η οποία επεκτείνει την ιδέα επιλογής μεταβλητών σε προβλήματα ταξινόμησης ή πρόβλεψης, λαμβάνοντας ταυτόχρονα υπόψη τη μεμονωμένη προβλεπτική δύναμη της κάθε μεταβλητής. Η προτεινόμενη μέθοδος βασίζεται στη μετρική AUC και σε μια τροποποίηση (ϵBIC) του κριτηρίου mBIC (modified Bayesian Information Criterion). Η σύγκριση της απόδοσης του νέου αυτού κριτηρίου ϵBIC με κλασικά κριτήρια πληροφορίας (όπως το AIC, BIC και mBIC) γίνεται αρχικά εκτελώντας εκτενή πειράματα προσομοίωσης και έπειτα αναλύοντας μία πραγματική βάση δεδομένων.

Στο Κεφάλαιο 9 μελετάται το πρόβλημα της επιλογής μεταβλητών σε προβλήματα παλινδρόμησης στην περίπτωση που οι παρατηρήσεις εισόδου και οι κλάσεις εξόδου ενός πραγματικού μεγάλου συνόλου δεδομένων είναι διαθέσιμες. Οι στατιστικές μέθοδοι οι οποίες εφαρμόζονται είναι οι διαδικασίες μη κοίλης ποινικοποιημένης πιθανοφάνειας (SCAD, LASSO και Hard), η λογιστική παλινδρόμηση, και τεχνικές επιλογής του καλύτερου υποσυνόλου μεταβλητών (με κριτήρια πληροφορίας τα AIC και BIC). Επιπλέον, παρουσιάζεται μια νέα προσέγγιση επιλογής μεταβλητών εμπνευσμένη από τους υπερκορεσμένους σχεδιασμούς, δοθέντος ενός μεγάλου συνόλου δεδομένων παρατηρήσεων.

Στο Κεφάλαιο 10 και τελευταίο κεφάλαιο του δεύτερου μέρους, μελετώνται η εφαρμογή κάποιων τεχνικών εξόρυξης δεδομένων, ενός γενετικού αλγορίθμου καθώς επίσης και η χρήση των πειραματικών σχεδιασμών σε βάσεις δεδομένων. Προτείνεται ένα κρυπτογραφικό σχήμα καθοδηγούμενο από την ίδια τη βάση των δεδομένων, με στόχο την κρυπτογράφηση συγκεκριμένων πεδίων της βάσης δεδομένων και την επιλογή ενός βέλτιστου σχεδιασμού αποτελούμενου από τις μεταβλητές της βάσης δεδομένων που ανιχνεύθηκαν να επηρεάζουν σημαντικά τη μεταβλητή απόκρισης.

Στο Κεφάλαιο 11 μελετάται η ύπαρξη των αλληλεπιδράσεων στους υπερκορεσμένους σχεδιασμούς και παρουσιάζεται μια σύντομη ιστορική αναδρομή σχετικά με το τι έχει προταθεί βιβλιογραφικά έως τώρα για την κατασκευή και την ανάλυση των υπερκορεσμένων σχεδιασμών συμπεριλαμβανομένων των αλληλεπιδράσεων. Επίσης, παρουσιάζονται οι βασικές αρχές οι οποίες διέπουν τις μεθόδους κρησαρίσματος των υπερκορεσμένων σχεδιασμών με αλληλεπιδράσεις.

Στο Κεφάλαιο 12 μελετάται η στατιστική ανάλυση υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων δεύτερης τάξης. Το πρόβλημα εύρεσης και σχηματισμού του καταλληλότερου μοντέλου αντιμετωπίζεται ως ένα πρόβλημα βελτιστοποίησης, του οποίου η επίλυση οδηγεί στο βέλτιστο υπόδειγμα. Η προτεινόμενη ποινικοποιημένη μέθοδος βελτιστοποίησης διατυπωμένη μέσω συναρτήσεων απώλειας και συναρτήσεων ποινής, οδηγεί σε μια πιο άμεση και συνεκτική μέθοδο κρησαρίσματος, η συμπεριφορά της οποίας μελετάται χρησιμοποιώντας το λογιστικό μοντέλο παλινδρόμησης, με στόχο την εύρεση και εκτίμηση των σημαντικών παραγόντων των υπερκορεσμένων σχεδιασμών.

Στο Κεφάλαιο 13 μελετάται η δομή ταυτόσημων επιδράσεων στους υπερκορεσμένους σχεδιασμούς συμπεριλαμβανομένων των αλληλεπιδράσεων. Παρουσιάζεται μία νέα μέθοδος βελτιστοποίησης, η οποία αποτελεί μία τεχνική προς τα πίσω απαλοιφής καθοδηγούμενη από ένα σχεδιασμό. Για την υλοποίηση της προτεινόμενης μεθόδου BEDDO (Backward Elimination Design-Driven Optimization), ενσωματώνονται και συνδυάζονται στατιστικά μέτρα βασισμένα στη συσχέτιση και παρουσιάζεται ένα νέο κριτήριο βελτιστότητας το οποίο αποτελεί μία τροποποιημένη μορφή του Cronbach άλφα συντελεστή. Με αυτό τον τρόπο και με τη βοήθεια του υπολογιστή, παρέχεται μία νέα κλάση μη ισορροπημένων υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις, οι οποίοι απορρέουν άμεσα από τη μέθοδο βελτιστοποίησης BEDDO.

Στο Κεφάλαιο 14 μελετώνται οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί. Το πλαίσιο ανάλυσης μέχρι σήμερα για τους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς βασίζεται στην κεντρική υπόθεση ότι τα πειραματικά δεδομένα προέρχονται από ένα γραμμικό μοντέλο. Στο κεφάλαιο αυτό, οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί συγκρίνονται με ισορροπημένους ομόλογούς τους, ως προς την απόδοσή τους, κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων.

Στο Κεφάλαιο 15 και τελευταίο κεφάλαιο της διατριβής, μελετώνται φράγματα για το κριτήριο βελτιστοποίησης $E(s^2)$. Παρουσιάζεται ένα νέο φράγμα του κριτηρίου βελτιστοποίησης $E(s^2)$ για μερικώς ισορροπημένους σχεδιασμούς. Επιπλέον, παρέχονται νέοι $E(s^2)$-βέλτιστοι μερικώς ισορροπημένοι σχεδιασμοί οι οποίοι προκύπτουν από την εφαρμογή του νέου αυτού κάτω φράγματος.

Αυτή η διδακτορική διατριβή στοιχειοθετήθηκε με το πρόγραμμα LaTeX (διανομές MiKTeX και BibTeX). Η συγγραφή έγινε με τη βοήθεια του προγράμματος WinEdt (στο λειτουργικό σύστημα Microsoft Windows). Η τελική ηλεκτρονική μορφή (Portable Document Format – PDF) δημιουργήθηκε με το πρόγραμμα PDFLaTeX. Για την ανάπτυξη των μεθόδων, των αλγορίθμων και των προσομοιώσεων, καθώς επίσης και τον έλεγχο των αποτελεσμάτων, χρησιμοποιήθηκε πρωτίστως το λογισμικό MATLAB και δευτερευόντως το λογισμικό IBM SPSS Clementine. Οι γραφικές παραστάσεις έγιναν με τη βοήθεια των προγραμμάτων MATLAB και Microsoft Office Excel, και η επεξεργασία των σχημάτων με το πρόγραμμα Adobe Photoshop.

Κατάλογος Σχημάτων

3.1 Σύγκριση των ROC καμπυλών 33

5.1 Σφάλματα Τύπου I και Τύπου II για την κατανομή Poisson 63
5.2 Σφάλματα Τύπου I και Τύπου II για τη γεωμετρική κατανομή 64
5.3 Σφάλματα Τύπου I και Τύπου II για την αρνητική διωνυμική κατανομή . . . 64
5.4 Σύγκριση σφαλμάτων Τύπου I για τα Σενάρια I, II και III 65
5.5 Σύγκριση σφαλμάτων Τύπου II για τα Σενάρια I, II και III 65

12.1 Σύγκριση σφαλμάτων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων . . 149
12.2 Ακρίβεια μοντέλων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων 149
12.3 Ανάλυση ευαισθησίας και ειδικότητας υπερκορεσμένων σχεδιασμών κύριων επιδράσεων 149
12.4 Σύγκριση σφαλμάτων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων 153
12.5 Ακρίβεια μοντέλων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων 154
12.6 Ανάλυση ευαισθησίας και ειδικότητας υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων 154

Κατάλογος Πινάκων

1.1 Εκτέλεση πειράματος ενός μόνο παράγοντα 9
1.2 Εκτέλεση πειράματος με παραγοντικό σχεδιασμό 10

2.1 Μοντέλα 1ης μελέτης προσομοίωσης 22
2.2 Απόδοση της 1ης προτεινόμενης μεθόδου για τα μοντέλα 1-22 με χρήση 1000 προσομοιώσεων 23

3.1 Πίνακας συνάφειας 28
3.2 Μοντέλα 2ης μελέτης προσομοίωσης 30
3.3 Απόδοση της 2ης προτεινόμενης μεθόδου για όλες τις τιμές κατωφλιού για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων 31
3.4 Συγκριτική απόδοση της 2ης προτεινόμενης μεθόδου για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων 32

4.1 Μοντέλα 3ης μελέτης προσομοίωσης 41
4.2 Απόδοση της 3ης προτεινόμενης μεθόδου για τα μοντέλα 1-57 με χρήση 1000 προσομοιώσεων 42
4.3 Συγκριτική απόδοση της 3ης προτεινόμενης μεθόδου για διαφορετικές εντροπίες για τα μοντέλα 1-57 με χρήση 1000 προσομοιώσεων 45
4.4 Συγκριτική απόδοση της 3ης προτεινόμενης μεθόδου για διαφορετικές επιλογές παραμέτρων για τα μοντέλα 1-57 με χρήση 1000 προσομοιώσεων . . 46

5.1 Μοντέλα 4ης μελέτης προσομοίωσης 57
5.2 Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο IV με χρήση 1000 προσομοιώσεων 60
5.3 Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο I με χρήση 1000 προσομοιώσεων 62
5.4 Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο II με χρήση 1000 προσομοιώσεων 62
5.5 Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο III με χρήση 1000 προσομοιώσεων 63
5.6 Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο I με χρήση 1000 προσομοιώσεων 68
5.7 Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο II με χρήση 1000 προσομοιώσεων 69
5.8 Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο III με χρήση 1000 προσομοιώσεων 70
5.9 Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο IV με χρήση 1000 προσομοιώσεων 71

7.1 Αποτελέσματα 10,000 Monte Carlo δοκιμών για τα λογιστικά μοντέλα παλινδρόμησης 92
7.2 Μοντέλα μελέτης προσομοίωσης 95
7.3 Απόδοση της προτεινόμενης μεθόδου για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων 96

7.4 Απόδοση της προτεινόμενης μεθόδου για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων αν θεωρήσουμε ότι δεν υπάρχουν ενεργοί παράγοντες . . . 97

8.1 Η τυχαία επιλεγμένη δομή για τις x_j . . . 104
8.2 Απόδοση της μεθόδου με χρήση 1000 προσομοιώσεων για το Σενάριο I ($0 < p < 0.5$) . . . 106
8.3 Απόδοση της μεθόδου με χρήση 1000 προσομοιώσεων για το Σενάριο II ($0.5 < p < 1$) . . . 106
8.4 Χρόνοι εκτέλεσης (sec) . . . 106
8.5 Χρόνοι εκτέλεσης (sec) . . . 106
8.6 Μελέτη Τραύματος . . . 109
8.7 Οι 44 σημαντικές επιλεγμένες μεταβλητές . . . 109
8.8 Εκτιμώμενοι συντελεστές β και τυπικά σφάλματα (σε παρενθέσεις) . . . 110
8.9 Χρόνοι εκτέλεσης (sec) . . . 110
8.10 Αξιολόγηση της απόδοσης . . . 111

9.1 Εκτιμώμενοι συντελεστές και τυπικά σφάλματα (σε παρανθέσεις) για το *Model III* . . . 117
9.2 Μελέτη Τραύματος . . . 118
9.3 Εκτιμώμενοι συντελεστές και τυπικά σφάλματα (σε παρανθέσεις) για το *Final* $Model_{SSDs}$ . . . 120

10.1 Μελέτη Τραύματος . . . 128
10.2 Κανόνες συσχέτισης . . . 130
10.3 Ο 1ος βέλτιστος υπερκορεσμένος σχεδιασμός . . . 130
10.4 Ο 2ος βέλτιστος υπερκορεσμένος σχεδιασμός . . . 130
10.5 Συγκριτική απόδοση μοντέλων . . . 131
10.6 Διαστήματα εμπιστοσύνης (95%) . . . 132
10.7 Συγκριτική απόδοση μοντέλων . . . 133
10.8 Διαστήματα εμπιστοσύνης (95%) . . . 133

12.1 Αξιολόγηση της απόδοσης της L_1-norm SVM για ΥΣ κύριων επιδράσεων . 148
12.2 Αξιολόγηση της απόδοσης της L_1-norm Huberized SVM για ΥΣ κύριων επιδράσεων . . . 148
12.3 Αξιολόγηση της απόδοσης της L_1-norm SVM για ΥΣ κύριων επιδράσεων και αλληλεπιδράσεων . . . 152
12.4 Αξιολόγηση της απόδοσης της L_1-norm Huberized SVM για ΥΣ κύριων επιδράσεων και αλληλεπιδράσεων . . . 153
12.5 Ο βέλτιστος υπερκορεσμένος σχεδιασμός . . . 156
12.6 Συγκριτική απόδοση μοντέλων . . . 157
12.7 Οι ενεργές επιδράσεις . . . 157

13.1 Τιμές κριτηρίων για ισορροπημένους $E(s^2)$-βέλτιστους ΥΣ . . . 165
13.2 Παράδειγμα 1: Ο αρχικός ισορροπημένος $\text{SSD}_{(6,10)}$ . . . 169
13.3 Παράδειγμα 1: Ο νέος μη ισορροπημένος $\text{SSD}_{(6,7)}$ . . . 169
13.4 Παράδειγμα 2: Ο αρχικός ισορροπημένος $\text{SSD}_{(8,14)}$ . . . 170
13.5 Παράδειγμα 2: Ο νέος μη ισορροπημένος $\text{SSD}_{(8,35)}$ . . . 170
13.6 Παράδειγμα 3: Ο αρχικός ισορροπημένος $\text{SSD}_{(10,18)}$ . . . 171
13.7 Παράδειγμα 3: Ο νέος μη ισορροπημένος $\text{SSD}_{(10,27)}$ . . . 171
13.8 Συγκριτική απόδοση ισορροπημένων ΥΣ και των αντίστοιχων μη ισορροπημένων ΥΣ . . . 173

14.1 Συγκριτική απόδοση των μη ισορροπημένων Bayesian D-βέλτιστων ΥΣ με τους ισορροπημένους ομόλογούς τους 180
14.2 Κριτήρια βελτιστότητας μη ισορροπημένων Bayesian D-βέλτιστων ΥΣ και των αντίστοιχων ισορροπημένων ΥΣ 181
14.3 Συγκριτική απόδοση των ισορροπημένων και μη ισορροπημένων $\mathrm{SSDs}_{(6,7)}$. 183
14.4 Συγκριτική απόδοση των ισορροπημένων και μη ισορροπημένων $\mathrm{SSDs}_{(10,27)}$ 183
14.5 Συγκριτική απόδοση των ισορροπημένων και μη ισορροπημένων $\mathrm{SSDs}_{(8,35)}$. 183
14.6 Κριτήρια βελτιστότητας μη ισορροπημένων ΥΣ με αλληλεπιδράσεις και των αντίστοιχων ισορροπημένων ΥΣ 184

15.1 Τιμές κριτηρίων για μερικώς ισορροπημένους (PB) και μη ισορροπημένους (U) υπερκορεσμένους σχεδιασμούς 195
15.2 Τιμές κριτηρίων για ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς 195

Μέρος I

Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Μέτρα της Θεωρίας Πληροφοριών

ΚΕΦΑΛΑΙΟ **1**

Εισαγωγικά Στοιχεία και Βασικές Έννοιες

The true logic of this world is
in the calculus of probabilities.

—*James C. Maxwell (1831–1879)*

Στο πρώτο αυτό κεφάλαιο παρουσιάζονται βασικοί ορισμοί, έννοιες, ιδιότητες, διάφορα στατιστικά εργαλεία και μέτρα από τη Θεωρία Πληροροφοριών, τα οποία χρησιμοποιούνται ή ενσωματώνονται στις προτεινόμενες μεθόδους ανάλυσης των υπερκορεσμένων σχεδιασμών που παρουσιάζονται σε επόμενα κεφάλαια. Επιπλέον, παρουσιάζεται μια σύντομη ιστορική αναδρομή σχετικά με το τι έχει προταθεί μέχρι σήμερα στη διεθνή βιβλιογραφία για την κατασκευή και την ανάλυση των υπερκορεσμένων σχεδιασμών. Τέλος, αναφέρονται τα ερευνητικά προβλήματα τα οποία καλούμαστε να αντιμετωπίσουμε στην παρούσα διατριβή σχετικά με την ανάλυση των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων.

1.1 Μέτρα της Θεωρίας Πληροφοριών

Σε αυτή την ενότητα παρουσιάζονται εν συντομία κάποιες βασικές έννοιες και ορισμοί από τη Θεωρία Πληροφοριών. Η Θεωρία Πληροφοριών είναι η επιστήμη η οποία ασχολείται με την έννοια της "πληροφορίας", τη μέτρηση αυτής καθώς και τις ποικίλες εφαρμογές της. Το 1924 ο Nyquist [175] δημοσίευσε μία μέθοδο με την οποία μπορούσαν τα μηνύματα να σταλούν κάνοντας χρήση τηλέγραφου με τη μέγιστη δυνατή ταχύτητα και χωρίς καμία παραμόρφωση. Εντούτοις, στην περιγραφή του αυτή δεν χρησιμοποίησε καθόλου τον όρο πληροφορία. Το 1928 ο Hartley [107] ήταν ο πρώτος που προσπάθησε να ορίσει ένα μέτρο πληροφορίας, με τον παρακάτω τρόπο. Ο Hartley όρισε το ποσό πληροφορίας ως το λογάριθμο του αριθμού των διακεκριμένων μηνυμάτων. Συνεπώς, ένα μήνυμα που αποτελείται από ℓ σύμβολα, περιέχει ℓ φορές περισσότερη πληροφορία από ένα μήνυμα που αποτελείται από ένα και μόνο σύμβολο. Σε αυτή την προσέγγισή του ο Hartley δεν λαμβάνει υπόψη του τις διάφορες πιθανότητες εμφάνισης των συμβόλων ή το γεγονός ότι μπορεί να υπάρχει μία εξάρτηση μεταξύ των ℓ διαδοχικών συμβόλων.

Ο Shannon [208] επέκτεινε τις θεωρίες των Nyquist και Hartley, δημοσιεύοντας το 1948 μία εργασία η οποία θεωρείται η πρώτη ολοκληρωμένη μαθηματική απόπειρα θεμελίωσης της Θεωρίας Πληροφοριών. Με αναφορά στο μέτρο του Hartley, ο Shannon πρότεινε ότι αυτό μπορεί πραγματικά να ερμηνευτεί ως το μέτρο για το ποσό πληροφορίας, με την υπόθεση ότι όλα τα σύμβολα έχουν ίση πιθανότητα εμφάνισης. Για τη γενική περίπτωση, ο Shannon εισήγαγε ένα μέτρο πληροφορίας βασισμένο στην έννοια της πιθανότητας, το οποίο συμπεριλαμβάνει το μέτρο του Hartley σαν μία ειδική περίπτωση. Ο Shannon γενικά θεωρείται ο ιδρυτής της Θεωρίας Πληροφοριών από τη στιγμή που αυτός πραγματικά οδήγησε στη θεμελίωση της Θεωρίας Πληροφοριών όπως είναι σήμερα γνωστή, συνδέοντας την πληροφορία με την αβεβαιότητα χρησιμοποιώντας την έννοια της πιθανότητας. Ο Shannon συνέδεσε την πληροφορία με την έννοια της πιθανότητας ως εξής. Αν θεωρήσουμε ένα δειγματικό χώρο όπου όλα τα ενδεχόμενα έχουν την ίδια πιθανότητα εμφάνισης (ισοπίθανα ενδεχόμενα), υπάρχει μία μεγάλη αβεβαιότητα σχετικά με το ποιό από τα ενδεχόμενα θα εμφανιστεί. Με άλλα λόγια, όταν εμφανιστεί ένα από αυτά τα ενδεχόμενα θα δώσει πολύ περισσότερη πληροφορία απ' ότι στις περιπτώσεις όπου ο δειγματικός χώρος είναι δομημένος με τέτοιον τρόπο ώστε ένα ενδεχόμενο να έχει μεγάλη πιθανότητα εμφάνισης. Συμπερασματικά, η πληροφορία συνδέεται με την έννοια της πιθανότητας μέσω της αβεβαιότητας. Το μέτρο πληροφορίας του Shannon δίνεται από τον ακόλουθο ορισμό.

Ορισμός 1.1. *Έστω ένα πιθανοθεωρητικό πείραμα με δειγματικό χώρο X και κατανομή πιθανότητας P, όπου $p(x_i)$ ή p_i είναι η πιθανότητα του ενδεχομένου $x_i \in X$. Τότε το μέσο ποσό πληροφορίας δίνεται από τη σχέση*

$$H(P) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) = -\sum_{i=1}^{n} p_i \log p_i, \;\; i = 1, 2 \ldots, n. \tag{1.1}$$

Η *H(P)* ικανοποιεί τις εξής βασικές ιδιότητες: είναι συνεχής ως προς p, συμμετρική, προσθετική και μεγιστοποιείται όταν όλες οι πιθανότητες p_i είναι ίσες. Φυσικά, υπάρχουν και άλλες ιδιότητες για την *H(P)*. Για περισσότερες λεπτομέρειες σχετικά με τις ιδιότητες του μέτρου πληροφορίας του Shannon ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στην εργασία [208] καθώς επίσης στα βιβλία [198] και [132].

1.1.1 Μέτρα Εντροπίας

Προηγουμένως αναφερθήκαμε στο μέτρο πληροφορίας του Shannon έτσι όπως αυτό ορίζεται σε ένα γενικό πιθανοθεωρητικό πείραμα. Παρακάτω παραθέτουμε έναν εναλλακτικό ορισμό χρησιμοποιώντας την έννοια της "εντροπίας". Η εντροπία ως έννοια χρησιμοποιήθηκε

αρχικά στη Θερμοδυναμική. Εντροπία κατά την κλασική έννοια (ορισμός κατά Boltzmann (1872), ο οποίος γενικεύτηκε από τον Gibbs το 1878) ορίζεται να είναι το μέτρο του πλήθους των πιθανών μικροκαταστάσεων ενός συστήματος. Με αυτό τον ορισμό η εντροπία ως έννοια διεισδύει όχι μόνο στις διαδικασίες και τη σημασία του φαινομένου της ζωής, αλλά κατ'επέκταση και σε αυτό που ονομάζουμε πληροφορία. Όταν η πληροφορία προκύπτει από ένα στατιστικό σύνολο δεδομένων, τότε μπορούμε να την κατανοήσουμε μέσω της έννοιας της εντροπίας. Η έμπνευση για την υιοθέτηση του όρου εντροπία στη Θεωρία Πληροφοριών προήλθε από τη στενή ομοιότητα μεταξύ του τύπου του Shannon (1.1) και πολύ γνωστών τύπων από τη Θερμοδυναμική. Κατά τον Shannon, η εντροπία της πληροφορίας (information entropy) ή απλώς εντροπία, είναι ένα μέτρο της ποσότητας πληροφορίας που περιέχεται σε ένα μήνυμα ή εναλλακτικά είναι ένα μέτρο της αβεβαιότητας σε ένα σύστημα. Η εντροπία μπορεί να θεωρηθεί επίσης ως ένα μέτρο της αβεβαιότητας που σχετίζεται με μία τυχαία μεταβλητή, καθώς ορίζεται με όρους της κατανομής πιθανότητάς της. Η εντροπία κατά Shannon στη Θεωρία Πληροφοριών ορίζεται ως ακολούθως.

Ορισμός 1.2. *Έστω μία διακριτή τυχαία μεταβλητή X με n εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$, η εντροπία κατά Shannon της μεταβλητής X δίνεται από τη σχέση*

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \tag{1.2}$$

όπου $p(X = x_i) = p(x_i)$ συμβολίζει τη συνάρτηση μάζας πιθανότητας (probability mass function) στο x_i.

Ας σημειωθεί εδώ ότι η επιλογή της βάσης του λογαρίθμου (2 ή 10 ή e) συζητήθηκε εκτενώς από τον ίδιο τον Shannon, είναι αυθαίρετη και αντιστοιχεί στην επιλογή μιας μονάδας μέτρησης πληροφοριών. Αν χρησιμοποιήσουμε το φυσικό λογάριθμο, η μονάδα πληροφορίας ονομάζεται nat (natural unit). Συνήθως στην πράξη το ποσό της πληροφορίας υπολογίζεται χρησιμοποιώντας το 2 σαν βάση. Τότε το ποσό πληροφορίας εκφράζεται σε bits (binary units), και ισχύει ότι 1 nat=1.44 bits. Οποιαδήποτε επιλογή μιας λογαριθμικής βάσης θα μπορούσε να χρησιμοποιηθεί στις αποδείξεις μας με ισοδύναμα αποτελέσματα, εντούτοις στην παρούσα διατριβή, επιλέγεται το 2 σαν βάση του λογαρίθμου, αφού αυτή είναι η πιο δημοφιλής βάση σε εφαρμογές που αφορούν την εντροπία του Shannon [198].

Το 1961, ο Rényi γενίκευσε την εντροπία του Shannon σε μια μονοπαραμετρική οικογένεια εντροπιών, ορίζοντας την εντροπία τάξης α η οποία καλείται ως Rényi εντροπία [189]. Η Rényi εντροπία έχει εφαρμογές στη Θεωρία Κωδίκων, στη Στατιστική, στη Μηχανική και σε άλλα συναφή πεδία. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει για περισσότερες λεπτομέρειες σχετικά με την Rényi εντροπία και τις εφαρμογές της στις εργασίες [189], [14], [117] και [237]. Το 2009, οι Golshani et al. [95] όρισαν τη δεσμευμένη Rényi εντροπία, παρουσιάζοντας έναν ορισμό διαφορετικό από αυτόν του Cachin [37], και ως εκ τούτου αποδείχτηκε και η ισχύς του κανόνα της αλυσίδας (chain rule).

Ορισμός 1.3. *Έστω μία διακριτή τυχαία μεταβλητή X με n εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$, η Rényi εντροπία της μεταβλητής X δίνεται από τη σχέση*

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p(x_i)^\alpha, \quad \alpha > 0, \quad \alpha \neq 1, \tag{1.3}$$

όπου $p(X = x_i) = p(x_i)$ συμβολίζει τη συνάρτηση μάζας πιθανότητας στο x_i. Ας σημειωθεί εδώ ότι η Rényi εντροπία τείνει στην εντροπία κατά Shannon καθώς το $a \to 1$.

Το 1988, ο Tsallis [217] πρότεινε μία γενίκευση της Boltzmann-Gibbs (BG) εντροπίας σε μη εκτατικά (non-extensive) συστήματα, η οποία έχει θεμελιώδη σημασία στη Στατιστική Φυσική. Το 2006, ο Furuichi [85] απέδειξε την ισχύ του κανόνα της αλυσίδας για την Tsallis εντροπία.

Ορισμός 1.4. *Έστω μία διακριτή τυχαία μεταβλητή X με n εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$, η Tsallis εντροπία της μεταβλητής X δίνεται από τη σχέση*

$$S_q(X) = -\sum_{i=1}^{n} p(x_i)^q \ln_q p(x_i), \qquad q \neq 1 \tag{1.4}$$

όπου $p(X = x_i) = p(x_i)$ συμβολίζει τη συνάρτηση μάζας πιθανότητας στο x_i, q είναι η παράμετρος η οποία επεκτείνει την εντροπία Shannon και $\ln_q(x) = \frac{x^{1-q}-1}{1-q}$ για οποιοδήποτε μη αρνητικό πραγματικό αριθμό x και q.

Το 1967, οι Havrda και Charvát [108] εισήγαγαν τη διαρθρωτική α-εντροπία, γνωστή και ως εντροπία τύπου β όπως ορίστηκε το 1970 από τον Daróczy [85].

Ορισμός 1.5. *Έστω μία διακριτή τυχαία μεταβλητή X με n εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$, η Havrda και Charvát εντροπία ή η εντροπία τύπου β της μεταβλητής X δίνεται από τη σχέση*

$$H_\beta(X) = \frac{\sum_{i=1}^{n} (p(x_i)^\beta - p(x_i))}{2^{1-\beta} - 1}, \qquad \beta > 0, \qquad \beta \neq 1, \tag{1.5}$$

όπου $p(X = x_i) = p(x_i)$ συμβολίζει τη συνάρτηση μάζας πιθανότητας στο x_i.

1.1.2 Μέτρα Πληροφορίας

Σε αυτή την ενότητα παρουσιάζονται δύο ευρέως διαδεδομένα μέτρα της Θεωρίας Πληροφοριών, το κέρδος πληροφορίας και η συμμετρική αβεβαιότητα, πάνω στα οποία βασίζονται οι προτεινόμενες μέθοδοι ανάλυσης των υπερκορεσμένων σχεδιασμών. Τα δύο μέτρα αυτά υπολογίζονται χρησιμοποιώντας από κοινού, δεσμευμένα και αμοιβαία μέτρα πληροφορίας καθώς και την εντροπία όπως αυτή παρουσιάστηκε στην παραπάνω ενότητα.

Ορισμός 1.6. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Η από κοινού εντροπία κατά Shannon ορίζεται να είναι*

$$H(X, Y) = -\sum_{i,j} p(x_i, y_j) \log p(x_i, y_j), \tag{1.6}$$

όπου η από κοινού πιθανότητα $p(x_i, y_j)$ συμβολίζει την πιθανότητα του x_i και του y_j.

Ορισμός 1.7. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Το δεσμευμένο μέτρο πληροφορίας ή η δεσμευμένη εντροπία κατά Shannon μπορεί να θεωρηθεί ως ένα μέτρο για την αβεβαιότητα της X δοθείσης της Y και ορίζεται να είναι*

$$H(X|Y) = -\sum_{j=1}^{m} p(y_j) \sum_{i=1}^{n} p(x_i|y_j) \log p(x_i|y_j), \tag{1.7}$$

όπου η δεσμευμένη πιθανότητα $p(x_i|y_j)$ συμβολίζει την πιθανότητα να πραγματοποιηθεί το x_i με δεδομένο ότι έχει πραγματοποιηθεί το y_j.

Ορισμός 1.8. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Η από κοινού Rényi εντροπία ορίζεται να είναι*

$$H_\alpha(X, Y) = \frac{1}{1-\alpha} \log \sum_{i,j} p(x_i, y_j)^\alpha, \qquad \alpha > 0, \qquad \alpha \neq 1, \tag{1.8}$$

όπου η από κοινού πιθανότητα $p(x_i, y_j)$ συμβολίζει την πιθανότητα του x_i και του y_j.

Ορισμός 1.9. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Η δεσμευμένη Rényi εντροπία ορίζεται να είναι*

$$H_\alpha(Y|X) = \frac{1}{1-\alpha} \log \frac{\sum_{i,j} p(x_i, y_j)^\alpha}{\sum_i p(x_i)^\alpha}, \tag{1.9}$$

ή εναλλακτικά

$$H_\alpha(Y|X) = \frac{1}{1-\alpha}[\log \sum_{i,j} p(x_i, y_j)^\alpha - \log \sum_i p(x_i)^\alpha] = H_\alpha(X, Y) - H_\alpha(X), \tag{1.10}$$

όπου $p(x_i, y_j)$ συμβολίζει την από κοινού πιθανότητα του x_i και του y_j.

Ορισμός 1.10. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Η από κοινού Tsallis εντροπία ορίζεται να είναι*

$$S_q(X, Y) = -\sum_{i,j} p(x_i, y_j)^q \ln_q p(x_i, y_j), \quad q \neq 1, \tag{1.11}$$

ή εναλλακτικά

$$S_q(X, Y) = S_q(X) + S_q(Y|X), \tag{1.12}$$

όπου η από κοινού πιθανότητα $p(x_i, y_j)$ συμβολίζει την πιθανότητα του x_i και του y_j.

Ορισμός 1.11. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Η δεσμευμένη Tsallis εντροπία ορίζεται να είναι*

$$S_q(X|Y) = -\sum_{i,j} p(x_i, y_j)^q \ln_q p(x_i|y_j), \quad q \neq 1, \tag{1.13}$$

όπου $p(x_i, y_j)$ συμβολίζει την από κοινού πιθανότητα του x_i και του y_j και η δεσμευμένη πιθανότητα $p(x_i|y_j)$ συμβολίζει την πιθανότητα να πραγματοποιηθεί το x_i με δεδομένο ότι έχει πραγματοποιηθεί το y_j.

Ορισμός 1.12. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Η από κοινού Havrda και Charvát εντροπία ή η από κοινού εντροπία τύπου β ορίζεται να είναι*

$$H_\beta(X, Y) = \frac{\sum_{i,j}(p(x_i, y_j)^\beta - p(x_i, y_j))}{2^{1-\beta} - 1}, \quad \beta > 0, \quad \beta \neq 1, \tag{1.14}$$

ή εναλλακτικά

$$H_\beta(X, Y) = H_\beta(X) + H_\beta(Y|X), \tag{1.15}$$

όπου $p(x_i, y_j)$ συμβολίζει την από κοινού πιθανότητα του x_i και του y_j.

Ορισμός 1.13. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές με n και m εξαγόμενα αποτελέσματα $\{x_i : i = 1, \ldots, n\}$ και $\{y_j : j = 1, \ldots, m\}$, αντίστοιχα. Η δεσμευμένη Havrda και Charvát εντροπία ή η δεσμευμένη εντροπία τύπου β ορίζεται να είναι*

$$H_\beta(X|Y) = \sum_j p(y_j)^\beta H_\beta(X|y_j), \quad \beta > 0, \quad \beta \neq 1. \tag{1.16}$$

Το κέρδος πληροφορίας (information gain) δύο τυχαίων μεταβλητών X και Y (ορισμός κατά Quinlan [187]) μπορεί να ερμηνευθεί ως ένα γενικό μέτρο της εξάρτησης μεταξύ των X και Y, και εκφράζει την ποσότητα της πληροφορίας που αποκαλύπτει η Y για την X. Με άλλα λόγια, το κέρδος πληροφορίας ή αμοιβαίο μέτρο πληροφορίας κατά Shannon (mutual information) είναι η μείωση της αβεβαιότητας της X λόγω της γνώσης της Y.

Ορισμός 1.14. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές. Το αμοιβαίο μέτρο πληροφορίας ή κέρδος πληροφορίας ορίζεται να είναι*

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y). \quad (1.17)$$

Το κέρδος πληροφορίας είναι ένα από τα πιο δημοφιλή μέτρα συσχέτισης που χρησιμοποιούνται σήμερα στον τομέα της εξόρυξης δεδομένων. Ειδικότερα, χρησιμοποιείται ως κριτήριο διάσπασης σε αλγόριθμους δέντρων απόφασης, όπως οι ID-3 και C5.0, και χρησιμοποιείται σε μεθόδους διακριτοποίησης οι οποίες βασίζονται στην εντροπία για την προεπεξεργασία δεδομένων εισόδου [187].

Οι Press et al. [184] πρότειναν ένα εναλλακτικό μέτρο συσχέτισης, το οποίο ονομάζεται συμμετρική αβεβαιότητα (symmetrical uncertainty) το οποίο αντισταθμίζει την προδιάθεση (bias) του κέρδους πληροφορίας υπέρ των μεταβλητών με τις περισσότερες τιμές. Κάτι τέτοιο επιτυγχάνεται με κανονικοποίηση της τιμής του στο εύρος [0,1].

Ορισμός 1.15. *Έστω X και Y είναι διακριτές τυχαίες μεταβλητές. Το μέτρο της συμμετρικής αβεβαιότητας ορίζεται να είναι*

$$SU(X,Y) = 2 \times [\frac{I(X;Y)}{H(X) + H(Y)}], \quad (1.18)$$

όπου $I(X;Y)$ είναι το κέρδος πληροφορίας και $H(X)$, $H(Y)$ η εντροπία των X και Y, αντίστοιχα.

Για περισσότερες λεπτομέρειες σχετικά με τα προαναφερθέντα μέτρα πληροφορίας και τις εφαρμογές τους σε κατηγορικά δεδομένα, ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στις εργασίες [53] και [113].

1.2 Σχεδιασμοί Πειραμάτων: Ορισμοί και Ιδιότητες

Σε αυτή την ενότητα παρατίθενται ο ορισμός του παραγοντικού, κλασματικού παραγοντικού και υπερκορεσμένου σχεδιασμού καθώς και βασικές ιδιότητές τους, όπως αυτά αναφέρονται διεξοδικά στο επιστημονικό άρθρο [22], καθώς επίσης στα βιβλία [124] και [125]. Επιπρόσθετα, παρουσιάζεται η κινητήριος δύναμη που οδηγεί έναν πειραματιστή να χρησιμοποιήσει τους παραγοντικούς, κλασματικούς παραγοντικούς και τους υπερκορεσμένους σχεδιασμούς στη μελέτη του.

1.2.1 Παραγοντικοί Σχεδιασμοί

Ορισμός 1.16. *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.*

Ορισμός 1.17. *Η επίδραση ενός παράγοντα ορίζεται να είναι η αλλαγή που γίνεται στην απόκριση (response) από την αλλαγή στο επίπεδο του παράγοντα. Αυτή η επίδραση συχνά συναντάται με τον όρο κύρια επίδραση (main effect), επειδή αναφέρεται στους παράγοντες που είναι πρωταρχικής σημασίας στο πείραμα.*

Οι παραγοντικοί σχεδιασμοί χρησιμοποιούνται ευρύτατα σε πειράματα που περιλαμβάνουν αρκετούς παράγοντες, όπου είναι αναγκαία η μελέτη της κοινής επίδρασης των παραγόντων στην απόκριση. Γενικά οι παραγοντικοί σχεδιασμοί είναι ιδιαίτερα αποδοτικοί για πειράματα αυτού του τύπου. Ας σημειωθεί εδώ 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 μία τέτοια περίπτωση, συμπεραίνουμε ότι υπάρχει αλληλεπίδραση (interaction) μεταξύ των παραγόντων. Στο τρίτο μέρος της παρούσας διατριβής αναφερόμαστε εκτενώς στην πιθανή ύπαρξη σημαντικών αλληλεπιδράσεων και μελατάμε διεξοδικά το ρόλο που παίζουν στην ανάλυση των σχεδιασμών πειραμάτων. Σε αυτό το πρώτο μέρος της παρούσας διατριβής θεωρούμε μονάχα μοντέλα κύριων επιδράσεων στις προτεινόμενες μεθόδους ανάλυσης, αγνοώντας την πιθανή ύπαρξη αλληλεπιδράσεων.

Υπάρχουν αρκετές ειδικές περιπτώσεις του γενικού παραγοντικού σχεδιασμού που είναι πολύ σημαντικές επειδή είτε χρησιμοποιούνται ευρύτατα σε ερευνητικές εργασίες είτε αποτελούν τη βάση άλλων σχεδιασμών με μεγάλη πρακτική αξία. Η πιο σημαντική από αυτές τις περιπτώσεις, είναι όταν έχουμε k παράγοντες κάθε έναν σε δύο μόνο στάθμες. Οι στάθμες αυτές μπορεί να είναι ποσοτικές ή ποιοτικές ή ακόμα μπορεί να είναι η παρουσία και απουσία ενός παράγοντα.

Ορισμός 1.18. *Μία πλήρης επανάληψη ενός σχεδιασμού με k παράγοντες κάθε έναν σε δύο μόνο στάθμες απαιτεί 2^k παρατηρήσεις και ονομάζεται 2^k παραγοντικός σχεδιασμός.*

Ακόμα και για ένα μέτριο αριθμό παραγόντων, ο συνολικός αριθμός των συνδυασμών αγωγών σε έναν 2^k παραγοντικό σχεδιασμό είναι μεγάλος. Σε μία τέτοια περίπτωση, αφού τα μέσα αντιμετώπισης είναι συνήθως περιορισμένα, ο αριθμός των επαναλήψεων τις οποίες μπορεί να κάνει ο πειραματιστής πρέπει να περιορισθεί. Συχνά, τα διαθέσιμα μέσα επιτρέπουν να γίνει μόνο μία επανάληψη του σχεδιασμού, εκτός αν ο πειραματιστής έχει τη βούληση να παραλείψει μερικούς από τους αρχικούς παράγοντες. Μία επανάληψη ενός 2^k παραγοντικού σχεδιασμού καλείται αρκετές φορές ως ένας μη-επαναλαμβανόμενος (unreplicated) παραγοντικός σχεδιασμός. Ο 2^k παραγοντικός σχεδιασμός είναι ιδιαίτερα χρήσιμος στα αρχικά στάδια της πειραματικής εργασίας όταν υπάρχουν πολλοί παράγοντες που πρέπει να εξεταστούν, από τη στιγμή που είναι αυτός που μας εφοδιάζει με το μικρότερο αριθμό εκτελέσεων (runs) με τις οποίες οι k παράγοντες πρέπει να μελετηθούν σε έναν πλήρη παραγοντικό σχεδιασμό.

Παρακάτω παρουσιάζουμε τα πλεονεκτήματα των παραγοντικών σχεδιασμών, εξηγώντας με αυτό τον τρόπο ποιο είναι το βασικό κίνητρο που έχει ένας πειραματιστής και αποφασίζει να χρησιμοποιήσει τους στατιστικούς σχεδιασμούς πειραμάτων στη μελέτη του. Σχετικά έχουμε το ακόλουθο παράδειγμα.

Παράδειγμα 1.1. *Θεωρούμε αρχικά τον απλούστερο τύπο παραγοντικών σχεδιασμών που είναι αυτοί με δύο μόνο παράγοντες, καθέναν σε δύο επίπεδα. Υποθέτουμε ότι έχουμε δύο παράγοντες A και B, καθέναν σε δύο επίπεδα. Συμβολίζουμε με A_1, A_2 τα δύο επίπεδα που αντιστοιχούν στον παράγοντα A, και με B_1, B_2 τα δύο επίπεδα που αντιστοιχούν στον παράγοντα B. Μεταβάλλοντας τους παράγοντες A και B, έναν τη φορά, προκύπτει πληροφορία για τους δύο παράγοντες, όπως φαίνεται στον Πίνακα 1.1 που ακολουθεί.*

Πίνακας 1.1: *Εκτέλεση πειράματος ενός μόνο παράγοντα*

$A \backslash B$	B_1	B_2
A_1	A_1B_1	A_1B_2
A_2	A_2B_1	

Η επίδραση του παράγοντα A δίνεται από τη διαφορά $A_2B_1 - A_1B_1$. Η επίδραση του παράγοντα B δίνεται από τη διαφορά $A_1B_2 - A_1B_1$. Για να εκτιμήσουμε τις επιδράσεις

των παραγόντων χρησιμοποιώντας τις μέσες αποκρίσεις, αφού υπάρχει και το πειραματικό σφάλμα, είναι σκόπιμο να πάρουμε δύο παρατηρήσεις σε κάθε συνδυασμό αγωγών. Επομένως χρειαζόμαστε 6 παρατηρήσεις.

Αν τώρα αποφασίσουμε να εκτελέσουμε ένα παραγοντικό πείραμα θα πάρουμε έναν επιπρόσθετο συνδυασμό αγωγών, τον A_2B_2, όπως φαίνεται στον Πίνακα 1.2 που ακολουθεί. Τότε χρησιμοποιώντας ακριβώς 4 παρατηρήσεις μπορούν να γίνουν δύο εκτιμήσεις της επίδρασης του A, οι $A_2B_1 - A_1B_1$ και $A_2B_2 - A_1B_2$. Παρόμοια μπορούν να γίνουν δύο εκτιμήσεις της επίδρασης του B, οι $A_1B_2 - A_1B_1$ και $A_2B_2 - A_2B_1$.

Πίνακας 1.2: *Εκτέλεση πειράματος με παραγοντικό σχεδιασμό*

$A \backslash B$	B_1	B_2
A_1	A_1B_1	A_1B_2
A_2	A_2B_1	$\mathbf{A_2B_2}$

Αυτές οι δύο εκτιμήσεις κάθε κύριας επίδρασης μπορούν να σταθμιστούν για να πάρουμε τις μέσες κύριες επιδράσεις, που είναι τόσο ακριβείς, όσο αυτές που πήραμε από το πείραμα ενός μόνο παράγοντα. Διαπιστώνουμε λοιπόν ότι η σχετική αποδοτικότητα (relative efficiency) του παραγοντικού σχεδιασμού στο πείραμα του ενός παράγοντα τη φορά είναι $6/4 = 1.5$ και φυσικά αυτή η σχετική αποδοτικότητα θα αυξάνει όσο αυξάνει και ο αριθμός των παραγόντων ενός πειράματος.

Ανακεφαλαιώνοντας, σημειώνουμε ότι οι παραγοντικοί σχεδιασμοί υπερέχουν των πειραμάτων ενός παράγοντα τη φορά για πολλούς λόγους. Αρχικά, είναι περισσότερο αποδοτικοί, στη συνέχεια επιτρέπουν οι επιδράσεις του παράγοντα να εκτιμώνται σε αρκετά επίπεδα των άλλων παραγόντων, και παρέχουν έγκυρα συμπεράσματα από τη στιγμή που δίνουν τη δυνατότητα στον πειραματιστή να εκτιμήσει τις μέσες κύριες επιδράσεις με την ίδια ακρίβεια συγκριτικά με αυτήν του πειράματος ενός μόνο παράγοντα τη φορά, αλλά χρησιμοποιώντας λιγότερες παρατηρήσεις (αντί για 6 μόνο 4 στο Παράδειγμα 1.1). Το επίτευγμα αυτό είναι ιδιαίτερα σημαντικό αν αναλογιστούμε το πόσο πολύ αυξάνεται ο απαιτούμενος αριθμός παρατηρήσεων, άρα και ο αριθμός των πειραματικών εκτελέσεων όσο αυξάνει ο αριθμός των παραγόντων ενός πειράματος.

Καθώς λοιπόν αυξάνει ο αριθμός των παραγόντων σε έναν 2^k παραγοντικό σχεδιασμό, ο αριθμός των εκτελέσεων που απαιτούνται για μία πλήρη επανάληψη του σχεδιασμού αυξάνει πάρα πολύ γρήγορα. Για παράδειγμα, ας θεωρήσουμε μία πλήρη επανάληψη ενός 2^5 σχεδιασμού η οποία απαιτεί 32 εκτελέσεις. Σε αυτόν το σχεδιασμό μόνο 5 από τους 31 βαθμούς ελευθερίας αντιστοιχούν στις κύριες επιδράσεις και μόνο 10 βαθμοί ελευθερίας αντιστοιχούν στις αλληλεπιδράσεις δύο παραγόντων. Οι υπόλοιποι 16 βαθμοί ελευθερίας αντιστοιχούν στις αλληλεπιδράσεις τριών και περισσοτέρων παραγόντων. Ας θεωρήσουμε τώρα την περίπτωση που ο πειραματιστής μπορεί λογικά να υποθέσει ότι ορισμένες αλληλεπιδράσεις υψηλής τάξης είναι αμελητέες. Τότε μπορεί να δοθεί πληροφορία για τις κύριες επιδράσεις και τις αλληλεπιδράσεις χαμηλής τάξης, από την εκτέλεση ενός μόνο κλάσματος του πλήρους παραγοντικού πειράματος στο οποίο αναφερόμαστε στην επόμενη ενότητα.
Για μια λεπτομερή ανασκόπηση των πειραματικών σχεδιασμών και των διαφόρων εφαρμογών τους, ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στην εργασία [69].

1.2.2 Κλασματικοί Παραγοντικοί Σχεδιασμοί

Σε αυτή την ενότητα παρατίθενται ο ορισμός, βασικές έννοιες και ιδιότητες του κλασματικού παραγοντικού σχεδιασμού, και εξηγείται γιατί οι κλασματικοί παραγοντικοί σχεδιασμοί είναι ιδιαίτερα χρήσιμοι όταν θέλουμε να εξετάσουμε ένα μεγάλο αριθμό παραγόντων αλλά έχουμε στη διάθεσή μας έναν περιορισμένο αριθμό πειραματικών εκτελέσεων.

Ορισμός 1.19. *Ένα μόνο κλάσμα ενός πλήρους παραγοντικού σχεδιασμού ονομάζεται κλασματικός παραγοντικός σχεδιασμός (fractional factorial design). Το ένα-δεύτερο (1/2) κλάσμα ενός 2^k παραγοντικού σχεδιασμού ονομάζεται συχνά και 2^{k-1} παραγοντικός σχεδιασμός. Τα δύο 1/2 κλάσματα αποτελούν έναν πλήρη 2^k παραγοντικό σχεδιασμό.*

Ορισμός 1.20. *Ένας σχεδιασμός είναι αναλυτικής τάξης R, αν καμία επίδραση p παραγόντων δεν είναι ταυτόσημη με άλλη επίδραση που περιέχει λιγότερους από R-p παράγοντες.*

Ορισμός 1.21. *Ορίζουσα σχέση ενός κλασματικού παραγοντικού σχεδιασμού καλείται το σύνολο όλων των στηλών που είναι ίσες με τη μοναδιαία στήλη I, η οποία περιέχει μόνο "+1".*

Ορισμός 1.22. *Η αναλυτική τάξη ενός κλασματικού παραγοντικού σχεδιασμού με δύο στάθμες είναι ίση με το μικρότερο αριθμό γραμμάτων σε οποιαδήποτε λέξη στην ορίζουσα σχέση.*

Συνήθως, είναι επιθυμητό να χρησιμοποιούνται κλασματικοί παραγοντικοί σχεδιασμοί που έχουν την υψηλότερη δυνατή αναλυτική τάξη (resolution), σύμφωνα με το βαθμό της κλασματικότητας που απαιτείται (ή με άλλα λόγια για δεδομένο αριθμό παρατηρήσεων). Η υψηλότερη αναλυτική τάξη βάζει λιγότερους περιορισμούς στις υποθέσεις που απαιτούνται όσον αφορά το ποιες αλληλεπιδράσεις είναι αμελητέες έτσι ώστε να ερμηνευτούν με αξιοπιστία τα δεδομένα. Ένα 1/2 κλάσμα του 2^k σχεδιασμού με την υψηλότερη αναλυτική τάξη μπορεί να κατασκευαστεί γράφοντας ένα βασικό σχεδιασμό που αποτελείται από τις εκτελέσεις για έναν πλήρη 2^{k-1} παραγοντικό σχεδιασμό, και στη συνέχεια προσθέτοντας τον k παράγοντα, αναγνωρίζοντας τις θετικές και αρνητικές στάθμες αυτού, με τα θετικά και αρνητικά πρόσημα της υψηλότερης τάξης αλληλεπίδρασης $ABC\ldots(K-1)$. Ένας άλλος τρόπος να δούμε την κατασκευή ενός 1/2 κλάσματος, είναι να χωρίσουμε τις εκτελέσεις σε δύο μπλοκ με την αλληλεπίδραση υψηλότερης τάξης αναμειγμένη. Τότε κάθε μπλοκ είναι ένας 2^{k-1} κλασματικός παραγοντικός σχεδιασμός με την υψηλότερη αναλυτική τάξη.

Όταν σε ένα 2^k σχεδιασμό δεν είναι εφικτό να εκτελεστούν και οι 2^k συνδυασμοί αγωγών, τότε χρησιμοποιείται το 1/2 κλάσμα του 2^k σχεδιασμού. Οι κλασματικοί παραγοντικοί σχεδιασμοί είναι ευρέως χρησιμοποιούμενοι τύποι σχεδιασμών σε πολλά πρακτικά προβλήματα. Για μια σωστή και επιτυχή χρήση των κλασματικών παραγοντικών σχεδιασμών, ο πειραματιστής μπορεί να βασιστεί σε τρεις ιδέες-κλειδιά οι οποίες περιγράφονται παρακάτω.

1. Η αρχή της σποραδικότητας των επιδράσεων (sparsity of effects principle). Στην περίπτωση ύπαρξης αρκετών παραγόντων, το σύστημα είναι πιθανό να οδηγείται αρχικά από μερικές από τις κύριες επιδράσεις και τις αλληλεπιδράσεις χαμηλής τάξης.

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

3. Ο ακολουθιακός πειραματισμός (sequential experimentation). Είναι δυνατόν να συνδυάσουμε τις εκτελέσεις δύο ή περισσότερων κλασματικών παραγοντικών σχεδιασμών έτσι ώστε να συγκεντρωθεί ακολουθιακά ένας μεγαλύτερος σχεδιασμός και να εκτιμηθούν οι επιδράσεις και οι αλληλεπιδράσεις των παραγόντων που ενδιαφέρουν.

Στο σημείο αυτό θα αναφερθούμε εν συντομία στην προβολική ιδιότητα των κλασματικών παραγοντικών σχεδιασμών. Η προβολική ιδιότητα είναι ιδιαίτερα σημαντική από τη στιγμή που μας επιτρέπει να δούμε ποιες πρόσθετες παρατηρήσεις μπορεί να έχουν αξία, εφόσον είναι διαθέσιμα τα αποτελέσματα του αρχικού κρησαρίσματος. Επιπλέον μας παρέχει γνώση για το πώς να αναθέτουμε τις μεταβλητές του σχεδιασμού στους παράγοντες που θεωρούνται εκ των προτέρων σημαντικοί.

Παρατήρηση 1.1. *Κάθε κλασματικός παραγοντικός σχεδιασμός αναλυτικής τάξης R περιέχει πλήρεις παραγοντικούς σχεδιασμούς σε οποιοδήποτε υποσύνολο R-1 παραγόντων.*

Η παραπάνω παρατήρηση είναι ιδιαίτερα χρήσιμη γιατί αν υποθέσουμε ότι ο πειραματιστής έχει αρκετούς παράγοντες που ενδεχομένως ενδιαφέρουν, αλλά πιστεύει ότι μόνο *R-1* από αυτούς έχουν σημαντικές επιδράσεις, τότε ένας κλασματικός παραγοντικός σχεδιασμός αναλυτικής τάξης R είναι η κατάλληλη επιλογή σχεδιασμού για το πείραμά του. Αν η υπόθεση του πειραματιστή (μόνο *R-1* παράγοντες έχουν σημαντικές επιδράσεις) είναι σωστή, τότε ο κλασματικός παραγοντικός σχεδιασμός αναλυτικής τάξης R θα προβάλλεται σε έναν πλήρη παραγοντικό σχεδιασμό στους *R-1* σημαντικούς παράγοντες.

Στο σημείο αυτό θα αναφερθούμε εν συντομία στη χρησιμότητα του ακολουθιακού πειραματισμού. Η χρήση κλασματικών παραγοντικών σχεδιασμών συχνά οδηγεί σε μεγάλη οικονομία και αποδοτικότητα στον πειραματισμό, ειδικότερα αν οι εκτελέσεις μπορούν να γίνουν ακολουθιακά. Ας υποθέσουμε ότι θέλουμε να ερευνήσουμε k παράγοντες (2^k εκτελέσεις). Είναι σχεδόν πάντα επιθυμητό να εκτελέσουμε έναν 2^{k-1} κλασματικό παραγοντικό σχεδιασμό ($2^k/2$ εκτελέσεις), να αναλύσουμε το αποτέλεσμα και τότε να αποφασίσουμε για το καλύτερο σύνολο των εκτελέσεων που θα κάνουμε στη συνέχεια. Αν είναι αναγκαίο να διαλύσουμε τις αμφιβολίες μπορούμε πάντοτε να εκτελέσουμε το εναλλακτικό κλάσμα και να ολοκληρώσουμε το 2^k σχεδιασμό. Όταν χρησιμοποιείται αυτή η μέθοδος για να ολοκληρώσουμε το σχεδιασμό, τα δύο $1/2$ κλάσματα παριστάνουν μπλοκ του πλήρους σχεδιασμού με την αλληλεπίδραση υψηλότερη τάξης να είναι αναμειγμένη με τα μπλοκ. Συνεπώς, ο ακολουθιακός πειραματισμός έχει σαν αποτέλεσμα την απώλεια πληροφορίας μόνο για την αλληλεπίδραση υψηλότερη τάξης. Εναλλακτικά, σε πολλές περιπτώσεις μαθαίνουμε αρκετά από το $1/2$ κλάσμα για να συνεχίσουμε στο επόμενο στάδιο του πειραματισμού, στο οποίο μπορεί να συμβαίνει πρόσθεση ή διαγραφή παραγόντων, αλλαγή αποκρίσεων ή μεταβολή κάποιων από τους παράγοντες σε νέα πεδία.

Τρία είναι τα πιο ευρέως χρησιμοποιούμενα κριτήρια για την ταξινόμηση ή το διαχωρισμό των κλασματικών παραγοντικών σχεδιασμών: το κριτήριο ελάχιστης απόκλισης (minimum aberration) [83], το κριτήριο βέλτιστων ροπών (optimal moments) [82] και το κριτήριο βαθμών ελευθερίας. Τα κριτήρια αναλυτικής τάξης και ελάχιστης απόκλισης βασίζονται σε δύο ιεραρχικές υποθέσεις που περιγράφονται ως ακολούθως.

1. Επιδράσεις μικρότερης τάξης θεωρούνται σημαντικότερες από επιδράσεις μεγαλύτερης τάξης.

2. Επιδράσεις ίδιας τάξης είναι εξίσου σημαντικές.

1.2.3 Υπερκορεσμένοι Σχεδιασμοί

Σε αυτή την ενότητα παρατίθενται ο ορισμός, βασικές έννοιες και οι ιδιότητες των υπερκορεσμένων σχεδιασμών, καθώς επίσης εξηγείται και η χρησιμότητά τους σε πειράματα κρησαρίσματος.

Με βάση αυτά που αναφέραμε στην προηγούμενη ενότητα οι κλασματικοί παραγοντικοί σχεδιασμοί χρησιμοποιούνται σε πειράματα που περιλαμβάνουν αρκετούς παράγοντες και είναι αναγκαία η μελέτη της κοινής επίδρασης των παραγόντων στην απόκριση με ένα σχετικά οικονομικό αριθμό εκτελέσεων. Με τον όρο κλασματικό παραγοντικό σχεδιασμό εννοούμε ότι σε κάθε πλήρη δοκιμή ή επανάληψη ενός πειράματος περιλαμβάνεται ένα υποσύνολο των δυνατών συνδυασμών των επιπέδων των παραγόντων. Οι υπερκορεσμένοι σχεδιασμοί (supersaturated designs) αποτελούν μία κλάση των κλασματικών παραγοντικών σχεδιασμών, και χρησιμοποιούνται στα αρχικά στάδια ενός επιστημονικού ή βιομηχανικού πειράματος για τον προσδιορισμό των ενεργών παραγόντων.

Οι υπερκορεσμένοι σχεδιασμοί μπορούν να διαχωριστούν σε δύο μεγάλες κλάσεις. Η πρώτη κλάση αποτελείται από τους υπερκορεσμένους σχεδιασμούς με δύο μόνο επίπεδα

(two-level). Η δεύτερη κλάση συνίσταται από τους υπερκορεσμένους σχεδιασμούς πολλών επιπέδων (multi-level), οι οποίοι περιλαμβάνουν παράγοντες που αποτελούνται όλοι από $s > 2$ επίπεδα. Οι υπερκορεσμένοι σχεδιασμοί ανάμεικτων επιπέδων (mixed-level) αποτελούν γενίκευση των υπερκορεσμένων σχεδιασμών πολλών επιπέδων και περιλαμβάνουν παράγοντες με διαφορετικούς αριθμούς επιπέδων μέσα στον ίδιο τον πίνακα του σχεδιασμού. Στην παρούσα διδακτορική διατριβή όταν θα αναφερόμαστε στους υπερκορεσμένους σχεδιασμούς θα εννοούμε αυτούς με δύο μόνο επίπεδα, οι οποίοι μελετήθηκαν για πολλά χρόνια και οι ιδιότητές τους προσέλκυσαν το ενδιαφέρον του μεγαλύτερου μέρους των ερευνητών.

Ορισμός 1.23. *Υπερκορεσμένοι σχεδιασμοί δύο επιπέδων ονομάζονται οι κλασματικοί παραγοντικοί σχεδιασμοί των οποίων ο αριθμός των παραγόντων m είναι μεγαλύτερος ή ίσος του αριθμού των πειραματικών εκτελέσεων, δηλαδή $m \geq n$ ή $m > n - 1$.*

Η ιδέα κλειδί για την κατασκευή και ανάλυση των υπερκορεσμένων σχεδιασμών, βασίζεται στην "αρχή της σποραδικότητας των επιδράσεων" (effect sparsity, Box and Meyer [23]). Υποθέτει ότι ενώ σε πολλές προκαταρκτικές μελέτες σε επιστημονικά και βιομηχανικά πειράματα περιλαμβάνεται ένας μεγάλος αριθμός από ενδεχόμενους κατάλληλους παράγοντες, λίγοι από αυτούς τους παράγοντες επηρεάζουν σημαντικά την απόκριση. Για τη στατιστική ανάλυση των υπερκορεσμένων σχεδιασμών, δεν μπορεί να χρησιμοποιηθεί η μέθοδος των ελαχίστων τετραγώνων γιατί ο αριθμός των παραγόντων είναι μεγαλύτερος από τον αριθμό των παρατηρήσεων. Μία από τις προκλήσεις του πειραματιστή στην ανάλυση των υπερκορεσμένων σχεδιασμών απορρέει από τις συσχετίσεις μεταξύ των στηλών του πίνακα του μοντέλου καθώς και από το γεγονός ότι οι κύριες επιδράσεις όλων των παραγόντων δεν μπορούν να εκτιμηθούν ταυτόχρονα. Θα δούμε σε επόμενη ενότητα τις εισηγήσεις και τους τρόπους που προτείνονται βιβλιογραφικά για την κατασκευή και ανάλυση των υπερκορεσμένων σχεδιασμών δύο επιπέδων.

Στα περισσότερα παραγοντικά πειράματα οι βαθμοί ελευθερίας που απαιτούνται για την εκτίμηση των επιδράσεων των παραγόντων δεν υπερβαίνουν τον αριθμό των εκτελέσεων. Αυτό εγγυάται ότι όλες οι επιδράσεις μπορούν να εκτιμηθούν ταυτόχρονα. Οι δύο επιπέδων σχεδιασμοί που χρησιμοποιούνται για τη μελέτη τέτοιων παραγόντων, όπου το πλήθος m των παραγόντων είναι ίσο με *n-1*, και n είναι το πλήθος των εκτελέσεων, λέγονται κορεσμένοι σχεδιασμοί (saturated designs). Σε κάποιες χαρακτηριστικές καταστάσεις όμως, η συνθήκη αυτή μπορεί να μην ικανοποιείται αν οι πειραματικές εκτελέσεις είναι υπολογιστικά ακριβές και χρονοβόρες και ειδικά αν πρόκειται να ερευνηθεί ένα μεγάλο πλήθος παραγόντων. Αναφέρουμε ένα παράδειγμα στο οποίο ένας κατασκευαστής αυτοκινήτων διεξάγει έναν έλεγχο πρόσκρουσης (crash test) σε μια νέα σειρά αυτοκινήτων. Το αντικείμενο είναι να βρεθεί ένα σύνολο από 54 χαρακτηριστικά ασφαλείας, όπως τροποποιημένοι αερόσακοι, πλευρικές μπάρες κ.α. που περιλαμβάνονται σε ένα σύστημα ασφαλείας ενός αυτοκινήτου. Ένας βολικός σχεδιασμός για αυτόν τον έλεγχο πιθανόν να είναι ένας πίνακας Hadamard τάξης 56, ο οποίος απαιτεί 56 εκτελέσεις που αντιστοιχούν σε 56 πρωτότυπα αυτοκίνητα στο παράδειγμα που θεωρήσαμε. Το εύλογο ερώτημα που τίθεται είναι ποιος τύπος σχεδιασμού θα χρησιμοποιηθεί τελικά από τη στιγμή που ο προϋπολογισμός του κατασκευαστή επιτρέπει το πολύ μισά από τα αυτοκίνητα που απαιτούνται για αυτό το τεστ.

Η κυριότερη χρήση των κλασματικών παραγοντικών σχεδιασμών είναι σε πειράματα κρησαρίσματος (screening experiments). Τα πειράματα κρησαρίσματος είναι πειράματα στα οποία συμμετέχουν πολλοί παράγοντες και έχουν σκοπό την αναγνώριση εκείνων των παραγόντων, αν υπάρχουν, που έχουν μεγάλες επιδράσεις. Τα πειράματα κρησαρίσματος εκτελούνται συνήθως στα αρχικά στάδια μιας έρευνας, όταν είναι πιθανό ότι πολλοί από τους παράγοντες που αρχικά θεωρούνται ως σημαντικοί, να έχουν μικρή ή μηδενική επίδραση στην απόκριση. Στη συνέχεια, οι παράγοντες που αναγνωρίζονται ως σημαντικοί, μελετώνται με περισσότερες λεπτομέρειες σε επόμενα πειράματα. Συνεπώς, επειδή το βασικό αντικείμενο ενός πειράματος κρησαρίσματος είναι να αναγνωριστούν λίγοι μόνο σημαντικοί παράγοντες οι οποίοι θα χρησιμοποιηθούν σε επόμενες μελέτες, οι επιστήμονες και οι πειραματιστές απαιτούν σχεδιασμούς

με το μικρότερο δυνατό αριθμό πειραματικών εκτελέσεων. Χρήσιμοι σχεδιασμοί σε τέτοιες περιπτώσεις είναι οι υπερκορεσμένοι σχεδιασμοί.

1.3 Ιστορική Αναδρομή Υπερκορεσμένων Σχεδιασμών

Σε αυτή την ενότητα παρουσιάζονται οι ποικίλες μέθοδοι κατασκευής και ανάλυσης των υπερκορεσμένων σχεδιασμών οι οποίες έχουν προταθεί μέχρι τώρα στη διεθνή βιβλιογραφία. Μία από τις προκλήσεις του πειραματιστή στο σημαντικό πρόβλημα της κατασκευής και ανάλυσης των υπερκορεσμένων σχεδιασμών απορρέει από τη δομή συσχετίσεων του σχεδιασμού η οποία επιδρά δραστικά στην ικανότητα εκτίμησης των επιδράσεων. Σε αυτή την ενότητα παρουσιάζουμε τις εισηγήσεις και τους τρόπους που προτείνονται βιβλιογραφικά για την κατασκευή και ανάλυση των υπερκορεσμένων σχεδιασμών δύο επιπέδων. Όπως αναμένεται η ιστορία των υπερκορεσμένων σχεδιασμών ξεκίνησε αρκετά χρόνια πριν με τους υπερκορεσμένους σχεδιασμούς δύο επιπέδων. Σε αυτή την ενότητα παραθέτουμε μια πολύ σύντομη ιστορική αναδρομή των μεθόδων κατασκευής των υπερκορεσμένων σχεδιασμών δύο επιπέδων, και αναφερόμαστε με περισσότερες λεπτομέρειες στην προηγούμενη συνεισφορά για την ανάλυση των υπερκορεσμένων σχεδιασμών δύο επιπέδων, από τη στιγμή που ένα από τα βασικά ερευνητικά αντικείμενα της παρούσας διδακτορικής διατριβής είναι η ανάλυση και όχι η κατασκευή των υπερκορεσμένων σχεδιασμών δύο επιπέδων.

1.3.1 Προηγούμενη Συνεισφορά στην Κατασκευή Υπερκορεσμένων Σχεδιασμών

Σε αυτή την ενότητα παρουσιάζεται μία σύντομη ιστορική ανασκόπηση των μεθόδων κατασκευής υπερκορεσμένων σχεδιασμών δύο επιπέδων σε χρονολογική σειρά. Το πρόβλημα κατασκευής των υπερκορεσμένων σχεδιασμών μελετήθηκε πρώτα από τον Satterthwaite [203]. Ο Satterthwaite το 1959 εισήγαγε την ιδέα του υπερκορεσμένου σχεδιασμού ως ενός τυχαίου ισορροπημένου σχεδιασμού. Οι Booth και Cox [21] ήταν οι πρώτοι που εξέτασαν συστηματικά τους υπερκορεσμένους σχεδιασμούς δύο επιπέδων. Οι Booth και Cox το 1962 γενίκευσαν τους σχεδιασμούς με στοιχειώδεις υπολογιστικές ερευνητικές μεθόδους και εξασφάλισαν συστηματικούς σχεδιασμούς, δηλαδή όχι τυχαίους. Επίσης, ανέπτυξαν βασικά κριτήρια βελτιστοποίησης ($ave(s^2)$ και s_{max}) τα οποία χρησιμοποιούνται μέχρι και σήμερα για τη σύγκριση και την κατασκευή υπερκορεσμένων σχεδιασμών. Για τριάντα περίπου χρόνια δεν συναντάμε στη διεθνή βιβλιογραφία εργασία που να αφορά τους σχεδιασμούς αυτούς, μέχρι το 1993, που ο Lin επανέφερε το θέμα και η ανάπτυξή τους από τότε μέχρι σήμερα είναι ραγδαία. Ο Lin [148] βασίστηκε στη δομή των ήδη υπαρχόντων ορθογώνιων σχεδιασμών, και συγκεκριμένα βασισμένος στο κλάσμα των πινάκων Hadamard πρότεινε μία νέα μέθοδο κατασκευής υπερκορεσμένων σχεδιασμών. Τον ίδιο χρόνο ο Wu [229] κατασκεύασε υπερκορεσμένους σχεδιασμούς χρησιμοποιώντας τις αλληλεπιδράσεις δύο παραγόντων των πινάκων Hadamard. Από τότε μέχρι και σήμερα αρκετοί ερευνητές έχουν ασχοληθεί με την κατασκευή και την εξέταση των ιδιοτήτων των υπερκορεσμένων σχεδιασμών.

Ποικίλες μέθοδοι για την κατασκευή των υπερκορεσμένων σχεδιασμών δύο επιπέδων έχουν προταθεί βιβλιογραφικά, όπως αυτές περιγράφονται στις επιστημονικές εργασίες [149], [172], [66], [47], [143], [215], [231], [151], [157], [34], [48], [4], [74], [30], [153], [35], [87], [31], [154], [202], [129], [99], [173], [32], [88], [126], [130], [131], [128], [90], [36], [119], [100] και [174].

1.3.2 Προηγούμενη Συνεισφορά στην Ανάλυση Υπερκορεσμένων Σχεδιασμών

Σε αυτή την ενότητα παρουσιάζονται αναλυτικά οι μέθοδοι ανάλυσης υπερκορεσμένων σχεδιασμών δύο επιπέδων που πρότειναν κατά καιρούς διάφοροι ερευνητές. Η ανάλυση των

υπερκορεσμένων σχεδιασμών παραμένει πρόκληση και ένα ιδιαίτερα δύσκολο πρόβλημα για τους ερευνητές δεδομένου ότι δεν υπάρχει βέλτιστη λύση και οι διάφορες προτεινόμενες μέθοδοι προσπαθούν να πάρουν το καλύτερο δυνατό αποτέλεσμα από τις λίγες πειραματικές εκτελέσεις που είναι διαθέσιμες στον πίνακα σχεδιασμού. Οι περισσότερες μέθοδοι που έχουν προταθεί στη βιβλιογραφία αποτελούν γενικεύσεις, επεκτάσεις και τροποποιήσεις γνωστών μεθόδων που αναπτύχθηκαν αρχικά για την ανάλυση των δεδομένων από κορεσμένους (saturated) ή μη κορεσμένους (unsaturated) σχεδιασμούς. Αυτές οι μέθοδοι μπορούν να ομαδοποιηθούν με διάφορους τρόπους. Για παράδειγμα, μπορεί κάποιος να τις ομαδοποιήσει σε μπεϋζιανές μεθόδους και μη μπεϋζιανές με την κλασική έννοια της θεωρίας αποφάσεων (frequentist). Ένας άλλος τρόπος για να ομαδοποιήσει κάποιος τις μεθόδους αυτές είναι σε μεθόδους ελαχίστων τετραγώνων, μεθόδους εκτίμησης μεροληψίας (bias) και μπεϋζιανές μεθόδους. Σε αυτή την ενότητα έχουμε επιλέξει να παρουσιάσουμε την ανασκόπηση των μεθόδων ανάλυσης υπερκορεσμένων σχεδιασμών δύο επιπέδων σε χρονολογική σειρά.

Η πρώτη μέθοδος σε αυτή την κατεύθυνση παρουσιάστηκε από τον Satterthwaite [203], ο οποίος πρότεινε μία γραφική μέθοδο ανάλυσης υπερκορεσμένων σχεδιασμών, ισοδύναμη με τη μέθοδο εκτίμησης ελαχίστων τετραγώνων για την προσαρμογή ενός απλού γραμμικού μοντέλου παλινδρόμησης. Στα αρχικά στάδια της δημιουργίας των υπερκορεσμένων σχεδιασμών, για την ανάλυσή τους προτάθηκαν διάφορες γραφικές μέθοδοι οι οποίες συντελούν στην αναγνώριση των σημαντικών παραγόντων. Ένα τέτοιο γράφημα που βοηθά στον εντοπισμό των ενεργών παραγόντων είναι η γραφική παράσταση της απόκρισης y με τη στάθμη του κάθε παράγοντα. Εντούτοις, αμφισβητείται το κατά πόσο η χρήση γραφικών μεθόδων προσφέρει χρήσιμες και αξιόπιστες πληροφορίες για τη σημαντικότητα των παραγόντων. Οι γραφικές μέθοδοι προσφέρουν ένα σύντομο οπτικό έλεγχο των ενεργών παραγόντων. Αν θέλουμε να έχουμε πιο ικανοποιητικά αποτελέσματα, τότε μπορούμε να συνδυάσουμε διάφορες γραφικές μεθόδους. Ο Srivastava [212] έδειξε ότι οποιοδήποτε σύνολο p ενεργών επιδράσεων μπορεί να εκτιμηθεί εάν όλα τα υποσύνολα των $2p$ μεταβλητών, μέσα στον πίνακα του μοντέλου, περιέχουν ανεξάρτητες στήλες. Ο Lin [148] εφάρμοσε την κατά βήματα μέθοδο επιλογής μεταβλητών σε 1/2 κλάσματα των Plackett-Burman σχεδιασμών. Ας σημειωθεί εδώ ότι η κατά βήματα επιλογή είναι μια διαδικασία που συνδυάζει την προς τα εμπρός επιλογή, με συνεχή έλεγχο των παραμέτρων που ήδη υπάρχουν στο μοντέλο. Επιπρόσθετα, όταν το μέγεθος των επιδράσεων ακολουθεί κατανομή Pareto (αυτό συμβαίνει όταν μόνο το 10-20% των παραγόντων είναι ενεργοί) τότε οι μέθοδοι ανάλυσης με χρήση του κανονικού διαγράμματος πιθανότητας και η κατά βήματα παλινδρόμηση είναι σε θέση να αναγνωρίσουν τους ενεργούς παράγοντες. Οι περισσότεροι ερευνητές που μελέτησαν το θέμα, πιστεύουν ότι όντως οι ενεργοί παράγοντες είναι στην πραγματικότητα ένα ποσοστό της τάξης του 10-20% (υπόθεση σποραδικότητας), οπότε μπορούν να χρησιμοποιηθούν και οι προαναφερθείσες μέθοδοι.

Ο Wu [229] εφάρμοσε την προς τα εμπρός μέθοδο επιλογής μεταβλητών και τη μέθοδο όλων των υποσυνόλων με στόχο να αναγνωρίσει τις κύριες επιδράσεις. Ο Lin [149] χρησιμοποίησε το κανονικό διάγραμμα πιθανότητας και κάποια τροποποίηση της μεθόδου ελαχίστων τετραγώνων, την παλινδρόμηση κορυφογραμμής (ridge) για την αναγνώριση των ενεργών επιδράσεων. Οι Chipman et al. [49] πρότειναν μία μερικώς μπεϋζιανή μέθοδο επιλογής μεταβλητών για την ανάλυση υπερκορεσμένων σχεδιασμών με πολύπλοκη δομή ταυτόσημων επιδράσεων. Οι Chen και Lin [46] διερεύνησαν την ικανότητα αναγνώρισης ενεργών επιδράσεων στους υπερκορεσμένους σχεδιασμούς κάνοντας υποθέσεις κανονικότητας. Οι Westfall et al. [227] βασισμένοι στην αρχή της σποραδικότητας των επιδράσεων πρότειναν μία τεχνική για τον έλεγχο του σφάλματος Τύπου I στη διαδικασία της προς τα εμπρός επιλογής μεταβλητών. Οι Abraham et al. [1] έδειξαν ότι εξαιτίας της δομής συσχετίσεων των υπερκορεσμένων σχεδιασμών μπορεί να προκύψουν και μη ασφαλή αποτελέσματα στη διευκρίνιση των επιδράσεων ανεξάρτητα από το ποια μέθοδος θα χρησιμοποιηθεί για την ανάλυση των υπερκορεσμένων σχεδιασμών.

Οι Beattie et al. [11] πρότειναν μία μπεϋζιανή τεχνική επιλογής μεταβλητών με δύο στάδια για την ανάλυση των υπερκορεσμένων σχεδιασμών. Οι Li και Lin [144] εισήγαγαν μία νέα μέθοδο επιλογής μεταβλητών για την αναγνώριση των ενεργών επιδράσεων στους υπερκορεσμένους σχεδιασμούς, βασισμένη στα μη κυρτά ποινικοποιημένα ελάχιστα τετράγωνα. Οι Li και Lin [145] σύγκριναν αυτήν τη νέα μέθοδο με την κατά βήματα μέθοδο και τη μπεϋζιανή μέθοδο επιλογής μεταβλητών των Beattie et al. [11]. Οι Holcomb et al. [111] πρότειναν τη μέθοδο μεταβλητότητας των αντιθέσεων και την τεχνική bootstrap για την ανάλυση των υπερκορεσμένων σχεδιασμών. Οι Lu και Wu [158] πρότειναν μία τροποποιημένη κατά βήματα μέθοδο επιλογής μεταβλητών η οποία βασίζεται στη σταδιακή μείωση των διαστάσεων του προβλήματος. Ο Yamata [232] χρησιμοποίησε την κατά βήματα παλινδρόμηση για την αναγνώριση των ενεργών επιδράσεων, αξιολογώντας την πιθανότητα επιλογής ενεργών παραγόντων μέσω μελέτης προσομοιώσεων και εξετάζοντας με ιδιαίτερη έμφαση τα σφάλματα Τύπου ΙΙ που προέκυψαν. Οι Koukouvinos και Stylianou [137] πρότειναν μία τροποποιημένη μέθοδο μεταβλητότητας των αντιθέσεων για την ανάλυση των υπερκορεσμένων σχεδιασμών. Οι Zhang et al. [235] εφάρμοσαν τη μερικώς μέθοδο επιλογής ελαχίστων τετραγώνων για την ανάλυση των υπερκορεσμένων σχεδιασμών. Οι Liu et al. [154] μελέτησαν τη σχέση που υπάρχει μεταξύ της μέγιστης επιτρεπτής συσχέτισης και του μεγέθους των συντελεστών των κύριων επιδράσεων των ενεργών παραγόντων. Οι Liu et al. [155] μελέτησαν τις δυσκολίες που αντιμετωπίζει ο πειραματιστής για την ανίχνευση των ενεργών παραγόντων στους υπερκορεσμένους σχεδιασμούς. Οι Holcomb et al. [112] σχεδίασαν ένα υπερκορεσμένο πείραμα χρησιμοποιώντας δεδομένα τα οποία προέκυψαν από μία κινητήριο μηχανή, και εφάρμοσαν τη μέθοδο που προτάθηκε στην εργασία [112] για να αναλύσουν τα δεδομένα αυτά και να αναδείξουν τη χρησιμότητα των υπερκορεσμένων σχεδιασμών στην πράξη.

Ο Georgiou [89] πρότεινε μία μέθοδο η οποία συνδυάζει την ανάλυση ιδιάζουσων τιμών (singular value decomposition) με την ανάλυση κύριων συνιστωσών και την ανάλυση παλινδρόμησης για να αναγνωρίσει τις ενεργές επιδράσεις στους υπερκορεσμένους σχεδιασμούς. Ο Cossari [52] εφάρμοσε την Box-Meyer μέθοδο για να αναγνωρίσει τους ενεργούς παράγοντες στους υπεκορεσμένους σχεδιασμούς. Οι Phoa et al. [182] πρότειναν μία νέα μέθοδο επιλογής μεταβλητών χησιμοποιώντας τη μέθοδο Dantzig selector [39]. Οι Marley και Woods [160] πρότειναν μία νέα επαναληπτική μέθοδο (model averaging) για την ανάλυση των υπερκορεσμένων σχεδιασμών. Οι Li et al. [147] πρότειναν μία νέα μέθοδο ανάλυσης συστάδων για την επιλογή μεταβλητών και την αναγνώριση των ενεργών επιδράσεων στους υπερκορεσμένους σχεδιασμούς. Οι Edwards και Mee [72] πρότειναν μία μέθοδο ανάλυσης των υπερκορεσμένων σχεδιασμών βασισμένη σε τεστ τυχαιοποίησης τα οποία εφαρμόζονται με στόχο τη μείωση του αριθμού των παραγόντων στα υποψήφια μοντέλα με μικρό καθολικό p-value.

Ο ενδιαφερόμενος αναγνώστης μπορεί επίσης να ανατρέξει εκτός από τα πρωτότυπα κείμενα, σε πρόσφατες χρήσιμες ανασκοπήσεις για την κατασκευή και την ανάλυση των υπερκορεσμένων σχεδιασμών, όπως είναι οι εργασίες [163], [92], [98] [123] και [91].

1.4 Ερευνητικά Προβλήματα στην Ανάλυση Υπερκορεσμένων Σχεδιασμών

Σε αυτή την ενότητα αναφέρονται τα δύο βασικά ερευνητικά προβλήματα που καλείται να αντιμετωπίσει ο πειραματιστής/ερευνητής στις μέρες μας όταν το θέμα το οποίο πραγματεύεται είναι η στατιστική ανάλυση των υπερκορεσμένων σχεδιασμών.

1.4.1 Ύπαρξη Μεροληψίας

Οι υπερκορεσμένοι σχεδιασμοί είναι πολύ σημαντικοί και χρήσιμοι, γι΄ αυτό και η ανάπτυξή τους όπως είδαμε είναι ραγδαία. Για την κατασκευή τους απαιτούνται πολλά μαθηματικά και στατιστικά εργαλεία. Για το δύσκολο πρόβλημα κατασκευής των υπερκορεσμένων

σχεδιασμών είδαμε στην ενότητα 1.3.1 ότι υπάρχει εκτενής βιβλιογραφία. Η ανάλυση των υπερκορεσμένων σχεδιασμών βρίσκεται ακόμα στα αρχικά στάδια της ανάπτυξής της, δεδομένου ότι δεν υπάρχει καθολικά βέλτιστη μέθοδος και για αυτόν το λόγο έχουν αναπτυχθεί όπως είδαμε στην προηγούμενη ενότητα 1.3.2 πολλές διαφορετικές μέθοδοι για το σκοπό αυτό.

Μέχρι στιγμής δεν έχει βρεθεί κάποια μέθοδος η οποία όταν χρησιμοποιείται να αναγνωρίζονται οι σημαντικοί παράγοντες και να εκτιμούνται με αμελητέο σφάλμα. Το πρόβλημα που υπάρχει στην ανάλυση των υπερκορεσμένων σχεδιασμών είναι η μη ορθογωνιότητα του πίνακα σχεδιασμού, καθιστώντας αναπόφευκτο το να γίνεται ανεκτή η ύπαρξη μιας μικρής μεροληψίας ανάμεσα σε όλες τις εκτιμημένες επιδράσεις, εφόσον οι ενεργοί παράγοντες μπορούν ακόμη να αναγνωριστούν. Ένας ενεργός παράγοντας μπορεί να αναγνωριστεί αν η επίδρασή του είναι αρκετά μεγάλη ώστε να μην επισκιάζεται από το πειραματικό σφάλμα και τη συνδυασμένη επίδραση μη σημαντικών παραγόντων. Όπως γνωρίζουμε, οι πειραματικές μέθοδοι περιέχουν κάποιο σφάλμα το οποίο διαφοροποιεί τις μετρήσεις από τα πραγματικά αποτελέσματα. Το πρόβλημα είναι να αναγνωριστούν οι σημαντικοί παράγοντες και όχι οι παράγοντες που έχουν "φουσκωμένη" επίδραση λόγω του σφάλματος. Με λίγα λόγια η οποιαδήποτε μέθοδος που θα επιλεγεί για την ανάλυση των υπερκορεσμένων σχεδιασμών θα πρέπει να λαμβάνει υπόψη της το πειραματικό σφάλμα και να προσπαθεί να διατηρεί τις τιμές του στο κατώτατο επιτρεπτό επίπεδο.

Συμπερασματικά, ο πρώτος βασικός μας στόχος είναι η ανάπτυξη αποτελεσματικών μεθόδων ανάλυσης των υπερκορεσμένων σχεδιασμών οι οποίες θα επιτυγχάνουν να αναγνωρίσουν τους ενεργούς παράγοντες με το μικρότερο δυνατό ποσοστό σφάλματος Τύπου I και Τύπου II. Το ποσοστό σφάλματος Τύπου I καλείται ο μέσος λόγος των μη ενεργών παραγόντων που δηλώθηκαν εσφαλμένα ως ενεργοί. Το ποσοστό σφάλματος Τύπου II καλείται ο μέσος λόγος των ενεργών παραγόντων που δηλώθηκαν εσφαλμένα ως μη ενεργοί.

1.4.2 Επέκταση από το Γραμμικό στο Γενικευμένο Γραμμικό Μοντέλο

Η χρήση γενικά των παραγοντικών σχεδιασμών κυρίως στηρίζεται στο γενικό γραμμικό μοντέλο, με τις γνωστές υποθέσεις περί κανονικής κατανομής των δεδομένων, ασυσχέτιστων σφαλμάτων και ομοιογένειας της διασποράς. Οι γραμμικές αυτές υποθέσεις δεν ικανοποιούνται σε πολλές εφαρμογές, όπως για παράδειγμα σε κλινικές ή επιδημιολογικές μελέτες ή σε περιπτώσεις που έχουμε δυαδικά δεδομένα ή δεδομένα συχνοτήτων. Σε αυτές τις περιπτώσεις συνίσταται η χρήση των γενικευμένων γραμμικών μοντέλων. Κατά την προσαρμογή ενός γενικευμένου γραμμικού μοντέλου, το σημαντικότερο στάδιο είναι η επιλογή του κατάλληλου σχεδιασμού. Μία πραγματική δυσκολία που δυσχεραίνει την κατασκευή ενός σχεδιασμού για ένα γενικευμένο γραμμικό μοντέλο είναι η εξάρτηση του πίνακα πληροφορίας από τις άγνωστες παραμέτρους του προσαρμοσμένου μοντέλου. Μία από τις πιο διαδεδομένες τεχνικές που έχουν προταθεί στη διεθνή βιβλιογραφία τα τελευταία 30 χρόνια για να επιλυθεί το παραπάνω πρόβλημα, είναι ο καθορισμός αρχικών (υποθετικών) τιμών των παραμέτρων και η επακόλουθη κατασκευή των λεγόμενων τοπικά βέλτιστων σχεδιασμών με χρήση οποιουδήποτε από τα γνωστά κριτήρια A-, D-, E-, G- βελτιστοποίησης [230] τα οποία και βασίζονται στον πίνακα πληροφορίας. Η βιβλιογραφία που σχετίζεται με τις διάφορες μεθόδους κατασκευής και επιλογής βέλτιστου σχεδιασμού θεωρώντας το γενικευμένο γραμμικό μοντέλο είναι εκτενής. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει για περισσότερες πληροφορίες στην πολύ χρήσιμη πρόσφατη ανασκόπηση [121] καθώς και στις αναφορές που εσωκλείει.

Για τη σύγκριση, την αξιολόγηση και την κατασκευή βέλτιστων υπερκορεσμένων σχεδιασμών αναπτύχθηκαν κατά καιρούς πολλά κριτήρια θεωρώντας πάντα το γενικό γραμμικό μοντέλο. Τα κριτήρια αυτά αναδεικνύουν κάποιο τρόπο μέτρησης για το κατά πόσο κοντά είναι ο σχεδιασμός στον ορθογώνιο ή το κατά πόσο κοντά είναι η πρόβλεψη της διασποράς ή αναφέρονται στην προβολικότητα του σχεδιασμού σε διάσταση χαμηλότερης τάξης. Όταν

λέμε προβολή του σχεδιασμού εννοούμε το σχεδιασμό που προκύπτει αν αφαιρέσουμε από τον αρχικό τις στήλες που αντιστοιχούν στους μη σημαντικούς παράγοντες. Σε τέτοιου τύπου κριτήρια αναφερόμαστε στο τρίτο μέρος της παρούσας διδακτορικής διατριβής. Όλες αυτές οι μέθοδοι κατασκευής και βελτιστοποίησης των υπερκορεσμένων σχεδιασμών προσπαθούν να μειώσουν τη μη ορθογωνιότητα του πίνακα σχεδιασμού X, ενώ παράλληλα η ορίζουσα του πίνακα πληροφορίας X^TX πρέπει να είναι διάφορη του μηδενός και στηρίζονται στο γενικό γραμμικό μοντέλο.

Επιπρόσθετα, όλες οι μέθοδοι ανάλυσης των υπερκορεσμένων σχεδιασμών, στις οποίες αναφερθήκαμε εκτενώς στην ενότητα 1.3.2, βασίζονται στο γενικό γραμμικό μοντέλο. Μέχρι τώρα βιβλιογραφικά δεν συναντάμε κάποια ερευνητική εργασία που να μελετά το πρόβλημα της ανάλυσης των υπερκορεσμένων σχεδιασμών θεωρώντας το γενικευμένο γραμμικό μοντέλο. Συμπερασματικά, ο δέυτερος βασικός μας στόχος είναι η ανάπτυξη αποτελεσματικών μεθόδων ανάλυσης των υπερκορεσμένων σχεδιασμών κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων.

ΚΕΦΑΛΑΙΟ **2**

Ανάλυση Υπερκορεσμένων Σχεδιασμών με Μέτρα Εντροπίας

Information: the negative
reciprocal value of probability.

—*Claude E. Shannon (1916–2001)*

Στο δεύτερο αυτό κεφάλαιο ασχολούμαστε με τη στατιστική ανάλυση μιας συγκεκριμένης κλάσης των παραγοντικών σχεδιασμών, τους υπερκορεσμένους σχεδιασμούς. Συγκεκριμένα παρουσιάζεται μία νέα μεθοδολογία επιλογής μεταβλητών σε έναν υπερκορεσμένο σχεδιασμό, όπου οι επεξηγηματικές μεταβλητές βρίσκονται σε δύο επίπεδα, με σκοπό τη γρήγορη και αποτελεσματική ανεύρεση των ενεργών κύριων επιδράσεων. Η έρευνα βασίζεται στην ανάπτυξη μιας μεθόδου ανάλυσης υπερκορεσμένων σχεδιασμών ώστε να αναγνωρίζονται οι σημαντικές κύριες επιδράσεις, έχοντας ως υπόβαθρο τη Θεωρία Πληροφοριών, καθώς εξετάζονται τα εξής μέτρα εντροπίας: Rényi εντροπία, Tsallis εντροπία και Havrda-Charvát εντροπία. Τα μέτρα αυτά σε συνδυασμό με το κέρδος πληροφορίας χρησιμοποιούνται για την ανάλυση των υπερκορεσμένων σχεδιασμών και την εύρεση των σημαντικών παραγόντων χρησιμοποιώντας γενικευμένα γραμμικά μοντέλα, και συγκεκριμένα το λογιστικό μοντέλο παλινδρόμησης στο οποίο η μεταβλητή απόκρισης είναι δίτιμη. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [127].

2.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Η έννοια της εντροπίας, όπως προτάθηκε από τον Shannon κατέχει κεντρικό ρόλο στη Θεωρία Πληροφοριών και χρησιμοποιείται ευρέως σε πολλές εφαρμογές, όπως για παράδειγμα στη Θεωρία Κωδίκων, σε μεθόδους διακριτοποίησης, σε μεθόδους κατασκευής δέντρων αποφάσεων και εφαρμόζεται στο αρχικό στάδιο προεπεξεργασίας των δεδομένων σε αναλύσεις εξόρυξης πληροφορίας από δεδομένα (data mining). Εκτός από την εντροπία του Shannon, στη βιβλιογραφία συναντάμε και άλλα μέτρα εντροπίας, όπως την Rényi εντροπία, Tsallis εντροπία και Havrda-Charvát εντροπία. Η ευρεία χρήση αυτών των μέτρων σε πολλά επιστημονικά πεδία, σε συνδυασμό με τη χρήση τους σε μεθόδους επιλογής μεταβλητών που εφαρμόζονται στην εξόρυξη δεδομένων [187], αποτέλεσε το κίνητρο για την περαιτέρω μελέτη αυτών των μέτρων και την ενσωμάτωσή τους σε μεθόδους ανάλυσης των υπερκορεσμένων σχεδιασμών.

Η προτεινόμενη μέθοδος που παρουσιάζεται διεξοδικά στην ενότητα που ακολουθεί, προσεγγίζει το πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών μέσω ποικίλων μέτρων εντροπίας, τα οποία συνδυάζονται με άλλα, ήδη υπάρχοντα μέτρα πληροφορίας, ώστε να χρησιμοποιηθούν στη συνέχεια για την εύρεση των ενεργών σημαντικών παραγόντων. Αυτή η προσέγγιση του προβλήματος ανάλυσης των υπερκορεσμένων σχεδιασμών είναι καινοτομική, καθώς επιτυγχάνεται ο συνδυασμός τριών διαφορετικών επιστημονικών πεδίων, της Θεωρίας Πληροφοριών, της εξόρυξης δεδομένων και των πειραματικών σχεδιασμών. Ας σημειωθεί εδώ ότι η πιο σημαντικη συνεισφορά της προτεινόμενης μεθόδου είναι ότι θεωρούμε τα γενικευμένα γραμμικά μοντέλα για την ανάλυση των υπερκορεσμένων σχεδιασμών, μέθοδος η οποία είναι μη τετριμμένη και μη αναπτυγμένη στην έως σήμερα διεθνή βιβλιογραφία.

2.2 Η Προτεινόμενη Μέθοδος Ανάλυσης

Στην ενότητα αυτή παρουσιάζεται η πρώτη προτεινόμενη μέθοδος ανάλυσης υπερκορεσμένων σχεδιασμών, η οποία έχει ως στόχο την αναγνώριση και επιλογή των σημαντικών ενεργών επιδράσεων, προσεγγίζοντας το πρόβλημα μέσω ποικίλων μέτρων της Θεωρίας Πληροφοριών. Η Rényi, Tsallis και Havrda-Charvát εντροπία σε συνδυασμό με το κέρδος πληροφορίας χρησιμοποιούνται προκειμένου να εντοπιστούν οι σημαντικοί παράγοντες μέσα από ένα μεγάλο σύνολο ενδεχόμενων ενεργών παραγόντων. Ένα αξιοσημείωτο πλεονέκτημα της προτεινόμενης μεθόδου είναι η χρήση των γενικευμένων γραμμικών μοντέλων, και συγκεκριμένα του λογιστικού μοντέλου παλινδρόμησης για την ανάλυση των δεδομένων από τους υπερκορεσμένους σχεδιασμούς. Η συμπεριφορά και η αποδοτικότητα της προτεινόμενης μεθόδου εξετάστηκαν και αξιολογήθηκαν μέσα από μία εμπειρική μελέτη η οποία περιελάμβανε εκτενή πειράματα προσομοίωσης των οποίων τα ενθαρρυντικά αποτελέσματα που προέκυψαν παρουσιάζονται σε επόμενη ενότητα.

2.2.1 Περιγραφή της Προτεινόμενης Μεθόδου

Η προτεινόμενη μέθοδος εκτιμάει το κέρδος πληροφορίας, χρησιμοποιώντας κάθε μία από τις τρεις προαναφερθείσες εντροπίες (Rényi, Tsallis και Havrda-Charvát). Για την Rényi εντροπία, θεωρήσαμε την παράμετρο α ίση με το 2, για την Tsallis εντροπία θεωρήσαμε την παράμετρο q πάλι ίση με το 2, ενώ για την Havrda-Charvát εντροπία θεωρήσαμε την παράμετρο β ίση με το 3. Ας σημειωθεί εδώ ότι όταν το $q = 2$, η Tsallis εντροπία ισοδυναμεί με το συντελεστή Gini (ή δείκτη Gini) [93]. Σύμφωνα με τη βιβλιογραφία που υπάρχει σχετικά με τα ανωτέρω μέτρα εντροπίας, οι επιλεγμένες τιμές παραμέτρων φαίνεται να αποτελούν την κλασική και ιδανική επιλογή, και αυτό αποτέλεσε το κριτήριο για μας ώστε να επιλέξουμε τις συγκεκριμένες τιμές παραμέτρων.

Η πρώτη προτεινόμενη μέθοδος ανάλυσης υπερκορεσμένων σχεδιασμών περιγράφεται αναλυτικά παρακάτω και εκτελεί τα ακόλουθα βήματα:

Βήμα 1. Δοθέντος ενός $n \times m$ υπερκορεσμένου πίνακα σχεδιασμού $X = [x_1, x_2, \ldots, x_m]$, όπου x_j, $j = 1, 2, \ldots, m$, αντιστοιχεί στην j-οστή στήλη του πίνακα, και ενός $n \times 1$ διανύσματος απόκρισης Y, υπολογίζουμε κάθε φορά την εντροπία Rényi, Tsallis και Havrda-Charvát. Επιπλέον, λαμβάνοντας υπόψη τις προαναφερθείσες μορφές στην ενότητα 1.1.2, υπολογίζουμε τη δεσμευμένη εντροπία σε σχέση με τη μεταβλητή απόκρισης, για κάθε περίπτωση αντίστοιχα.

Βήμα 2. Χρησιμοποιώντας κάθε μία από τις ανωτέρω εντροπίες, υπολογίζουμε το κέρδος πληροφορίας για κάθε μέτρο εντροπίας αντίστοιχα. Ως εκ τούτου, θα προκύψουν τρία διανύσματα με τις αντίστοιχες τιμές του κέρδους πληροφορίας **ig**=$(ig_1, ig_2, \ldots, ig_m)$ όπου ig_j, για $j = 1, 2, \ldots, m$, είναι η τιμή του κέρδους πληροφορίας (IG) που αντιστοιχεί στην j-οστή μεταβλητή. Σε κάθε μία από τις τρεις περιπτώσεις, ταξινομούμε τους παράγοντες σύμφωνα με τις τιμές του κέρδους πληροφορίας που προέκυψαν. Με άλλα λόγια ταξινομούμε σε φθίνουσα σειρά το διάνυσμα κέρδους πληροφορίας **ig**.

Βήμα 3. Κρατάμε τους παράγοντες που βρέθηκαν να έχουν τις w υψηλότερες τιμές κέρδους πληροφορίας. Η τιμή κατωφλιού w, η οποία καθορίζει το μέγιστο αριθμό σημαντικών παραγόντων που μπορούν να προσδιοριστούν έχει οριστεί να είναι ίση με $w = \frac{m}{2}$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου πίνακα σχεδιασμού X.

Βήμα 4. Στη συνέχεια, προσδιορίζουμε τους σημαντικότερους από αυτούς τους παράγοντες, διατηρώντας μόνο εκείνους των οποίων η τιμή του μέτρου κέρδους πληροφορίας βρέθηκε να είναι μεγαλύτερη από την προκαθορισμένη τιμή κατωφλιού που είναι ίση με 0.01. Εφόσον το **ig** μέτρο μιας μεταβλητής είναι μεγαλύτερο από το 0.01, τότε ο αντίστοιχος παράγοντας θεωρείται σημαντικός, σε αντίθετη περίπτωση θεωρείται μη σημαντικός.

Βήμα 5. Συμπερασματικά, κρατάμε τους παράγοντες με τις w υψηλότερες τιμές κέρδους πληροφορίας οι οποίες ταυτόχρονα είναι μεγαλύτερες από την προκαθορισμένη τιμή κατωφλιού. Αυτοί οι παράγοντες θεωρούνται σημαντικοί.

2.2.2 Αξιολόγηση της Προτεινόμενης Μεθόδου

Για τη διερεύνηση και την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου πραγματοποιήσαμε μία εμπειρική μελέτη, θεωρώντας ένα ευρύ φάσμα προσομειωμένων μοντέλων και υπερκορεσμένων σχεδιασμών.

Στη μελέτη προσομοίωσης χρησιμοποιήσαμε υπερκορεσμένους σχεδιασμούς οι οποίοι κατασκευάζονται από τα κλάσματα πινάκων Hadamard, σύμφωνα με τη μέθοδο που προτάθηκε από τον Lin [148]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 10$ παράγοντες και $n = 6$ πειραματικές εκτελέσεις. Επιπλέον, για τα πειράματα προσομοίωσης, χρησιμοποιήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς, όπως αυτοί κατασκευάστηκαν από τους Koukouvinos et al. [130]. Αυτοί οι υπερκορεσμένοι σχεδιασμοί έχουν n πειραματικές εκτελέσεις και $m = q \cdot (n-1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Στις προσομοιώσεις μας θεωρήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, m) τιμές: $(6, 10)$, $(8, 14)$, $(10, 18)$, $(12, 22)$, $(14, 26)$, $(16, 30)$, $(18, 34)$, $(20, 38)$ και $(22, 42)$. Εκτός από αυτούς τους σχεδιασμούς, θεωρήσαμε και τους s-block-ορθογώνιους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με n πειραματικές εκτελέσεις και $m = s(n-1)$ παράγοντες, όπως αυτοί κατασκευάστηκαν από τους Tang και Wu [215]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 22$ παράγοντες και $n = 12$ πειραματικές εκτελέσεις.

Η απόφαση σχετικά με την τιμή κατωφλιού w, η οποία στην πράξη καθορίζει το μέγιστο αριθμό των σημαντικών παραγόντων που μπορούν να αναγνωριστούν, καθώς και η απόφαση σχετικά με την τιμή κατωφλιού του κέρδους πληροφορίας είναι πολύ σημαντικές. Στην πειραματική μελέτη μας θέσαμε το $w = m/2$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου πίνακα σχεδιασμού, δηλαδή ο αριθμός των παραγόντων υπό εξέταση, και για το κέρδος πληροφορίας θέσαμε την τιμή κατωφλιού ίση με 0.01. Η απόφαση για το w έγινε βάση της αρχής της σποραδικότητας [23] ενώ η απόφαση για την τιμή κατωφλιού του κέρδους πληροφορίας ελήφθη μετά από εκτενείς δοκιμές προσομοιώσεων.

Προκειμένου να αξιολογήσουμε την προτεινόμενη μέθοδο και να συγκρίνουμε τα αποτελέσματα χρησιμοποιώντας κάθε φορά μία από τις τρεις θεωρούμενες εντροπίες, εκτιμήσαμε τα σφάλματα Τύπου Ι και Τύπου ΙΙ. Αυτό που μας ενδιαφέρει να ελέγξουμε στα πειράματα προσομείωσης είναι το κόστος του να δηλώσουμε μία ανενεργή επίδραση ως ενεργή (σφάλμα Τύπου Ι) καθώς και το κόστος του να δηλώσουμε μία ενεργή επίδραση ως ανενεργή (σφάλμα Τύπου ΙΙ).

Για τα πειράματα προσομοίωσης, αναπτύξαμε λογιστικά μοντέλα παλινδρόμησης με τους συντελεστές παλινδρόμησης να λαμβάνουν τυχαίες τιμές από το διάνυσμα β και θεωρήσαμε μοντέλα μόνο κύριων επιδράσεων τα οποία δεν συμπεριλαμβάνουν αλληλεπιδράσεις. Σημειώνουμε εδώ ότι από τη στιγμή που θεωρήσαμε το λογιστικό μοντέλο παλινδρόμησης στα προσαρμοσμένα προσομειωμένα μοντέλα, δεν ήταν δυνατή η πραγματοποίηση συγκρίσεων με άλλες μεθόδους στην υπάρχουσα βιβλιογραφία οι οποίες βασίζονται στο γενικό γραμμικό μοντέλο. Περισσότερες πληροφορίες σχετικά με το λογιστικό μοντέλο παλινδρόμησης παραθέτουμε στην ενότητα 5.2.1 της παρούσας διατριβής. Οι πραγματικά ενεργοί παράγοντες επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων και οι συντελεστές των μη ενεργών παραγόντων τέθηκαν ίσοι με μηδέν στο πραγματικό μοντέλο. Στον Πίνακα 2.1 περιγράφονται τα μοντέλα τα οποία θεωρήσαμε στη μελέτη προσομοίωσης.

Πίνακας 2.1: Μοντέλα 1ης μελέτης προσομοίωσης

Μοντέλο	ΥΣ από	m	n	β
1	[148]	10	6	[1, 0, 4, 0, 8, 0, -3, 0, 0, 0]$'$
2	[148]	10	6	[-1, -1, 0, 0, 0, -1, 0, 0, 0, 0]$'$
3	[130]	22	12	[-3, 0, 0, 0, 4, 0, 0, 0, 0, 9, 0, 0, 0, 17, 0, 0, 0, 0, -21, 0, 0, 0]$'$
4	[130]	22	12	[1, 2, 0, 0, 0, 16, 0, 0, 0, 0, 7, 0, 0, 0, -1, 0, 0, 5, 0, 18, 0, 0]$'$
5	[130]	18	10	[-4, 2, 0, 0, 0, -3, 0, 0, 7, 0, 0, 17, 0, 5, 0, 0, 21, 0]$'$
6	[130]	18	10	[1, 0, 0, 0, 4, 0, 2, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0]$'$
7	[130]	18	10	[0, -1, 1, -1, 0, 0, -1, 0, 0, -1, 0, -1, 0, 0, 0, -1, 0, 0]$'$
8	[130]	26	14	[2, 0, 0, 5, 0, 0, -7, 0, 0, -2, 1, 0, 0, 3, 0, 8, 0, 0, -9, 0, 0, -1, 0, 4, 0, 0]$'$
9	[130]	26	14	[2, 0, 0, 5, 0, 0, -7, 0, 0, 0, 1, 0, 0, 3, 0, 8, 0, 0, -9, 0, 0, -1, 0, 4, 0, 0]$'$
10	[130]	30	16	[1, 2, 0, 0, 5, 0, 0, 0, 9, 0, 0, 0, 11, 0, 0, 7, 0, 0, 23, 0, 0, 24, 0, 0, 15, 0, 0, 17, 0, 0]$'$
11	[130]	30	16	[-1, -2, -1, 0, 2, 0, 0, 0, 6, 0, 4, 0, 0, 0, 0, 0, -3, 2, 0, 0, 0, 0, 1, 0, 0, 7, 0, 0, 0, 0]$'$
12	[130]	34	18	[2, 0, -1, 3, 0, 0, 7, 0, 0, 8, 0, 0, 13, 0, 0, 17, 0, 0, 28, 0, 0, 25, 0, 0, -5, 0, 0, -7, 0, 0, -2, 0, -4, 0]$'$
13	[130]	34	18	[-1, -2, -3, -4, 0, 0, 0, 0, 1, 0, 8, 0, 0, 0, 0, -7, 0, 0, 0, -1, 0, 9, 0, 0, 0, -5, 0, 0, 0, 0, 0, 0, 9, 0]$'$
14	[130]	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]$'$
15	[130]	38	20	[1, 2, 3, 4, 5, 6, 7, 0, 0, 0, 1, 0, 0, 1, 0, -1, 0, 0, -1, 0, 0, 1, 0, 0, 0, -1, 0, 4, 0, 0, 0, 0, 0, -2, 0, 0, 0, 0]$'$
16	[130]	42	22	[0, 0, 0, 0, -3, 0, 0, -2, 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]$'$
17	[130]	10	6	[-4, 0, 0, 0, 7, 0, 0, 0, 8, 0]$'$
18	[130]	10	6	[-1, 0, 0, 4, 0, 0, -17, 0, 0, 0]$'$
19	[130]	10	6	[1, 0, -8, 0, -21, 0, -17, 0, 0, 0]$'$
20	[130]	14	8	[1, 0, 0, 2, 0, 0, -4, 0, 0, 7, 0, 0, 11, 0]$'$
21	[130]	14	8	[1, 4, -5, 2, 0, 0, 0, -1, 0, 0, -7, 0, 0, 0]$'$
22	[215]	22	12	[-1, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 8, 0, 0, 0, 0, 9, 0, 0]$'$

Στα πειράματά μας εκτελέσαμε τις προσομοιώσεις 1000 φορές και τα ληφθέντα αποτελέσματα συνοψίζονται στον Πίνακα 2.2 που ακολουθεί. Συγκεκριμένα, στον Πίνακα 2.2, η πρώτη στήλη αναφέρεται στον αριθμό που αντιστοιχεί σε κάθε μοντέλο που χρησιμοποιείται. Οι στήλες που ονομάζονται ως "Type I" και "Type II", αναφέρονται στις τιμές των σφαλμάτων Τύπου Ι και Τύπου ΙΙ που αντιστοιχούν σε κάθε περίπτωση εντροπίας.

Σύμφωνα με τον Πίνακα 2.2, τα ποσοστά σφάλματος Τύπου Ι και τα ποσοστά σφάλματος

Πίνακας 2.2: Απόδοση της 1ης προτεινόμενης μεθόδου για τα μοντέλα 1-22 με χρήση 1000 προσομοιώσεων

Εντροπία	Rényi		Tsallis		Havrda-Charvát	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II
1	0.33	0.25	0.17	0.00	0.33	0.25
2	0.29	0.33	0.29	0.00	0.29	0.00
3	0.35	0.00	0.35	0.00	0.35	0.00
4	0.27	0.00	0.27	0.00	0.27	0.00
5	0.00	0.14	0.27	0.14	0.27	0.14
6	0.36	0.00	0.36	0.00	0.36	0.00
7	0.18	0.00	0.18	0.00	0.18	0.00
8	0.25	0.10	0.25	0.10	0.25	0.10
9	0.29	0.11	0.29	0.11	0.29	0.11
10	0.30	0.10	0.30	0.10	0.30	0.10
11	0.30	0.10	0.30	0.10	0.30	0.10
12	0.29	0.15	0.29	0.15	0.29	0.15
13	0.26	0.00	0.26	0.00	0.26	0.00
14	0.33	0.09	0.33	0.09	0.33	0.09
15	0.26	0.13	0.26	0.13	0.26	0.13
16	0.31	0.08	0.31	0.08	0.31	0.08
17	0.14	0.00	0.29	0.00	0.29	0.00
18	0.29	0.00	0.29	0.00	0.29	0.00
19	0.17	0.25	0.17	0.00	0.17	0.00
20	0.22	0.00	0.22	0.00	0.22	0.00
21	0.25	0.17	0.25	0.17	0.25	0.17
22	0.35	0.00	0.35	0.00	0.35	0.00

Τύπου II, αντίστοιχα, είναι σχεδόν τα ίδια για κάθε μία από τις τρεις περιπτώσεις εντροπιών που θεωρήσαμε. Για τα 17 από τα 22 θεωρούμενα μοντέλα, τα ποσοστά σφάλματος Τύπου I και τα ποσοστά σφάλματος Τύπου II, αντίστοιχα, είναι ακριβώς τα ίδια για κάθε μία από τις τρεις εντροπίες. Με άλλα λόγια, η προτεινόμενη μέθοδος μπορεί να θεωρηθεί εύρωστη σε σχέση με την επιλογή της εντροπίας, δεδομένου ότι παρουσιάζει την ίδια ικανότητα αναγνώρισης των πραγματικά ενεργών και μη ενεργών παραγόντων για κάθε μία από τις τρεις εντροπίες (Rényi, Tsallis και Havrda-Charvát), και μόνο σε 5 από τα 22 θεωρούμενα μοντέλα παρατηρούμε ότι υπάρχουν πολύ μικρές οριακές διαφορές μεταξύ τους. Για παράδειγμα, στις περιπτώσεις που χρησιμοποιείται ο υπερκορεσμένος σχεδιασμός του Lin [148] βλέπουμε ότι υπάρχουν μικρές διαφορές υπέρ της Tsallis εντροπίας.

Γενικά παρατηρούμε ότι η προτεινόμενη μέθοδος επιτυγχάνει ικανοποιητικά χαμηλές τιμές για τα δύο υπό εκτίμηση σφάλματα και σχεδόν για όλα τα μοντέλα. Συγκεκριμένα, οι τιμές των σφαλμάτων Τύπου II είναι χαμηλότερες από τις αντίστοιχες τιμές των σφαλμάτων Τύπου I. Το μέσο ποσοστό σφάλματος Τύπου I των 22 μοντέλων σε 1000 προσομοιώσεις είναι της τάξης του 26% και το αντίστοιχο ποσοστό για τα σφάλματα Τύπου II είναι 9%. Το γεγονός αυτό υποδεικνύει ότι η προτεινόμενη μέθοδος παρουσιάζει την τάση να δηλώνει εσφαλμένα περισσότερους μη ενεργούς παραγόντες ως ενεργούς, και λιγότερους ενεργούς παραγόντες ως μη ενεργούς. Εκτός από αυτό, σημειώνουμε ότι οι τιμές των σφαλμάτων παραμένουν ικανοποιητικά χαμηλές ακόμα και για τα μοντέλα στα οποία η αρχή της σποραδικότητας είναι ασθενέστερη, δηλαδή στα μοντέλα των οποίων ο αριθμός των ενεργών παραγόντων p είναι μεγαλύτερος από $\frac{n}{2}$.

2.3 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάσαμε μία μέθοδο για τον προσδιορισμό των ενεργών επιδράσεων στους υπερκορεσμένους σχεδιασμούς για την ανάλυση δεδομένων που προέρχονται από ένα λογιστικό μοντέλο παλινδρόμησης. Όλα αυτά επιτυγχάνονται μέσω του συνδυασμού

θεμελιωδών μέτρων της Θεωρίας Πληροφοριών. Ο πρωταρχικός μας στόχος ήταν η ανάπτυξη μιας αποτελεσματικής μεθόδου ανάλυσης των υπερκορεσμένων σχεδιασμών η οποία θα επιτυγχάνει να αναγνωρίσει τους ενεργούς παράγοντες με το μικρότερο δυνατό σφάλμα Τύπου I και Τύπου II. Τα σφάλματα Τύπου I και Τύπου II είναι εξίσου σημαντικά και θα πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα, όμως δεδομένου ότι οι υπερκορεσμένοι σχεδιασμοί κυρίως χρησιμοποιούνται σε πειράματα κρησαρίσματος με στόχο να αναγνωρίσουν εκείνους τους παράγοντες που θα πρέπει να εξεταστούν περαιτέρω, οι χαμηλές τιμές σφάλματος Τύπου II είναι ιδιαίτερα επιθυμητές. Η προτεινόμενη μέθοδος φαίνεται να επιτυγχάνει κάτι τέτοιο και να συμπεριφέρεται ικανοποιητικά, δεδομένου ότι παρουσιάζει ιδιαίτερα χαμηλές τιμές για τα σφάλματα Τύπου II και διατηρεί τα σφάλματα Τύπου I σε ικανοποιητικά χαμηλά επίπεδα. Επιπλέον, η προτεινόμενη μέθοδος μπορεί να θεωρηθεί εύρωστη σε σχέση με την επιλογή της εντροπίας, δεδομένου ότι παρουσιάζει παρόμοια ικανότητα αναγνώρισης των πραγματικά ενεργών και μη ενεργών παραγόντων για κάθε μία από τις τρεις θεωρούμενες εντροπίες (Rényi, Tsallis και Havrda-Charvát).

ΚΕΦΑΛΑΙΟ **3**

Ανάλυση Υπερκορεσμένων Σχεδιασμών με το Κριτήριο της Συμμετρικής Αβεβαιότητας

Information is the
resolution of uncertainty.

—*Claude E. Shannon (1916–2001)*

Στο τρίτο αυτό κεφάλαιο ασχολούμαστε και πάλι με τη στατιστική ανάλυση των υπερκορεσμένων σχεδιασμών. Συγκεκριμένα, τροποποιώντας τον αλγόριθμο εξόρυξης δεδομένων Fast Correlation Based Filter παρουσιάζεται μία νέα μεθοδολογία επιλογής μεταβλητών σε έναν υπερκορεσμένο σχεδιασμό, όπου οι επεξηγηματικές μεταβλητές βρίσκονται σε δύο επίπεδα. Η μελέτη έχει και πάλι ως υπόβαθρο τη Θεωρία Πληροφοριών, καθώς η εντροπία Shannon σε συνδυασμό με το κέρδος πληροφορίας και το κριτήριο της συμμετρικής αβεβαιότητας χρησιμοποιούνται για την ανάλυση των υπερκορεσμένων σχεδιασμών και την εύρεση των ενεργών κύριων επιδράσεων, θεωρώντας γενικευμένα γραμμικά μοντέλα και συγκεκριμένα το λογιστικό μοντέλο παλινδρόμησης στο οποίο η μεταβλητή απόκρισης είναι δίτιμη. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [133].

3.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Η εκτεταμένη εφαρμογή διαφόρων μέτρων πληροφορίας σε πολλούς επιστημονικούς τομείς, όπως αναφέρθηκαν ανωτέρω, σε συνδυασμό με τη χρήση της έννοιας της εντροπίας σε μεθόδους επιλογής μεταβλητών και σε πολλούς αλγόριθμους οι οποίοι υλοποιούνται στην ανάλυση εξόρυξης δεδομένων, μας παρακίνησε να διευρύνουμε την εφαρμογή τους στη στατιστική αναπτύσσοντας μεθόδους ανάλυσης των υπερκορεσμένων σχεδιασμών.

Η μέθοδος που παρουσιάσαμε στο προηγούμενο Κεφάλαιο 2 αποτελεί την πρώτη μας προσπάθεια να αναλύσουμε υπερκορεσμένους σχεδιασμούς κάτω από την υπόθεση ενός γενικευμένου γραμμικού μοντέλου, συγκεκριμένα ενός λογιστικού μοντέλου παλινδρόμησης, μέθοδος η οποία είναι μη τετριμμένη και μη αναπτυγμένη στην έως σήμερα διεθνή βιβλιογραφία, όπου συναντάμε το πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών μόνο κάτω υπό τη σκέπη του κλασικού γραμμικού μοντέλου. Οι τιμές των σφαλμάτων Τύπου I και Τύπου II που προέκυψαν ήταν ικανοποιητικά χαμηλές, αλλά από τη στιγμή που το μέσο ποσοστό σφάλματος Τύπου I ήταν της τάξης του 26% και το αντίστοιχο ποσοστό για τα σφάλματα Τύπου II 9%, διαπιστώνουμε ότι υπάρχει σίγουρα κίνητρο περαιτέρω μελέτης και περιθώριο βελτίωσης.

Στην πρώτη προτεινόμενη μέθοδο ανάλυσης τα μέτρα εντροπίας σε συνδυασμό με το κέρδος πληροφορίας χρησιμοποιούνται για την ανάλυση των υπερκορεσμένων σχεδιασμών και την εύρεση των ενεργών κύριων επιδράσεων. Το κέρδος πληροφορίας αποτελεί ένα συμμετρικό μέτρο για δύο τυχαίες μεταβλητές X και Y. Η συμμετρία είναι μία ιδιαίτερα επιθυμητή ιδιότητα για ένα μέτρο εξάρτησης (συσχέτισης) μεταξύ των παραγόντων, αλλά προδιαθέτει το κέρδος πληροφορίας υπέρ των μεταβλητών με τις περισσότερες τιμές. Για το λόγο αυτό, οι τιμές αυτές θα πρέπει να κανονικοποιηθούν έτσι ώστε να εξασφαλιστεί ότι είναι συγκρίσιμες και έχουν την ίδια επίδραση.

Το γεγονός αυτό αποτέλεσε το κίνητρο για την περαιτέρω μελέτη των μέτρων πληροφορίας και την ιδέα για την ενσωμάτωση του κριτηρίου της συμμετρικής αβεβαιότητας σε μεθόδους ανάλυσης των υπερκορεσμένων σχεδιασμών. Οι Press et al. [184] πρότειναν ένα εναλλακτικό μέτρο συσχέτισης, το οποίο ονομάζεται συμμετρική αβεβαιότητα (symmetrical uncertainty-SU) το οποίο αντισταθμίζει την προδιάθεση (bias) του κέρδους πληροφορίας υπέρ των μεταβλητών με τις περισσότερες τιμές. Κάτι τέτοιο επιτυγχάνεται με κανονικοποίηση της τιμής του στο εύρος [0,1].

3.2 Η Προτεινόμενη Μέθοδος Ανάλυσης

Σε αυτή την ενότητα παρουσιάζεται η δεύτερη προτεινόμενη μέθοδος ανάλυσης υπερκορεσμένων σχεδιασμών, η οποία έχει ως στόχο την αναγνώριση και επιλογή των σημαντικών ενεργών επιδράσεων. Η προτεινόμενη μέθοδος αποτελεί τροποποίηση ενός ήδη υπάρχοντος αλγόριθμου επιλογής μεταβλητών, ο οποίος καλείται Fast Correlation Based Filter (FCBF) αλγόριθμος [234] και χρησιμοποιείται στην ανάλυση εξόρυξης δεδομένων. Προσεγγίζουμε το πρόβλημα μέσω ποικίλων μέτρων της Θεωρίας Πληροφοριών, συγκεκριμένα η Shannon εντροπία σε συνδυασμό με το κέρδος πληροφορίας, καθώς και το κριτήριο της συμμετρικής αβεβαιότητας χρησιμοποιούνται προκειμένου να εντοπιστούν οι σημαντικοί παράγοντες μέσα από ένα μεγάλο σύνολο ενδεχόμενων ενεργών παραγόντων. Ένα αξιοσημείωτο πλεονέκτημα της προτεινόμενης μεθόδου είναι η χρήση των γενικευμένων γραμμικών μοντέλων, και συγκεκριμένα του λογιστικού μοντέλου παλινδρόμησης για την ανάλυση των δεδομένων από τους υπερκορεσμένους σχεδιασμούς. Η συμπεριφορά και η αποδοτικότητα της προτεινόμενης μεθόδου εξετάστηκαν και αξιολογήθηκαν μέσα από μια εμπειρική και συγκριτική μελέτη που περιελάμβανε εκτενή πειράματα προσομοίωσης των οποίων τα ενθαρρυντικά αποτελέσματα που προέκυψαν παρουσιάζονται σε επόμενη ενότητα.

3.2.1 Περιγραφή της Προτεινόμενης Μεθόδου

Στη σημερινή εποχή που η συλλογή δεδομένων είναι επιτακτική ανάγκη και ο όγκος της πληροφορίας τεράστιος, αντιμετωπίζουμε σε πολλούς επιστημονικούς τομείς, όπως για παράδειγμα στη μηχανική εκμάθηση (machine learning) ή στην εξόρυξη δεδομένων, προβλήματα μοντελοποίησης υψηλής διάστασης, δηλαδή προβλήματα στα οποία ο αριθμός των μεταβλητών p είναι κατά πολύ μεγαλύτερος των διαθέσιμων παρατηρήσεων n (δηλ. $p \gg n$). Διάφορες μέθοδοι έχουν προταθεί βιβλιογραφικά με στόχο την εύρεση και επιλογή των σημαντικών ματαβλητών σε τέτοιου είδους προβλήματα. Οι υπερκορεσμένοι σχεδιασμοί εξ'ορισμού είναι σχεδιασμοί των οποίων ο αριθμός των παραγόντων m είναι μεγαλύτερος ή ίσος του αριθμού των πειραματικών εκτελέσεων n, δηλαδή $m \geq n$ ή $m > n - 1$. Βλέποντας λοιπόν πιο ανοιχτά το πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών θα μπορούσαμε να πούμε ότι αποτελεί μία υποπερίπτωση των προβλημάτων μοντελοποίησης υψηλής διάστασης. Η διαπίστωση αυτή μας παρακίνησε να τροποποιήσουμε έναν ήδη υπάρχον αλγόριθμο επιλογής μεταβλητών που χρησιμοποιείται σε τέτοιου τύπου προβλήματα, και να αναπτύξουμε κατ'επέκταση μία μέθοδο ανάλυσης των υπερκορεσμένων σχεδιασμών. Η προτεινόμενη μέθοδος συνδυάζει την εντροπία Shannon και το κέρδος πληροφορίας, και καταλήγει να επιλέγει τις ενεργές κύριες επιδράσεις μέσω του κριτηρίου της συμμετρικής αβεβαιότητας.

Το μέτρο της συμμετρικής αβεβαιότητας χρησιμοποιείται στον FCBF αλγόριθμο [234], και η προτεινόμενη μέθοδος επιλογής μεταβλητών αποτελεί τροποποίηση του FCBF αλγορίθμου, από τη στιγμή που χρησιμοποιεί το μέτρο της συμμετρικής αβεβαιότητας για να καθορίσει στο τελικό στάδιο τους σημαντικούς παράγοντες, ορίζοντας κάποια τιμή κατωφλιού της μορφής $SU > \delta$. Ας σημειωθεί εδώ ότι επιλέξαμε μεταξύ πολλών άλλων αλγορίθμων επιλογής μεταβλητών να τροποποιήσουμε τον FCBF αλγόριθμο, για τους εξής λόγους. Ο FCBF αλγόριθμος συγκρινόμενος με άλλους αντίστοιχους αλγόριθμους επιλογής μεταβλητών (ReliefF [122], CorrSF [103] και ConsSF [61]) τρέχει πολύ πιο γρήγορα, γεγονός που επαληθεύει την ανώτερη υπολογιστική αποδοτικότητά του, και είναι ο πιο αποτελεσματικός από την άποψη ότι μπορεί να αφαιρέσει ένα μεγαλύτερο αριθμό περιττών μεταβλητών. Επιπρόσθετα, ο Pearson Redundancy Based Filter (PRBF) αλγόριθμος [17] ο οποίος προτάθηκε πρόσφατα, μετά από την εφαρμογή του και την εμπειρική σύγκριση με τους προαναφερθέντες αλγόριθμους (FCBF, CorrSF, ReliefF και ConsSF) ναι μεν έδωσε πολύ ενθαρρυντικά αποτελέσματα αλλά η απόδοσή του δεν ήταν καλύτερη από αυτήν του FCBF αλγορίθμου.

Η δεύτερη προτεινόμενη μέθοδος επιλογής μεταβλητών για την ανάλυση των υπερκορεσμένων σχεδιασμών περιγράφεται αναλυτικά παρακάτω και εκτελεί τα ακόλουθα βήματα:

Βήμα 1. Δοθέντος ενός $n \times m$ υπερκορεσμένου πίνακα σχεδιασμού $X = [x_1, x_2, \ldots, x_m]$, όπου x_j, $j = 1, 2, \ldots, m$, αντιστοιχεί στην j-οστή στήλη του πίνακα, και ενός $n \times 1$ διανύσματος απόκρισης Y, υπολογίζουμε βήμα-βήμα τα παρακάτω μέτρα πληροφορίας.

Βήμα 2. Αρχικά, υπολογίζουμε το διάνυσμα των τιμών της εντροπίας $H(X) = (H(x_1), H(x_2), \ldots, H(x_m))$, όπου $H(x_j)$, για $j = 1, 2, \ldots, m$, είναι η αντίστοιχη τιμή του μέτρου εντροπίας $H(X)$ για την j-οστή μεταβλητή. Επιπλέον, υπολογίζουμε το διάνυσμα των τιμών της δεσμευμένη εντροπίας $H(X|Y) = (H(x_1|y), H(x_2|y), \ldots, H(x_m|y))$, όπου $H(x_j|y)$, για $j = 1, 2, \ldots, m$, είναι η αντίστοιχη τιμή του μέτρου της δεσμευμένης εντροπίας $H(X|Y)$ για την j-οστή μεταβλητή.

Βήμα 3. Έπειτα, υπολογίζουμε το διάνυσμα των τιμών του κέρδους πληροφορίας $I(X|Y) = (I(x_1|y), I(x_2|y), \ldots, I(x_m|y))$, όπου $I(x_j|y) = H(x_j) - (H(x_j|y))$, για $j = 1, 2, \ldots, m$, είναι η αντίστοιχη τιμή του μέτρου του κέρδους πληροφορίας $I(X|Y)$ για την j-οστή μεταβλητή σε σχέση με τη μεταβλητή απόκρισης.

Βήμα 4. Τέλος, υπολογίζουμε το διάνυσμα των τιμών της συμμετρικής αβεβαιότητας $SU = (su_1, su_2, \ldots, su_m)$, όπου $su_j = 2 \times \left[\frac{I(x_j|y)}{H(x_j)+H(y)}\right]$, για $j = 1, 2, \ldots, m$, είναι η αντί-

στοιχη τιμή του μέτρου της συμμετρικής αβεβαιότητας για την j-οστή μεταβλητή σε σχέση με τη μεταβλητή απόκρισης.

Βήμα 5. Αν η τιμή του SU μέτρου μιας μεταβλητής είναι τουλάχιστον τόσο μεγάλη όσο η προκαθορισμένη τιμή κατωφλιού, τότε η αντίστοιχη μεταβλητή θεωρείται ότι είναι σημαντική, σε διαφορετική περίπτωση δηλώνεται ως μη σημαντική.

3.2.2 Κριτήρια Αξιολόγησης

Προκειμένου να αξιολογήσουμε την προτεινόμενη μέθοδο, χρησιμοποιήσαμε δύο κριτήρια απόδοσης. Το πρώτο κριτήριο που χρησιμοποιήσαμε είναι τα σφάλματα Τύπου I και Τύπου II. Αυτό που μας ενδιαφέρει να ελέγξουμε στα πειράματα προσομείωσης είναι το κόστος του να δηλώσουμε μία ανενεργή επίδραση ως ενεργή (σφάλμα Τύπου I) καθώς και το κόστος του να δηλώσουμε μία ενεργή επίδραση ως ανενεργή (σφάλμα Τύπου II). Το δεύτερο κριτήριο που χρησιμοποιήσαμε είναι οι καμπύλες λειτουργικού χαρακτηριστκού δέκτη (Receiver Operating Characteristic-ROC) και το εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve-AUC). Μια καμπύλη λειτουργικού χαρακτηριστικού δέκτη ή απλά μια ROC καμπύλη είναι μια γραφική αναπαράσταση της ευαισθησίας (sensitivity) έναντι του 1-ειδικότητα (1-specificity) για ένα σύστημα δυαδικής ταξινόμησης. Μια ROC καμπύλη μπορεί ισοδύναμα να εκπροσωπηθεί με τη γραφική αναπαράσταση του ποσοστού των αληθώς θετικών (True Positive Rate-TPR= TP/(TP + FN)) έναντι του ποσοστού των ψευδώς θετικών (False Positive Rate-FPR=FP/(TN + FP)). Είναι επίσης γνωστή ως καμπύλη σχετικού λειτουργικού χαρακτηριστικού αφού εκτελεί τη σύγκριση δύο λειτουργικών χαρακτηριστικών, των TPR και FPR.

Σε ένα λογιστικό μοντέλο πρόβλεψης διπλής κλάσης, το αποτέλεσμα χαρακτηρίζεται από τη θετική (P) ή αρνητική (N) κλάση. Υπάρχουν τέσσερις πιθανές εκβάσεις για ένα δυαδικό ταξινομητή. Αν το αποτέλεσμα της πρόβλεψης είναι P και η πραγματική τιμή είναι επίσης P, αυτό ονομάζεται αληθώς θετικό (TP). Ωστόσο, εάν η πραγματική τιμή είναι N, λέγεται ψευδώς θετικό (FP). Αντίθετα, ένα αληθώς αρνητικό (TN) προκύπτει όταν τόσο το αποτέλεσμα της πρόβλεψης όσο και η πραγματική τιμή είναι N, και ένα ψευδώς αρνητικό (FN) προκύπτει όταν το αποτέλεσμα της πρόβλεψης είναι N, ενώ η πραγματική τιμή είναι P. Ας ορίσουμε ένα πείραμα με P θετικές και N αρνητικές περιπτώσεις. Τα τέσσερα πιθανά αποτελέσματα μπορούν να παρουσιαστούν σε έναν 2×2 πίνακα συνάφειας (Πίνακας 3.1).

Πίνακας 3.1: Πίνακας συνάφειας

True Class		P	N
Hypothesized Class	P	TP	FP
Hypothesized Class	N	FN	TN

Ο ενδιαφερόμενος ανγνώστης μπορεί να ανατρέξει στις επιστημονικές εργασίες [179], [180] και [181] για περισσότερες πληροφορίες σχετικά με τις καμπύλες λειτουργικού χαρακτηριστικού δέκτη, και στην εργασία [25] για το εμβαδόν κάτωθεν αυτής.

Η ROC καμπύλη αποτελεί ένα μέτρο της ακρίβειας ενός πειράματος πρόβλεψης διπλής κλάσης καθώς παρέχει τη δυνατότητα οπτικής και ποσοτικής εκτίμησης της συνολικής απόδοσης και αποτελεσματικότητας μιας πειραματικής δοκιμασίας. Το εμβαδόν κάτωθεν της καμπύλης αποτελεί ένα μέτρο της ποιότητας διαχωρισμού δύο κλάσεων και χρησιμοποιείται ευρέως στη στατιστική συμπερασματολογία των ROC καμπυλών. Η περιοχή κάτω από την καμπύλη ROC (AUC) ισούται με την πιθανότητα ένας ταξινομητής να κατατάξει ένα τυχαία επιλεγμένο θετικό υπόδειγμα (instance) υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό υπόδειγμα

Εδώ ενδιαφέρον έχει ο έλεγχος της υπόθεσης:

$$H_0 : AUC = 0.5,$$

έναντι της εναλλακτικής
$H_1 : AUC > 0.5$.

Η τιμή $AUC = 0.5$ αντιστοιχεί σε ένα πειραματικό τεστ που μαντεύει τυχαία (random guess) και ουσιαστικά δεν έχει αξιόπιστη προβλεπτική ικανότητα. Όσο υψηλότερη είναι η τιμή του AUC, τόσο το καλύτερο, με την τιμή 0.50 να δείχνει τυχαία απόδοση και την τιμή 1.00 να δηλώνει τέλεια απόδοση.

Το σημείο απόφασης αντιστοιχεί στο σημείο που ορίζει το επίπεδο σημαντικότητας α το οποίο είναι ίσο με α=P(Σφάλμα Τύπου I)=P(Απορρίπτεται η $H_0|H_0$ αληθινή). Ομοίως, για την εναλλακτική υπόθεση H_1 μπορεί να οριστεί το β=P(Σφάλμα Τύπου II)=P(Δεκτή η $H_0|H_0$ λανθασμένη). Είναι φανερό ότι θέλουμε τα α και β όσο το δυνατόν μικρότερα. Το $\gamma = 1 - \beta$ λέγεται ισχύς ή δύναμη του τεστ με $1 - \beta$=P(Δεκτή η $H_1|H_1$ αληθινή) και η τιμή του θέλουμε να είναι όσο το δυνατόν μεγαλύτερη. Μεταβάλλοντας το επίπεδο σημαντικότητας και παίρνοντας το γράφημα που ορίζουν τα $(\alpha, 1 - \beta)$ ορίζουμε την ROC καμπύλη για τον έλεγχο της μηδενικής υπόθεσης H_0 έναντι της εναλλακτικής υπόθεσης H_1. Διαπιστώνουμε λοιπόν ότι υπάρχει αντιστοίχιση των δύο κριτηρίων τα οποία επιλέξαμε για την αξιολόγηση της προτεινόμενης μεθοδολογίας.

3.3 Αξιολόγηση της Προτεινόμενης Μεθόδου

Για τη διερεύνηση και την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου πραγματοποιήσαμε μια εμπειρική και συγκριτική μελέτη, θεωρώντας ένα ευρύ φάσμα προσομειωμένων μοντέλων και υπερκορεσμένων σχεδιασμών. Στη μελέτη προσομοίωσης θεωρήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς, όπως αυτοί κατασκευάστηκαν από τους Koukouvinos et al. [130]. Αυτοί οι υπερκορεσμένοι σχεδιασμοί έχουν n πειραματικές εκτελέσεις και $m = q \cdot (n - 1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Στις προσομοιώσεις μας χρησιμοποιήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, m) τιμές: $(6, 10)$, $(8, 14)$, $(10, 18)$, $(12, 22)$, $(14, 26)$, $(16, 30)$, $(18, 34)$, $(20, 38)$ και $(22, 42)$. Εκτός από αυτούς τους σχεδιασμούς, θεωρήσαμε και τους s-block-ορθογώνιους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με n πειραματικές εκτελέσεις και $m = s(n - 1)$ παράγοντες, όπως αυτοί κατασκευάστηκαν από τους Tang και Wu [215]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 22$ παράγοντες και $n = 12$ πειραματικές εκτελέσεις.

Η επιλογή της τιμής κατωφλιού του SU μέτρου, η οποία καθορίζει το κατά πόσο ένας παράγοντας είναι σημαντικός ή όχι, διαδραματίζει σημαντικό ρόλο. Για αυτόν το λόγο, πολλές διαφορετικές τιμές κατωφλιού (0.001, 0.01, 0.05, 0.1, 0.15, 0.2, median(SU)) εξετάστηκαν με στόχο να βρεθούν οι βέλτιστες τιμές για την προτεινόμενη μέθοδο. Η τελική απόφαση ελήφθη σύμφωνα με τις τιμές κατωφλιού που παρατηρήθηκαν να αποτελούν καλές επιλογές μετά από δοκιμές προσομοίωσης. Έτσι επιλέξαμε τρεις τιμές κατωφλιού προκειμένου να μελετήσουμε τα αποτελέσματα, την απόδοση και τη σταθερότητα της μεθόδου μας. Οι πρώτες δύο τιμές κατωφλιού ήταν σταθεροί αριθμοί, ενώ η τρίτη τιμή κατωφλιού βασίστηκε στις εκτιμώμενες τιμές του SU. Ως εκ τούτου, οι τρεις τιμές κατωφλιού που εφαρμόστηκαν ήταν οι 0.01, 0.05 και η διάμεσος του SU (median(SU)).

3.3.1 Εμπειρική Μελέτη

Για τα πειράματα προσομοίωσης, αναπτύξαμε λογιστικά μοντέλα παλινδρόμησης με τους συντελεστές παλινδρόμησης να λαμβάνουν τυχαίες τιμές από το διάνυσμα β και θεωρήσαμε μοντέλα μόνο κύριων επιδράσεων τα οποία δεν συμπεριλαμβάνουν αλληλεπιδράσεις. Περισσότερες πληροφορίες σχετικά με το λογιστικό μοντέλο παλινδρόμησης παραθέτουμε στην ενότητα 5.2.1 της παρούσας διατριβής. Να σημειώσουμε εδώ ότι η κατανομή των αντιθέσεων δεν

επηρεάζεται από το μέτρο (magnitude) των συντελεστών αλλά εξαρτάται από το σχετικό μέγεθος (relative size) των συντελεστών. Δεδομένου ότι οι συνθήκες στην πράξη είναι συνήθως διαφορετικές από εκείνες των προσομοιώσεων και ότι ο πειραματιστής δεν μπορεί να γνωρίζει εκ των προτέρων πόσοι και ποιοί παράγοντες μπορεί να είναι ενεργοί, προκειμένου να εξετάσουμε το πόσο ευαίσθητα είναι τα αποτελέσματα στην επιλογή καθώς και τον αριθμό των ενεργών στηλών-παραγόντων, μελετήσαμε ένα ευρύ φάσμα προσομειωμένων μοντέλων στα πειράματά μας στα οποία θεωρούμε για κάθε υπερκορεσμένο σχεδιασμό διαφορετικό αριθμό ενεργών παραγόντων, διαφορετική επιλογή ενεργών στηλών και διαφορετικούς συντελεστές β. Οι πραγματικά ενεργοί παράγοντες επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων και οι συντελεστές των μη ενεργών παραγόντων τέθηκαν ίσοι με μηδέν στο πραγματικό μοντέλο. Στον Πίνακα 3.2 περιγράφονται τα μοντέλα τα οποία θεωρήσαμε στη μελέτη προσομοίωσης.

Πίνακας 3.2: Μοντέλα 2ης μελέτης προσομοίωσης

Μοντέλο	ΥΣ από	m	n	β
1	[215]	22	12	[1, 0, 0, -13, 0, 0, 4, -3, 0, 0, -6, 0, 0, -7, 0, -24, 0, -5, 0, -21, 0, 0]$'$
2	[130]	22	12	[1, 0, 0, 0, -1, 0, 1, 0, 0, 0, 0, 0, -2, 0, 0, -4, 0, 0, 0, -1, 0, 0]$'$
3	[130]	18	10	[0, 0, -5, 0, -3, 4, 0, 0, 7, 0, 0, 1, 2, 0, 0, -7, 0, 3]$'$
4	[130]	18	10	[0, 0, -7, 0, 0, 6, 4, 0, 0, -3, 0, 0, 1, 1, 0, 0, 0, 1]$'$
5	[130]	18	10	[0, -4, 1, 0, -2, 0, 3, -4, 0, 0, 0, -6, 0, 0, -2, 0, 2, 0]$'$
6	[130]	18	10	[1, 0, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0]$'$
7	[130]	18	10	[0, 0, 0, -13, 0, -24, 0, 2, 0, 0, 11, 0, 0, 21, 0, 0, -6, 0]$'$
8	[130]	26	14	[-2, -1, 0, 0, 2, -3, 0, 0, -3, 0, 0, 2, 0, -4, 0, 0, 0, -6, 0, 0, -4, 0, 0, 0, -5, 0]$'$
9	[130]	26	14	[1, 0, 1, 0, 2, -4, 0, 0, 5, 0, 0, -1, 0, 0, 2, 1, 0, 7, 3, 1, 0, 0, -2, 0, 1, 0]$'$
10	[130]	30	16	[0, -1, 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]$'$
11	[130]	30	16	[-2, -1, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, -1, 0, 0, -7, 0, 0, -3, 0, 0, 0, 0, 0, 0, 0, -4, 0]$'$
12	[130]	34	18	[0, 0, 0, 2, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0]$'$
13	[130]	34	18	[2, 0, 0, 3, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]$'$
14	[130]	34	18	[2, 3, 4, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$'$
15	[130]	38	20	[0, 0, 0, 0, -17, 0, 0 ,0, 0, 0, -23, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 26, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$'$
16	[130]	38	20	[0, -5, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0 ,2 ,0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0]$'$
17	[130]	38	20	[0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0]$'$
18	[130]	38	20	[0, 0, 0, 0, -17, 0, 0, 0, 0, 0, -23, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 26, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -21, 0, 0, 0, 0, 0]$'$
19	[130]	42	22	[0, 0, 0, 0, -2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, -4, 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]$'$
20	[130]	42	22	[0, 0, 3, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0]$'$
21	[130]	10	6	[-7, 0, 0, 0, 0, -7, 0, 0, 2, 0]$'$
22	[130]	10	6	[-17, -13, 0, 0, 0, -2, 0, 0, 6, 0]$'$
23	[130]	14	8	[1, 0, -3 ,0 ,0, -1, 0, 0, 0, 0, 7, 0, 0, 0]$'$
24	[130]	14	8	[-6, 0, 0 ,0 ,2, 0, 0, 4, 0, 0, 0, -8, 0, 0]$'$

Για κάθε ένα από αυτά τα μοντέλα, 1000 σύνολα δεδομένων παρήχθησαν, και τα ληφθέντα αποτελέσματα μετά την εφαρμογή της προτεινόμενης μεθόδου, παρουσιάζονται στον Πίνακα 3.3, σύμφωνα με τις αντίστοιχες τιμές κατωφλιού. Συγκεκριμένα στον Πίνακα 3.3, στην πρώτη στήλη αναφέρεται ο αριθμός που αντιστοιχεί σε κάθε μοντέλο που χρησιμοποιείται, και στην πρώτη γραμμή παρουσιάζεται η επιλεγμένη τιμή κατωφλιού. Οι στήλες που ονομάζονται ως "Type I" και "Type II", αναφέρονται στις τιμές των σφαλμάτων Τύπου I και Τύπου II που αντιστοιχούν σε κάθε τιμή κατωφλιού.

Παρατηρούμε από τον Πίνακα 3.3 ότι η προτεινόμενη μέθοδος παρουσιάζει ικανοποιητικά χαμηλές τιμές και για τα δύο σφάλματα σχεδόν για όλα τα μοντέλα, καθώς επίσης και ότι τα σφάλματα Τύπου II είναι σημαντικά μικρότερα από τα σφάλματα Τύπου I για την πλειονότητα των μοντέλων, συγκεκριμένα για 16 από τα 24 μοντέλα. Εκτός από αυτό, σημειώνουμε ότι οι τιμές των σφαλμάτων παραμένουν χαμηλές ακόμα και για τα μοντέλα στα οποία η αρχή της σποραδικότητας είναι ασθενέστερη, δηλαδή στα μοντέλα των οποίων ο αριθμός των ενεργών παραγόντων p είναι μεγαλύτερος από $\frac{n}{2}$. Γενικά, ο Πίνακας 3.3 υποδεικνύει ότι η προτεινόμενη μέθοδος τείνει να δηλώνει με υψηλότερο ποσοστό τις ανενεργές επιδράσεις ως ενεργές και με πολύ χαμηλότερο ποσοστό τις ενεργές επιδράσεις ως ανενεργές. Υπό αυτή την έννοια, η προτεινόμενη μέθοδος θα μπορούσε πράγματι να χαρακτηριστεί ως συντηρητική (conservative).

Πίνακας 3.3: Απόδοση της 2ης προτεινόμενης μεθόδου για όλες τις τιμές κατωφλιού για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων

Τιμές κατωφλιού	0.01		0.05		median(**su**)	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II
1	0.15	0.11	0.15	0.11	0.15	0.11
2	0.25	0.17	0.25	0.17	0.25	0.17
3	0.20	0.12	0.20	0.12	0.20	0.12
4	0.27	0.14	0.27	0.14	0.27	0.14
5	0.30	0.25	0.30	0.25	0.30	0.25
6	0.36	0.00	0.36	0.00	0.36	0.00
7	0.33	0.17	0.33	0.17	0.33	0.17
8	0.37	0.10	0.37	0.10	0.37	0.30
9	0.30	0.15	0.30	0.15	0.30	0.23
10	0.36	0.12	0.36	0.12	0.36	0.50
11	0.18	0.25	0.18	0.25	0.18	0.36
12	0.27	0.00	0.27	0.00	0.41	0.00
13	0.41	0.00	0.41	0.00	0.40	0.10
14	0.24	0.20	0.24	0.20	0.45	0.20
15	0.31	0.00	0.31	0.00	0.46	0.00
16	0.18	0.25	0.18	0.25	0.47	0.25
17	0.23	0.25	0.23	0.25	0.47	0.25
18	0.09	0.25	0.09	0.25	0.47	0.25
19	0.19	0.00	0.19	0.00	0.43	0.00
20	0.21	0.40	0.21	0.40	0.46	0.20
21	0.14	0.00	0.14	0.00	0.14	0.00
22	0.17	0.25	0.17	0.25	0.17	0.25
23	0.20	0.25	0.20	0.25	0.20	0.25
24	0.20	0.25	0.20	0.25	0.20	0.25

Παρατηρούμε επίσης από τον Πίνακα 3.3 ότι η προτεινόμενη μέθοδος επιτυγχάνει ακριβώς τις ίδιες τιμές για τα σφάλματα Τύπου I και Τύπου II, αντίστοιχα, για όλα τα μοντέλα, στις περιπτώσεις που η τιμή κατωφλιού ισούται με 0.01 ή 0.05. Η προτεινόμενη μέθοδος για 13 από τα 24 μοντέλα, φαίνεται να έχει ένα από αυτά τα σφάλματα σημαντικά υψηλότερο, όταν η τιμή κατωφλιού είναι ίση με τη διάμεσο του SU. Για αυτόν το λόγο αποφασίσαμε να χρησιμοποιήσουμε τις τιμές κατωφλιού 0.01 και 0.05 για περαιτέρω συγκριτική μελέτη, η οποία ακολουθεί παρακάτω.

3.3.2 Συγκριτική Μελέτη

Προκειμένου να εξεταστεί η αποτελεσματικότητα της μεθόδου και από τη στιγμή που η προτεινόμενη μέθοδος βασίζεται στο κριτήριο της συμμετρικής αβεβαιότητας, το οποίο αποτελεί μέτρο συσχέτισης και ανεξαρτησίας, εναλλακτικά σκεφτήκαμε να τη συγκρίνουμε με δύο χ^2 τεστ ανεξαρτησίας, το Pearson χ^2 τεστ και το Likelihood ratio χ^2 τεστ. Το Pearson χ^2 τεστ αποτελεί ένα τεστ της ανεξαρτησίας μεταξύ των X και Y, το οποίο περιλαμβάνει τη διαφορά μεταξύ των παρατηρούμενων και αναμενόμενων συχνοτήτων [177]. Το Likelihood ratio χ^2 τεστ αποτελεί ένα τεστ της ανεξαρτησίας μεταξύ των X και Y το οποίο περιλαμβάνει την αναλογία μεταξύ των παρατηρούμενων και αναμενόμενων συχνοτήτων [54]. Οι αναμενόμενες συχνότητες εκτιμώνται κάτω από τη μηδενική υπόθεση της ανεξαρτησίας και για τα δύο τεστ.

Αρχικά υπολογίσαμε τα p-values για κάθε ένα από τα χ^2 τεστ. Ταξινομήσαμε τις μεταβλητές βάση των p-values που προέκυψαν, κατά αύξουσα διάταξη. Έπειτα, επιλέξαμε τις σημαντικές μεταβλητές βάση αυτής της κατάταξης και υπολογίσαμε τα σφάλματα Τύπου I και Τύπου II για τις δύο τιμές αποκοπής, 0.01 και 0.05. Ακολουθήσαμε τα βήματα αυτά για το Pearson χ^2 τεστ καθώς και το Likelihood ratio χ^2 τεστ. Στο τέλος, συγκρίναμε αυτές τις δύο μεθόδους με την προτεινόμενη, χρησιμοποιώντας τα σφάλματα Τύπου I και Τύπου II τα

οποία προέκυψαν για την κάθε μία υπό εξέταση μέθοδο και για την κάθε μία τιμή κατωφλιού.

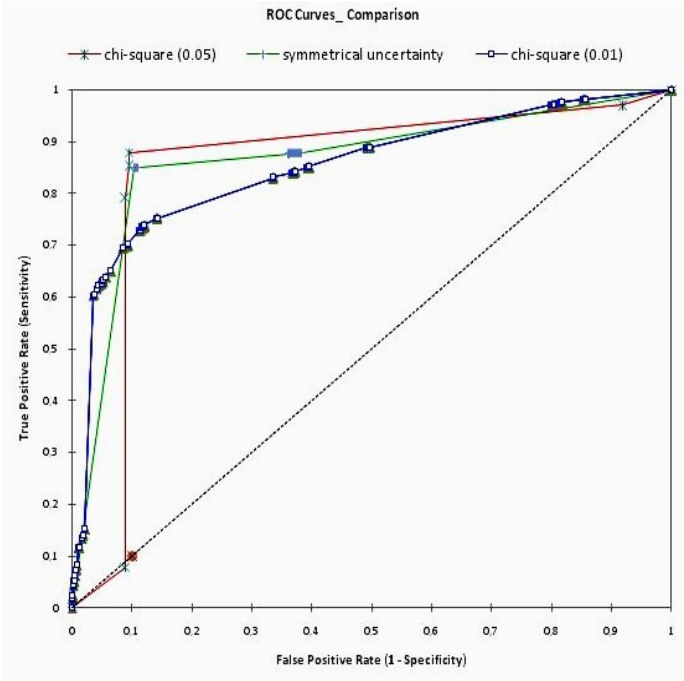
Πίνακας 3.4: Συγκριτική απόδοση της 2ης προτεινόμενης μεθόδου για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων

Μέθοδοι	Symmetrical Uncertainty		Pearson & Likelihood		Pearson & Likelihood	
Τιμές κατωφλιού	(0.01 & 0.05)		(0.01)		(0.05)	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II
1	0.15	0.11	0.19	0.28	0.21	0.26
2	0.25	0.17	0.38	0.31	0.42	0.28
3	0.20	0.12	0.22	0.17	0.25	0.14
4	0.27	0.14	0.28	0.21	0.30	0.19
5	0.30	0.25	0.34	0.32	0.37	0.28
6	0.36	0.00	0.34	0.34	0.38	0.28
7	0.33	0.17	0.35	0.25	0.36	0.22
8	0.37	0.10	0.45	0.18	0.47	0.14
9	0.30	0.15	0.43	0.29	0.45	0.27
10	0.36	0.12	0.46	0.15	0.48	0.12
11	0.18	0.25	0.26	0.30	0.31	0.28
12	0.27	0.00	0.35	0.08	0.38	0.05
13	0.41	0.00	0.58	0.06	0.59	0.02
14	0.24	0.20	0.33	0.25	0.36	0.23
15	0.31	0.00	0.37	0.04	0.41	0.02
16	0.18	0.25	0.22	0.28	0.25	0.26
17	0.23	0.25	0.26	0.33	0.27	0.28
18	0.09	0.25	0.15	0.27	0.19	0.25
19	0.19	0.00	0.22	0.09	0.24	0.04
20	0.21	0.40	0.28	0.67	0.31	0.66
21	0.14	0.00	0.25	0.07	0.27	0.04
22	0.17	0.25	0.38	0.31	0.39	0.28
23	0.20	0.25	0.24	0.34	0.27	0.31
24	0.20	0.25	0.25	0.32	0.26	0.30

Παρατηρούμε από τον Πίνακα 3.4 ότι η προτεινόμενη μέθοδος επιτυγχάνει ακριβώς τις ίδιες τιμές για τα σφάλματα Τύπου I και Τύπου II, αντίστοιχα, για όλα τα μοντέλα, στις περιπτώσεις που η τιμή κατωφλιού ισούται με 0.01 ή 0.05. Επίσης, παρατηρούμε ότι από την εφαρμογή του Pearson χ^2 τεστ καθώς και του Likelihood ratio χ^2 τεστ προέκυψαν τα ίδια ακριβώς σφάλματα Τύπου I και Τύπου II, αντίστοιχα, λαμβάνοντας υπόψη πρώτα την τιμή κατωφλιού 0.01 και στη συνέχεια την τιμή κατωφλιού 0.05. Η προτεινόμενη μέθοδος επιτυγχάνει ταυτόχρονα χαμηλότερα σφάλματα Τύπου I και Τύπου II σε σύγκριση με αυτά των άλλων δύο χ^2 τεστ. Συμπεραίνουμε λοιπόν ότι η προτεινόμενη μέθοδος η οποία χρησιμοποιεί το κριτήριο της συμμετρικής αβεβαιότητας υπερτερεί των Pearson και Likelihood ratio χ^2 τεστ και για τις δύο τιμές κατωφλιού, δηλαδή τις τιμές 0.01 και 0.05.

Εκτός από τον υπολογισμό των σφαλμάτων Τύπου I και Τύπου II για την κάθε μία υπό εξέταση μέθοδο, επιπροσθέτως κατασκευάσαμε την ROC καμπύλη της κάθε μεθόδου και υπολογίσαμε την περιοχή κάτωθεν της ROC καμπύλης για την κάθε μέθοδο. Ας σημειωθεί εδώ ότι από τη στιγμή που το Pearson καθώς και το Likelihood ratio χ^2 τεστ είχαν την ίδια ακριβώς απόδοση, συνδυάσαμε τα αποτελέσματά τους τα οποία προέκυψαν από τα 24 θεωρούμενα μοντέλα και κατασκευάσαμε το Σχήμα 3.1. Το Σχήμα 3.1 αναπαριστά γραφικά τη σύγκριση των μέσων ROC καμπυλών στις 1000 προσομειώσεις, οι οποίες παρήχθησαν με τη χρήση του κριτηρίου της συμμετρικής αβεβαιότητας και των χ^2 τεστ και για τις δύο τιμές αποκοπής, δηλαδή τις τιμές 0.01 και 0.05. Η μέση ROC καμπύλη για την προτεινόμενη μέθοδο (με χρήση του κριτηρίου της συμμετρικής αβεβαιότητας) είναι η ίδια και για τις δύο τιμές κατωφλιού από τη στιγμή που τα σφάλματα Τύπου I και Τύπου II, αντίστοιχα, ήταν ακριβώς τα ίδια για την τιμή 0.01 και την τιμή 0.05.

Μια ROC καμπύλη η οποία βρίσκεται πιο κοντά στην επάνω αριστερή γωνία του γραφή-



Σχήμα 3.1: Σύγκριση των ROC καμπυλών

ματος είναι πιο επιθυμητή και περισσότερο ακριβής. Παρατηρούμε από το Σχήμα 3.1 ότι όλες οι ROC καμπύλες των υπό εξέταση μεθόδων βρίσκονται πάνω από τη διαγώνιο και είναι κοντά στο σημείο (0,1), δηλαδή, στην πάνω αριστερή γωνία, γεγονός το οποίο υποδεικνύει ότι και οι τρεις μέθοδοι παρουσιάζουν καλή απόδοση. Παρ'όλα αυτά η προτεινόμενη μέθοδος επιτυγχάνει την υψηλότερη τιμή για το εμβαδόν κάτωθεν της καμπύλης (AUC). Το AUC (συμμετρική αβεβαιότητα)= 0.865 > AUC (χ^2 τεστ 0.01)= 0.853 > AUC (χ^2 τεστ 0.05) = 0.851. Να σημειώσουμε εδώ ότι υπάρχει μικρή διαφορά στην τιμή AUC των χ^2 τεστ για τις δύο διαφορετικές τιμές αποκοπής. Το AUC αποτελεί ένα μέτρο της προβλεπτικής ικανότητας και εκτιμά την πιθανότητα οι προβλέψεις και τα αποτελέσματα να συμφωνούν αρμονικά. Όσο υψηλότερη είναι η τιμή του AUC, τόσο το καλύτερο. Η τιμή του AUC της προτεινόμενης μεθόδου ισούται με 0.865 και ήταν η υψηλότερη που προέκυψε, οπότε η ανάλυση των ROC καμπυλών έδειξε ότι η προτεινόμενη μέθοδος είναι πιο αποτελεσματική σε σύγκριση με τις χ^2 μεθόδους.

Τέλος, συμπεραίνουμε ότι τα κριτήρια που χρησιμοποιήσαμε για την αξιολόγηση της απόδοσης των μεθόδων, δηλαδή, τα σφάλματα Τύπου I και Τύπου II καθώς και η ανάλυση των ROC καμπυλών, όλα μας υπέδειξαν ότι η προτεινόμενη μέθοδος υπερτερεί των Pearson και Likelihood ratio χ^2 τεστ για την ανάλυση των υπερκορεσμένων σχεδιασμών.

3.4 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάσαμε μία μέθοδο για τον προσδιορισμό των ενεργών επιδράσεων στους υπερκορεσμένους σχεδιασμούς με δεδομένα απόκρισης τα οποία προέρχονται από ένα λογιστικό μοντέλο παλινδρόμησης. Προσεγγίσαμε το πρόβλημα συνδυάζοντας θεμελιώδη μέτρα της Θεωρίας Πληροφοριών τα οποία μας οδήγησαν στην επιλογή του κριτηρίου της συμμετρικής αβεβαιότητας για την εύρεση των σημαντικών μεταβλητών. Η εμπειρική απόδοση της προτεινόμενης μεθόδου εξετάστηκε μέσω εκτενών πειραμάτων προσομοίωσης για διαφορετικές θεωρούμενες τιμές κατωφλιού της συμμετρικής αβεβαιότητας. Όπως με οποιαδήποτε μέθοδο ανάλυσης υπερκορεσμένων σχεδιασμών, τα σφάλματα Τύπου I και Τύπου

ΙΙ είναι εξίσου σημαντικά και θα πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα. Όμως δεδομένου ότι οι υπερκορεσμένοι σχεδιασμοί κυρίως χρησιμοποιούνται σε πειράματα κρησαρίσματος με στόχο να αναγνωρίσουν εκείνους τους παράγοντες που θα πρέπει να εξεταστούν περαιτέρω, οι χαμηλές τιμές σφάλματος Τύπου ΙΙ είναι ιδιαίτερα επιθυμητές.

Συγκεκριμένα, το μέσο ποσοστό σφάλματος Τύπου Ι των 24 μοντέλων σε 1000 προσομοιώσεις ήταν της τάξης του 25% και το αντίστοιχο ποσοστό για τα σφάλματα Τύπου ΙΙ ήταν της τάξης του 15%. Η προτεινόμενη μέθοδος φαίνεται να συμπεριφέρεται ικανοποιητικά, δεδομένου ότι επιτυγχάνει χαμηλές τιμές για τα σφάλματα Τύπου ΙΙ, το οποίο είναι πολύ σημαντικό, προκειμένου να μην παραλείπονται οι σημαντικοί παράγοντες του μοντέλου. Επιπλέον, η προτεινόμενη μέθοδος διατηρεί τα σφάλματα Τύπου Ι σε ικανοποιητικά χαμηλά επίπεδα. Όταν μια τιμή κατωφλιού συνεπάγεται χαμηλό σφάλμα Τύπου Ι, αυτό σημαίνει ότι έχει την ικανότητα να αποκλείει τους περιττούς παράγοντες, και ο ερευνητής βασισμένος στους παράγοντες που επιλέχθηκαν είναι σε θέση πλέον να μειώσει το υπολογιστικό κόστος των επιπλέον πειραμάτων που θα εκτελέσει. Από τη μελέτη προσομοίωσης, κατάλληλες για αυτόν το σκοπό βρέθηκαν οι τιμές κατωφλιού 0.01 και 0.05. Σε γενικά πλαίσια, όλα τα κριτήρια που χρησιμοποιήσαμε για την αξιολόγηση της απόδοσης των μεθόδων, δηλαδή, τα σφάλματα Τύπου Ι και Τύπου ΙΙ και η ανάλυση των ROC καμπυλών, μας υπέδειξαν ότι η προτεινόμενη μέθοδος υπερτερεί των Pearson και Likelihood ratio χ^2 τεστ για την ανάλυση των υπερκορεσμένων σχεδιασμών.

ΚΕΦΑΛΑΙΟ **4**

Αλγοριθμική Προσέγγιση του Προβλήματος Ανάλυσης Υπερκορεσμένων Σχεδιασμών

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

—*John Johnson (1977)*

Στο τέταρτο αυτό κεφάλαιο ασχολούμαστε με τη στατιστική ανάλυση των υπερκορεσμένων σχεδιασμών, προσεγγίζοντας το πρόβλημα αλγοριθμικά. Συγκεκριμένα, αναπτύσσουμε έναν αλγόριθμο επιλογής μεταβλητών βασισμένο σε θεμελιώδη μέτρα της Θεωρίας Πληροφοριών. Μέτρα όπως η Shannon, Rényi, Tsallis και Havrda-Charvát εντροπία σε συνδυασμό με το κέρδος πληροφορίας και το κριτήριο της συμμετρικής αβεβαιότητας χρησιμοποιούνται για την ανάλυση των υπερκορεσμένων σχεδιασμών, όπου οι επεξηγηματικές μεταβλητές βρίσκονται σε δύο επίπεδα και τα δεδομένα απόκρισης προέρχονται από το λογιστικό μοντέλο παλινδρόμησης στο οποίο η μεταβλητή απόκρισης είναι δίτιμη. Στόχος αυτής της αλγοριθμικής προσέγγισης του προβλήματος ανάλυσης υπερκορεσμένων σχεδιασμών είναι η γρήγορη και αποτελεσματική εύρεση των ενεργών κύριων επιδράσεων. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [7].

4.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Οι μέθοδοι ανάλυσης που παρουσιάσαμε στα προηγούμενα Κεφάλαια 2 και 3, αποτελούν την πρώτη και τη δεύτερή μας προσπάθεια, αντίστοιχα, έτσι ώστε να αναλύσουμε υπερκορεσμένους σχεδιασμούς κάτω από την υπόθεση ενός γενικευμένου γραμμικού μοντέλου, συγκεκριμένα ενός λογιστικού μοντέλου παλινδρόμησης. Οι τιμές των σφαλμάτων Τύπου I και Τύπου II που προέκυψαν από τις δύο αυτές μεθόδους ήταν ικανοποιητικά χαμηλές, αλλά από τη στιγμή που το μέσο ποσοστό σφάλματος Τύπου I ήταν της τάξης του 26% και 25% για την 1η και 2η μέθοδο αντίστοιχα, καθώς και το μέσο ποσοστό σφάλματος Τύπου II ήταν της τάξης του 9% και 15% για την 1η και 2η μέθοδο αντίστοιχα, διαπιστώνουμε ότι υπάρχει σίγουρα κίνητρο περαιτέρω μελέτης και περιθώριο βελτίωσης.

Όπως προαναφέραμε στην ενότητα 1.4, μέχρι στιγμής δεν έχει βρεθεί κάποια μέθοδος ανάλυσης των υπερκορεσμένων σχεδιασμών η οποία όταν χρησιμοποιείται να αναγνωρίζονται οι σημαντικοί παράγοντες και να εκτιμούνται με αμελητέο σφάλμα. Το πρόβλημα που υπάρχει στην ανάλυση των υπερκορεσμένων σχεδιασμών είναι η μη ορθογωνιότητα του πίνακα σχεδιασμού, καθιστώντας αναπόφευκτο το να γίνεται ανεκτή η ύπαρξη μιας μικρής μεροληψίας ανάμεσα σε όλες τις εκτιμημένες επιδράσεις. Στόχος μας λοιπόν είναι να προσπαθήσουμε να διατηρήσουμε τις τιμές των σφαλμάτων στο κατώτατο επιτρεπτό επίπεδο. Σκεφτήκαμε λοιπόν να συνδυάσουμε τα μέτρα πληροφορίας καθώς επίσης και τα βήματα υλοποίησης των μεθόδων ανάλυσης που παρουσιάστηκαν στις εργασίες [127] και [133], με στόχο να αναπτύξουμε έναν συνεκτικό αλγόριθμο της Θεωρίας Πληροφοριών για την ανάλυση των υπερκορεσμένων σχεδιασμών, έτσι ώστε να αναγνωρίζονται γρήγορα και αποτελεσματικά οι σημαντικές επιδράσεις κάτω από την υπόθεση ενός λογιστικού μοντέλου παλινδρόμησης.

4.2 Η Προτεινόμενη Μέθοδος Ανάλυσης

Σε αυτή την ενότητα παρουσιάζεται η τρίτη προτεινόμενη μέθοδος ανάλυσης υπερκορεσμένων σχεδιασμών, η οποία έχει ως στόχο την αναγνώριση και επιλογή των σημαντικών ενεργών επιδράσεων. Η προτεινόμενη μέθοδος αποτελεί μία αλγοριθμική προσέγγιση του προβλήματος ανάλυσης υπερκορεσμένων σχεδιασμών. Τα μέτρα πληροφορίας τα οποία συνθέτουν τον προτεινόμενο αλγόριθμο (Information Theoretical Algorithm-ITA), είναι η Shannon, Rényi, Tsallis και Havrda-Charvát εντροπία σε συνδυασμό με το κέρδος πληροφορίας και τη συμμετρική αβεβαιότητα. Σημαντικό πλεονέκτημα της προτεινόμενης μεθόδου είναι η χρήση των γενικευμένων γραμμικών μοντέλων, και συγκεκριμένα του λογιστικού μοντέλου παλινδρόμησης για την ανάλυση των δεδομένων από τους υπερκορεσμένους σχεδιασμούς. Η συμπεριφορά και η αποδοτικότητα της προτεινόμενης μεθόδου εξετάστηκαν και αξιολογήθηκαν μέσα από μια εμπειρική και συγκριτική μελέτη, η οποία περιελάμβανε εκτενή πειράματα προσομοίωσης των οποίων τα ενθαρρυντικά αποτελέσματα που προέκυψαν παρουσιάζονται σε επόμενη ενότητα.

4.2.1 Περιγραφή της Προτεινόμενης Μεθόδου

Σε αυτή την ενότητα παρουσιάζεται αναλυτικά ο προτεινόμενος αλγόριθμος επιλογής μεταβλητών για τους υπερκορεσμένους σχεδιασμούς. Ο προτεινόμενος αλγόριθμος καλείται ως Information Theoretical Algorithm (ITA), βασίζεται σε ποικίλα μέτρα πληροφορίας και εκτελεί τα ακόλουθα βήματα:

Βήμα 1. Δοθέντος ενός $n \times m$ υπερκορεσμένου πίνακα σχεδιασμού $X = [x_1, x_2, \ldots, x_m]$, όπου x_j, $j = 1, 2, \ldots, m$, αντιστοιχεί στην j-οστή στήλη του πίνακα, και ενός $n \times 1$ διανύσματος απόκρισης Y, υπολογίζουμε το διάνυσμα των τιμών της Shannon εντροπίας $H(X) = (H(x_1), H(x_2), \ldots, H(x_m))$, όπου $H(x_j)$, για $j = 1, 2, \ldots, m$, είναι η αντίστοιχη τιμή του μέτρου εντροπίας $H(X)$ για την j-οστή μεταβλητή. Επιπρόσθε-

τα, υπολογίζουμε το διάνυσμα των τιμών της δεσμευμένη εντροπίας σε σχέση με τη μεταβλητή απόκρισης $H(X|Y) = (H(x_1|y), H(x_2|y), \ldots, H(x_m|y))$, όπου $H(x_j|y)$, για $j = 1, 2, \ldots, m$, είναι η αντίστοιχη τιμή του μέτρου της δεσμευμένης εντροπίας $H(X|Y)$ για την j-οστή μεταβλητή.

Βήμα 2. Έπειτα, υπολογίζουμε το διάνυσμα των τιμών του κέρδους πληροφορίας **ig**=$(ig_1, ig_2, \ldots, ig_m)$, όπου ig_j, για $j = 1, 2, \ldots, m$, είναι η τιμή του μέτρου κέρδους πληροφορίας (IG) που αντιστοιχεί στην j-οστή μεταβλητή.

Βήμα 3. Ταξινομούμε τους παράγοντες σύμφωνα με τις τιμές του κέρδους πληροφορίας που προέκυψαν. Με άλλα λόγια ταξινομούμε σε φθίνουσα σειρά το διάνυσμα κέρδους πληροφορίας **ig**. Κρατάμε τους παράγοντες που βρέθηκαν να έχουν τις w υψηλότερες τιμές κέρδους πληροφορίας. Η τιμή κατωφλιού w, η οποία καθορίζει το μέγιστο αριθμό σημαντικών παραγόντων που μπορούν να προσδιοριστούν έχει οριστεί να είναι ίση με $w = \frac{m}{2}$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου πίνακα σχεδιασμού X.

Βήμα 4. Δοθέντος του υπερκορεσμένου πίνακα σχεδιασμού X, υπολογίζουμε τη συμμετρική αβεβαιότητα SU κάθε μεταβλητής σε σχέση με τη μεταβλητή απόκρισης. Υπολογίζουμε δηλαδή το διάνυσμα των τιμών $SU = (su_1, su_2, \ldots, su_m)$, όπου su_j, για $j = 1, 2, \ldots, m$, είναι η αντίστοιχη τιμή του μέτρου της συμμετρικής αβεβαιότητας για την j-οστή μεταβλητή. Αν η τιμή του SU μέτρου μιας μεταβλητής είναι τουλάχιστον τόσο μεγάλη όσο η προκαθορισμένη τιμή κατωφλιού (= median(**su**)), τότε η αντίστοιχη μεταβλητή θεωρείται ότι είναι σημαντική, σε διαφορετική περίπτωση δηλώνεται ως μη σημαντική.

Βήμα 5. Τέλος, κρατάμε μόνο τους παράγοντες που βρέθηκαν να έχουν τις $w = \frac{m}{2}$ υψηλότερες τιμές κέρδους πληροφορίας και ταυτόχρονα η τιμή της συμμετρικής αβεβαιότητάς τους βρέθηκε να είναι μεγαλύτερη από την προκαθορισμένη τιμή κατωφλιού (= median(**su**)). Αυτοί οι παράγοντες θεωρούνται σημαντικοί.

4.2.2 Κριτήρια Αξιολόγησης

Προκειμένου να αξιολογήσουμε την απόδοση του προτεινόμενου αλγορίθμου επιλογής μεταβλητών, χρησιμοποιήσαμε δύο κριτήρια απόδοσης. Το πρώτο κριτήριο που χρησιμοποιήσαμε είναι τα σφάλματα Τύπου I και Τύπου II. Όπως συμβαίνει σε πολλά προβλήματα λήψης αποφάσεων, οποιοδήποτε σφάλμα πρέπει να εξισορροπείται σε σχέση με το κόστος. Στους σχεδιασμούς κρησαρίσματος, αυτό που μας ενδιαφέρει να ελέγξουμε είναι το κόστος του να δηλώσουμε μία ανενεργή επίδραση ως ενεργή (σφάλμα Τύπου I) καθώς και το κόστος του να δηλώσουμε μία ενεργή επίδραση ως ανενεργή (σφάλμα Τύπου II). Αξίζει να σημειωθεί εδώ ότι τα σφάλματα Τύπου II δημιουργούν αρκετά προβλήματα, όπως διαπιστώθηκε από τον Lin [149], όπως άλλωστε και τα σφάλματα Τύπου I, δεδομένου ότι μπορεί να οδηγήσουν σε περιττό υπολογιστικό κόστος στα πειράματα που ακολουθούν, και κάτι τέτοιο μπορεί να είναι αρκετά επιζήμιο σε ένα πείραμα το οποίο έχει άμεσο αντίκτυπο στην πράξη. Σε καταστάσεις που ισχύει η αρχή της σποραδικότητας των επιδράδεων είναι πολύ πιθανό να προκύψουν σφάλματα Τύπου I [149].

Το δεύτερο κριτήριο που χρησιμοποιήσαμε είναι οι καμπύλες λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic-ROC). Εφαρμόσαμε τη μεθοδολογία των ROC καμπυλών [179] σε κάθε ένα από τα θεωρούμενα μοντέλα ως ένα πρόσθετο στατιστικό εργαλείο για τη σύγκριση των αποτελεσμάτων και για την αξιολόγηση της αποτελεσματικότητας της προτεινόμενης μεθόδου. Μία ROC καμπύλη, εξ'ορισμού, χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός συστήματος με διχοτομικές εκβάσεις (0 ή 1) και χρησιμοποιείται ευρέως για την επιλογή των ενδεχόμενων βέλτιστων μοντέλων, καθώς και

για τη σύγκριση μοντέλων. Ένα ψευδώς θετικό σφάλμα παρουσιάζεται όταν μια αρνητική τιμή δείχνει θετικό αποτέλεσμα δοκιμής, και αντιστρόφως, ένα ψευδώς αρνητικό σφάλμα παρουσιάζεται όταν μια θετική τιμή δείχνει ένα αρνητικό αποτέλεσμα δοκιμής. Τα ποσοστά με τα οποία αυτά τα σφάλματα συμβαίνουν, καλούνται ως ψευδώς θετικό ποσοστό (False Positive Rate (FPR) ή σφάλμα Τύπου I) και ψευδώς αρνητικό ποσοστό (False Negative Rate (FNR) ή σφάλμα Τύπου II). Σε ένα λογιστικό μοντέλο πρόβλεψης διπλής κλάσης, μια ROC καμπύλη, είναι μια γραφική αναπαράσταση της ευαισθησίας (sensitivity=1-FNR) έναντι του 1- ειδικότητα (1-specificity=FPR) για όλα τα πιθανά διαχωριστικά όρια [180].

Συγκεκριμένα, ισχύουν οι παρακάτω ισοδυναμίες:

- Sensitivity=$P(\hat{y} = 1|y = 1)$= 1-FNR= 1-Type II error,
- Specificity=$P(\hat{y} = 0|y = 0)$=1-FPR=1-Type I error.

4.3 Αξιολόγηση της Προτεινόμενης Μεθόδου

Για τη διερεύνηση και την αξιολόγηση της απόδοσης του προτεινόμενου αλγορίθμου πραγματοποιήσαμε μια εμπειρική και συγκριτική μελέτη, θεωρώντας ένα ευρύ φάσμα προσομειωμένων μοντέλων και υπερκορεσμένων σχεδιασμών. Στη μελέτη προσομοίωσης χρησιμοποιήσαμε τους s-block-ορθογώνιους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με n πειραματικές εκτελέσεις και $m = s(n-1)$ παράγοντες, όπως αυτοί κατασκευάστηκαν από τους Tang και Wu [215]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 22$ παράγοντες και $n = 12$ πειραματικές εκτελέσεις. Επιπλεόν, στη μελέτη προσομοίωσης χρησιμοποιήσαμε υπερκορεσμένους σχεδιασμούς οι οποίοι κατασκευάζονται από τα κλάσματα πινάκων Hadamard, σύμφωνα με τη μέθοδο που προτάθηκε από τον Lin [148]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 10$ παράγοντες και $n = 6$ πειραματικές εκτελέσεις. Εκτός από αυτούς τους σχεδιασμούς, θεωρήσαμε και τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς, όπως αυτοί κατασκευάστηκαν από τους Koukouvinos et al. [130]. Αυτοί οι υπερκορεσμένοι σχεδιασμοί έχουν n πειραματικές εκτελέσεις και $m = q \cdot (n-1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Στις προσομοιώσεις μας θεωρήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, m) τιμές: $(6, 10)$, $(8, 14)$, $(10, 18)$, $(12, 22)$, $(14, 26)$, $(16, 30)$, $(18, 34)$, $(20, 38)$ και $(22, 42)$. Οι $E(s^2)$-βέλτιστοι υπερκορεσμένοι σχεδιασμοί είναι ισορροπημένοι σχεδιασμοί. Χαλαρώνοντας αυτό τον περιορισμό, θεωρήσαμε και μη ισορροπημένους υπερκορεσμένους σχεδιασμούς. Συγκεκριμένα θεωρήσαμε τους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς όπως αυτοί πινακοποιούνται από τους Jones et al. στον Πίνακα 1 που εσωκλείει η επιστημονική εργασία [118]. Οι σχεδιασμοί αυτοί έχουν n πειραματικές εκτελέσεις και $m \neq 2(n-1)$ παράγοντες. Στη μελέτη προσομοίωσης θεωρήσαμε τους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, m) τιμές: $(12, 16)$, $(12, 18)$, $(12, 24)$, $(18, 24)$, $(18, 30)$, $(24, 30)$, $(18, 36)$. Επιπροσθέτως, συμπεριλαμβάνουμε στη μελέτη προσομοίωσης τον Bayesian D-βέλτιστο υπερκορεσμένο σχεδιασμό με $n = 15$ πειραματικές εκτελέσεις και $m = 20$ παράγοντες σε 3 μπλοκ, όπως αυτός πινακοποιείται από τους Jones et al. στον Πίνακα 2 που εσωκλείει η επιστημονική εργασία [118].

Η απόφαση σχετικά με την τιμή κατωφλιού w, η οποία στην πράξη καθορίζει το μέγιστο αριθμό των σημαντικών παραγόντων που μπορούν να αναγνωριστούν, παίζει σημαντικό ρόλο στην ανάπτυξη της μεθόδου. Στη πειραματική μελέτη μας θέσαμε το $w = m/2$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου πίνακα σχεδιασμού, δηλαδή ο αριθμός των παραγόντων υπό εξέταση. Η απόφαση για το w έγινε βάση της αρχής της σποραδικότητας [23]. Για να ποσοτικοποιήσουμε τη σποραδικότητα στο σύνολο των δεδομένων κατά τη διάρκεια των πειραμάτων προσομοίωσης, χρησιμοποιήσαμε την ακόλουθη διαδικασία: Οι πραγματικά ενεργές μεταβλητές επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών

παραγόντων και μόνο οι κύριες επιδράσεις ελήφθησαν υπόψη. Για κάθε πείραμα, εκτιμήσαμε όλες τις κύριες επιδράσεις, οι οποίες στη συνέχεια επισημάνθηκαν ως ενεργές ή μη ενεργές. Έπειτα υπολογίζοντας το ποσοστό των εκτιμώμενων ενεργών επιδράσεων που εμφανίζονται σε κάθε πείραμα, εκτιμήσαμε ότι ο αριθμός των ενεργών επιδράσεων που μπορεί να αναγνωριστεί, δεν υπερβαίνει το $w = \frac{m}{2}$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου σχεδιασμού.

Η απόφαση σχετικά με την τιμή κατωφλιού του SU μέτρου, η οποία καθορίζει το κατά πόσο ένας παράγοντας είναι σημαντικός ή όχι, παίζει επίσης σημαντικό ρόλο στην ανάπτυξη της μεθόδου. Για αυτόν το λόγο, πολλές διαφορετικές τιμές κατωφλιού (0.001, 0.01, 0.05, 0.1, 0.25, 0.4, median(**su**)) εξετάστηκαν με στόχο να βρεθεί η βέλτιστη τιμή κατωφλιού του SU για την προτεινόμενη μέθοδο. Οι πρώτες έξι τιμές κατωφλιού ήταν σταθεροί αριθμοί, οι οποίοι επιλέχθηκαν σύμφωνα με τις τιμές SU που παρατηρήθηκαν στα περισσότερα από τα πειράματά μας, ενώ η τιμή κατωφλιού median(**su**) βασίστηκε στις εκτιμώμενες τιμές του SU, συγκεκριμένα στο εύρος των τιμών του SU οι οποίες παρατηρήθηκαν στα πειράματά μας. Τελικά, καθορίσαμε την τιμή κατωφλιού του SU να είναι ίση με το median(**su**), δεδομένου ότι η επιλογή αυτή παρατηρήθηκε να είναι η βέλτιστη μετά από δοκιμές προσομοίωσης, δίνοντας ως αποτέλεσμα χαμηλότερες τιμές σφάλματος.

4.3.1 Εμπειρική Μελέτη

Για κάθε μία από τις 1000 επαναλήψεις στα πειράματα προσομοίωσης ακολουθήσαμε τα εξής βήματα:

1. Τα δεδομένα παράχθηκαν από το λογιστικό μοντέλο παλινδρόμησης όπως αυτό περιγράφεται στην ενότητα 5.2.1.

2. Οι πραγματικά ενεργές μεταβλήτες επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων. Από τις $1, \ldots, m$ στήλες του X, p στήλες ανατέθηκαν τυχαία στους ενεργούς παράγοντες με το $p \leq \frac{m}{2}$.

3. Οι συντελεστές των ενεργών παραγόντων ελήφθησαν ως εξής: ένα δείγμα μεγέθους p προήλθε από την κανονική κατανομή $N(\mu, 0.2)$, και τα ± 1 πρόσημα κατανεμήθηκαν τυχαία σε κάθε αριθμό.

4. Για τις μη ενεργές μεταβλητές, στο πραγματικό μοντέλο, οι συντελεστές τους ελήφθησαν τυχαία από την κανονική κατανομή $N(0, 0.2)$.

Να σημειώσουμε εδώ ότι μας ενδιαφέρει να μελετήσουμε μοντέλα που περιλαμβάνουν μόνο κύριες επιδράσεις, οπότε θεωρούμε ότι τα μοντέλα μας δεν συμπεριλαμβάνουν αλληλεπιδράσεις. Επιπλέον, η κατανομή των αντιθέσεων δεν επηρεάζεται από το το μέτρο (magnitude) των συντελεστών, αλλά εξαρτάται από το σχετικό μέγεθος (relative size) των συντελεστών. Δεδομένου ότι οι συνθήκες στην πράξη είναι συνήθως διαφορετικές από εκείνες των προσομοιώσεων και ότι ο πειραματιστής δεν μπορεί να γνωρίζει εκ των προτέρων πόσοι και ποιοί παράγοντες μπορεί να είναι ενεργοί, προκειμένου να εξετάσουμε το πόσο ευαίσθητα είναι τα αποτελέσματα στην επιλογή και τον αριθμό των ενεργών στηλών-παραγόντων, θεωρήσαμε την τυχαία ανάθεση των ενεργών παραγόντων στις στήλες, από τη στιγμή που είναι σημαντικό να αποφύγουμε την οποιαδήποτε μεροληψία επιλογής. Επίσης, η επιλογή των κατανομών στα προηγούμενα βήματα 3 και 4, εξασφαλίζει γενικά το διαχωρισμό μεταξύ των συντελεστών των ενεργών και ανενεργών παραγόντων. Για το λόγο αυτό, θεωρήσαμε ένα ευρύ φάσμα προσομειωμένων μοντέλων που ήταν διαφορετικά από την άποψη αυτή.

Το μέγεθος των συντελεστών παλινδρόμησης για κάθε έναν από τους p ενεργούς παράγοντες προήλθε τυχαία από την κανονική κατανομή $N(\mu, 0.2)$ για τα ακόλουθα σενάρια:

- Σενάριο Ι: οι συντελεστές παίρνουν τυχαίες τιμές από το διάνυσμα $\beta \sim N(\mu = 4, 0.2)$,

- Σενάριο II: οι συντελεστές παίρνουν τυχαίες τιμές από το διάνυσμα $\beta \sim N(\mu = 5, 0.2)$,
- Σενάριο III: οι συντελεστές παίρνουν τυχαίες τιμές από το διάνυσμα $\beta \sim N(\mu = 8, 0.2)$.

Στην εμπειρική μελέτη μας χρησιμοποιήσαμε όλα τα μοντέλα τα οποία απαριθμούνται στον Πίνακα 4.1.

Στα πειράματά μας εκτελέσαμε τις προσομοιώσεις 1000 φορές και τα ληφθέντα αποτελέσματα συνοψίζονται στον Πίνακα 4.2. Συγκεκριμένα, στον Πίνακα 4.2, η πρώτη στήλη αναφέρεται στον αριθμό που αντιστοιχεί σε κάθε μοντέλο που χρησιμοποιείται. Οι στήλες που ονομάζονται ως "Type I" και "Type II", αναφέρονται στις τιμές των σφαλμάτων Τύπου I και Τύπου II που αντιστοιχούν στην κάθε μέθοδο που εξετάσαμε.

Αρχικά, από τον Πίνακα 4.2 προκύπτει ότι η προτεινόμενη μέθοδος έχει μικρές τιμές για τα σφάλματα Τύπου I για σχεδόν όλα τα μοντέλα. Επίσης, από τον Πίνακα 4.2 μπορούμε να παρατηρήσουμε ότι η προτεινόμενη μέθοδος επιτυγχάνει για όλα τα μοντέλα τα σφάλματα Τύπου II να είναι μηδέν, γεγονός το οποίο μας υποδεικνύει ότι καμία από τις ενεργές επιδράσεις δεν ανιχνεύεται λανθασμένα ως ανενεργή, ή με άλλα λόγια όλες οι ενεργές επιδράσεις ανιχνεύονται σωστά με 100% επιτυχία. Από τον Πίνακα 4.2 προκύπτει ότι η προτεινόμενη μέθοδος παρουσιάζει την τάση να δηλώνει με ένα υψηλότερο ποσοστό τις ανενεργές επιδράσεις ως ενεργές ενώ με μηδενικό ποσοστό τις ενεργές επιδράσεις ως ανενεργές. Υπό αυτή την έννοια, η προτεινόμενη μέθοδος θα μπορούσε πράγματι να χαρακτηριστεί ως συντηρητική.

Επιπλέον, παρατηρούμε ότι η ROC μέθοδος φαίνεται να έχει και τα δύο αυτά ποσοστά σφάλματος σημαντικά υψηλότερα σε σύγκριση με την προτεινόμενη μέθοδο. Αμφότερα τα σφάλματα τα οποία προέκυψαν από την ROC μεθοδολογία είναι πολύ υψηλά για όλα τα θεωρούμενα μοντέλα. Τα αποτελέσματα των προσομοιώσεων από την ROC μεθοδολογία υποδεικνύουν μοντέλα τυχαίας πρόβλεψης (random guess), και ως εκ τούτου όχι βέλτιστα, όπως θα μπορούσε να δει κάποιος εύκολα από τα δύο ποσοστά σφάλματος τα οποία είναι περίπου κοντά στο 50%. Ως εκ τούτου, η προτεινόμενη μέθοδος παρουσιάζει πολύ καλύτερη απόδοση στην εύρεση των σημαντικών παραγόντων στους υπερκορεσμένους σχεδιασμούς θεωρώντας δίτιμες αποκρίσεις, και κατέχει σημαντικά μικρότερα ποσοστά σφάλματος σε σύγκριση με τη μέθοδο ROC ανάλυσης.

Πίνακας 4.1: Μοντέλα 3ης μελέτης προσομοίωσης

Μοντέλο	ΥΣ από	m	n	β
1	[148]	10	6	$\beta \sim N(\mu = 4, 0.2)$
2	[148]	10	6	$\beta \sim N(\mu = 5, 0.2)$
3	[148]	10	6	$\beta \sim N(\mu = 8, 0.2)$
4	[130]	22	12	$\beta \sim N(\mu = 4, 0.2)$
5	[130]	22	12	$\beta \sim N(\mu = 5, 0.2)$
6	[130]	22	12	$\beta \sim N(\mu = 8, 0.2)$
7	[130]	18	10	$\beta \sim N(\mu = 4, 0.2)$
8	[130]	18	10	$\beta \sim N(\mu = 5, 0.2)$
9	[130]	18	10	$\beta \sim N(\mu = 8, 0.2)$
10	[130]	26	14	$\beta \sim N(\mu = 4, 0.2)$
11	[130]	26	14	$\beta \sim N(\mu = 5, 0.2)$
12	[130]	26	14	$\beta \sim N(\mu = 8, 0.2)$
13	[130]	30	16	$\beta \sim N(\mu = 4, 0.2)$
14	[130]	30	16	$\beta \sim N(\mu = 5, 0.2)$
15	[130]	30	16	$\beta \sim N(\mu = 8, 0.2)$
16	[130]	34	18	$\beta \sim N(\mu = 4, 0.2)$
17	[130]	34	18	$\beta \sim N(\mu = 5, 0.2)$
18	[130]	34	18	$\beta \sim N(\mu = 8, 0.2)$
19	[130]	38	20	$\beta \sim N(\mu = 4, 0.2)$
20	[130]	38	20	$\beta \sim N(\mu = 5, 0.2)$
21	[130]	38	20	$\beta \sim N(\mu = 8, 0.2)$
22	[130]	42	22	$\beta \sim N(\mu = 4, 0.2)$
23	[130]	42	22	$\beta \sim N(\mu = 5, 0.2)$
24	[130]	42	22	$\beta \sim N(\mu = 8, 0.2)$
25	[130]	10	6	$\beta \sim N(\mu = 4, 0.2)$
26	[130]	10	6	$\beta \sim N(\mu = 5, 0.2)$
27	[130]	10	6	$\beta \sim N(\mu = 8, 0.2)$
28	[130]	14	8	$\beta \sim N(\mu = 4, 0.2)$
29	[130]	14	8	$\beta \sim N(\mu = 5, 0.2)$
30	[130]	14	8	$\beta \sim N(\mu = 8, 0.2)$
31	[215]	22	12	$\beta \sim N(\mu = 4, 0.2)$
32	[215]	22	12	$\beta \sim N(\mu = 5, 0.2)$
33	[215]	22	12	$\beta \sim N(\mu = 8, 0.2$
34	[118]	16	12	$\beta \sim N(\mu = 4, 0.2)$
35	[118]	16	12	$\beta \sim N(\mu = 5, 0.2)$
36	[118]	16	12	$\beta \sim N(\mu = 8, 0.2)$
37	[118]	18	12	$\beta \sim N(\mu = 4, 0.2)$
38	[118]	18	12	$\beta \sim N(\mu = 5, 0.2)$
39	[118]	18	12	$\beta \sim N(\mu = 8, 0.2)$
40	[118]	24	12	$\beta \sim N(\mu = 4, 0.2)$
41	[118]	24	12	$\beta \sim N(\mu = 5, 0.2)$
42	[118]	24	12	$\beta \sim N(\mu = 8, 0.2)$
43	[118]	24	18	$\beta \sim N(\mu = 4, 0.2)$
44	[118]	24	18	$\beta \sim N(\mu = 5, 0.2)$
45	[118]	24	18	$\beta \sim N(\mu = 8, 0.2)$
46	[118]	30	18	$\beta \sim N(\mu = 4, 0.2)$
47	[118]	30	18	$\beta \sim N(\mu = 5, 0.2)$
48	[118]	30	18	$\beta \sim N(\mu = 8, 0.2)$
49	[118]	30	24	$\beta \sim N(\mu = 4, 0.2)$
50	[118]	30	24	$\beta \sim N(\mu = 5, 0.2)$
51	[118]	30	24	$\beta \sim N(\mu = 8, 0.2)$
52	[118]	36	18	$\beta \sim N(\mu = 4, 0.2)$
53	[118]	36	18	$\beta \sim N(\mu = 5, 0.2)$
54	[118]	36	18	$\beta \sim N(\mu = 8, 0.2)$
55	[118]	20	15	$\beta \sim N(\mu = 4, 0.2)$
56	[118]	20	15	$\beta \sim N(\mu = 5, 0.2)$
57	[118]	20	15	$\beta \sim N(\mu = 8, 0.2)$

Πίνακας 4.2: Απόδοση της 3ης προτεινόμενης μεθόδου για τα μοντέλα 1-57 με χρήση 1000 προσομοιώσεων

Μέθοδος	Algorithm		ROC	
Μοντέλο	Type I	Type II	Type I	Type II
1	0.27	0.00	0.40	0.38
2	0.28	0.00	0.40	0.40
3	0.32	0.00	0.40	0.40
4	0.26	0.00	0.45	0.44
5	0.29	0.00	0.45	0.44
6	0.35	0.00	0.45	0.45
7	0.17	0.00	0.42	0.42
8	0.21	0.00	0.44	0.42
9	0.25	0.00	0.44	0.44
10	0.26	0.00	0.46	0.44
11	0.29	0.00	0.46	0.46
12	0.31	0.00	0.46	0.46
13	0.27	0.00	0.47	0.46
14	0.31	0.00	0.47	0.47
15	0.31	0.00	0.47	0.47
16	0.24	0.00	0.47	0.47
17	0.26	0.00	0.47	0.47
18	0.29	0.00	0.47	0.47
19	0.26	0.00	0.47	0.46
20	0.27	0.00	0.47	0.46
21	0.31	0.00	0.48	0.47
22	0.30	0.00	0.47	0.45
23	0.31	0.00	0.47	0.47
24	0.33	0.00	0.47	0.47
25	0.17	0.00	0.40	0.40
26	0.22	0.00	0.40	0.45
27	0.29	0.00	0.40	0.45
28	0.21	0.00	0.43	0.43
29	0.22	0.00	0.43	0.43
30	0.25	0.00	0.45	0.43
31	0.30	0.00	0.45	0.43
32	0.33	0.00	0.45	0.45
33	0.35	0.00	0.45	0.45
34	0.17	0.00	0.40	0.40
35	0.22	0.00	0.40	0.45
36	0.29	0.00	0.40	0.45
37	0.21	0.00	0.43	0.43
38	0.22	0.00	0.43	0.43
39	0.25	0.00	0.45	0.43
40	0.22	0.00	0.40	0.38
41	0.24	0.00	0.40	0.40
42	0.28	0.00	0.40	0.40
43	0.24	0.00	0.47	0.47
44	0.26	0.00	0.47	0.47
45	0.29	0.00	0.47	0.47
46	0.26	0.00	0.46	0.44
47	0.29	0.00	0.46	0.46
48	0.31	0.00	0.46	0.46
49	0.27	0.00	0.47	0.46
50	0.31	0.00	0.47	0.47
51	0.31	0.00	0.47	0.47
52	0.30	0.00	0.47	0.45
53	0.31	0.00	0.47	0.47
54	0.33	0.00	0.47	0.47
55	0.30	0.00	0.45	0.43
56	0.33	0.00	0.45	0.45
57	0.37	0.00	0.45	0.45

4.3.2 Συγκριτική Μελέτη

Ο προτεινόμενος αλγόριθμος που παρουσιάσαμε παραπάνω, εκτιμά το κέρδος πληροφορίας χρησιμοποιώντας την εντροπία κατά Shannon. Προκειμένου να εξετάσουμε το αν η προτεινόμενη μέθοδος μπορεί να γενικευτεί και να θεωρηθεί εύρωστη σε σχέση με την επιλογή της εντροπίας, εξετάσαμε περαιτέρω την απόδοση της προτεινόμενης μεθόδου, χρησιμοποιώντας τα μέτρα εντροπίας Rényi, Tsallis και Havrda-Charvát όπως αυτά παρουσιάστηκαν λεπτομερώς στις ενότητες 1.1.1 και 1.1.2. Στα αποτελέσματα που παρουσιάζονται παρακάτω η προτεινόμενη μέθοδος εκτιμά το κέρδος πληροφορίας χρησιμοποιώντας κάθε μία από τις τρεις προαναφερθείσες εντροπίες. Για την Rényi εντροπία, θεωρήσαμε την παράμετρο α ίση με το 2, για την Tsallis εντροπία θεωρήσαμε την παράμετρο q πάλι ίση με το 2, ενώ για την Havrda-Charvát εντροπία θεωρήσαμε την παράμετρο β ίση με το 3. Ας σημειωθεί εδώ ότι όταν το $q = 2$, η Tsallis εντροπία ισοδυναμεί με το συντελεστή Gini (ή δείκτη Gini) [93]. Σύμφωνα με τη βιβλιογραφία που υπάρχει σχετικά με τα ανωτέρω μέτρα εντροπίας, οι επιλεγμένες τιμές παραμέτρων φαίνεται να αποτελούν την κλασική και ιδανική επιλογή, και αυτό αποτέλεσε το κριτήριο για μας ώστε να επιλέξουμε τις συγκεκριμένες τιμές παραμέτρων.

Στα πειράματά μας εκτελέσαμε τις προσομοιώσεις 1000 φορές και τα ληφθέντα αποτελέσματα συνοψίζονται στον Πίνακα 4.3. Συγκεκριμένα, στον Πίνακα 4.3, η πρώτη στήλη αναφέρεται στον αριθμό που αντιστοιχεί σε κάθε μοντέλο που χρησιμοποιείται. Οι στήλες που ονομάζονται ως "Type I" και "Type II", αναφέρονται στις τιμές των σφαλμάτων Τύπου I και Τύπου II που αντιστοιχούν σε κάθε περίπτωση εντροπίας, δηλαδή την Rényi, Tsallis και Havrda-Charvát εντροπία.

Εξετάσαμε τη μέθοδο ξεχωριστά για κάθε μία εντροπία και από τον Πίνακα 4.3 παρατηρούμε ότι τα σφάλματα Τύπου I και Τύπου II, αντίστοιχα, είναι τα ίδια για κάθε μία από τις τρεις διαφορετικές περιπτώσεις εντροπίας, καθώς και για την Shannon εντροπία (όπως φαίνεται από τα αποτελέσματα τα οποία παρουσιάζονται στη στήλη "Algorithm" του Πίνακα 4.2). Με άλλα λόγια, η προτεινόμενη μέθοδος μπορεί να θεωρηθεί εύρωστη σε σχέση με την επιλογή της εντροπίας, δεδομένου ότι παρουσιάζει την ίδια ικανότητα αναγνώρισης των πραγματικά ενεργών και μη ενεργών παραγόντων για κάθε μία από τις εντροπίες, δηλαδή την Shannon, Rényi, Tsallis και Havrda-Charvát εντροπία.

Η απόφαση σχετικά με την τιμή κατωφλιού του SU μέτρου, η οποία καθορίζει το κατά πόσο ένας παράγοντας είναι σημαντικός ή όχι, διαδραματίζει όπως προαναφέραμε σημαντικό ρόλο στην ανάπτυξη της μεθόδου. Για αυτόν το λόγο, πολλές διαφορετικές τιμές κατωφλιού (0.001, 0.01, 0.05, 0.1, 0.25, 0.4, median(**su**)) εξετάστηκαν για κάθε μία από τις θεωρούμενες εντροπίες, με στόχο να βρεθεί η βέλτιστη τιμή κατωφλιού του SU για την προτεινόμενη μέθοδο και να εξεταστεί η ευρωστία της μεθόδου υπό αυτή την έννοια.

Για κάθε ένα από τα θεωρούμενα μοντέλα, 1000 σύνολα δεδομένων παρήχθησαν και τα ληφθέντα αποτελέσματα μετά την εφαρμογή της μεθόδου, παρουσιάζονται στον Πίνακα 4.4 σε αντιστοιχία με τις τιμές κατωφλιού. Στην πρώτη γραμμή του Πίνακα 4.4, παρουσιάζεται η επιλεγμένη τιμή κατωφλιού. Η πρώτη στήλη αναφέρεται στον αριθμό που αντιστοιχεί σε κάθε μοντέλο που χρησιμοποιείται. Οι στήλες που ονομάζονται ως "Type I" και "Type II", αναφέρονται στις τιμές των σφαλμάτων Τύπου I και Τύπου II που αντιστοιχούν σε κάθε τιμή κατωφλιού.

Από τον Πίνακα 4.4, μπορούμε εύκολα να παρατηρήσουμε ότι τα σφάλματα Τύπου I είναι τα ίδια για όλα τα μοντέλα και για τις επτά διαφορετικές τιμές κατωφλιού. Παρατηρούμε επίσης από τον Πίνακα 4.4 ότι στα περισσότερα μοντέλα (30/57), τα σφάλματα Τύπου II βρέθηκαν να είναι ίσα με το 0.00 για τις επτά διαφορετικές τιμές κατωφλιού, και αυτά τα μοντέλα αποτελούν περιπτώσεις στις οποίες όλες οι ενεργές επιδράσεις ανιχνεύονται σωστά. Συμπεραίνουμε από τα αποτελέσματα που προέκυψαν ότι η προτεινόμενη μέθοδος θα μπορούσε να χαρακτηριστεί εύρωστη σε σχέση με την επιλογή της τιμής κατωφλιού του SU μέτρου. Τελικά, καθορίσαμε την τιμή κατωφλιού του SU για την προτεινόμενη μέθοδο να είναι η median(**su**), δεδομένου ότι η τιμή αυτή παρατηρήθηκε να είναι η βέλτιστη επιλογή η οποία

είχε ως αποτέλεσμα μηδενικές τιμές σφάλματος Τύπου II για όλα τα θεωρούμενα μοντέλα. Σημειώνουμε εδώ ότι όταν θεωρήσαμε ως τιμές κατωφλιού τις 0.001, 0.01, 0.05, 0.1, 0.25 και 0.4, τότε στα υπόλοιπα 27 μοντέλα (από τα 57 στο σύνολο, βλέπε Πίνακα 4.4), τα σφάλματα Τύπου II που προέκυψαν ήταν υψηλότερα από τα αντίστοιχα μηδενικά σφάλματα τα οποία προέκυψαν για την τιμή κατωφλιού ίση με median(**su**).

Πίνακας 4.3: Συγκριτική απόδοση της 3ης προτεινόμενης μεθόδου για διαφορετικές εντροπίες για τα μοντέλα 1-57 με χρήση 1000 προσομοιώσεων

Εντροπία	Rényi		Tsallis		Havrda-Charvát	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II
1	0.27	0.00	0.27	0.00	0.27	0.00
2	0.28	0.00	0.28	0.00	0.28	0.00
3	0.32	0.00	0.32	0.00	0.32	0.00
4	0.26	0.00	0.26	0.00	0.26	0.00
5	0.29	0.00	0.29	0.00	0.29	0.00
6	0.35	0.00	0.35	0.00	0.35	0.00
7	0.17	0.00	0.17	0.00	0.17	0.00
8	0.21	0.00	0.21	0.00	0.21	0.00
9	0.25	0.00	0.25	0.00	0.25	0.00
10	0.26	0.00	0.26	0.00	0.26	0.00
11	0.29	0.00	0.29	0.00	0.29	0.00
12	0.31	0.00	0.31	0.00	0.31	0.00
13	0.27	0.00	0.27	0.00	0.27	0.00
14	0.31	0.00	0.31	0.00	0.31	0.00
15	0.31	0.00	0.31	0.00	0.31	0.00
16	0.24	0.00	0.24	0.00	0.24	0.00
17	0.26	0.00	0.26	0.00	0.26	0.00
18	0.29	0.00	0.29	0.00	0.29	0.00
19	0.26	0.00	0.26	0.00	0.26	0.00
20	0.27	0.00	0.27	0.00	0.27	0.00
21	0.31	0.00	0.31	0.00	0.31	0.00
22	0.30	0.00	0.30	0.00	0.30	0.00
23	0.31	0.00	0.31	0.00	0.31	0.00
24	0.33	0.00	0.33	0.00	0.33	0.00
25	0.17	0.00	0.17	0.00	0.17	0.00
26	0.22	0.00	0.22	0.00	0.22	0.00
27	0.29	0.00	0.29	0.00	0.29	0.00
28	0.21	0.00	0.21	0.00	0.21	0.00
29	0.22	0.00	0.22	0.00	0.22	0.00
30	0.25	0.00	0.25	0.00	0.25	0.00
31	0.30	0.00	0.30	0.00	0.30	0.00
32	0.33	0.00	0.33	0.00	0.33	0.00
33	0.35	0.00	0.35	0.00	0.35	0.00
34	0.17	0.00	0.17	0.00	0.17	0.00
35	0.22	0.00	0.22	0.00	0.22	0.00
36	0.29	0.00	0.29	0.00	0.29	0.00
37	0.21	0.00	0.21	0.00	0.21	0.00
38	0.22	0.00	0.22	0.00	0.22	0.00
39	0.25	0.00	0.25	0.00	0.25	0.00
40	0.22	0.00	0.22	0.00	0.22	0.00
41	0.24	0.00	0.24	0.00	0.24	0.00
42	0.28	0.00	0.28	0.00	0.28	0.00
43	0.24	0.00	0.24	0.00	0.24	0.00
44	0.26	0.00	0.26	0.00	0.26	0.00
45	0.29	0.00	0.29	0.00	0.29	0.00
46	0.26	0.00	0.26	0.00	0.26	0.00
47	0.29	0.00	0.29	0.00	0.29	0.00
48	0.31	0.00	0.31	0.00	0.31	0.00
49	0.27	0.00	0.27	0.00	0.27	0.00
50	0.31	0.00	0.31	0.00	0.31	0.00
51	0.31	0.00	0.31	0.00	0.31	0.00
52	0.30	0.00	0.30	0.00	0.30	0.00
53	0.31	0.00	0.31	0.00	0.31	0.00
54	0.33	0.00	0.33	0.00	0.33	0.00
55	0.30	0.00	0.30	0.00	0.30	0.00
56	0.33	0.00	0.33	0.00	0.33	0.00
57	0.37	0.00	0.37	0.00	0.37	0.00

Πίνακας 4.4: Συγκριτική απόδοση της 3ης προτεινόμενης μεθόδου για διαφορετικές επιλογές παραμέτρων για τα μοντέλα 1-57 με χρήση 1000 προσομοιώσεων

$SU \geq$	$median(\mathbf{su})$		0.001		0.01		0.05		0.1		0.25		0.4	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
1	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00
2	0.28	0.00	0.28	0.00	0.28	0.00	0.28	0.00	0.28	0.25	0.28	0.00	0.28	0.00
3	0.32	0.00	0.32	0.25	0.32	0.25	0.32	0.25	0.32	0.25	0.32	0.25	0.32	0.25
4	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00
5	0.29	0.00	0.29	0.03	0.29	0.03	0.29	0.03	0.29	0.03	0.29	0.03	0.29	0.03
6	0.35	0.00	0.35	0.29	0.35	0.29	0.35	0.29	0.35	0.29	0.35	0.29	0.35	0.29
7	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00
8	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00
9	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00
10	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00
11	0.29	0.00	0.29	0.30	0.29	0.30	0.29	0.30	0.29	0.30	0.29	0.30	0.29	0.30
12	0.31	0.00	0.31	0.33	0.31	0.33	0.31	0.33	0.31	0.33	0.31	0.33	0.31	0.33
13	0.27	0.00	0.27	0.10	0.27	0.10	0.27	0.10	0.27	0.10	0.27	0.10	0.27	0.10
14	0.31	0.00	0.31	0.10	0.31	0.10	0.31	0.10	0.31	0.10	0.31	0.10	0.31	0.10
15	0.31	0.00	0.31	0.10	0.31	0.10	0.31	0.10	0.31	0.10	0.31	0.10	0.31	0.10
16	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00
17	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00
18	0.29	0.00	0.29	0.15	0.29	0.15	0.29	0.15	0.29	0.15	0.29	0.15	0.29	0.15
19	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00
20	0.27	0.00	0.27	0.24	0.27	0.24	0.27	0.24	0.27	0.24	0.27	0.24	0.27	0.24
21	0.31	0.00	0.31	0.27	0.31	0.27	0.31	0.27	0.31	0.27	0.31	0.27	0.31	0.27
22	0.30	0.00	0.30	0.15	0.30	0.15	0.30	0.15	0.30	0.15	0.30	0.15	0.30	0.15
23	0.31	0.00	0.31	0.15	0.31	0.15	0.31	0.15	0.31	0.15	0.31	0.15	0.31	0.15
24	0.33	0.00	0.33	0.17	0.33	0.17	0.33	0.17	0.33	0.17	0.33	0.17	0.33	0.17
25	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00
26	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00
27	0.29	0.00	0.29	0.25	0.29	0.25	0.29	0.25	0.29	0.25	0.29	0.25	0.29	0.25
28	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00
29	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00
30	0.25	0.00	0.25	0.17	0.25	0.17	0.25	0.17	0.25	0.17	0.25	0.17	0.25	0.17
31	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00
32	0.33	0.00	0.33	0.00	0.33	0.00	0.33	0.00	0.33	0.00	0.33	0.00	0.33	0.00
33	0.35	0.00	0.35	0.10	0.35	0.10	0.35	0.10	0.35	0.10	0.35	0.10	0.35	0.10
34	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00	0.17	0.00
35	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00
36	0.29	0.00	0.29	0.07	0.29	0.07	0.29	0.07	0.29	0.07	0.29	0.07	0.29	0.07
37	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00	0.21	0.00
38	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00
39	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00
40	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00	0.22	0.00
41	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00
42	0.28	0.00	0.28	0.14	0.28	0.14	0.28	0.14	0.28	0.14	0.28	0.14	0.28	0.14
43	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00	0.24	0.00
44	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00
45	0.29	0.00	0.29	0.18	0.29	0.18	0.29	0.18	0.29	0.18	0.29	0.18	0.29	0.18
46	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00	0.26	0.00
47	0.29	0.00	0.29	0.21	0.29	0.21	0.29	0.21	0.29	0.21	0.29	0.21	0.29	0.21
48	0.31	0.00	0.31	0.24	0.31	0.24	0.31	0.24	0.31	0.24	0.31	0.24	0.31	0.24
49	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00	0.27	0.00
50	0.31	0.00	0.31	0.22	0.31	0.22	0.31	0.22	0.31	0.22	0.31	0.22	0.31	0.22
51	0.31	0.00	0.31	0.00	0.31	0.00	0.31	0.00	0.31	0.00	0.31	0.00	0.31	0.00
52	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00
53	0.31	0.00	0.31	0.09	0.31	0.09	0.31	0.09	0.31	0.09	0.31	0.09	0.31	0.09
54	0.33	0.00	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09	0.33	0.09
55	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00	0.30	0.00
56	0.33	0.00	0.33	0.24	0.33	0.24	0.24	0.24	0.33	0.24	0.24	0.24	0.33	0.24
57	0.37	0.00	0.37	0.28	0.37	0.28	0.37	0.28	0.37	0.28	0.37	0.28	0.37	0.28

4.4 Συμπεράσματα

Σε αυτό το κεφάλαιο, παρουσιάσαμε έναν συνεκτικό αλγόριθμο της Θεωρίας Πληροφοριών (Information Theoretical Algorithm-ITA) για την εύρεση και επιλογή των σημαντικών παραγόντων, αναλύοντας δεδομένα από υπερκορεσμένους σχεδιασμούς και θεωρώντας το λογιστικό μοντέλο παλινδρόμησης. Η εμπειρική απόδοση του προτεινόμενου αλγορίθμου, βάση της εκτενούς μελέτης προσομοιώσεων, υποδεικνύει ότι η προτεινόμενη μέθοδος έχει χαμηλά ποσοστά σφάλματος, αποδίδει πολύ καλά, και ότι είναι εύρωστη από τη στιγμή που είτε με τη χρήση διαφορετικών μέτρων εντροπίας είτε με τη χρήση διαφορετικών τιμών κατωφλιού προέκυψαν παρόμοια αποτελέσματα.

Όπως προαναφέραμε στην ενότητα 1.4, μέχρι στιγμής δεν μπορεί να βρεθεί κάποια μέθοδος ανάλυσης των υπερκορεσμένων σχεδιασμών η οποία όταν χρησιμοποιείται να αναγνωρίζονται οι σημαντικοί παράγοντες και να εκτιμούνται με αμελητέο σφάλμα, πρόβλημα το οποίο απορρέει εξαιτίας της μη ορθογωνιότητας του πίνακα σχεδιασμού η οποία καθιστά αναπόφευκτο το να γίνεται ανεκτή η ύπαρξη μιας μικρής μεροληψίας ανάμεσα σε όλες τις εκτιμημένες επιδράσεις. Βάση των παραπάνω είναι αξιοσημείωτο το γεγονός ότι η προτεινόμενη μέθοδος επιτυγχάνει να ανιχνεύσει όλες τις πραγματικά ενεργές κύριες επιδράσεις με αμελητέο σφάλμα. Το γεγονός ότι η προτεινόμενη μέθοδος επιτυγχάνει μηδενικές τιμές για τα σφάλματα Τύπου II είναι ιδιαίτερα σημαντικό προκειμένου να μην παραλείπονται οι σημαντικοί παράγοντες του μοντέλου.

Όπως προαναφέραμε τα σφάλματα Τύπου II δημιουργούν αρκετά προβλήματα όπως άλλωστε και τα σφάλματα Τύπου I, δεδομένου ότι μπορεί να οδηγήσουν σε περιττό υπολογιστικό κόστος στα πειράματα που ακολουθούν, και κάτι τέτοιο μπορεί να είναι αρκετά επιζήμιο σε ένα πείραμα το οποίο έχει άμεσο αντίκτυπο στην πράξη. Όπως με οποιαδήποτε μέθοδο ανάλυσης υπερκορεσμένων σχεδιασμών, τα σφάλματα Τύπου I και Τύπου II είναι εξίσου σημαντικά και θα πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα. Η προτεινόμενη μέθοδος φαίνεται να επιτυγχάνει κάτι τέτοιο και να συμπεριφέρεται ιδιαίτερα ικανοποιητικά, δεδομένου ότι το μέσο ποσοστό σφάλματος Τύπου I των 57 μοντέλων σε 1000 προσομοιώσεις ήταν της τάξης του 27% και το αντίστοιχο ποσοστό για τα σφάλματα Τύπου II ήταν μηδενικό. Το μέσο ποσοστό σφάλματος Τύπου I της τάξης του 27% το οποίο προέκυψε, δεν μας προβληματίζει ιδιαίτερα, δεδομένου ότι σε καταστάσεις που ισχύει η αρχή της σποραδικότητας των επιδράσεων, είναι πολύ πιθανό να προκύψουν σφάλματα Τύπου I, όπως διαπιστώθηκε από τον Lin [149]. Η προτεινόμενη μέθοδος διατηρεί τα σφάλματα Τύπου I σε ικανοποιητικά χαμηλά επίπεδα, το οποίο σημαίνει ότι έχει την ικανότητα να αποκλείει τους περιττούς παράγοντες, και ο ερευνητής βασισμένος μόνο στους παράγοντες που επιλέχθηκαν ως σημαντικοί είναι σε θέση πλέον να μειώσει το υπολογιστικό κόστος των επιπλέον πειραμάτων που θα εκτελέσει. Όμως, δεδομένου ότι οι υπερκορεσμένοι σχεδιασμοί κυρίως χρησιμοποιούνται σε πειράματα κρησαρίσματος με στόχο να αναγνωρίσουν εκείνους τους παράγοντες που θα πρέπει να εξεταστούν περαιτέρω, οι μηδενικές τιμές σφάλματος Τύπου II της προτεινόμενης μεθόδου είναι αξιοσημείωτες και πόσο μάλλον κάτω από την υπόθεση ενός γενικευμένου γραμμικού μοντέλου, συγκεκριμένα ενός λογιστικού μοντέλου παλινδρόμησης, μέθοδος η οποία είναι μη τετριμμένη και μη αναπτυγμένη στην έως σήμερα διεθνή βιβλιογραφία, όπου συναντάμε το πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών μόνο κάτω υπό τη σκέπη του κλασικού γραμμικού μοντέλου.

ΚΕΦΑΛΑΙΟ **5**

Επέκταση του Προβλήματος σε Γενικευμένα Γραμμικά Μοντέλα με Διακριτή Απόκριση

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

—*George B. Dantzig (1914–2005)*

Στο πέμπτο αυτό κεφάλαιο παρουσιάζεται η γενίκευση του προβλήματος ανάλυσης των υπερκορεσμένων σχεδιασμών, για τους οποίους οι επεξηγηματικές μεταβλητές βρίσκονται σε δύο επίπεδα και τα δεδομένα απόκρισης είναι πλέον διακριτά. Ο αλγόριθμος της Θεωρίας Πληροφοριών, όπως αυτός παρουσιάστηκε στο Κεφάλαιο 4 καθώς επίσης και στην επιστημονική εργασία [7], εφαρμόζεται μονάχα όταν τα δεδομένα απόκρισης είναι δίτιμα, δηλαδή προέρχονται από το λογιστικό μοντέλο παλινδρόμησης. Εδώ επεκτείνουμε την αλγοριθμική προσέγγιση του προβλήματος ανάλυσης υπερκορεσμένων σχεδιασμών, θεωρώντας γενικευμένα γραμμικά μοντέλα των οποίων τα δεδομένα απόκρισης προέρχονται από την Poisson, γεωμετρική και αρνητική διωνυμική κατανομή. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [8].

5.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Οι μέθοδοι ανάλυσης που παρουσιάσαμε στα προηγούμενα Κεφάλαια 2, 3 και 4 αποτελούν μεθόδους ανάλυσης υπερκορεσμένων σχεδιασμών κάτω από την υπόθεση ενός λογιστικού μοντέλου παλινδρόμησης. Οι τιμές των σφαλμάτων που προέκυψαν από τις δύο πρώτες μεθόδους ήταν ικανοποιητικά χαμηλές, αλλά οι τιμές των σφαλμάτων Τύπου I και Τύπου II που προέκυψαν από την τρίτη μέθοδο ήταν οι καλύτερες μέχρι στιγμής, συγκεκριμένα για τα σφάλματα Τύπου I το ποσοστό ήταν της τάξης του 27% και το αντίστοιχο ποσοστό για τα σφάλματα Τύπου II ήταν μηδενικό. Η τρίτη μέθοδος ανάλυσης [7] ανιχνεύει όλες τις πραγματικά ενεργές κύριες επιδράσεις με αμελητέο σφάλμα, δεδομένου ότι επιτυγχάνει μηδενικές τιμές για τα σφάλματα Τύπου II, το οποίο είναι ιδιαίτερα σημαντικό, προκειμένου να μην παραλείπονται οι σημαντικοί παράγοντες του μοντέλου. Σκεφτήκαμε λοιπόν να επεκτείνουμε τον αλγόριθμο επιλογής μεταβλητών [7] για την ανάλυση των υπερκορεσμένων σχεδιασμών, με στόχο τη γρήγορη και αποτελεσματική αναγνώριση των σημαντικών επιδράσεων, κάτω από την υπόθεση όχι μόνο ενός λογιστικού μοντέλου παλινδρόμησης, αλλά γενικά ενός γενικευμένου γραμμικού μοντέλου με διακριτή απόκριση.

5.2 Γενικευμένα Γραμμικά Μοντέλα

Γενικευμένα γραμμικά μοντέλα (generalized linear models) χρησιμοποιήσαμε για την ανάπτυξη της προτεινόμενης μεθόδου ανάλυσης των υπερκορεσμένων σχεδιασμών. Τα γενικευμένα γραμμικά μοντέλα αποτελούν μια κλάση στατιστικών μοντέλων που είναι φυσική γενίκευση και επέκταση των κλασικών γραμμικών μοντέλων [164]. Τα γενικευμένα γραμμικά μοντέλα αναπτύχθηκαν το 1972 από τους Nelder και Wedderburn [171]. Ιδιαίτερα σημαντική στη θεωρία που αναπτύχθηκε γύρω από τα γενικευμένα γραμμικά μοντέλα είναι και η συμβολή των McCullagh και Nelder [161]. Τα γενικευμένα γραμμικά μοντέλα αναπτύχθηκαν για να επιτρέψουν στον πειραματιστή να προσαρμόσει μοντέλα παλινδρόμησης των οποίων τα δεδομένα απόκρισης ακολουθούν μια πολύ γενική κατανομή, η οποία καλείται μονοπαραμετρική εκθετική οικογένεια κατανομών. Η εκθετική οικογένεια κατονομών περιλαμβάνει κατανομές όπως η κανονική, η διωνυμική, η Poisson, η γεωμετρική, η αρνητική διωνυμική, η εκθετική, η γάμμα και η αντίστροφη κανονική κατανομή. Γενικό και σημαντικό χαρακτηριστικό των γενικευμένων γραμμικών μοντέλων είναι ότι η διακύμανση της απόκρισης σχετίζεται μέσω μιας συνάρτησης με τη μέση τιμή της απόκρισης, δηλαδή την αναμενόμενη τιμή της απόκρισης σε κάθε σημείο παρατήρησης.

Ας υποθέσουμε ότι τα $y_i, i = 1, 2, \ldots, n$ αναπαριστούν τις τιμές της μεταβλητής απόκρισης (response values), τότε το γενικευμένο γραμμικό μοντέλο δίνεται από τη σχέση

$$g(\mu_i) = g[E(y_i)] = \mathbf{x}_i'\boldsymbol{\beta}, \tag{5.1}$$

όπου $\mathbf{x}_i$ αποτελεί το διάνυσμα των μεταβλητών παλινδρόμησης για την i-οστή παρατήρηση και $\boldsymbol{\beta}$ είναι το διάνυσμα των άγνωστων παραμέτρων ή των συντελεστών παλινδρόμησης. Κάθε γενικευμένο γραμμικό μοντέλο αποτελείται από τρεις όρους, την κατανομή της μεταβλητής απόκρισης, τη γραμμική προβλέπουσα (linear predictor) η οποία περιέχει τις μεταβλητές παλινδρόμησης και τη συνάρτηση σύνδεσης (link function) g η οποία συνδέει τη γραμμική προβλέπουσα με τη μέση τιμή της μεταβλητής απόκρισης.

Ας θεωρήσουμε για παράδειγμα το κλασικό γραμμικό μοντέλο παλινδρόμησης

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon, \tag{5.2}$$

όπου ε η συνιστώσα του τυχαίου σφάλματος. Η κατανομή της απόκρισης είναι κανονική, η γραμμική προβλέπουσα είναι η

$$\mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m,$$

και η συνάρτηση σύνδεσης είναι η ταυτοτική, δηλαδή g(a)=a και δίνεται από τον τύπο

$$E(y) = \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m.$$

Διαπιστώνουμε ότι το γραμμικό μοντέλο παλινδρόμησης στην κλασική του μορφή (5.2) αποτελεί ένα γενικευμένο γραμμικό μοντέλο. Εντούτοις, ένα γενικευμένο γραμμικό μοντέλο μπορεί να περιλαμβάνει και ένα μη γραμμικό μοντέλο, και αυτό εξαρτάται από την επιλογή της συνάρτησης σύνδεσης g. Για περισσότερες λεπτομέρειες σχετικά με την επιλογή της συνάρτησης σύνδεσης στα γενικευμένα γραμμικά μοντέλα ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στην εργασία [58]. Για παράδειγμα, αν χρησιμοποιήσουμε ως συνάρτηση σύνδεσης την g(a)=lna θα έχουμε

$$E(y) = \mu = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m}.$$

Γενικότερα, τα γενικευμένα γραμμικά μοντέλα μπορούν να θεωρηθούν ως μια ενοποίηση και επέκταση των γραμμικών και μη γραμμικών μοντέλων παλινδρόμησης, τα οποία ενσωματώνουν μια μεγάλη οικογένεια κανονικών και μη κανονικών κατανομών για τα δεδομένα απόκρισης. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στα πολύ χρήσιμα βιβλία [171], [161], [167] και [2] για περισσότερες πληροφορίες σχετικά με τη θεωρία και τις εφαρμογές των γενικευμένων γραμμικών μοντέλων στις διάφορες επιστήμες.

5.2.1 Το Λογιστικό Μοντέλο Παλινδρόμησης

Το λογιστικό μοντέλο παλινδρόμησης (logistic regression model) άρχισε να χρησιμοποιείται ευρέως κατά τη δεκαετία του 1950, κυρίως σε εφαρμογές στη Βιοστατιστική και ανήκει στην κατηγορία των γενικευμένων γραμμικών μοντέλων. Η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί σε περιπτώσεις όπου η μεταβλητή απόκρισης είναι δίτιμη, με άλλα λόγια η πρόβλεψη είναι το αποτέλεσμα μιας διαδικασίας Bernoulli, όπως επιτυχία/αποτυχία ή π.χ. σε ένα πείραμα που εξετάζει αν το φάρμακο ενεργεί ή όχι σε έναν ασθενή. Η μεταβλητή απόκρισης στο λογιστικό μοντέλο παλινδρόμησης έχει μόνο δύο πιθανά εξαγόμενα, τα οποία συμβολίζονται με 0 και 1.

Εάν τα y_i, i=1,2,...,n, αναπαριστούν τις τιμές απόκρισης, τότε ένα γενικευμένο γραμμικό μοντέλο, όπως είδαμε παραπάνω, δίνεται από τη σχέση (5.1). Ας θεωρήσουμε τώρα την περίπτωση στην οποία, η απόκριση στο i-οστό σημείο είναι μια τυχαία μεταβλητή y_i, η οποία ακολουθεί την κατανομή Bernoulli και λαμβάνει μόνο δύο τιμές, 0 και 1. Τότε

$$E(y_i) = P_i = P(\mathbf{x}_i),\ i=1,2,\ldots,n$$

και

$$Var(y_i) = P_i(1 - P_i),$$

όπου το P_i συμβολίζει την πιθανότητα μιας διαδικασίας Bernoulli και το $\mathbf{x}_i$ είναι το διάνυσμα των μεταβλητών παλινδρόμησης. Η παράμετρος P_i, και συνεπώς η διασπορά, είναι συνάρτηση των μεταβλητών παλινδρόμησης $\mathbf{x}_i$.

Τώρα θα παρουσιάσουμε λεπτομερώς το λογιστικό μοντέλο παλινδρόμησης το οποίο θεωρήσαμε στα πειράματα προσομοίωσης που εκτελέσαμε στα Κεφάλαια 2, 3, 4 και φυσικά στο παρόν Κεφάλαιο 5.

Ας υποθέσουμε ότι έχουμε n πειραματικές εκτελέσεις, έναν $n \times m$ υπερκορεσμένο πίνακα σχεδιασμού $X = [x_1, x_2, \ldots, x_m]$ και μια δίτιμη μεταβλητή απόκρισης y (0 ή 1) η οποία εξαρτάται από το σύνολο μεταβλητών παλινδρόμησης x_1,x_2,...,x_m. Εάν δηλώσουμε αυθαίρετα ότι το y=1 αναπαριστά την επιτυχία και το y=0 αναπαριστά την αποτυχία, τότε μπορούμε

να μοντελοποιήσουμε τη μέση τιμή απόκρισης $P(x_i)$, όπου $P(x_i)$ είναι η πιθανότητα επιτυχίας και το x_i αναπαριστά το διάνυσμα των μεταβλητών παλινδρόμησης για την i-οστή παρατήρηση. Το λογιστικό μοντέλο παλινδρόμησης για τη μέση απόκριση $P(x_i)$ δίνεται από τη σχέση

$$P(x_i) = 1/(1 + e^{-x_i'\beta}), \tag{5.3}$$

όπου ο όρος $x_i'\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}$ είναι η γραμμική προβλέπουσα. Ας σημειωθεί εδώ ότι $0 \leq P(x_i) \leq 1$.

Στα πειράματα προσομοίωσης που εκτελέσαμε, τα λογιστικά μοντέλα παλινδρόμησης παράχθηκαν ακολουθώντας κάθε φορά το παρακάτω πρωτόκολλο προσομοίωσης.

Υποθέτουμε ότι έχουμε ένα λογιστικό μοντέλο παλινδρόμησης το οποίο έχει τη μορφή

$$y_i = P(x_i) + \varepsilon, \tag{5.4}$$

όπου $P(x_i)$ αναπαριστά την ποσότητα όπως αυτή ορίστηκε στη σχέση (5.3), και ότι η μεταβλητή απόκρισης y_i λαμβάνει μόνο δύο τιμές, 0 ή 1. Εδώ, η ποσότητα ε αποτελεί τη συνιστώσα του σφάλματος και μπορεί να λαμβάνει μόνο μία από τις δύο πιθανές τιμές. Εάν το $y=1$, τότε $\varepsilon = 1 - P(x_i)$ με πιθανότητα $P(x_i)$, και εάν το $y=0$ τότε $\varepsilon = -P(x_i)$ με πιθανότητα $1 - P(x_i)$. Συμπερασματικά, η συνιστώσα του σφάλματος ε ακολουθεί μια κατανομή με μέση τιμή μηδέν και διασπορά ίση με $P(x_i)[1 - P(x_i)]$. Άρα, η δεσμευμένη κατανομή της μεταβλητής απόκρισης ακολουθεί μια Bernoulli κατανομή με πιθανότητα επιτυχίας η οποία δίνεται από τη δεσμευμένη μέση τιμή της $P(x_i)$. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στα πολύ χρήσιμα βιβλία [171], [161], [167] και [2] για περισσότερες πληροφορίες σχετικά με το λογιστικό μοντέλο παλινδρόμησης.

5.2.2 Μοντέλα Παλινδρόμησης Poisson, Αρνητικής Διωνυμικής και Γεωμετρικής Κατανομής

Αρχικά, θεωρούμε ένα *Poisson* μοντέλο παλινδρόμησης με τα δεδομένα απόκρισης να ακολουθούν την κατανομή Poisson ή με άλλα λόγια να είναι Poisson μετρήσεις (counts). Οι Poisson μετρήσεις μπορεί να είναι για παράδειγμα, μη αρνητικές ακέραιες τιμές οι οποίες μπορεί να περιγράφουν οτιδήποτε εμπλέκεται στην επίδοση ενός βιολογικού ή κοινωνικού συστήματος ή την παραγωγική διαδικασία μιας βιομηχανικής εφαρμογής.

Τώρα θα παρουσιάσουμε λεπτομερώς το μοντέλο και τις υποθέσεις, τα οποία διαμορφώνουν το Poisson μοντέλο παλινδρόμησης το οποίο θεωρήσαμε στα πειράματα προσομοίωσης που εκτελέσαμε παρακάτω.

Ας υποθέσουμε ότι έχουμε n πειραματικές εκτελέσεις, έναν $n \times m$ υπερκορεσμένο πίνακα σχεδιασμού $X = [x_1, x_2, \ldots, x_m]$ και ότι οι αποκρίσεις $y_1, y_2, \ldots, y_n$ στη δομή παλινδρόμησης αποτελούν μετρήσεις οι οποίες ακολουθούν ανεξάρτητες Poisson κατανομές με $E(y_i) = \mu_i$ και $Var(y_i) = \mu_i$. Το σύνολο των μεταβλητών παλινδρόμησης $x_1, x_2, \ldots, x_m$ επηρεάζει το μ μέσω της εκθετικής συνάρτησης σύνδεσης

$$\mu_i = e^{x_i'\beta}, \quad i = 1, 2, \ldots, n, \tag{5.5}$$

όπου $x_i' = [1, x_{i1}, \ldots, x_{im}]$ και $x_i'\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}$, με $\beta_0, \beta_1, \ldots, \beta_m$ να είναι οι άγνωστοι συντελεστές παλινδρόμησης. Ο όρος $x_i'\beta$ είναι η γραμμική προβλέπουσα για το μοντέλο παλινδρόμησης Poisson. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στα πολύ χρήσιμα βιβλία [38], [167] και [2] για περισσότερες πληροφορίες σχετικά με το μοντέλο παλινδρόμησης Poisson.

Εάν τα y_i, $i=1,2,\ldots,n$, αναπαριστούν τις τιμές απόκρισης, τότε ένα γενικευμένο γραμμικό μοντέλο, όπως είδαμε παραπάνω, δίνεται από τη σχέση (5.1). Ας θεωρήσουμε τώρα την περίπτωση στην οποία η απόκριση στο i-οστό σημείο είναι μια τυχαία μεταβλητή y_i, η οποία ακολουθεί την κατανομή Poisson και λαμβάνει ως τιμές Poisson μετρήσεις. Τότε για το

μοντέλο παλινδρόμησης Poisson ισχύει μια πολύ ισχυρή υπόθεση, ότι η διακύμανση είναι ίση με τη μέση τιμή

$$Var(y_i) = E(y_i) = \mu_{x_i}.$$

Εντούτοις, στην πράξη παρατηρούμε πολλές φορές το φαινόμενο της υπερ-διασποράς (overdispersion), δηλαδή η διασπορά να υπερβαίνει τη μέση τιμή, το οποίο συναντάται συχνά στη μοντελοποίηση με τα δεδομένα απόκρισης να είναι μετρήσεις (counts). Το μοντέλο παλινδρόμησης *αρνητικής διωνυμικής κατανομής* είναι η κατάλληλη επιλογή και μπορεί να χρησιμοποιηθεί για δεδομένα μετρήσεων τα οποία εμφανίζουν το φαινόμενο της υπερ-διασποράς. Το μοντέλο παλινδρόμησης αρνητικής διωνυμικής κατανομής μπορεί να θεωρηθεί ως μια επέκταση του μοντέλου παλινδρόμησης Poisson από τη στιγμή που έχει την ίδια δομή με αυτή του μοντέλου παλινδρόμησης Poisson για τη μέση τιμή και θεωρεί μία επιπλέον παράμετρο α για να μοντελοποιήσει την υπερ-διασπορά. Η γενική κλάση ενός μοντέλου αρνητικής διωνυμικής κατανομής μπορεί να εκφραστεί ως μία Γάμμα ανάμειξη μιας Poisson τυχαίας μεταβλητής η οποία αντιπροσωπεύει την υπερ-διασπορά προσθέτοντας έναν λανθάνων όρο ετερογένειας για την παρατήρηση i στο μοντέλο δεδομένων καταμέτρησης, και έχει δεσμευμένη μέση τιμή $\mu_i = e^{x_i'\beta}$ και δεσμευμένη διακύμανση $\mu_i + \alpha\mu_i^p$, όπου ισχύει γενικά $-\infty < p < \infty$. Οι πιο κλασικές παραλλαγές της αρνητικής διωνυμικής κατανομής, οι οποίες συνήθως χρησιμοποιούνται σε εφαρμογές παλινδρόμησης, αντιστοιχούν στις τιμές $p = 1$ ή $p = 2$. Στην εμπειρική μελέτη που πραγματοποιήθηκε εδώ, θεωρούμε την τιμή $p = 2$, η οποία είναι η τετραγωνική συνάρτηση διακύμανσης.

Τώρα θα παρουσιάσουμε λεπτομερώς το μοντέλο και τις υποθέσεις, τα οποία διαμορφώνουν το μοντέλο παλινδρόμησης αρνητικής διωνυμικής κατανομής το οποίο θεωρήσαμε στα πειράματα προσομοίωσης που εκτελέσαμε παρακάτω.

Ας υποθέσουμε ότι έχουμε n πειραματικές εκτελέσεις, έναν $n \times m$ υπερκορεσμένο πίνακα σχεδιασμού $X = [x_1, x_2, \ldots, x_m]$ και ότι οι αποκρίσεις $y_1, y_2, \ldots, y_n$ στη δομή παλινδρόμησης αποτελούν μετρήσεις οι οποίες ακολουθούν την αρνητική διωνυμική κατανομή και εξαρτώνται από το σύνολο μεταβλητών παλινδρόμησης x_1,x_2,...,x_m. Τότε, το μοντέλο της αρνητικής διωνυμικής κατανομής είναι της μορφής

$$E(y_i|x_i, \epsilon_i) = \mu_i\epsilon_i = e^{x_i'\beta}\epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{5.6}$$

όπου τα ϵ_i υποθέτουμε ότι είναι ανεξάρτητα και ακολουθούν μια μονοπαραμετρική κατανομή gamma(θ, θ) με $E(\epsilon_i) = 1$ και $Var(\epsilon_i) = \frac{1}{\theta}$. Σημειώνουμε εδώ ότι όταν η απόκριση εξαρτάται από τα x_i και ϵ_i, τότε η εξαρτημένη μεταβλητή απόκρισης των μετρήσεων y_i ακολουθεί ακόμα την κατανομή Poisson. Εντούτοις, όταν η απόκριση εξαρτάται μόνο από τα x_i και y_i, τότε ακολουθεί την αρνητική διωνυμική κατανομή, και αυτή η κατανομή έχει δεσμευμένη μέση τιμή μ_i και δεσμευμένη διακύμανση $\mu_i(1 + (\frac{1}{\theta})\mu_i)$.

Στην εμπειρική μελέτη που πραγματοποιήθηκε εδώ, τα σύνολα δεδομένων τα οποία ακολουθούν την Poisson και αρνητική διωνυμική κατανομή, παράχθηκαν χρησιμοποιώντας τον ίδιο φυσικό μέσο $\mu_i = e^{x_i'\beta}, i = 1, 2, \ldots, n$, με τη διαφορά ότι το μοντέλο της αρνητικής διωνυμικής κατανομής χρησιμοποιεί περαιτέρω την παράμετρο $\theta = 1$. Η παράμετρος θ είναι ανάλογη προς το μ_i στη γενική της μορφή ($\theta_i \propto \mu_i^\lambda$, για κάποιο $\lambda > 0$). Η παράμετρος υπερ-διασποράς θ στην παλινδρόμηση αρνητικής διωνυμικής κατανομής δεν επηρεάζει τις αναμενόμενες μετρήσεις και επηρεάζει μόνο την εκτιμώμενη διακύμανση των αναμενόμενων μετρήσεων. Στην εμπειρική μελέτη που πραγματοποιήθηκε εδώ, η παράμετρος θ θεωρούμε ότι είναι σταθερή (αυτή η φόρμουλα έχει ληφθεί από την εργασία [73]), δεδομένου ότι συνιστάται η χρήση συντηρητικών τιμών της παραμέτρου υπερ-διασποράς σε εφαρμογές παλινδρόμησης (βλέπε [38]). Ως εκ τούτου, η Poisson και η αρνητική διωνυμική κατανομή συνδέουν τη γραμμική προβλέπουσα με τη φυσική μέση τιμή της μεταβλητής απόκρισης μέσω των παραμέτρων μ_i και $\frac{1}{1+\frac{\mu_i}{\theta}}$, αντίστοιχα. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στα βιβλία [38] και [109] για περισσότερες πληροφορίες σχετικά με το μοντέλο παλινδρόμησης της αρνητικής διωνυμικής

κατανομής.

Η *γεωμετρική κατανομή* αποτελεί ειδική περίπτωση της αρνητικής διωνυμικής κατανομής. Όπως η αρνητική διωνυμική κατανομή, έτσι και η γεωμετρική κατανομή βασίζεται σε μία αλληλουχία ανεξάρτητων και πανομοιότυπων δοκιμών Bernoulli. Ας υποθέσουμε ότι ο αριθμός των δοκιμών Bernoulli δεν έχει καθοριστεί εκ των προτέρων, αλλά αντίθετα, η αλληλουχία των δοκιμών Bernoulli συνεχίζεται μέχρις ότου παρατηρηθεί ένας συγκεκριμένος αριθμός r επιτυχιών. Τότε, ο αριθμός των δοκιμών που απαιτούνται για την παρατήρηση της r επιτυχίας ακολουθεί την αρνητική διωνυμική κατανομή. Ας θεωρήσουμε τώρα την περίπτωση στην οποία το $r = 1$, δηλαδή, μια αλληλουχία δοκιμών Bernoulli η οποία διεξάγεται μέχρι να παρατηρηθεί η πρώτη επιτυχία, με πιθανότητα επιτυχίας p. Σε μια τέτοια περίπτωση, ο αριθμός των δοκιμών που απαιτούνται μέχρις ότου παρατηρηθεί η πρώτη επιτυχία ακολουθεί τη γεωμετρική κατανομή.

Τώρα θα παρουσιάσουμε λεπτομερώς το μοντέλο και τις υποθέσεις, τα οποία διαμορφώνουν το μοντέλο παλινδρόμησης γεωμετρικής κατανομής το οποίο θεωρήσαμε στα πειράματα προσομοίωσης που εκτελέσαμε παρακάτω.

Ας υποθέσουμε ότι έχουμε n πειραματικές εκτελέσεις, έναν $n \times m$ υπερκορεσμένο πίνακα σχεδιασμού $X = [x_1, x_2, \ldots, x_m]$ και ότι οι αποκρίσεις $y_1, y_2, \ldots, y_n$ στη δομή παλινδρόμησης αποτελούν μετρήσεις οι οποίες ακολουθούν τη γεωμετρική κατανομή, οι οποίες εξαρτώνται από το σύνολο μεταβλητών παλινδρόμησης x_1,x_2,...,x_m. Για τα πειράματα προσομοίωσης της γεωμετρική κατανομής με παράμετρο p, μοντελοποιούμε τη μέση τιμή απόκρισης $P(x_i)$, όπου $P(x_i)$ είναι η πιθανότητα επιτυχίας (=p) και το x_i αναπαριστά τη μεταβλητή παλινδρόμησης για την i-οστή παρατήρηση. Τότε, το μοντέλο για την $P(x_i)$ δίνεται από τη λογιστική συνάρτηση σύνδεσης

$$P(x_i) = 1/(1 + e^{-x_i'\beta}), \tag{5.7}$$

όπου $x_i'\beta$= β_0+$\beta_1 x_{i1}$+...+ $\beta_m x_{im}$ είναι η γραμμική προβλέπουσα και ισχύει ότι $0 \leq P(x_i) \leq 1$. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στο βιβλίο [2] για περισσότερες πληροφορίες σχετικά με το μοντέλο παλινδρόμησης της γεωμετρικής κατανομής.

5.3 Η Προτεινόμενη Μέθοδος Ανάλυσης

Σε αυτή την ενότητα παρουσιάζεται η τέταρτη προτεινόμενη μέθοδος ανάλυσης υπερκορεσμένων σχεδιασμών, η οποία αποτελεί επέκταση-γενίκευση του αλγορίθμου Θεωρίας Πληροφοριών που παρουσιάστηκε στην εργασία [7] και έχει ως στόχο την αναγνώριση και επιλογή των σημαντικών παραγόντων μέσα από ένα μεγάλο σύνολο ενδεχόμενων ενεργών παραγόντων. Αξιοσημείωτο πλεονέκτημα της προτεινόμενης μεθόδου είναι η χρήση των γενικευμένων γραμμικών μοντέλων με διακριτή απόκριση για την ανάλυση δεδομένων από υπερκορεσμένους σχεδιασμούς, μέθοδος η οποία είναι μη τετριμμένη και μη αναπτυγμένη στην έως σήμερα διεθνή βιβλιογραφία όπου συναντάμε το πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών μόνο κάτω υπό τη σκέπη του κλασικού γραμμικού μοντέλου. Η συμπεριφορά και η αποδοτικότητα της προτεινόμενης μεθόδου εξετάστηκαν και αξιολογήθηκαν μέσα από μια εμπειρική και συγκριτική μελέτη, η οποία περιελάμβανε εκτενή πειράματα προσομοίωσης των οποίων τα ενθαρρυντικά αποτελέσματα που προέκυψαν παρουσιάζονται σε επόμενη ενότητα. Ας σημειωθεί εδώ, ότι στα πειράματα προσομοίωσης εξετάσαμε εκ νέου διαφορετικές τιμές κατωφλιού για το w και το SU, με σκοπό να εξετάσουμε και τελικά να αναδείξουμε την ευρωστία καθώς και την καταλληλότητα των τιμών κατωφλιού που επιλέχθηκαν αρχικά στην εργασία [7] ακόμα και σε περιπτώσεις που η απόκριση δεν είναι δίτιμη αλλά γενικά διακριτή, και με αυτό τον τρόπο να επιτύχουμε την καλύτερη δυνατή απόδοση του αλγορίθμου.

5.3.1 Περιγραφή της Προτεινόμενης Μεθόδου

Ο προτεινόμενος αλγόριθμος επιλογής μεταβλητών στους υπερκορεσμένους σχεδιασμούς εφαρμόζεται κάτω από την υπόθεση ενός γενικευμένου γραμμικού μοντέλου του οποίου τα δεδομένα απόκρισης προέρχονται από τη διωνυμική, Poisson, γεωμετρική και αρνητική διωνυμική κατανομή. Ο προτεινόμενος Αλγόριθμος 5.1 δίνεται παρακάτω σε μορφή ψευδοκώδικα.

Αλγόριθμος 5.1.

Require: $X_{n\times m} = [X_1, X_2, \ldots, X_m]$ {*where* X_j *is the j-th column of the supersaturated design matrix,* $j = 1, 2, \ldots, m$}
Require: $Y_{n\times 1}$ {$Y_{n\times 1}$ *is the common discrete response vector,* $i = 1, 2, \ldots, n$}
Ensure: $w = \frac{m}{2}$ {*where m is the number of columns in the supersaturated design matrix*}
for 1 *to* 1000 *iterations* ***do***
  repeat
    the following processing
  until *all the following conditions are met*
Ensure: $Y_k \neq Y_l$, $\forall k \neq l$, *for* $k, l = 1, \ldots, 1000$ {Y_k *(or* Y_l*) is the common discrete response vector of the k-th (or the l-th) iteration*}
  for $j = 1$ *to* m ***do***
    compute the entropy value $H(X_j)$ *for each j-th factor* X_j *of* X *(*$j = 1, 2, \ldots, m$*), i.e.,* $H(X_1)$,...,$H(X_m)$ {*The entropy of* X_j *with n outcomes* $\{x_{ij} : i = 1, \ldots, n\}$ *is given by*

$$H(X_j) = -\sum_{i=1}^{n} p(x_{ij}) log_2(p(x_{ij})).$$

For $j = 1$,

$$H(X_1) = -p(x_{11}) log_2(p(x_{11})) - \cdots - p(x_{n1}) log_2(p(x_{n1}));$$

for $j = 2$,

$$H(X_2) = -p(x_{12}) log_2(p(x_{12})) - \cdots - p(x_{n2}) log_2(p(x_{n2}));$$

$$\ldots\ldots;$$

for $j = k$,

$$H(X_k) = -p(x_{1k}) log_2(p(x_{1k})) - \cdots - p(x_{nk}) log_2(p(x_{nk})).$$

The $H(X_j)$ *value quantifies the uncertainty of* X_j*; that is, how much information we gain from the j-th factor* X_j *of* X}
  end for
  for $j = 1$ *to* m ***do***
    compute each conditional entropy value $H(X_j|Y)$*, for* $j = 1, 2, \ldots, m$*, i.e.,* $H(X_j|Y) = (H(X_1|y), \ldots, H(X_m|y))$ {*the* $H(X_j|Y)$ *value quantifies the remaining uncertainty of the j-th factor* X_j *of* X*, when* Y *is known*}
  end for
  for $j = 1$ *to* m ***do***
    compute the information gain value for each j-th factor X_j *of the supersaturated design matrix* X*, i.e.,* $I(X_j; Y) = H(X_j) - H(X_j|Y)$ *(*$j = 1, 2, \ldots, m$*)* {*the* $I(X_j; Y)$ *value quantifies the reduction in the uncertainty of the j-th factor* X_j *of* X*, due to the knowledge of* Y}
    place all $I(X_j; Y)$ *values for* $j = 1, 2, \ldots, m$ *into a vector* $(I(X_1; Y), \ldots, I(X_m; Y))$.
  end for
  for $j = 1$ *to* m ***do***
    sort in descending order the information gain vector $\boldsymbol{ig} = (ig_1, ig_2, \ldots, ig_m)$ {*where* $ig_j = I(X_j; Y)$ *is the information gain value for the j-th factor* X_j *of* X *(*$j = 1, 2, \ldots, m$*)*}
  end for
  for $j = 1$ *to* m ***do***
    compute the symmetrical uncertainty value $su_j = 2 \times \left[\frac{I(X_j;Y)}{H(X_j)+H(Y)}\right]$ *for each j-th factor* X_j *of* X *(*$j = 1, 2, \ldots, m$*), with respect to the response variable* {*the entropy value* $H(Y)$ *quantifies the uncertainty of the count response variable* Y}
    place all su_j *values for* $j = 1, 2, \ldots, m$ *into a vector* $\boldsymbol{su} = (su_1, su_2, \ldots, su_m)$
  end for
  for all j, $j = 1, 2, \ldots, m$ *such that* ***do***
    have the w highest information gain values in ***ig***
    AND
    their position in the sorted ***ig*** *vector is at most m/2*
    AND
    if $su_j \geq median(\boldsymbol{su})$ ***then***
      declare X_j *of* X *as active* {*INCLUDE*}
    else if $su_j < median(\boldsymbol{su})$ ***then***
      declare X_j *of* X *as inactive* {*EXCLUDE*}
    end if
  end for
end for

5.3.2 Αξιολόγηση της Προτεινόμενης Μεθόδου

Για τη διερεύνηση και την αξιολόγηση της απόδοσης του προτεινόμενου αλγορίθμου πραγματοποιήσαμε μια εμπειρική μελέτη, θεωρώντας ένα ευρύ φάσμα προσομειωμένων μοντέλων και υπερκορεσμένων σχεδιασμών. Στη μελέτη προσομοίωσης χρησιμοποιήσαμε τους s-block-ορθογώνιους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με n πειραματικές εκτελέσεις και $m = s(n-1)$ παράγοντες, όπως αυτοί κατασκευάστηκαν από τους Tang και Wu [215]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 22$ παράγοντες και $n = 12$ πειραματικές εκτελέσεις. Επιπλεόν, στη μελέτη προσομοίωσης χρησιμοποιήσαμε υπερκορεσμένους σχεδιασμούς οι οποίοι κατασκευάζονται από τα κλάσματα πινάκων Hadamard, σύμφωνα με τη μέθοδο που προτάθηκε από τον Lin [148]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 10$ παράγοντες και $n = 6$ πειραματικές εκτελέσεις. Εκτός από αυτούς τους σχεδιασμούς, θεωρήσαμε και τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχέδιασμούς, όπως αυτοί κατασκευάστηκαν από τους Koukouvinos et al. [130]. Αυτοί οι υπερκορεσμένοι σχεδιασμοί έχουν n πειραματικές εκτελέσεις και $m = q \cdot (n-1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Στις προσομοιώσεις μας θεωρήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, m) τιμές: $(6, 10)$, $(8, 14)$, $(10, 18)$, $(12, 22)$, $(14, 26)$, $(16, 30)$, $(18, 34)$, $(20, 38)$ και $(22, 42)$.

Να σημειώσουμε εδώ ότι και πάλι μας ενδιαφέρει να μελετήσουμε μοντέλα μόνο κύριων επιδράσεων, οπότε θεωρούμε ότι τα μοντέλα μας δεν συμπεριλαμβάνουν αλληλεπιδράσεις. Οι πραγματικά ενεργοί παράγοντες επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων και οι συντελεστές των μη ενεργών παραγόντων τέθηκαν ίσοι με μηδέν στο πραγματικό μοντέλο. Επιπρόσθετα, προκειμένου να εξετάσουμε το πόσο ευαίσθητα είναι τα αποτελέσματα στην επιλογή και τον αριθμό των ενεργών στηλών-παραγόντων, μελετήσαμε ένα ευρύ φάσμα προσομειωμένων μοντέλων στα πειράματά μας στα οποία θεωρούμε για κάθε υπερκορεσμένο σχεδιασμό διαφορετικό αριθμό ενεργών παραγόντων, διαφορετική επιλογή ενεργών στηλών και διαφορετικά β. Για τα πειράματα προσομοίωσης, αναπτύξαμε γενικευμένα γραμμικά μοντέλα με τους συντελεστές παλινδρόμησης να λαμβάνουν τυχαίες τιμές από το διάνυσμα β. Στον Πίνακα 5.1 περιγράφονται τα μοντέλα τα οποία θεωρήσαμε στη μελέτη προσομοίωσης.

Πίνακας 5.1: Μοντέλα 4ης μελέτης προσομοίωσης

Μοντέλο	ΥΣ από	m	n	β
1	[130]	10	6	[-17,-13,0,0,0,-2,0,0,6,0] $'$
2	[130]	10	6	[1,0,-8,0,-21,0,-17,0,0,0]$'$
3	[130]	10	6	[-1,0,0,4,0,0,-17,0,0,0]$'$
4	[130]	10	6	[-9,0,0,3,10,0,0,-9,0,0]$'$
5	[130]	14	8	[1,0,0,2,0,0,-4,0,0,7,0,0,11,0]$'$
6	[130]	14	8	[5,5,0,0,-5,0,0,3,0,-2,0,0,2,0]$'$
7	[130]	14	8	[-6,0,0,0,0,-2,0,0,0,0,-15,0,0,0]$'$
8	[130]	14	8	[-4,0,0,0,0,0,0,0,-9,0,4,-5,0,0]$'$
9	[130]	18	10	[0,0,-5,0,-3,4,0,0,7,0,0,1,2,0,0,-7,0,3]$'$
10	[130]	18	10	[0,0,0,0,0,-3,0,-8,0,0,0,0,2,0,0,-9,0,0]$'$
11	[130]	18	10	[0,0,0,0,0,0,2,0,6,10,0,0,0,0,0,0,11,20]$'$
12	[130]	18	10	[0,8,0,0,0,0,7,0,0,4,0,-5,0,4,7,0,0,0]$'$
13	[130]	22	12	[1,2,0,0,0,16,0,0,0,0,7,0,0,0,-1,0,0,5,0,18,0,0]$'$
14	[130]	22	12	[1,0,0,-13,0,0,4,-3,0,0,-6,0,0,-7,0,-24,0,-5,0,-21,0,0]$'$
15	[130]	22	12	[-3,0,0,0,4,0,0,0,0,9,0,0,0,17,0,0,0,0,-21,0,0,0]$'$
16	[130]	22	12	[0,0,0,0,20,-16,0,0,0,0,0,0,0,0,0,0,20,0,0,-7,0,0]$'$
17	[130]	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] $'$
18	[130]	26	14	[2,0,0,5,0,0,-7,0,0,-2,1,0,0,3,0,8,0,0,-9,0,0,-1,0,4,0,0]$'$
19	[130]	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] $'$
20	[130]	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] $'$
21	[130]	30	16	[1,2,0,0,5,0,0,0,9,0,0,0,11,0,0,7,0,0,23,0,0,24,0,0,15,0,0,17,0,0] $'$
22	[130]	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] $'$
23	[130]	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] $'$
24	[130]	30	16	[0,0,-4,0,2,-15,-17,0,0,0,19,0,0,5,4,-12,0,0,0,0,0,0,0,9,15,-10,0,0,0,-4,0] $'$
25	[130]	34	18	[-1,-2,-3,-4,0,0,0,0,1,0,8,0,0,0,0,-7,0,0,0,-1,0,9,0,0,0,-5,0,0,0,0,0,0,9,0]$'$
26	[130]	34	18	[2,0,-1,3,0,0,7,0,0,8,0,0,13,0,0,17,0,0,28,0,0,25,0,0,-5,0,0,-7,0,0,-2,0,-4,0] $'$
27	[130]	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]$'$
28	[130]	34	18	[5,0,4,0,0,0,0,0,0,-6,0,0,0,0,0,0,0,0,0,-12,10,0,0,-14,0,0,0,0,0,0,0,-15,0,0]$'$
29	[130]	38	20	[1,2,3,4,5,6,7,0,0,0,1,0,0,1,0,-1,0,0,-1,0,0,1,0,0,0,-1,0,4,0,0,0,0,0,-2,0,0,0,0] $'$
30	[130]	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]$'$
31	[130]	38	20	[0,0,0,0,-17,0,0,0,0,0,-23,0,0,0,0,0,0,0,0,0,0,0,26,0,0,0,0,0,0,0,0,0,0,-21,0,0,0,0,0]$'$
32	[130]	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]$'$
33	[130]	42	22	[0,0,0,0,-3,0,0,-2,0,0,0,0,1,-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]$'$
34	[130]	42	22	[0,0,1,1,3,0,0,4,0,0,0,0,0,4,0,0,0,1,0,2,0,0,0,6,-3,0,0,0,2,0,0,0,0,0,0,0,3,0,0,0,0,0,0]$'$
35	[130]	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]$'$
36	[130]	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,-10,0,0,0,0,0,0,0,0,0,14,0,7]$'$
37	[215]	22	12	[0,0,2,3,0,0,13,1,2,0,5,0,0,1,8,0,0,0,0,0,-1,0]$'$
38	[215]	22	12	[0,0,0,0,20,-16,0,0,0,0,0,0,0,0,0,0,20,0,0,-7,0,0]$'$
39	[215]	22	12	[2,-1,1,0,0,0,0,0,0,0,0,0,0,-1,1,0,2,0,0,0,0]$'$
40	[215]	22	12	[0,-2,-2,1,2,0,0,4,0,0,0,4,0,0,0,0,0,0,-3,0,2,2]$'$
41	[148]	10	6	[0,5,0,0,-8,1,0,0,0,0]$'$
42	[148]	10	6	[0,0,0,0,0,10,0,0,0,17]$'$
43	[148]	10	6	[0,0,0,0,0,0,0,7,-1,0]$'$
44	[148]	10	6	[0,9,0,0,0,-6,0,0,0,7]$'$

5.4 Συγκριτική Μελέτη

Προκειμένου να αξιολογήσουμε την απόδοση του προτεινόμενου αλγορίθμου επιλογής μεταβλητών χρησιμοποιήσαμε τα σφάλματα Τύπου I και Τύπου II (όσο μικρότερα τόσο καλύτερα) και στα πειράματά μας εκτελέσαμε τις προσομοιώσεις 1000 φορές για κάθε ένα από τα ακόλουθα σενάρια:

- Σενάριο I: 1000 σύνολα δεδομένων με την απόκριση Y να προέρχεται από την Poisson κατανομή παράχθηκαν τυχαία για κάθε ένα από τα μοντέλα,
- Σενάριο II: 1000 σύνολα δεδομένων με την απόκριση Y να προέρχεται από τη γεωμετρική κατανομή παράχθηκαν τυχαία για κάθε ένα από τα μοντέλα,
- Σενάριο III: 1000 σύνολα δεδομένων με την απόκριση Y να προέρχεται από την αρνητική διωνυμική κατανομή παράχθηκαν τυχαία για κάθε ένα από τα μοντέλα,
- Σενάριο IV: 1000 σύνολα δεδομένων με την απόκριση Y να προέρχεται από την Bernoulli κατανομή παράχθηκαν τυχαία για κάθε ένα από τα μοντέλα.

5.4.1 Συγκριτική Μελέτη: Bernoulli Απόκριση

Σε αυτή την ενότητα, παρουσιάζεται μια συγκριτική μελέτη προσομοίωσης, προκειμένου να αξιολογηθεί η απόδοση και τα πλεονεκτήματα της προτεινόμενης μεθόδου. Υποθέτουμε ότι τα δεδομένα απόκρισης ακολουθούν Bernoulli κατανομή, και συγκρίνουμε τον προτεινόμενο αλγόριθμο της Θεωρίας Πληροφοριών (SU), με δύο ευρέως γνωστούς και χρησιμοποιούμενους αλγόριθμους επιλογής μεταβλητών, τον Conditional Mutual Information Maximization (CMIM) αλγόριθμο ο οποίος προτάθηκε από τον Fleuret [81] και τον minimal-redundancy-maximal-relevance (mRMR) αλγόριθμο επιλογής μεταβλητών ο οποίος προτάθηκε από τους Peng et al. [178]. Η επιλογή των CMIM και mRMR αλγορίθμων μεταξύ άλλων υπαρχόντων αλγορίθμων επιλογής μεταβλητών έγινε με βάση το γεγονός ότι ο CMIM και ο mRMR αλγόριθμος, καθώς και ο προτεινόμενος αλγόριθμος (SU), όλοι τους βασίζονται σε μέτρα της Θεωρίας Πληροφοριών. Συγκεκριμένα, ο CMIM αλγόριθμος αποτελεί μία τεχνική επιλογής μεταβλητών βασισμένη στο μέτρο της δεσμευμένης αμοιβαίας πληροφορίας, η οποία επιλέγει επαναληπτικά τις μεταβλητές που μεγιστοποιούν την αμοιβαία πληροφορία τους, με την κλάση πρόβλεψης να εξαρτάται από οποιαδήποτε μεταβλητή έχει ήδη επιλεγεί. Ας σημειωθεί εδώ ότι η CMIM διαδικασία μπορεί να εφαρμοστεί σε προβλήματα ταξινόμησης δύο μόνο κλάσεων, και για αυτόν το λόγο εφαρμόζουμε τον CMIM αλγόριθμο και τον συγκρίνουμε με τον προτεινόμενο (SU) μόνο στην περίπτωση που έχουμε δίτιμα δεδομένα απόκρισης (Bernoulli). Από την άλλη μεριά, ο αλγόριθμος mRMR μεγιστοποιεί την αμοιβαία πληροφορία μεταξύ των επιλεγμένων μεταβλητών και της επιθυμητής τιμής της απόκρισης (relevance) και ελαχιστοποιεί την αμοιβαία πληροφορία μεταξύ των επιλεγμένων μεταβλητών (redundancy). Ας σημειωθεί εδώ ότι το εξαγόμενο αποτέλεσμα του αλγορίθμου mRMR πρέπει να είναι μια μεταβλητή με στόχο την ταξινόμηση σε κλάσεις, και για αυτόν το λόγο εφαρμόζουμε τον mRMR αλγόριθμο και τον συγκρίνουμε με τον προτεινόμενο (SU) μόνο στην περίπτωση που έχουμε δίτιμα δεδομένα απόκρισης (Bernoulli).

Τα ληφθέντα αποτελέσματα μετά την εφαρμογή του προτεινόμενου αλγορίθμου (SU), καθώς και των CMIM και mRMR αλγορίθμων, συνοψίζονται στον Πίνακα 5.2. Στον Πίνακα 5.2, η πρώτη στήλη αναφέρει τον αριθμό που αντιστοιχεί στο μοντέλο που χρησιμοποιήθηκε, ενώ οι δύο στήλες με τίτλο "Type I" και "Type II" παρουσιάζουν τις μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II, αντίστοιχα, που προέκυψαν μετά από 1000 προσομοιώσεις, σύμφωνα κάθε φορά με τον υπό εξέταση αλγόριθμο.

Όσον αφορά τον αλγόριθμο CMIM, οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II είναι 0.53 και 0.10, αντίστοιχα. Παρατηρούμε από τον Πίνακα 5.2, ότι η μέθοδος CMIM παράγει πολύ υψηλά σφάλματα Τύπου I στην πλειοψηφία των περιπτώσεων με αποτέλεσμα την επιλογή πολλών ανενεργών παραγόντων. Συγκεκριμένα, ο αλγόριθμος CMIM παρουσιάζει τα υψηλότερα σφάλματα Τύπου I συγκρινόμενος με την προτεινόμενη μέθοδο (SU) και την mRMR. Ωστόσο, η μέθοδος CMIM παράγει χαμηλά σφάλματα Τύπου II. Η μέθοδος CMIM έχει σημαντικά χαμηλότερα σφάλματα Τύπου II σε σύγκριση με την mRMR, αλλά υψηλότερα σφάλματα Τύπου II σε σύγκριση με την προτεινόμενη μέθοδο (SU).

Όσον αφορά τον αλγόριθμο mRMR, οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II είναι 0.13 και 0.31, αντίστοιχα. Παρατηρούμε από τον Πίνακα 5.2, ότι η mRMR μέθοδος παράγει τα πιο χαμηλά σφάλματα Τύπου I αλλά πολύ υψηλά σφάλματα Τύπου II στην πλειοψηφία των περιπτώσεων. Συγκεκριμένα, ο αλγόριθμος mRMR έχει τις υψηλότερες τιμές σφάλματος Τύπου II σε σύγκριση με την προτεινόμενη μέθοδο (SU) και την CMIM.

Όσον αφορά την προτεινόμενη μέθοδο (SU), οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II είναι 0.23 και 0.06, αντίστοιχα. Συγκεκριμένα, τα σφάλματα Τύπου I είναι σε χαμηλό επίπεδο, και τα σφάλματα Τύπου II της προτεινόμενης μεθόδου είναι πολύ χαμηλά και σημαντικά μικρότερα από τα σφάλματα Τύπου I για όλα τα θεωρούμενα μοντέλα. Επιπλέον, παρατηρούμε από τον Πίνακα 5.2 ότι σε μερικές από τις περιπτώσεις που εξετάστηκαν, τα σφάλματα Τύπου II είναι 0.00, και αυτές αντιστοιχούν σε περιπτώσεις όπου όλες οι ενεργές επιδράσεις ανιχνεύθηκαν σωστά (για 20 από τα 44 μοντέλα για το Σενάριο IV). Επιπλέον, η

προτεινόμενη μέθοδος επιτυγχάνει τις χαμηλότερες τιμές σφαλμάτων Τύπου II σε σχέση με τη μέθοδο CMIM και την mRMR. Έτσι καταλήγουμε στο συμπέρασμα ότι η προτεινόμενη μέθοδος (SU) έχει την τάση να δηλώνει σε ένα υψηλότερο ποσοστό μη ενεργές επιδράσεις ως ενεργές και σε πολύ χαμηλότερο ποσοστό τις ενεργές επιδράσεις ως ανενεργές. Ως εκ τούτου, η προτεινόμενη μέθοδος μπορεί να θεωρηθεί πράγματι συντηρητική υπό αυτή την έννοια.

Συμπερασματικά, η προτεινόμενη μέθοδος (SU) υπερτερεί σαφώς από τη CMIM μέθοδο, έχοντας ως αποτέλεσμα χαμηλότερες τιμές σφαλμάτων Τύπου I και Τύπου II. Επιπλέον, η προτεινόμενη μέθοδος (SU) φαίνεται να υπερέχει της μεθόδου mRMR, δεδομένου ότι έχει σημαντικά χαμηλότερα σφάλματα Τύπου II, και μικρή διαφορά στις τιμές των σφαλμάτων Τύπου I συγκρινόμενη με την mRMR. Η προτεινόμενη μέθοδος (SU) μπορεί να θεωρηθεί συντηρητική και γενικά επιτυγχάνει μια σταθερή απόδοση, δεδομένου ότι είναι η μόνη μέθοδος από την οποία προέκυψαν ταυτόχρονα χαμηλές τιμές και για τα δύο σφάλματα. Οι άλλοι δύο αλγόριθμοι που εξετάστηκαν φαίνεται να παρουσιάζουν σημαντικά υψηλότερα σφάλματα Τύπου I (CMIM) ή Τύπου II (mRMR). Επιπλέον, από την προτεινόμενη μέθοδο (SU) προέκυψαν τα χαμηλότερα σφάλματα Τύπου II, γεγονός που είναι ζωτικής σημασίας για την επιλογή μεταβλητών στην ανάλυση υπερκορεσμένων σχεδιασμών. Τα σφάλματα Τύπου I και Τύπου II είναι εξίσου σημαντικά και θα πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα, όμως δεδομένου ότι οι υπερκορεσμένοι σχεδιασμοί κυρίως χρησιμοποιούνται σε πειράματα κρησαρίσματος με στόχο να αναγνωρίσουν εκείνους τους παράγοντες που θα πρέπει να εξεταστούν περαιτέρω, οι χαμηλές τιμές σφάλματος Τύπου II είναι ιδιαίτερα επιθυμητές. Η προτεινόμενη μέθοδος κατορθώνει κάτι τέτοιο επιτυχώς σε σύγκριση με τις άλλες δύο μεθόδους που εξετάστηκαν.

Πίνακας 5.2: Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο IV με χρήση 1000 προσομοιώσεων

Μοντέλο	Type I (SU)	Type I (CMIM)	Type I (mRMR)	Type II (SU)	Type II (CMIM)	Type II (mRMR)
1	0.16	0.33	0.17	0.00	0.25	0.25
2	0.16	0.34	0.16	0.00	0.25	0.25
3	0.28	0.71	0.14	0.00	0.33	0.33
4	0.00	0.17	0.00	0.00	0.00	0.00
5	0.27	0.53	0.12	0.10	0.01	0.21
6	0.12	0.25	0.13	0.00	0.00	0.16
7	0.18	0.09	0.09	0.00	0.66	0.33
8	0.19	0.33	0.20	0.02	0.00	0.50
9	0.23	0.38	0.16	0.17	0.10	0.20
10	0.14	0.21	0.07	0.00	0.02	0.25
11	0.30	0.30	0.15	0.00	0.00	0.40
12	0.16	0.29	0.16	0.00	0.00	0.33
13	0.27	0.64	0.14	0.02	0.07	0.30
14	0.23	0.30	0.15	0.11	0.22	0.22
15	0.35	0.41	0.11	0.00	0.00	0.40
16	0.38	0.44	0.17	0.00	0.00	0.75
17	0.26	0.29	0.27	0.18	0.17	0.36
18	0.25	0.98	0.18	0.10	0.00	0.30
19	0.23	0.58	0.20	0.00	0.00	0.38
20	0.20	0.69	0.11	0.01	0.00	0.49
21	0.30	1.00	0.10	0.10	0.00	0.20
22	0.24	0.49	0.22	0.12	0.08	0.45
23	0.23	0.94	0.09	0.12	0.00	0.27
24	0.12	0.90	0.13	0.17	0.01	0.19
25	0.30	0.49	0.12	0.08	0.10	0.26
26	0.28	0.55	0.19	0.15	0.21	0.31
27	0.14	0.86	0.11	0.17	0.03	0.23
28	0.22	0.40	0.07	0.00	0.14	0.42
29	0.27	0.92	0.24	0.15	0.03	0.37
30	0.33	1.00	0.18	0.09	0.00	0.45
31	0.44	0.78	0.04	0.04	0.03	0.41
32	0.15	0.95	0.02	0.00	0.00	0.18
33	0.18	0.91	0.11	0.09	0.00	0.25
34	0.36	0.80	0.14	0.11	0.00	0.41
35	0.05	0.56	0.02	0.00	0.00	0.14
36	0.24	0.70	0.08	0.11	0.00	0.27
37	0.25	0.46	0.24	0.14	0.09	0.36
38	0.11	1.00	0.11	0.00	0.00	0.50
39	0.27	0.52	0.12	0.07	0.01	0.33
40	0.15	0.09	0.08	0.00	0.10	0.11
41	0.14	0.16	0.14	0.00	0.64	0.33
42	0.37	0.00	0.12	0.00	0.50	0.50
43	0.25	0.12	0.12	0.00	0.49	0.00
44	0.42	0.57	0.14	0.33	0.00	0.34
Average	**0.23**	**0.53**	**0.13**	**0.06**	**0.10**	**0.31**

5.4.2 Συγκριτική Μελέτη: Count Απόκριση

Από όσο καλά γνωρίζουμε, υπάρχουν αλγόριθμοι επιλογής μεταβλητών που έχουν σχεδιαστεί για δυαδικά, κατηγορικά ή συνεχή δεδομένα απόκρισης, αλλά υπάρχει έλλειψη στη βιβλιογραφία για δεδομένα απόκρισης τα οποία αποτελούν μετρήσεις (counts). Για αυτόν το λόγο, η αποδοτικότητα της προτεινόμενης μεθόδου για δεδομένα απόκρισης τα οποία προέρχονται από την Poisson, γεωμετρική και αρνητική διωνυμική κατανομή, εξετάστηκε και αξιολογήθηκε μέσα από μια εμπειρική μελέτη η οποία περιελάμβανε εκτενή πειράματα προσομοίωσης και δεν ήταν δυνατή η πραγματοποίηση συγκρίσεων με άλλες μεθόδους στην

υπάρχουσα βιβλιογραφία.

Τα αποτελέσματα που προκύπτουν από την εφαρμογή της προτεινόμενης μεθόδου για δεδομένα απόκρισης τα οποία αποτελούν μετρήσεις και συγκεκριμένα προέρχονται από την Poisson, γεωμετρική και αρνητική διωνυμική κατανομή συνοψίζονται στους Πίνακες 5.3, 5.4 και 5.5, αντιστοίχως. Σε κάθε πίνακα, η πρώτη στήλη αναφέρεται στον αριθμό που αντιστοιχεί στο μοντέλο που χρησιμοποιήθηκε, ενώ οι δύο στήλες με τίτλο "Type I" και "Type II" παρουσιάζουν τις μέσες τιμές των σφαλμάτων Τύπου I και Τύπου II, αντίστοιχα, που προέκυψαν μετά από 1000 προσομοιώσεις για κάθε θεωρούμενο Σενάριο.

Παρατηρούμε από τους Πίνακες 5.3, 5.4 και 5.5, ότι τα σφάλματα Τύπου I είναι υψηλότερα από τα σφάλματα Τύπου II για όλα τα θεωρούμενα μοντέλα. Για την πλειονότητα των θεωρούμενων μοντέλων, οι τιμές των σφαλμάτων Τύπου I βρίσκονται στο εύρος [0.15, 0.35], όπως άλλωστε φαίνεται και από τα Σχήματα 5.1, 5.2 και 5.3. Παρ' όλα αυτά, οι πιο φουσκωμένες τιμές των σφαλμάτων Τύπου I δεν αποτελούν πρόβλημα στη μελέτη μας, δεδομένου ότι σε καταστάσεις που ισχύει η αρχή της σποραδικότητας, είναι πολύ πιθανό να προκύψουν σφάλματα Τύπου I, όπως διαπιστώθηκε από τον Lin στην εργασία [149].

Επιπλέον, παρατηρούμε από τους Πίνακες 5.3, 5.4 και 5.5, ότι τα σφάλματα Τύπου II είναι σημαντικά χαμηλότερα από τα σφάλματα Τύπου I για όλα τα θεωρούμενα μοντέλα. Για την πλειονότητα των θεωρούμενων μοντέλων, οι τιμές των σφαλμάτων Τύπου II βρίσκονται στο εύρος [0, 0.15], όπως άλλωστε φαίνεται και από τα Σχήματα 5.1, 5.2 και 5.3. Επίσης, από τους Πίνακες 5.3, 5.4 και 5.5 προκύπτει ότι σε μερικές από τις περιπτώσεις που εξετάστηκαν, τα σφάλματα Τύπου II είναι 0.00, και αυτές αντιστοιχούν σε περιπτώσεις όπου όλες οι ενεργές επιδράσεις ανιχνεύθηκαν σωστά (25 από τα 44 μοντέλα για το Σενάριο I, 23 από τα 44 μοντέλα για το Σενάριο II, και 16 από τα 44 μοντέλα για το Σενάριο III).

Μια γενική εικόνα που προκύπτει από τα Σχήματα 5.1, 5.2 και 5.3 είναι ότι η προτεινόμενη μέθοδος έχει την τάση να δηλώνει σε ένα υψηλότερο ποσοστό μη ενεργές επιδράσεις ως ενεργές και σε πολύ χαμηλότερο ποσοστό τις ενεργές επιδράσεις ως ανενεργές. Ως εκ τούτου, η προτεινόμενη μέθοδος μπορεί να θεωρηθεί πράγματι συντηρητική υπό αυτή την έννοια. Επιπλέον, σημειώνουμε εδώ ότι τα σφάλματα παραμένουν χαμηλά ακόμα και για τα μοντέλα στα οποία η αρχή της σποραδικότητας είναι ασθενέστερη, δηλαδή, μοντέλα στα οποία ο αριθμός των ενεργών παραγόντων είναι κοντά στο $\frac{m}{2}$.

Επιπροσθέτως, τα Σχήματα 5.4 και 5.5 δείχνουν ότι η προτεινόμενη μέθοδος έχει παρόμοια απόδοση για οποιοδήποτε θεωρούμενο Σενάριο (I, II και III). Από τους Πίνακες 5.3, 5.4 και 5.5, μπορούμε εύκολα να παρατηρήσουμε ότι σε μερικές από τις περιπτώσεις που εξετάστηκαν, οι τιμές για τα σφάλματα είναι ταυτόσημες, δηλαδή, για 30 από τα 44 μοντέλα, τα σφάλματα Τύπου I και Τύπου II της κατανομής Poisson είναι πανομοιότυπα με εκείνα της γεωμετρικής κατανομής, για 21 από τα 44 μοντέλα, τα σφάλματα Τύπου I και Τύπου II της κατανομής Poisson είναι πανομοιότυπα με εκείνα της αρνητικής διωνυμικής κατανομής, για 19 από τα 44 μοντέλα, τα σφάλματα Τύπου I και Τύπου II της γεωμετρικής κατανομής είναι πανομοιότυπα με εκείνα της αρνητικής διωνυμικής κατανομής. Επιπλέον, από τους Πίνακες 5.3, 5.4 και 5.5, προκύπτει ότι οι μέσες τιμές για τα σφάλματα Τύπου I για τα 44 θεωρούμενα μοντέλα είναι 0.29, 0.29 και 0.30 για την Poisson, γεωμετρική και αρνητική διωνυμική κατανομή, αντίστοιχα. Περαιτέρω, από τα αποτελέσματα τα οποία παρουσιάζονται στους Πίνακες 5.3, 5.4 και 5.5, επιβεβαιώνεται η καλή απόδοση της προτεινόμενης μεθόδου, δεδομένου ότι η τιμή της ισχύος (1 -Type II error rate) είναι ιδιαίτερα υψηλή, συγκεκριμένα είναι περίπου ίση με 0.95, 0.94 και 0.92 για την Poisson, γεωμετρική και αρνητική διωνυμική κατανομή, αντίστοιχα, γεγονός το οποίο υποδεικνύει ότι σχεδόν όλες οι ενεργές επιδράσεις ανιχνεύονται πράγματι σωστά. Με βάση τα παραπάνω, η προτεινόμενη μέθοδος θα μπορούσε να χαρακτηριστεί ως εύρωστη για τη μοντελοποίηση δεδομένων απόκρισης τα οποία αποτελούν μετρήσεις.

Πίνακας 5.3: Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο I με χρήση 1000 προσομοιώσεων

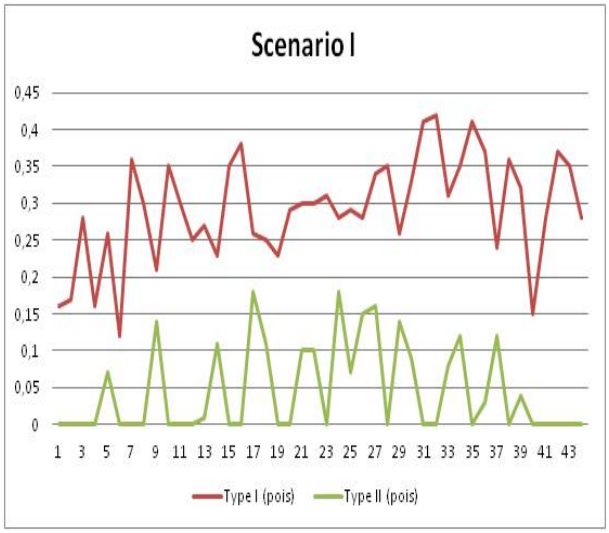
Μοντέλο	Type I	Type II	Μοντέλο	Type I	Type II
1	0.16	0.00	23	0.31	0.00
2	0.17	0.00	24	0.28	0.18
3	0.28	0.00	25	0.29	0.07
4	0.16	0.00	26	0.28	0.15
5	0.26	0.07	27	0.34	0.16
6	0.12	0.00	28	0.35	0.00
7	0.36	0.00	29	0.26	0.14
8	0.30	0.00	30	0.33	0.09
9	0.21	0.14	31	0.41	0.00
10	0.35	0.00	32	0.42	0.00
11	0.30	0.00	33	0.31	0.08
12	0.25	0.00	34	0.35	0.12
13	0.27	0.01	35	0.41	0.00
14	0.23	0.11	36	0.37	0.03
15	0.35	0.00	37	0.24	0.12
16	0.38	0.00	38	0.36	0.00
17	0.26	0.18	39	0.32	0.04
18	0.25	0.11	40	0.15	0.00
19	0.23	0.00	41	0.28	0.00
20	0.29	0.00	42	0.37	0.00
21	0.30	0.10	43	0.35	0.00
22	0.30	0.10	44	0.28	0.00

Πίνακας 5.4: Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο II με χρήση 1000 προσομοιώσεων

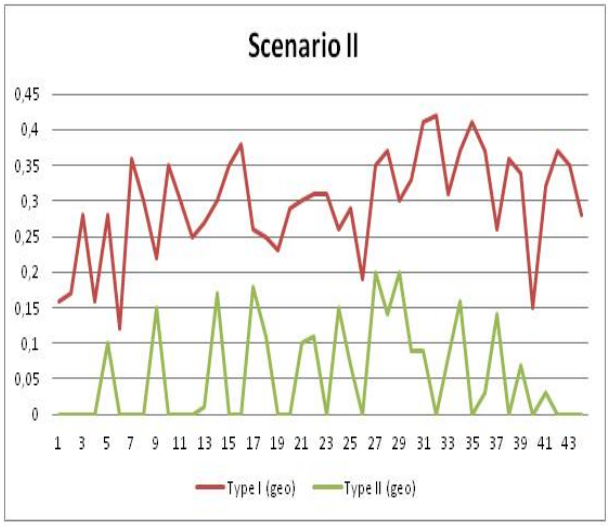
Μοντέλο	Type I	Type II	Μοντέλο	Type I	Type II
1	0.16	0.00	23	0.31	0.00
2	0.17	0.00	24	0.26	0.15
3	0.28	0.00	25	0.29	0.07
4	0.16	0.00	26	0.19	0.00
5	0.28	0.10	27	0.35	0.20
6	0.12	0.00	28	0.37	0.14
7	0.36	0.00	29	0.30	0.20
8	0.30	0.00	30	0.33	0.09
9	0.22	0.15	31	0.41	0.09
10	0.35	0.00	32	0.42	0.00
11	0.30	0.00	33	0.31	0.08
12	0.25	0.00	34	0.37	0.16
13	0.27	0.01	35	0.41	0.00
14	0.30	0.17	36	0.37	0.03
15	0.35	0.00	37	0.26	0.14
16	0.38	0.00	38	0.36	0.00
17	0.26	0.18	39	0.34	0.07
18	0.25	0.11	40	0.15	0.00
19	0.23	0.00	41	0.32	0.03
20	0.29	0.00	42	0.37	0.00
21	0.30	0.10	43	0.35	0.00
22	0.31	0.11	44	0.28	0.00

Πίνακας 5.5: Εμπειρική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο III με χρήση 1000 προσομοιώσεων

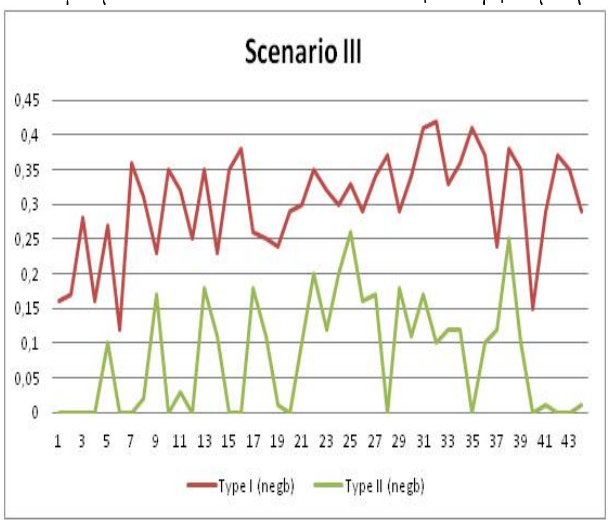
Μοντέλο	Type I	Type II	Μοντέλο	Type I	Type II
1	0.16	0.00	23	0.32	0.12
2	0.17	0.00	24	0.30	0.20
3	0.28	0.00	25	0.33	0.26
4	0.16	0.00	26	0.29	0.16
5	0.27	0.10	27	0.34	0.17
6	0.12	0.00	28	0.37	0.00
7	0.36	0.00	29	0.29	0.18
8	0.31	0.02	30	0.34	0.11
9	0.23	0.17	31	0.41	0.17
10	0.35	0.00	32	0.42	0.10
11	0.32	0.03	33	0.33	0.12
12	0.25	0.00	34	0.35	0.12
13	0.35	0.18	35	0.41	0.00
14	0.23	0.11	36	0.37	0.10
15	0.35	0.00	37	0.24	0.12
16	0.38	0.00	38	0.38	0.25
17	0.26	0.18	39	0.35	0.10
18	0.25	0.11	40	0.15	0.00
19	0.24	0.01	41	0.29	0.01
20	0.29	0.00	42	0.37	0.00
21	0.30	0.10	43	0.35	0.00
22	0.35	0.20	44	0.29	0.01



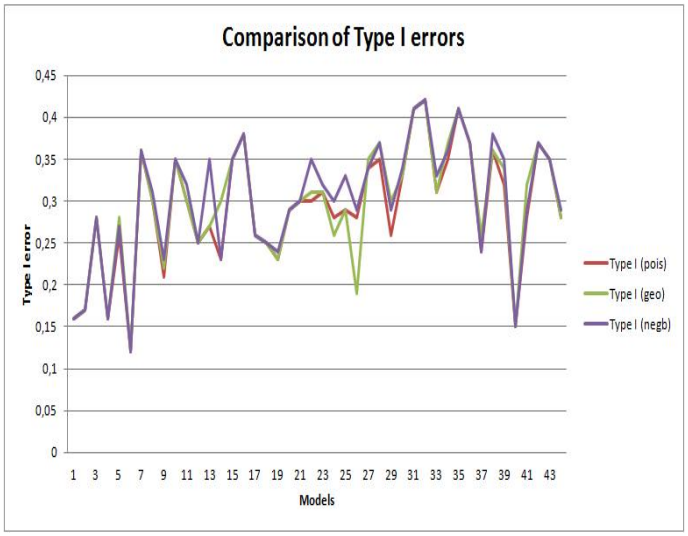
Σχήμα 5.1: Σφάλματα Τύπου I και Τύπου II για την κατανομή Poisson



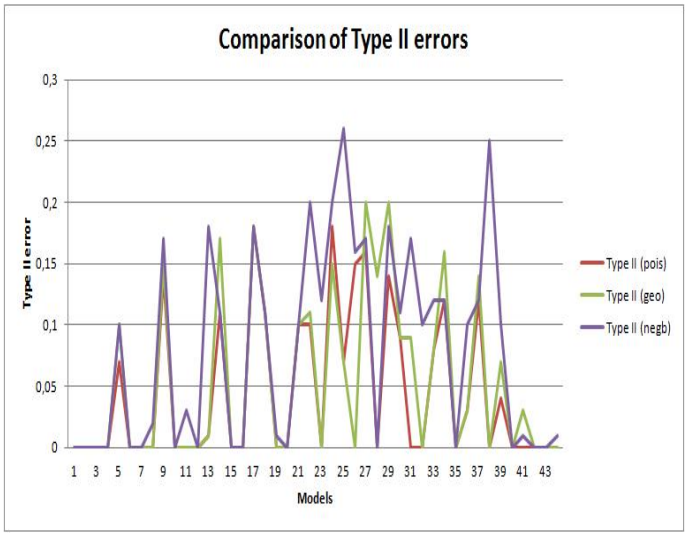
Σχήμα 5.2: Σφάλματα Τύπου I και Τύπου II για τη γεωμετρική κατανομή



Σχήμα 5.3: Σφάλματα Τύπου I και Τύπου II για την αρνητική διωνυμική κατανομή



Σχήμα 5.4: Σύγκριση σφαλμάτων Τύπου I για τα Σενάρια I, II και III



Σχήμα 5.5: Σύγκριση σφαλμάτων Τύπου II για τα Σενάρια I, II και III

5.4.3 Συγκριτική Μελέτη: Επιλογή Τιμών Κατωφλιού

Η απόφαση σχετικά με την τιμή κατωφλιού w, η οποία στην πράξη καθορίζει το μέγιστο αριθμό των σημαντικών παραγόντων που μπορούν να αναγνωριστούν, παίζει σημαντικό ρόλο στην ανάπτυξη της μεθόδου. Οι Box και Meyer [23] και πρόσφατα οι Marley και Woods [160], διεπίστωσαν ότι η αποτελεσματικότητα των υπερκορεσμένων σχεδιασμών για την εύρεση των ενεργών παραγόντων έγκειται στην απαίτηση του πειραματιστή "ο αριθμός των ενεργών παραγόντων να είναι μικρός" (η γνωστή αρχή της σποραδικότητας). Στην επιστημονική εργασία [7] θέσαμε το $w = m/2$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου πίνακα σχεδιασμού, δηλαδή ο αριθμός των παραγόντων υπό εξέταση. Η απόφαση για το w έγινε βάση της αρχής της σποραδικότητας [23]. Για να ποσοτικοποιήσουμε τη σποραδικότητα στο σύνολο των δεδομένων κατά τη διάρκεια των πειραμάτων προσομοίωσης, χρησιμοποιήσαμε την ακόλουθη διαδικασία: Οι πραγματικά ενεργές μεταβλητές επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων και μόνο οι κύριες επιδράσεις ελήφθησαν υπόψη. Για κάθε πείραμα, εκτιμήσαμε όλες τις κύριες επιδράσεις, οι οποίες στη συνέχεια επισημάνθηκαν ως ενεργές ή μη ενεργές. Έπειτα υπολογίζοντας το ποσοστό των εκτιμώμενων ενεργών επιδράσεων που εμφανίζονται σε κάθε πείραμα, εκτιμήσαμε ότι ο αριθμός των ενεργών επιδράσεων που μπορεί να αναγνωριστεί, δεν υπερβαίνει το $w = \frac{m}{2}$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου σχεδιασμού. Αυτή η επιλογή της τιμής κατωφλιού στους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς μπορεί να θεωρηθεί ως συντηρητική. Για αυτόν το λόγο, θεωρήσαμε εδώ μία νέα τιμή κατωφλιού προς εξέταση ($w < \frac{m}{2}$) προκειμένου να εξασφαλίσουμε την καλύτερη δυνατή επιλογή της τιμής κατωφλιού w.

Η απόφαση σχετικά με την τιμή κατωφλιού του SU μέτρου, η οποία καθορίζει το κατά πόσο ένας παράγοντας είναι σημαντικός ή όχι, παίζει σημαντικό ρόλο στην ανάπτυξη της μεθόδου. Για αυτόν το λόγο, πολλές διαφορετικές τιμές κατωφλιού (0.001, 0.01, 0.05, 0.1, 0.25, 0.4, median(**su**)) εξετάστηκαν στην εργασία [7] με στόχο να βρεθεί η βέλτιστη τιμή κατωφλιού του SU για την προτεινόμενη μέθοδο όταν η απόκριση είναι δίτιμη. Οι πρώτες έξι τιμές κατωφλιού ήταν σταθεροί αριθμοί, οι οποίοι επιλέχθηκαν σύμφωνα με τις τιμές SU που παρατηρήθηκαν στα περισσότερα από τα πειράματά μας, ενώ η τιμή κατωφλιού median(**su**) βασίστηκε στις εκτιμώμενες τιμές του SU, συγκεκριμένα στο εύρος των τιμών του SU οι οποίες παρατηρήθηκαν στα πειράματά μας. Τελικά, καθορίσαμε στην εργασία [7] την τιμή κατωφλιού του SU να είναι ίση με το median(**su**), δεδομένου ότι η επιλογή αυτή παρατηρήθηκε να είναι η βέλτιστη επιλογή μετά από δοκιμές προσομοίωσης, έχοντας ως αποτέλεσμα χαμηλότερες τιμές σφάλματος.

Σε αυτή τη μελέτη προσομοίωσης, εξετάσαμε επιπλέον νέες διαφορετικές τιμές κατωφλιού με στόχο να βρεθεί η βέλτιστη τιμή κατωφλιού του SU για τη μοντελοποίση σε διακριτά δεδομένα απόκρισης. Οι νέες τιμές κατωφλιού που εξετάσαμε είναι οι $SU \geqslant$ median(**su**), $SU >$ median(**su**), $SU \geqslant$ mean(**su**), $SU >$ mean(**su**), $SU \geqslant$ geometric mean(**su**) και $SU >$ geometric mean(**su**).

Σημειώνουμε εδώ ότι παρόλο που εξετάστηκαν εκ νέου πολλές και διαφορετικές τιμές κατωφλιού, τελικά, καθορίσαμε την τιμή κατωφλιού του SU και του w να είναι ίση με το median(**su**) και το $\frac{m}{2}$, αντίστοιχα, δεδομένου ότι αυτές οι τιμές παρατηρήθηκαν να είναι οι βέλτιστες επιλογές μετά από εκτενείς δοκιμές προσομοίωσης, έχοντας ως αποτέλεσμα χαμηλότερες τιμές σφάλματος συγκριτικά με άλλες τιμές κατωφλιού (βλέπε Πίνακες 5.6, 5.7, 5.8 και 5.9 για περισσότερες λεπτομέρειες).

Για κάθε ένα από τα 44 θεωρούμενα μοντέλα που παρουσιάζονται στον Πίνακα 5.1, 1000 σύνολα δεδομένων παράχθηκαν για κάθε εξεταζόμενο Σενάριο, και τα αποτελέσματα που ελήφθησαν μετά την εφαρμογή της προτεινόμενης μεθόδου παρουσιάζονται στους Πίνακες 5.6, 5.7, 5.8 και 5.9 για το Σενάριο I, II, III και IV, αντίστοιχα, σύμφωνα πάντα με τις τιμές κατωφλιού που εξετάστηκαν. Στην πρώτη στήλη, δίνεται ο αριθμός που αντιστοιχεί σε κάθε μοντέλο που χρησιμοποιείται. Οι στήλες με τίτλο "Type I" και "Type II" παρουσιάζουν

τις μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II που προέκυψαν μετά από 1000 προσομοιώσεις, σε αντιστοιχία με κάθε εξεταζόμενη τιμή κατωφλιού. Στην τελευταία γραμμή του κάθε πίνακα δίνονται οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II για τα 44 θεωρούμενα μοντέλα.

Τα αποτελέσματα τα οποία προέκυψαν για κάθε ένα από τα θεωρούμενα Σενάρια μετά από τη σύγκριση των διάφορων τιμών κατωφλιού συνοψίζονται παρακάτω.

- Σενάριο I: Από τον Πίνακα 5.3, παρατηρούμε ότι οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II για τα 44 θεωρούμενα μοντέλα είναι 0.29 και 0.04, αντίστοιχα, με θεωρούμενες τιμές κατωφλιού για το SU την τιμή median(**su**) και για το w την τιμή $w = \frac{m}{2}$. Με αυτές τις τιμές κατωφλιού, η προτεινόμενη μέθοδος έχει σημαντικά χαμηλότερα σφάλματα Τύπου II, και ελαφρώς υψηλότερα σφάλματα Τύπου I (βλέπε Πίνακα 5.6 για σύγκριση).

- Σενάριο II: Από τον Πίνακα 5.4, παρατηρούμε ότι οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II για τα 44 θεωρούμενα μοντέλα είναι 0.29 και 0.05, αντίστοιχα, με θεωρούμενες τιμές κατωφλιού για το SU την τιμή median(**su**) και για το w την τιμή $w = \frac{m}{2}$. Με αυτές τις τιμές κατωφλιού, η προτεινόμενη μέθοδος έχει σημαντικά χαμηλότερα σφάλματα Τύπου II, και ελαφρώς υψηλότερα σφάλματα Τύπου I (βλέπε Πίνακα 5.7 για σύγκριση).

- Σενάριο III: Από τον Πίνακα 5.5, παρατηρούμε ότι οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II για τα 44 θεωρούμενα μοντέλα είναι 0.30 και 0.07, αντίστοιχα, με θεωρούμενες τιμές κατωφλιού για το SU την τιμή median(**su**) και για το w την τιμή $w = \frac{m}{2}$. Με αυτές τις τιμές κατωφλιού, η προτεινόμενη μέθοδος έχει σημαντικά χαμηλότερα σφάλματα Τύπου II, και ελαφρώς υψηλότερα σφάλματα Τύπου I (βλέπε Πίνακα 5.8 για σύγκριση).

- Σενάριο IV: Από τον Πίνακα 5.2, παρατηρούμε ότι οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II για τα 44 θεωρούμενα μοντέλα είναι 0.23 και 0.06, αντίστοιχα, με θεωρούμενες τιμές κατωφλιού για το SU την τιμή median(**su**) και για το w την τιμή $w = \frac{m}{2}$. Με αυτές τις τιμές κατωφλιού, η προτεινόμενη μέθοδος έχει σημαντικά χαμηλότερα σφάλματα Τύπου II, και ελαφρώς υψηλότερα σφάλματα Τύπου I (βλέπε Πίνακα 5.9 για σύγκριση).

Τα παραπάνω συγκριτικά αποτελέσματα υποδεικνύουν ότι η προτεινόμενη μέθοδος, επιλέγοντας ως τιμή κατωφλιού για το SU την τιμή median(**su**) και για το w την τιμή $w = \frac{m}{2}$, είναι αρκετά εύρωστη για τη μοντελοποίηση δεδομένων απόκρισης τα οποία αποτελούν μετρήσεις. Οι μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II για τα 44 θεωρούμενα μοντέλα είναι σχεδόν πανομοιότυπες για τα Σενάρια I, II και III.

Σε γενικές γραμμές, καταλήγουμε στο συμπέρασμα ότι για κάθε ένα από τα εξεταζόμενα Σενάρια, η προτεινόμενη μέθοδος φαίνεται να έχει ικανοποιητική απόδοση επιλέγοντας ως τιμή κατωφλιού για το SU την τιμή median(**su**) και για το w την τιμή $w = \frac{m}{2}$. Με αυτή την επιλογή των τιμών κατωφλιού, η προτεινόμενη μέθοδος επιτυγχάνει τις χαμηλότερες τιμές για τα σφάλματα Τύπου II για όλα τα Σενάρια που εξετάστηκαν. Το γεγονός ότι η επιλογή αυτών των τιμών κατωφλιού οδηγεί σε ελαφρώς υψηλότερες τιμές για τα σφάλματα Τύπου I δεν μας ενοχλεί, δεδομένου ότι σε καταστάσεις που ισχύει η αρχή της σποραδικότητας, είναι πολύ πιθανό να προκύψουν σφάλματα Τύπου I [149]. Άλλωστε, οι υπερκορεσμένοι σχεδιασμοί κυρίως χρησιμοποιούνται σε πειράματα κρησαρίσματος με στόχο να αναγνωρίσουν εκείνους τους παράγοντες που θα πρέπει να εξεταστούν περαιτέρω, και οι χαμηλές τιμές σφάλματος Τύπου II είναι περισσότερο επιθυμητές. Η προτεινόμενη μέθοδος επιλέγοντας ως τιμή κατωφλιού για το SU την τιμή median(**su**) και για το w την τιμή $w = \frac{m}{2}$ κατορθώνει κάτι τέτοιο επιτυχώς, σε σύγκριση με τις άλλες τιμές κατωφλιού που εξετάστηκαν.

Πίνακας 5.6: Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο I με χρήση 1000 προσομοιώσεων

$(w < \frac{k}{2})$	$SU \geqslant$ median(**su**)		$SU >$ median(**su**)		$SU \geqslant$ mean(**su**)		$SU >$ mean(**su**)		$SU \geqslant$ geomean(**su**)		$SU >$ geomean(**su**)	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
1	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.26
2	0.16	0.25	0.16	0.26	0.16	0.25	0.16	0.26	0.16	0.26	0.16	0.25
3	0.28	0.33	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.25	0.25	0.01	0.96	0.20	0.26	0.20	0.24	0.19	0.28	0.19	0.32
6	0.12	0.16	0.00	0.66	0.12	0.16	0.12	0.16	0.12	0.16	0.12	0.16
7	0.27	0.00	0.18	0.00	0.18	0.00	0.18	0.00	0.27	0.00	0.18	0.00
8	0.20	0.00	0.19	0.02	0.20	0.08	0.19	0.02	0.19	0.01	0.19	0.03
9	0.23	0.29	0.20	0.25	0.20	0.27	0.20	0.26	0.22	0.30	0.20	0.27
10	0.28	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.14	0.00
11	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20
12	0.16	0.00	0.16	0.00	0.16	0.00	0.16	0.00	0.16	0.00	0.16	0.00
13	0.22	0.06	0.11	0.27	0.14	0.25	0.15	0.26	0.15	0.22	0.16	0.22
14	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.22
15	0.29	0.00	0.00	0.60	0.29	0.00	0.29	0.00	0.29	0.00	0.29	0.00
16	0.38	0.25	0.05	0.75	0.38	0.25	0.38	0.25	0.38	0.25	0.38	0.25
17	0.26	0.27	0.26	0.27	0.26	0.27	0.26	0.27	0.26	0.27	0.26	0.27
18	0.27	0.23	0.14	0.30	0.15	0.31	0.13	0.30	0.14	0.30	0.13	0.30
19	0.23	0.11	0.01	0.62	0.21	0.13	0.21	0.12	0.23	0.11	0.22	0.12
20	0.34	0.02	0.18	0.00	0.22	0.08	0.21	0.09	0.20	0.08	0.21	0.05
21	0.25	0.10	0.05	0.30	0.05	0.30	0.05	0.30	0.05	0.30	0.05	0.30
22	0.30	0.20	0.00	0.68	0.00	0.68	0.00	0.66	0.28	0.20	0.25	0.22
23	0.31	0.12	0.22	0.14	0.21	0.15	0.20	0.19	0.23	0.12	0.23	0.12
24	0.24	0.19	0.09	0.23	0.10	0.19	0.10	0.23	0.10	0.19	0.08	0.26
25	0.25	0.06	0.10	0.43	0.10	0.45	0.11	0.41	0.13	0.40	0.12	0.40
26	0.23	0.15	0.08	0.36	0.08	0.40	0.08	0.39	0.12	0.32	0.11	0.35
27	0.31	0.19	0.07	0.23	0.08	0.25	0.07	0.24	0.08	0.25	0.06	0.28
28	0.33	0.00	0.22	0.14	0.22	0.14	0.22	0.14	0.22	0.14	0.22	0.14
29	0.26	0.21	0.13	0.41	0.13	0.36	0.12	0.38	0.12	0.40	0.12	0.35
30	0.29	0.09	0.22	0.27	0.22	0.27	0.22	0.27	0.22	0.27	0.22	0.27
31	0.41	0.00	0.22	0.24	0.23	0.22	0.23	0.24	0.28	0.16	0.23	0.21
32	0.39	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.14	0.00
33	0.31	0.16	0.20	0.21	0.19	0.17	0.20	0.18	0.23	0.23	0.19	0.18
34	0.34	0.15	0.23	0.34	0.22	0.37	0.20	0.37	0.18	0.41	0.20	0.35
35	0.38	0.00	0.02	0.00	0.02	0.00	0.02	0.00	0.02	0.00	0.02	0.00
36	0.34	0.10	0.07	0.19	0.09	0.18	0.08	0.18	0.08	0.19	0.07	0.19
37	0.23	0.23	0.00	0.76	0.20	0.28	0.21	0.26	0.21	0.25	0.21	0.26
38	0.33	0.00	0.11	0.00	0.11	0.00	0.11	0.00	0.11	0.00	0.11	0.00
39	0.28	0.08	0.20	0.10	0.21	0.10	0.20	0.08	0.24	0.16	0.19	0.12
40	0.15	0.11	0.07	0.11	0.07	0.11	0.07	0.11	0.07	0.11	0.07	0.11
41	0.14	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.15	0.01
42	0.25	0.00	0.12	0.50	0.12	0.50	0.12	0.50	0.12	0.50	0.12	0.50
43	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00
44	0.28	0.33	0.14	0.33	0.28	0.33	0.28	0.33	0.28	0.33	0.28	0.33
Average	**0.26**	**0.12**	**0.12**	**0.27**	**0.16**	**0.20**	**0.16**	**0.20**	**0.17**	**0.19**	**0.16**	**0.19**

Πίνακας 5.7: Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο II με χρήση 1000 προσομοιώσεων

$(w < \frac{k}{2})$	$SU \geqslant$ median(**su**)		$SU >$ median(**su**)		$SU \geqslant$ mean(**su**)		$SU >$ mean(**su**)		$SU \geqslant$ geomean(**su**)		$SU >$ geomean(**su**)	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
1	0.16	0.25	0.16	0.25	0.16	0.26	0.16	0.25	0.16	0.25	0.16	0.26
2	0.16	0.25	0.16	0.25	0.16	0.26	0.16	0.25	0.16	0.26	0.16	0.25
3	0.28	0.33	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.25	0.26	0.01	0.92	0.18	0.34	0.20	0.32	0.19	0.25	0.15	0.33
6	0.12	0.16	0.00	1.00	0.12	0.16	0.12	0.16	0.12	0.16	0.12	0.16
7	0.27	0.00	0.18	0.00	0.18	0.00	0.18	0.00	0.18	0.00	0.18	0.00
8	0.20	0.01	0.19	0.06	0.19	0.02	0.19	0.01	0.19	0.05	0.20	0.00
9	0.21	0.27	0.16	0.32	0.16	0.29	0.17	0.30	0.17	0.30	0.15	0.35
10	0.28	0.00	0.13	0.01	0.14	0.00	0.14	0.00	0.14	0.00	0.14	0.00
11	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20
12	0.16	0.00	0.15	0.01	0.16	0.00	0.16	0.00	0.16	0.00	0.16	0.00
13	0.24	0.08	0.12	0.27	0.17	0.22	0.14	0.25	0.15	0.23	0.19	0.20
14	0.30	0.33	0.30	0.33	0.30	0.33	0.30	0.33	0.30	0.33	0.30	0.33
15	0.41	0.40	0.41	0.40	0.41	0.40	0.41	0.40	0.41	0.40	0.41	0.40
16	0.38	0.25	0.05	0.75	0.38	0.25	0.38	0.25	0.38	0.25	0.38	0.25
17	0.39	0.45	0.20	0.72	0.20	0.72	0.20	0.71	0.20	0.72	0.20	0.71
18	0.25	0.20	0.12	0.30	0.13	0.30	0.12	0.30	0.14	0.29	0.14	0.30
19	0.24	0.13	0.03	0.55	0.19	0.15	0.20	0.17	0.21	0.13	0.19	0.14
20	0.34	0.04	0.19	0.00	0.06	0.29	0.07	0.31	0.21	0.03	0.23	0.04
21	0.30	0.20	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30
22	0.30	0.21	0.00	0.65	0.02	0.66	0.00	0.67	0.00	0.64	0.00	0.68
23	0.27	0.01	0.22	0.01	0.09	0.48	0.09	0.48	0.22	0.01	0.22	0.01
24	0.19	0.12	0.14	0.14	0.10	0.31	0.07	0.33	0.09	0.33	0.10	0.33
25	0.25	0.08	0.12	0.43	0.09	0.44	0.08	0.46	0.09	0.43	0.12	0.44
26	0.14	0.00	0.08	0.15	0.00	0.29	0.00	0.28	0.00	0.28	0.00	0.29
27	0.30	0.19	0.11	0.32	0.08	0.30	0.12	0.27	0.12	0.26	0.09	0.31
28	0.40	0.28	0.40	0.28	0.40	0.28	0.40	0.28	0.40	0.28	0.40	0.28
29	0.29	0.25	0.12	0.47	0.12	0.39	0.13	0.37	0.13	0.37	0.13	0.40
30	0.33	0.18	0.33	0.18	0.18	0.27	0.18	0.27	0.18	0.27	0.18	0.27
31	0.42	0.14	0.25	0.12	0.24	0.14	0.23	0.12	0.33	0.16	0.32	0.15
32	0.39	0.00	0.15	0.02	0.14	0.00	0.14	0.00	0.13	0.00	0.14	0.00
33	0.31	0.17	0.14	0.24	0.16	0.22	0.14	0.26	0.15	0.19	0.15	0.22
34	0.34	0.16	0.13	0.48	0.19	0.42	0.20	0.37	0.19	0.41	0.24	0.42
35	0.38	0.00	0.18	0.00	0.16	0.00	0.15	0.00	0.16	0.00	0.16	0.00
36	0.46	0.50	0.46	0.50	0.46	0.49	0.46	0.50	0.46	0.49	0.46	0.48
37	0.25	0.25	0.00	0.76	0.21	0.32	0.21	0.29	0.20	0.30	0.23	0.32
38	0.33	0.00	0.11	0.00	0.11	0.00	0.11	0.00	0.11	0.00	0.11	0.00
39	0.27	0.07	0.19	0.22	0.18	0.17	0.22	0.15	0.17	0.14	0.21	0.18
40	0.15	0.11	0.07	0.11	0.07	0.13	0.07	0.11	0.07	0.11	0.07	0.11
41	0.28	0.33	0.28	0.33	0.00	0.33	0.00	0.33	0.27	0.34	0.28	0.33
42	0.25	0.02	0.25	0.00	0.00	0.50	0.00	0.50	0.25	0.00	0.25	0.00
43	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.24	0.00	0.24	0.00
44	0.29	0.34	0.28	0.35	0.27	0.39	0.28	0.33	0.29	0.35	0.28	0.33
Average	**0.27**	**0.16**	**0.16**	**0.29**	**0.16**	**0.26**	**0.16**	**0.26**	**0.18**	**0.23**	**0.18**	**0.23**

Πίνακας 5.8: Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο III με χρήση 1000 προσομοιώσεων

$(w < \frac{k}{2})$	$SU \geqslant$ median(**su**)		$SU >$ median(**su**)		$SU \geqslant$ mean(**su**)		$SU >$ mean(**su**)		$SU \geqslant$ geomean(**su**)		$SU >$ geomean(**su**)	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
1	0.16	0.25	0.16	0.25	0.16	0.26	0.16	0.25	0.15	0.28	0.16	0.25
2	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25
3	0.28	0.33	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.25	0.26	0.00	0.88	0.15	0.34	0.16	0.30	0.14	0.36	0.15	0.33
6	0.12	0.16	0.00	1.00	0.12	0.16	0.12	0.16	0.12	0.18	0.12	0.16
7	0.27	0.00	0.18	0.00	0.18	0.00	0.18	0.00	0.18	0.00	0.18	0.00
8	0.20	0.00	0.20	0.00	0.19	0.01	0.19	0.04	0.18	0.04	0.19	0.02
9	0.25	0.32	0.16	0.33	0.17	0.30	0.19	0.26	0.15	0.33	0.17	0.30
10	0.28	0.00	0.13	0.01	0.13	0.03	0.14	0.00	0.14	0.00	0.13	0.03
11	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20
12	0.17	0.00	0.14	0.03	0.16	0.00	0.15	0.01	0.15	0.01	0.16	0.00
13	0.38	0.25	0.33	0.28	0.33	0.28	0.33	0.28	0.33	0.28	0.33	0.28
14	0.38	0.44	0.38	0.44	0.38	0.44	0.38	0.44	0.38	0.44	0.38	0.44
15	0.35	0.20	0.35	0.20	0.35	0.20	0.35	0.20	0.35	0.20	0.35	0.20
16	0.38	0.25	0.05	0.75	0.38	0.25	0.38	0.25	0.38	0.25	0.38	0.25
17	0.26	0.27	0.20	0.27	0.13	0.45	0.12	0.46	0.20	0.27	0.20	0.27
18	0.25	0.20	0.12	0.30	0.15	0.31	0.15	0.30	0.13	0.30	0.12	0.30
19	0.23	0.11	0.04	0.56	0.20	0.14	0.19	0.16	0.21	0.13	0.20	0.14
20	0.34	0.04	0.22	0.01	0.21	0.01	0.20	0.04	0.22	0.02	0.22	0.06
21	0.25	0.10	0.25	0.10	0.10	0.30	0.10	0.30	0.10	0.30	0.10	0.30
22	0.30	0.09	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20
23	0.36	0.25	0.36	0.25	0.19	0.25	0.18	0.25	0.19	0.25	0.18	0.25
24	0.43	0.49	0.14	0.49	0.17	0.49	0.14	0.49	0.15	0.49	0.17	0.50
25	0.42	0.44	0.41	0.45	0.43	0.45	0.42	0.44	0.43	0.45	0.42	0.44
26	0.23	0.15	0.17	0.30	0.09	0.42	0.09	0.43	0.09	0.45	0.09	0.45
27	0.30	0.19	0.07	0.30	0.09	0.21	0.07	0.25	0.08	0.28	0.07	0.28
28	0.33	0.00	0.22	0.14	0.22	0.14	0.22	0.14	0.22	0.14	0.22	0.14
29	0.27	0.22	0.13	0.40	0.12	0.47	0.13	0.48	0.13	0.45	0.12	0.38
30	0.37	0.27	0.37	0.27	0.18	0.45	0.18	0.45	0.18	0.45	0.18	0.45
31	0.42	0.11	0.34	0.25	0.25	0.22	0.25	0.22	0.25	0.22	0.25	0.22
32	0.39	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.15	0.02	0.14	0.00
33	0.32	0.19	0.16	0.22	0.17	0.23	0.18	0.22	0.13	0.26	0.18	0.30
34	0.35	0.17	0.16	0.47	0.20	0.43	0.24	0.42	0.24	0.40	0.22	0.48
35	0.38	0.00	0.04	0.00	0.03	0.00	0.05	0.00	0.04	0.00	0.03	0.00
36	0.40	0.29	0.39	0.29	0.40	0.29	0.40	0.29	0.40	0.29	0.40	0.29
37	0.24	0.24	0.04	0.75	0.21	0.29	0.22	0.25	0.19	0.33	0.21	0.33
38	0.38	0.25	0.38	0.25	0.38	0.25	0.38	0.25	0.38	0.25	0.38	0.25
39	0.30	0.14	0.17	0.18	0.20	0.17	0.20	0.25	0.20	0.21	0.20	0.24
40	0.15	0.11	0.07	0.11	0.07	0.11	0.07	0.11	0.07	0.11	0.07	0.11
41	0.14	0.00	0.13	0.04	0.14	0.00	0.14	0.01	0.13	0.03	0.13	0.04
42	0.25	0.00	0.12	0.50	0.12	0.50	0.13	0.50	0.12	0.50	0.12	0.50
43	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00
44	0.28	0.33	0.14	0.34	0.28	0.33	0.26	0.38	0.28	0.34	0.28	0.33
Average	**0.28**	**0.17**	**0.18**	**0.28**	**0.19**	**0.23**	**0.19**	**0.24**	**0.19**	**0.24**	**0.19**	**0.22**

Πίνακας 5.9: Συγκριτική απόδοση της 4ης προτεινόμενης μεθόδου για τα μοντέλα 1-44 για το Σενάριο IV με χρήση 1000 προσομοιώσεων

$(w < \frac{k}{2})$	$SU \geqslant$ median(**su**)		$SU >$ median(**su**)		$SU \geqslant$ mean(**su**)		$SU >$ mean(**su**)		$SU \geqslant$ geomean(**su**)		$SU >$ geomean(**su**)	
Μοντέλο	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
1	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.26
2	0.16	0.25	0.16	0.26	0.16	0.26	0.16	0.26	0.16	0.25	0.16	0.25
3	0.28	0.33	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66	0.00	0.66
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.32	0.38	0.10	0.72	0.25	0.46	0.24	0.46	0.22	0.49	0.24	0.48
6	0.12	0.16	0.00	0.66	0.12	0.16	0.12	0.16	0.12	0.16	0.12	0.18
7	0.27	0.00	0.18	0.00	0.18	0.00	0.18	0.00	0.27	0.00	0.18	0.00
8	0.20	0.02	0.19	0.02	0.19	0.03	0.18	0.08	0.18	0.10	0.19	0.02
9	0.22	0.27	0.15	0.37	0.16	0.34	0.16	0.33	0.16	0.34	0.16	0.34
10	0.28	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.27	0.01	0.14	0.00
11	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20	0.30	0.20
12	0.16	0.00	0.15	0.01	0.16	0.00	0.16	0.00	0.16	0.01	0.15	0.02
13	0.25	0.11	0.14	0.25	0.16	0.22	0.15	0.23	0.15	0.23	0.20	0.21
14	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.22	0.23	0.22
15	0.29	0.00	0.00	0.60	0.29	0.00	0.29	0.00	0.29	0.00	0.29	0.00
16	0.38	0.25	0.05	0.75	0.38	0.25	0.38	0.25	0.38	0.25	0.38	0.25
17	0.26	0.27	0.26	0.27	0.26	0.27	0.26	0.27	0.26	0.27	0.26	0.27
18	0.25	0.00	0.12	0.30	0.12	0.30	0.13	0.30	0.14	0.30	0.13	0.30
19	0.20	0.05	0.11	0.14	0.16	0.09	0.16	0.06	0.16	0.05	0.17	0.02
20	0.33	0.02	0.18	0.00	0.05	0.29	0.06	0.22	0.17	0.00	0.20	0.02
21	0.23	0.17	0.10	0.24	0.09	0.27	0.09	0.31	0.16	0.20	0.14	0.18
22	0.24	0.19	0.00	0.68	0.00	0.68	0.00	0.66	0.28	0.20	0.25	0.22
23	0.31	0.12	0.22	0.14	0.21	0.15	0.20	0.19	0.23	0.12	0.23	0.12
24	0.23	0.18	0.11	0.19	0.08	0.30	0.10	0.23	0.12	0.17	0.16	0.21
25	0.28	0.14	0.26	0.29	0.09	0.43	0.07	0.46	0.25	0.29	0.27	0.29
26	0.23	0.15	0.08	0.36	0.08	0.40	0.08	0.39	0.12	0.32	0.11	0.35
27	0.30	0.18	0.11	0.19	0.08	0.25	0.07	0.24	0.14	0.17	0.15	0.17
28	0.33	0.00	0.22	0.14	0.22	0.14	0.22	0.14	0.22	0.14	0.22	0.14
29	0.28	0.23	0.12	0.38	0.12	0.44	0.12	0.39	0.18	0.34	0.20	0.34
30	0.29	0.09	0.22	0.27	0.22	0.27	0.22	0.27	0.22	0.27	0.22	0.27
31	0.41	0.03	0.18	0.33	0.18	0.28	0.14	0.31	0.38	0.15	0.35	0.17
32	0.39	0.00	0.15	0.00	0.14	0.00	0.13	0.00	0.15	0.00	0.16	0.00
33	0.28	0.09	0.18	0.17	0.14	0.23	0.15	0.21	0.20	0.16	0.21	0.16
34	0.35	0.17	0.31	0.21	0.18	0.38	0.18	0.39	0.29	0.22	0.30	0.20
35	0.38	0.00	0.30	0.00	0.04	0.00	0.05	0.00	0.36	0.00	0.35	0.00
36	0.34	0.10	0.24	0.16	0.10	0.18	0.12	0.19	0.33	0.10	0.32	0.11
37	0.23	0.22	0.03	0.70	0.21	0.29	0.22	0.25	0.19	0.33	0.20	0.33
38	0.33	0.00	0.11	0.00	0.11	0.00	0.11	0.00	0.11	0.00	0.11	0.00
39	0.27	0.06	0.15	0.18	0.16	0.18	0.15	0.14	0.17	0.20	0.17	0.10
40	0.15	0.11	0.07	0.11	0.07	0.11	0.07	0.11	0.15	0.12	0.07	0.11
41	0.14	0.00	0.14	0.00	0.14	0.01	0.14	0.02	0.14	0.00	0.15	0.01
42	0.25	0.00	0.13	0.50	0.12	0.50	0.12	0.50	0.25	0.01	0.25	0.00
43	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00
44	0.29	0.34	0.28	0.33	0.27	0.38	0.28	0.34	0.28	0.33	0.28	0.33
Average	**0.26**	**0.12**	**0.14**	**0.25**	**0.15**	**0.22**	**0.15**	**0.22**	**0.20**	**0.17**	**0.19**	**0.17**

5.5 Συμπεράσματα

Η ανάλυση των δεδομένων για τους υπερκορεσμένους σχεδιασμούς παραμένει μια πρόκληση για τον πειραματιστή και ένας πολλά υποσχόμενος ερευνητικός τομέας για μελλοντική δουλειά. Οι μέθοδοι ανάλυσης που παρουσιάσαμε στα προηγούμενα Κεφάλαια 2, 3 και 4 αποτελούν μεθόδους ανάλυσης υπερκορεσμένων σχεδιασμών κάτω από την υπόθεση ενός λογιστικού μοντέλου παλινδρόμησης. Σε αυτό το κεφάλαιο επεκτείναμε τον αλγόριθμο της Θεωρίας Πληροφοριών που δόθηκε στην εργασία [7] για την ανάλυση των υπερκορεσμένων σχεδιασμών θεωρώντας όχι μόνο το λογιστικό μοντέλο παλινδρόμησης, αλλά γενικά ένα γενικευμένο γραμμικό μοντέλο με διακριτή απόκριση. Συγκεκριμένα επεκτείναμε τη χρήση του αλγορίθμου θεωρώντας γενικευμένα γραμμικά μοντέλα, των οποίων τα δεδομένα απόκρισης προέρχονται από την Poisson, γεωμετρική και αρνητική διωνυμική κατανομή.

Η συγκριτική απόδοση του προτεινόμενου αλγορίθμου, με βάση την εκτενή μελέτη προσομοιώσεων, υποδεικνύει ότι η προτεινόμενη μέθοδος έχει χαμηλά ποσοστά σφάλματος, αποδίδει καλά, και είναι εύρωστη από τη στιγμή που με τη χρήση διαφορετικών τιμών κατωφλιού προέκυψαν παρόμοια αποτελέσματα. Ας σημειωθεί εδώ ότι τα σφάλματα Τύπου ΙΙ δημιουργούν αρκετά προβλήματα, όπως διαπιστώθηκε από τον Lin στην εργασία [149], όπως άλλωστε και τα σφάλματα Τύπου Ι, δεδομένου ότι μπορεί να οδηγήσουν σε περιττό υπολογιστικό κόστος στα πειράματα που ακολουθούν, και κάτι τέτοιο μπορεί να είναι αρκετά επιζήμιο σε ένα πείραμα το οποίο έχει άμεσο αντίκτυπο στην πράξη. Όπως με οποιαδήποτε μέθοδο ανάλυσης υπερκορεσμένων σχεδιασμών, τα σφάλματα Τύπου Ι και Τύπου ΙΙ είναι εξίσου σημαντικά και θα πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα. Η προτεινόμενη μέθοδος φαίνεται να επιτυγχάνει κάτι τέτοιο και να συμπεριφέρεται ιδιαίτερα ικανοποιητικά, δεδομένου ότι το μέσο ποσοστό σφάλματος Τύπου Ι και Τύπου ΙΙ των 44 μοντέλων σε 1000 προσομοιώσεις, είναι της τάξης του 29% και 4% αντίστοιχα για το μοντέλο παλινδρόμησης Poisson, της τάξης του 29% και 5% αντίστοιχα για το μοντέλο παλινδρόμησης της γεωμετρικής, της τάξης του 30% και 7% αντίστοιχα για το μοντέλο παλινδρόμησης της αρνητικής διωνυμικής, και της τάξης του 23% και 6% αντίστοιχα για το λογιστικό μοντέλο παλινδρόμησης. Ας σημειωθεί εδώ ότι η προτεινόμενη μέθοδος υπερτερεί δύο γνωστών μεθόδων επιλογής μεταβλητών (CMIM και mRMR) κάτω από την υπόθεση ενος λογιστικού μοντέλου παλινδρόμησης.

Η προτεινόμενη μέθοδος επιτυγχάνει να ανιχνεύσει όλες τις πραγματικά ενεργές κύριες επιδράσεις σχεδόν με αμελητέο σφάλμα, δεδομένου ότι επιτυγχάνει σχεδόν μηδενικές τιμές για τα σφάλματα Τύπου ΙΙ, το οποίο είναι ιδιαίτερα σημαντικό, προκειμένου να μην παραλείπονται οι σημαντικοί παράγοντες του μοντέλου. Το γεγονός ότι το μέσο ποσοστό σφάλματος Τύπου Ι κυμαίνεται μεταξύ του 23% έως 30% δεν μας προβληματίζει ιδιαίτερα, δεδομένου ότι σε καταστάσεις που ισχύει η αρχή της σποραδικότητας, είναι πολύ πιθανό να προκύψουν σφάλματα Τύπου Ι. Η προτεινόμενη μέθοδος διατηρεί τα σφάλματα Τύπου Ι σε τέτοια επίπεδα έτσι ώστε να διατηρεί την ικανότητα να αποκλείει τους περιττούς παράγοντες, και ο ερευνητής βασισμένος μόνο στους παράγοντες που επιλέχθηκαν ως σημαντικοί είναι σε θέση πλέον να μειώσει το υπολογιστικό κόστος των επιπλέον πειραμάτων που θα εκτελέσει. Όμως δεδομένου ότι οι υπερκορεσμένοι σχεδιασμοί κυρίως χρησιμοποιούνται σε πειράματα κρησαρίσματος με στόχο να αναγνωρίσουν εκείνους τους παράγοντες που θα πρέπει να εξεταστούν περαιτέρω, οι σχεδόν μηδενικές τιμές σφάλματος Τύπου ΙΙ της προτεινόμενης μεθόδου είναι αξιοσημείωτες και πόσο μάλλον κάτω από την υπόθεση ενός γενικευμένου γραμμικού μοντέλου με διακριτά δεδομένα απόκρισης, μέθοδος η οποία είναι μη τετριμμένη και μη αναπτυγμένη στην έως σήμερα διεθνή βιβλιογραφία όπου συναντάμε το πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών μόνο κάτω υπό τη σκέπη του κλασικού γραμμικού μοντέλου.

Μέρος II

Επιλογή Μεταβλητών και Επιλογή Μοντέλου

ΚΕΦΑΛΑΙΟ **6**

Εισαγωγικά Στοιχεία και Βασικές Έννοιες

Avoid complexities.
Make everything as simple as possible.

—*Henry H. Maudslay (1771–1831)*

Στο έκτο αυτό κεφάλαιο αναφερόμαστε συνοπτικά στο πρόβλημα εξεύρεσης των πραγματικά σημαντικών μεταβλητών, με άλλα λόγια στο πρόβλημα εξεύρεσης του βέλτιστου μοντέλου ή υποδείγματος που είναι και ο βασικός στόχος μιας στατιστικής μοντελοποίησης σε ένα σύνολο δεδομένων. Επιπλέον, παρουσιάζουμε κριτήρια και ελέγχους που χρησιμοποιούνται για το σκοπό αυτό, αναφερόμαστε στην έννοια της πολυπλοκότητας ενός μοντέλου, και τέλος παρουσιάζουμε μια σύντομη ιστορική αναδρομή σχετικά με το τι έχει προταθεί βιβλιογραφικά μέχρι τώρα για την καλύτερη δυνατή ανάπτυξη ενός μοντέλου.

6.1 Ανάπτυξη Βέλτιστου Μοντέλου

Στις πρώτες φάσεις της διερεύνησης ενός προβλήματος έχουν συλλεχθεί στοιχεία για πολλές υποψήφιες επεξηγηματικές μεταβλητές και ενδέχεται ο πειραματιστής να μην γνωρίζει ποιες από αυτές τις μεταβλητές πράγματι συντελούν στην πρόβλεψη της εξαρτημένης μεταβλητής. Όταν μας δίνονται δεδομένα, προσπαθούμε να βρούμε ένα μοντέλο που να προσαρμόζεται όσο το δυνατόν καλύτερα στα δεδομένα. Επειδή το πραγματικό μοντέλο είναι άγνωστο και έχουμε ένα πλήθος από ανεξάρτητες μεταβλητές που μπορούν να συμπεριληφθούν στο μοντέλο, το πρόβλημά μας είναι να επιλέξουμε μόνο αυτές τις ανεξάρτητες μεταβλητές που μας δίνουν το καλύτερο μοντέλο. Σε αρκετές περιπτώσεις, η συμπεριφορά της εξαρτημένης μεταβλητής μπορεί να περιγραφεί σε εξίσου ικανοποιητικό βαθμό κάνοντας χρήση μόνο μερικών από τις αρχικές διαθέσιμες επεξηγηματικές μεταβλητές, και το φαινόμενο αυτό, δηλαδή να καταλήγουμε σε μοντέλα που περιέχουν μόνο μερικές από τις αρχικές μεταβλητές, είναι αρκετά σύνηθες στην πράξη.

Η εξεύρεση των πραγματικά σημαντικών μεταβλητών, με άλλα λόγια η εξεύρεση του βέλτιστου μοντέλου ή υποδείγματος είναι ο βασικός στόχος μιας στατιστικής μοντελοποίησης σε ένα σύνολο δεδομένων. Οι πρακτικοί λόγοι είναι προφανείς, αφού με αυτό τον τρόπο δεν θα ξοδέψουμε χρόνο ή χρήμα στην άσκοπη μέτρηση ασήμαντων μεταβλητών. Επιπρόσθετα, προκύπτει και ένα τεχνικό θέμα, το οποίο αφορά την ποιότητα των μελλοντικών προβλέψεων. Είναι γνωστό ότι η πρόσθεση μιας οποιασδήποτε μεταβλητής σε ένα μοντέλο βελτιώνει ποσοτικά, έστω και οριακά, την προσαρμογή του στο υπό εξέταση σύνολο δεδομένων. Ωστόσο, δεν συμβαίνει το ίδιο και για τις μελλοντικές προβλέψεις χρησιμοποιώντας αυτό το εκτιμηθέν μοντέλο. Η παρουσία μη σημαντικών μεταβλητών στο μοντέλο μπορεί να μειώσει την ακρίβεια της μελλοντικής πρόβλεψης. Γενικά, υπάρχει μια γενική προτίμηση για απλά μοντέλα με όσο το δυνατόν λιγότερους όρους. Για όλους αυτούς τους λόγους είναι σημαντικό να παρουσιάσουμε μεθόδους επιλογής των καταλληλότερων μεταβλητών από το αρχικό διαθέσιμο σύνολο, για την προσαρμογή ενός μοντέλου το οποίο θα εξακολουθεί να επεξηγεί τη συμπεριφορά της μεταβλητής απόκρισης σε ικανοποιητικό βαθμό. Η επιλογή των καλύτερων μεταβλητών γίνεται με τη βοήθεια κατάλληλων ελέγχων και κριτηρίων που μας εξασφαλίζουν την ορθότητα της επιλογής μας.

6.2 Έλεγχοι και Κριτήρια

Οι έλεγχοι και τα κριτήρια που αφορούν την επιλογή των καταλληλότερων μεταβλητών για την ανάπτυξη ενός βέλτιστου μοντέλου, μπορούν να ταξινομηθούν σε κατηγορίες που αφορούν στατιστικούς ελέγχους (όπως έλεγχος t, έλεγχος F), γραφικές παραστάσεις και μέτρα καταλληλότητας (όπως συντελεστής προσδιορισμού R^2, κριτήριο C_p-Mallows, κριτήρια πληροφορίας). Τα μέτρα καταλληλότητας αποτελούν σημαντικές αριθμητικές ποσότητες, οι οποίες χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου αλλά και τη σύγκριση διαφορετικών μοντέλων ως προς τη σπουδαιότητα τους. Πολλές φορές για την αξιολόγηση της μεταβλητής απόκρισης έχουμε στη διάθεσή μας πολλές επεξηγηματικές μεταβλητές χωρίς αυτές να είναι όλες στατιστικά σημαντικές. Εύκολα μπορούμε να καταλάβουμε ότι η μέτρηση παραπάνω μεταβλητών από αυτές που πραγματικά επεξηγούν το φαινόμενο σε μελλοντικές μελέτες, ή ο καθορισμός και ο έλεγχος μη σημαντικών ποσοτήτων για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής, μπορεί να είναι δύσκολος, χρονοβόρος αλλά και οικονομικά ασύμφορος. Γεννιέται λοιπόν το ερώτημα αν μπορούμε να εντοπίσουμε το μικρότερο υποσύνολο μεταβλητών από το αρχικό μας σύνολο, αναπτύσσοντας έτσι ένα μοντέλο, το οποίο όμως θα εξακολουθεί να επεξηγεί τη συμπεριφορά της ανεξάρτητης μεταβλητής μας σε ικανοποιητικό βαθμό. Η απάντηση στο παραπάνω ερώτημα είναι καταφατική και μάλιστα η επιλογή των καλύτερων μεταβλητών γίνεται με τη βοήθεια κατάλληλων ελέγχων και κριτηρίων τα οποία συναντάμε στις μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών.

Τα κριτήρια που συνήθως χρησιμοποιούνται στις μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών ταξινομούνται σε τέσσερις κατηγορίες:

1. Κριτήρια πρόβλεψης,

2. Κριτήρια Πληροφορίας,

3. Διαδικασίες επαναχρησιμοποίησης των δεδομένων και καθοδήγησης από τα δεδομένα,

4. Μπεϋζιανή επιλογή μεταβλητών.

Εμείς θα χρησιμοποιήσουμε στην παρούσα διατριβή κριτήρια πληροφορίας τα οποία συνδέονται με μέτρα πιθανοφάνειας ή απόκλισης. Κάθε ένα από αυτά τα κριτήρια μπορεί να θεωρηθεί ένας προσεγγιστικά αμερόληπτος εκτιμητής της αναμενόμενης συνολικής απόκλισης, μία μη αρνητική ποσότητα που μετρά την απόσταση μεταξύ του πραγματικού μοντέλου και του προσαρμοσμένου μοντέλου. Τα κριτήρια αυτά συνήθως αποτελούνται από δύο όρους, ο πρώτος όρος αποτελεί το μεροληπτικό εκτιμητή της αναμενόμενης συνολικής απόκλισης και ο δεύτερος όρος είναι ο κατάλληλος διορθωτικός όρος που καθιστά τον εκτιμητή αμερόληπτο ασυμπτωτικά.

6.3 Μέθοδοι Επιλογής των Καταλληλότερων Μεταβλητών

Η επιλογή μεταβλητών σε πολυδιάστατα προβλήματα είναι μία σημαντική διαδικασία που αποσκοπεί στην αναγνώριση των σημαντικών μεταβλητών που έχουν σημαντική επίδραση στην εξαρτημένη μεταβλητή. Ποικίλες μέθοδοι έχουν προταθεί κατά καιρούς βιβλιογραφικά για το σκοπό αυτό, στις οποίες αναφερόμαστε συνοπτικά παρακάτω, κάνοντας ταυτόχρονα και μία σύντομη ιστορική ανασκόπηση.

6.3.1 Μέθοδοι Επιλογής του Καλύτερου Υποσυνόλου Μεταβλητών

Η εξεύρεση των πραγματικά σημαντικών μεταβλητών, με άλλα λόγια η εξεύρεση του βέλτιστου μοντέλου ή υποδείγματος είναι θεμελιώδης σε προβλήματα στατιστικής μοντελοποίησης μεγάλης διάστασης, σε διάφορους επιστημονικούς τομείς. Εκτεταμένη έρευνα στο χώρο αυτό έχει γίνει τις τελευταίες δεκαετίες, ενώ αρκετές μελέτες σχετίζονται με την Ιατρική και τη Βιολογία, όπως είναι οι εργασίες [76], [77], [78], [79], [86], [209] και [214]. Η μέθοδος επιλογής με τη χρήση καλύτερων υποσυνόλων χρησιμοποιείται με ορισμένα κριτήρια τα οποία συνδυάζουν στατιστικά μέτρα με συναρτήσεις ποινής για την αύξηση του αριθμού των μεταβλητών πρόβλεψης στο μοντέλο. Τα πιο δημοφιλή από τα εν λόγω κριτήρια είναι τα κριτήρια πληροφορίας, και συγκεκριμένα τα κριτήρια AIC (Akaike Information Criteria) και BIC (Bayesian Information Criteria). Το κριτήριο AIC προτάθηκε από τον Akaike [3] και επιλέγει το μοντέλο που ελαχιστοποιεί την ποσότητα $AIC = -2l + 2q$ και το BIC προτάθηκε από τον Schwarz [205] και έχει παρόμοια μορφή με το AIC εκτός του ότι η λογαριθμοπιθανοφάνεια ποινικοποιείται από τον όρο $qlog(n)$ αντί του $2q$, επιλέγοντας το μοντέλο που ελαχιστοποιεί την ποσότητα $BIC = -2l + qlog(n)$, όπου l είναι η λογαριθμοπιθανοφάνεια του μοντέλου, q είναι ο αριθμός των παραμέτρων του μοντέλου και n είναι ο αριθμός των παρατηρήσεων. Σημαντικές εργασίες πάνω στις μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών είναι οι [110] και [188]. Επίσης, μία σύντομη ιστορική ανασκόπηση, αναλυτική περιφραφή των μεθόδων επιλογής του καλύτερου υποσυνόλου μεταβλητών στην παλινδρόμηση, καθώς επίσης και η περιγραφή διαφόρων κριτηρίων πληροφορίας που έχουν προταθεί κατά καιρούς βιβλιογραφικά παρουσιάζονται στη σημαντική επιστημονική εργασία [162].

Οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών αποτελούν κλασικές τεχνικές επιλογής των καταλληλότερων μεταβλητών, εντούτοις χαρακτηρίζονται από αρκετά μειονεκτήματα. Οποιαδήποτε μέθοδος επιλογής του καλύτερου υποσυνόλου μεταβλητών αποτελεί

μία εξαντλητική διαδικασία, αφού αναζητάει κάθε δυνατό υποσύνολο μεταβλητών ώστε να επιλέξει το καλύτερο δυνατό. Οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών είναι εξαιρετικά χρονοβόρες όταν ο αριθμός των μεταβλητών είναι μεγάλος και αυτή η υπολογιστική δυσκολία αποτρέπει την ευρεία χρήση τους για το σκοπό αυτό. Γενικά οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών παρουσιάζουν αδυναμίες όπως είναι η έλλειψη σταθερότητας ή το γεγονός ότι αγνοούν στοχαστικά σφάλματα που δημιουργούνται κατά τη διαδικασία, με αποτέλεσμα τη δύσκολη κατανόηση των θεωρητικών τους ιδιοτήτων [27].

6.3.2 Μέθοδοι Επιλογής Μοντέλου με Βήματα

Όταν το πλήθος των ανεξάρτητων μεταβλητών είναι p, πρέπει να εξετάσουμε 2^p μοντέλα και να επιλέξουμε το καλύτερο σύμφωνα με το κριτήριο που χρησιμοποιούμε. Για $p = 4, 5, 10$, όλα τα δυνατά μοντέλα είναι 16, 32, 1024, αντίστοιχα, και ο απαιτούμενος χρόνος ακόμη και με ηλεκτρονικό υπολογιστή είναι απαγορευτικός. Για την ενίσχυση της προβλεψιμότητας και την απόκτηση του καλύτερου μοντέλου, όπως αυτό προκύπτει από τη διαδικασία μοντελοποίησης, άλλες παραδοσιακές τεχνικές επιλογής μεταβλητών που έχουν αναπτυχθεί για την επιλογή του καλύτερου υποσυνόλου μεταβλητών από ένα αρχικό σύνολο, βασίζονται στην αφαίρεση και την πρόσθεση μεταβλητών σε ένα μοντέλο. Τέτοιες μέθοδοι επιλογής μοντέλου με βήματα, είναι η διαδικασία της προς τα πίσω απαλοιφής ή διαδοχικής αφαίρεσης (backward elimination), η προς τα εμπρός επιλογή ή διαδοχική πρόσθεση (forward selection) και η κατά βήματα εμπρός-πίσω επιλογή (stepwise selection). Αυτές οι μέθοδοι επιλογής καλού μοντέλου δεν εξετάζουν όλες τις δυνατές περιπτώσεις, αλλά επιλέγουν μόνο αυτές τις μεταβλητές που θεωρούνται σημαντικές. Η επιλογή του καλύτερου μοντέλου με αυτές τις μεθόδους δεν δίνει πάντα το βέλτιστο αποτέλεσμα που δίνει η εξέταση όλων των δυνατών μοντέλων, αλλά δίνει αξιόπιστα μοντέλα εξετάζοντας ορισμένα μόνο καλά μοντέλα και έτσι οι υπολογισμοί είναι λιγότεροι και ο απαιτούμενος χρόνος είναι ικανοποιητικός ακόμη και για μεγάλες τιμές του p. Σε αυτές τις μεθόδους μία εναλλακτική διαδικασία είναι όταν βρούμε το καλύτερο μοντέλο, μπορούμε να το συγκρίνουμε με όλα τα μοντέλα που έχουν το ίδιο πλήθος ερμηνευτικών μεταβλητών.

Περιληπτικά, η μέθοδος της προς τα πίσω απαλοιφής ξεκινάει συμπεριλαμβάνοντας στο μοντέλο όλες τις διαθέσιμες μεταβλητές και αφαιρεί μία μία τις μεταβλητές, ξεκινώντας από τη λιγότερο σημαντική, αν ο έλεγχος που πραγματοποιείται με τη βοήθεια του F-ελέγχου είναι στατιστικά σημαντικός. Η διαδικασία επαναλαμβάνεται μέχρι η αφαίρεση μιας οποιασδήποτε μεταβλητής να είναι στατιστικά σημαντική, οπότε και σταματάει. Η μέθοδος της προς τα εμπρός επιλογής ξεκινάει από το μοντέλο που περιέχει μόνο το σταθερό όρο και προσθέτει τη μεταβλητή εκείνη, που μας δίνει τη μεγαλύτερη στατιστικά σημαντική τιμή της F του F-ελέγχου. Η διαδικασία επαναλαμβάνεται μέχρι η τιμή του F-ελέγχου για την πρόσθεση οποιασδήποτε από τις εναπομείναντες μεταβλητές να μην είναι στατιστικά σημαντική, οπότε και σταματάει. Η κατά βήματα εμπρός-πίσω επιλογή αποτελεί διόρθωση-βελτίωση της μεθόδου της προς τα εμπρός επιλογής και μία από τις πιο ευρέως χρησιμοποιούμενες διαδικασίες επιλογής μοντέλου με βήματα. Η διόρθωση αυτή αφορά την παρεμβολή ενός επιπλέον ελέγχου σε κάθε επανάληψη της διαδικασίας πρόσθεσης μιας μεταβλητής. Ο έλεγχος αυτός εξετάζει την περίπτωση που η προσθήκη μιας νέας μεταβλητής στο μοντέλο οδηγεί στην εξασθένηση της σημαντικότητας κάποιας άλλης, που είχε εισαχθεί νωρίτερα, με συνέπεια να πρέπει να εξεταστεί η αξία παραμονής της στο μοντέλο. Με άλλα λόγια σε κάθε βήμα, όταν μία μεταβλητή είναι σημαντική και εισάγεται στο μοντέλο, εξετάζουμε μήπως υπάρχουν στο μοντέλο μεταβλητές που είχαν εισαχθεί σε προηγούμενα βήματα και δεν είναι πλέον σημαντικές. Αυτό δεν είναι περίεργο διότι η συμβολή κάθε μεταβλητής εξαρτάται από τη σειρά που εισάγεται στο μοντέλο. Το κριτήριο εξόδου είναι το ίδιο όπως στη μέθοδο της προς τα πίσω απαλοιφής. Για να μην αντιμετωπίσουμε την περίπτωση κυκλικής εναλλαγής, οπότε η διαδικασία δεν τελειώνει, λαμβάνεται τιμή εισόδου μεγαλύτερη ή ίση της τιμής εξόδου. Σε αντίθετη περίπτωση μπορεί να απορρίπτουμε μεταβλητή που μόλις κάναμε δεκτή.

Για περισσότερες λεπτομέρειες σχετικά με τις μεθόδους επιλογής μοντέλου με βήματα ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στο βιβλίο [71].

6.3.3 Ποινικοποιημένες Μέθοδοι Επιλογής Μεταβλητών

Πρόσφατα, προτάθηκαν καινοτόμες μέθοδοι επιλογής μεταβλητών οι οποίες χρησιμοποιούν έννοιες όπως η ποινικοποιημένη πιθανοφάνεια και τα ποινικοποιημένα ελάχιστα τετράγωνα. Οι Fan και Li [76] πρότειναν μία κλάση διαδικασιών επιλογής μεταβλητών μέσω της μη κοίλης ποινικοποιημένης πιθανοφάνειας (nonconcave penalized likelihood), σε μία προσπάθειά τους να επιλέξουν μεταβλητές και ταυτόχρονα να εκτιμήσουν τους συντελεστές παλινδρόμησης. Για αυτόν το λόγο πρότειναν μία ενιαία προσέγγιση με τη χρήση ποινικοποιημένων ελαχίστων τετραγώνων, διατηρώντας τις καλές ιδιότητες της μεθόδου επιλογής του καλύτερου υποσυνόλου μεταβλητών και της παλινδρόμησης διάσελου (ridge regression). Με την παλινδρόμηση διάσελου λαμβάνουμε μεροληπτικές εκτιμήτριες για τις παραμέτρους του μοντέλου, με την ελπίδα ότι το μέσο τετραγωνικό τους σφάλμα είναι μικρότερο από αυτό που δίνει η μέθοδος ελαχίστων τετραγώνων. Οι Li και Lin [144] εισήγαγαν μια επέκταση αυτής της μεθόδου. Επεκτείνουν τις προσεγγίσεις της μη κοίλης ποινικοποιημένης πιθανοφάνειας στα ελάχιστα τετράγωνα, δηλαδή τα μη κοίλα ποινικοποιημένα ελάχιστα τετράγωνα, και εστιάζουν στην περίπτωση, όπου ο πίνακας σχεδιασμού δεν είναι πλήρους βαθμού. Επί της ουσίας, χρησιμοποίησαν μία επαναληπτική παλινδρόμηση κορυφογραμμής για να βρούν τη λύση των ποινικοποιημένων ελαχίστων τετραγώνων, επιλέγοντας μία αρχική τιμή των αγνώστων συντελεστών.

Η διαδικασία της ποινικοποίησης συνίσταται στην εισαγωγή συναρτήσεων ποινής και η μεθοδολογία αυτή επεκτείνεται σε μοντέλα βασισμένα στην πιθανοφάνεια, όπως είναι τα γενικευμένα γραμμικά μοντέλα, γεγονός που αποτελεί και τη σημαντικότερη διαφορά σε σχέση με τις παραδοσιακές μεθόδους οι οποίες είναι εφαρμόσιμες αποκλειστικά στο γραμμικό μοντέλο. Στην περίπτωση των γενικευμένων γραμμικών μοντέλων τα στατιστικά συμπεράσματα εξάγονται σύμφωνα με τις βασικές συναρτήσεις πιθανοφάνειας και για την επιλογή των σημαντικών μεταβλητών χρησιμοποιείται ο εκτιμητής ποινικοποιημένης μεγίστης πιθανοφάνειας.

Στις παραπάνω εργασίες μία μορφή των ποινικοποιημένων ελαχίστων τετραγώνων ορίζεται ως

$$Q(\beta) \equiv \sum_{i=1}^{n} l_i(g(x_i^T\beta), y_i) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{6.1}$$

όπου (x_i, y_i) τα δεδομένα τα οποία συλλέγονται ανεξάρτητα, και δεδομένων των x_i, τα y_i έχουν συνάρτηση πυκνότητας πιθανότητας $f_i(g(x_i^T\beta), y_i)$, όπου g είναι μια γνωστή συνάρτηση σύνδεσης (link function), το $l_i = \log f_i$ συμβολίζει τη δεσμευμένη λογαριθμοπιθανοφάνεια των y_i, το $p_{\lambda_n}(.)$ είναι η συνάρτηση ποινής και το λ είναι μια άγνωστη παράμετρος κατωφλιού η οποία μπορεί να επιλεχθεί με προσεγγίσεις καθοδηγούμενες από τα δεδομένα (data-driven approaches), όπως είναι η διασταυρωμένη επικύρωση (Cross Validation-CV) και η γενικευμένη διασταυρωμένη επικύρωση (Generalized Cross Validation-GCV) [55].

Η μεγιστοποίηση της συνάρτησης ποινικοποιημένης πιθανοφάνειας ισοδυναμεί με την ελαχιστοποίηση της ποσότητας

$$-\sum_{i=1}^{n} l_i(g(x_i^T\beta), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{6.2}$$

ως προς β, ενώ ένας εκτιμητής ποινικοποιημένης μεγίστης πιθανοφάνειας του β προκύπτει ελαχιστοποιώντας τη σχέση (6.2) ως προς β, για κάποια παράμετρο κατωφλιού λ.

Εκτιμητές ποινικοποιημένης μεγίστης πιθανοφάνειας του β προκύπτουν με τη χρήση διάφορων συναρτήσεων ποινής, όπως η L_1 ποινή $p_\lambda(|\beta|) = \lambda|\beta|$ η οποία οδηγεί στη μέθοδο του ελάχιστου απόλυτου τελεστή συρρίκνωσης και επιλογής (Least Absolute Shrinkage

and Selection Operator-LASSO) [216], η ποινή "σκληρού" κατωφλιού (Hard) [5], $p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, όπου $I(.)$ είναι μια δείκτρια συνάρτηση, και η ποινή ομαλά περικομμένης απόλυτης απόκλισης (Smoothly Clipped Absolute Deviation-SCAD) [75] της οποίας η πρώτη παράγωγος ορίζεται ως $p'_\lambda(|\beta|) = \lambda\Big\{I(|\beta| \leq \lambda) + \frac{(\alpha\lambda-\beta)_+}{(\alpha-1)\lambda} I(|\beta| > \lambda)\Big\}$, για κάποιο $\beta > 0$ και $\alpha > 2$, με $p_\lambda(0) = 0$. Για την καλύτερη επιλογή του α, σύμφωνα με πρόσφατη βιβλιογραφία [76], η τιμή $\alpha \approx 3.7$ φαίνεται να αποδίδει αρκετά ικανοποιητικά σε πολλά προβλήματα επιλογής μεταβλητών.

Τα πλεονεκτήματα της νέας αυτής προσέγγισης είναι πολλαπλά και αναφέρονται παρακάτω.

1. Η μέθοδος των ποινικοποιημένων ελαχίστων τετραγώνων διαφοροποιείται από τις παραδοσιακές τεχνικές επιλογής μεταβλητών, εξαλείφοντας τις ασήμαντες μεταβλητές μέσω του υπολογισμού των συντελεστών παλινδρόμησης οι οποίοι συρρικνώνονται ακριβώς στο μηδέν.

2. Με τον τρόπο αυτό επιτυγχάνεται η επιλογή των σημαντικών μεταβλητών και η ταυτόχρονη εκτίμηση των συντελεστών παλινδρόμησης.

3. Έχουν ισχυρό θεωρητικό υπόβαθρο και δίνουν πάντα καλύτερα και ορθότερα αποτελέσματα.

4. Είναι εφαρμόσιμες στο γραμμικό μοντέλο, αλλά μπορούν να επεκταθούν και στο γενικευμένο γραμμικό μοντέλο.

5. Δημοφιλή κριτήρια επιλογής μεταβλητών μπορούν να αποκομιστούν από τα ποινικοποιημένα ελάχιστα τετράγωνα.

Τα ποινικοποιημένα ελάχιστα τετράγωνα κάνοντας χρήση της ποινής της εντροπίας είναι δυνατόν να ξαναγραφτούν ως

$$\sum_{i=1}^{n} l_i(g(x_i^T\beta), y_i) - \frac{1}{2}\lambda_n^2|M|, \tag{6.3}$$

όπου $|M|$ το μέγεθος ή η διάσταση του υποκείμενου υποψήφιου μοντέλου που ισούται με τον αριθμό των μη μηδενικών συντελεστών παλινδρόμησης του μοντέλου. Πολλά κριτήρια επιλογής μεταβλητών είναι συνδεδεμένα με αυτά τα ποινικοποιημένα ελάχιστα τετράγωνα της μορφής (6.3), θεωρώντας διαφορετικές τιμές του λ_n. Για παράδειγμα, το Akaike κριτήριο πληροφορίας και το Μπεϋζιανό κριτήριο πληροφορίας, αντιστοιχούν σε $\lambda_n = \sqrt{2}(\sigma/\sqrt{n})$ και $\lambda_n = \sqrt{\log n}(\sigma/\sqrt{n})$, αντίστοιχα, παρόλο που αυτά τα δύο κριτήρια παρακινούνται από διαφορετικές αρχές.

6.4 Πολυπλοκότητα Μοντέλου

Το κύριο πρόβλημα σε οποιαδήποτε διαδικασία κατασκευής μοντέλου είναι να επιλέξουμε από ένα μεγάλο σύνολο μεταβλητών εκείνες που πραγματικά πρέπει να συμπεριληφθούν στο "καλύτερο" μοντέλο. Η εύρεση του "καλύτερου" μοντέλου πρόβλεψης είναι η διαδικασία με την οποία ένα μοντέλο έχει δημιουργηθεί ή επιλεγεί στην προσπάθειά του να προβλέπει καλύτερα την πιθανότητα ενός εξαγόμενου αποτελέσματος [207]. Σε πολλές περιπτώσεις, το καλύτερο μοντέλο επιλέγεται με βάση τη θεωρία ανίχνευσης (detection theory) για να προσπαθήσει να μαντέψει την πιθανότητα ενός εξαγόμενου αποτελέσματος δοθέντος ένος συνόλου δεδομένων εκπαίδευσης ή δεδομένων εισόδου.

Η επιλογή του καλύτερου μοντέλου αποτελεί πρόκληση για τον πειραματιστή και ένα δύσκολο έργο στις μέρες μας που το μέγεθος και η πολυπλοκότητα των μοντέλων αυξάνεται συνεχώς. Υπάρχει μια γενική προτίμηση για απλά μοντέλα με όσο το δυνατόν λιγότερους

όρους. Αυτό οφείλεται σε μια γενική αίσθηση ότι κανένα στατιστικό μοντέλο δεν είναι σωστό και το μόνο που μπορούμε να καταφέρουμε είναι να περιγράψουμε τα κύρια χαρακτηριστικά ενός φαινομένου. Η Φύση μπορεί να είναι σύνθετη και πολύπλοκη, αλλά γενικώς λίγοι είναι οι παράγοντες που κυριαρχούν. Εάν δύο μοντέλα προσαρμόζονται στα δεδομένα εξίσου καλά, προτιμάται γενικά το απλούστερο μοντέλο (Occam's razor αρχή) [10]. Ένας κύριος λόγος για αυτό είναι ότι συνήθως το απλούστερο μοντέλο επιτρέπει την καλύτερη γενίκευση από ότι θα μπορούσε το πιο σύνθετο μοντέλο.

Αρκετά μέτρα πολυπλοκότητας έχουν προταθεί για την αντικειμενική ποσοτικοποίηση της πολυπλοκότητας ενός μοντέλου, και παρ' όλα αυτά δεν καλύπτουν όλες τις πτυχές της. Ενδεικτικά αναφέρουμε κάποια πολύ σημαντικά μέτρα πολυπλοκότητας που συναντάμε στη διεθνή βιβλιογραφία κάνοντας ταυτόχρονα μία σύντομη ιστορική ανασκόπηση της έννοιας της πολυπλοκότητας ενός μοντέλου. Η παραμετρική πολυπλοκότητα (parametric complexity) [168] μετρά τον αριθμό των παραμέτρων στο μοντέλο (όσες περισσότερες παραμέτρους ένα μοντέλο περιλαμβάνει, τόσο πιο πολύπλοκο είναι). Η γεωμετρική πολυπλοκότητα (geometric complexity) [170] μετρά τον αριθμό των διαφορετικών κατανομών πιθανότητας που λαμβάνει υπόψη ένα μοντέλο. Η Kolmogorov-Chaitin πολυπλοκότητα [142] χαρακτηρίζει ακολουθίες συμβόλων θεωρώντας την αντιπροσώπευσή τους σε σχέση με τις Turing μηχανές. Μια ποσότητα που συνδέεται στενότερα με την έννοια της στατιστικής πολυπλοκότητας είναι το λογικό βάθος (logical depth) του Bennett [13]. Ενώ η Kolmogorov-Chaitin πολυπλοκότητα μιας συμβολοσειράς ορίζεται ως το μήκος του μικρότερου προγράμματος μιας Turing μηχανής ικανό να αναπαράγει ακριβώς αυτήν τη συμβολοσειρά, το λογικό βάθος του Bennett ορίζεται ως ο χρόνος εκτέλεσης που απαιτείται για την υλοποίηση του αλγορίθμου. Εντούτοις, η Kolmogorov-Chaitin πολυπλοκότητα στην αλγοριθμική Θεωρία Πληροφοριών, σημαίνει κάτι αρκετά διαφορετικό από την υπολογιστική πολυπλοκότητα στην ανάλυση των αλγορίθμων. Αυτά, με τη σειρά τους, είναι και τα δύο διαφορετικά από τη στοχαστική πολυπλοκότητα (stochastic complexity) που χρησιμοποιείται στην εκτίμηση της τάξης ενός μοντέλου (model-order) στη Στατιστική [80], [29].

Ο Rissanen (βλ. [190], [191], [192], [193], [194]) μελέτησε και πρότεινε στατιστικές μεθόδους κατασκευής μοντέλου, οι οποίες βασίζονται στη στοχαστική πολυπλοκότητα και το μήκος περιγραφής. Το κάθε ένα από αυτά τα κριτήρια βασίζονται στην ιδέα της κωδικοποίησης των δεδομένων με τον ελάχιστο αριθμό δυαδικών ψηφίων [102]. Η αρχή του ελάχιστου μήκους περιγραφής (minimum description length) αναφέρει ότι οποιαδήποτε κανονικότητα σε ένα δοθέν σύνολο παρατηρούμενων δεδομένων μπορεί να χρησιμοποιηθεί για τη συμπίεση των δεδομένων, αλλά η βέλτιστη κωδικοποίηση των παρατηρούμενων δεδομένων είναι η συντομότερη δυνατή. Ως βασικό στοιχείο της αρχής του ελάχιστου μήκους περιγραφής (βλ. [106], [96] και [97]) η στοχαστική πολυπλοκότητα μετρά και ελέγχει την καλή προσαρμογή που έχει η κλάση ενός μοντέλου πιθανοτήτων ως προς το να εξηγήσει τη στατιστική κανονικότητα που υπάρχει στα δεδομένα. Ορίζεται ως το μήκος του συντομότερου προθέματος (prefix) κώδικα που μπορεί να επιτευχθεί για τα δεδομένα από ένα βέλτιστο σχήμα κωδικοποίησης το οποίο καθορίζεται από την κλάση του μοντέλου. Οι Bialek et al. [16] όρισαν την πολυπλοκότητα ως ένα μέτρο δυσκολίας εκμάθησης ενός υποδείγματος.

Πρόσφατα, η πολυπλοκότητα του μοντέλου χρησιμοποιείται σε πολλούς επιστημονικούς τομείς και εξυπηρετεί διαφορετικούς σκοπούς, όπως για παράδειγμα στις εργασίες [70] και [140]. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στις πολύ σημαντικές εργασίες [195], [196] και [197] για τα σύγχρονα πρότυπα της Θεωρίας Πληροφοριών με βάση την αλγοριθμική θεωρία της πολυπλοκότητας. Επίσης, τα μέτρα πολυπλοκότητας και προσαρμογής μοντέλου μελετήθηκαν και από Μπεϋζιανή σκοπιά [211], [218]. Ένα στατιστικό μέτρο της πολυπλοκότητας γνωστό στον τομέα της φυσικής ως LMC πολυπλοκότητα [156], [40] προέρχεται θεωρώντας τον τέλειο κρύσταλλο και το απομονωμένο ιδανικό αέριο ως παραδείγματα απλών μοντέλων, και ως εκ τούτου, ως συστήματα με μηδενική πολυπλοκότητα. Μια μαθηματική έκφραση για τον υπολογισμό της ποσότητας της LMC πολυπλοκότητας (C) είναι

η $C = H \cdot D$, όπου το H αναπαριστά το περιεχόμενο των πληροφοριών του συστήματος και το D περιγράφει πόσο απότομη είναι η χωρική κατανομή του. Η απλή εκθετική εντροπία Shannon θεωρείται ως ένας ποσοδείκτης του H. Με άλλα λόγια, η LMC πολυπλοκότητα είναι η αλληλεπίδραση μεταξύ της πληροφορίας H που αποθηκεύεται σε ένα σύστημα και της απόστασής της από την ισοκατανομή, που ονομάζεται ανισορροπία D (disequilibrium).

ΚΕΦΑΛΑΙΟ 7

Στατιστική Πολυπλοκότητα και Περιθώρια Κατανομή σε Πειράματα Bernoulli

Can the existence of a mathematical entity
be proved without defining it?

—*Jacques S. Hadamard (1865–1963)*

Στο έβδομο αυτό κεφάλαιο παρουσιάζουμε ένα νέο μέτρο στατιστικής πολυπλοκότητας (Entropy Prior Complexity-EPC), μελετάμε τις ιδιότητες του νέου αυτού μέτρου, και γενικά τις διάφορες πτυχές της πολυπλοκότητας ενός μοντέλου. Συγκεκριμένα, μελετάμε την παράμετρο prior και τη σύνδεσή της με την έννοια της πολυπλοκότητας ενός μοντέλου, καθώς επίσης και με την έννοια της εντροπίας. Επιπλέον, ενσωματώνουμε επιτυχώς το μέτρο πολυπλοκότητας EPC στο πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών, με στόχο την αναζήτηση και εύρεση του βέλτιστου μοντέλου μέσα από πολλά υποψήφια λογιστικά μοντέλα παλινδρόμησης. Επιπλέον, μελετάμε την έννοια και τον υπολογισμό της εκ των προτέρων περιθώριας κατανομής (Marginal Prior Distribution-MPD) ενός πειράματος Bernoulli. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [6] καθώς επίσης και στην επιστημονική εργασία [9].

7.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Αρκετά μέτρα πολυπλοκότητας έχουν προταθεί για την αντικειμενική ποσοτικοποίηση της πολυπλοκότητας ενός μοντέλου, και παρ' όλα αυτά δεν καλύπτουν όλες τις πτυχές της. Στο Κεφάλαιο 6 αναφερθήκαμε σε κάποια πολύ σημαντικά μέτρα πολυπλοκότητας που συναντάμε στη διεθνή βιβλιογραφία κάνοντας ταυτόχρονα μία σύντομη ιστορική ανασκόπηση της έννοιας της πολυπλοκότητας ενός μοντέλου. Ένα εξαιρετικά πολύπλοκο μοντέλο μπορεί να φαίνεται ότι προσαρμόζεται καλά, αλλά χωρίς να δίνεται προσοχή στην έννοια της πολυπλοκότητας του μοντέλου, θα μπορούσε να μας οδηγήσει σε λανθασμένες εκτιμήσεις των παραμέτρων και λανθασμένες προβλέψεις. Προκειμένου να βελτιωθεί η διαδικασία επιλογής μοντέλου, πρέπει αρχικά να ορίσουμε την έννοια της πολυπλοκότητας μοντέλου. Η πολυπλοκότητα αποτελεί γενικά ένα μέτρο της δυσκολίας που εμπλέκεται στην περιγραφή ενός μοντέλου ή μιας διαδικασίας, και δεν φαίνεται να υπάρχει ακριβής ορισμός της πολυπλοκότητας ενός μοντέλου βιβλιογραφικά.

Το πιο κοινό χαρακτηριστικό που συνήθως αγνοείται από πολλά μέτρα πολυπλοκότητας είναι η παράμετρος prior, η οποία αποτελεί αναπόσπαστο μέρος του μοντέλου όπως θα δείξουμε παρακάτω και θα μπορούσε να επηρεάσει την πολυπλοκότητα ενός μοντέλου σημαντικά. Οι παραπάνω λόγοι μας παρακίνησαν να μελετήσουμε την παράμετρο prior και τη σύνδεσή της με την έννοια της πολυπλοκότητας ενός μοντέλου καθώς επίσης και με την έννοια της εντροπίας. Επιπρόσθετα, στην περίπτωση που ο πειραματιστής θέλει να γνωρίζει εκ των προτέρων την περιθώρια πιθανότητα των δεδομένων δοθέντος του μοντέλου, τότε σε ένα τέτοιο μοντέλο βασισμένο στην πιθανοφάνεια, η πρόβλεψη παίρνει τη μορφή κατανομής (εκ των προτέρων περιθώρια κατανομή) η οποία στην πράξη είναι δύσκολο να υπολογιστεί αναλυτικά. Για το λόγο αυτό, προτείνουμε έναν εναλλακτικό τρόπο υπολογισμού της, μέσω παραμέτρων που αποτελούν τις εμπειρικές εκτιμήτριες των λόγων πιθανοφάνειας.

7.2 Η Παράμετρος Prior σε Ένα Πείραμα Bernoulli

Σε αυτή την ενότητα μελετάμε τον τρόπο με τον οποίο συνδέεται η παράμετρος prior μέσω της εντροπίας με την έννοια της πολυπλοκότητας ενός μοντέλου στη διακριτή καθώς επίσης και στη συνεχή περίπτωση. Σε ένα μοντέλο βασισμένο στην πιθανοφάνεια, η πρόβλεψη παίρνει τη μορφή κατανομής (εκ των προτέρων περιθώρια κατανομή) η οποία στην πράξη είναι δύσκολο να υπολογιστεί αναλυτικά. Για το λόγο αυτό, προτείνουμε έναν εναλλακτικό τρόπο υπολογισμού της σε ένα πείραμα Bernoulli.

7.2.1 Σύνδεση της Παραμέτρου Prior με την Πολυπλοκότητα ενός Μοντέλου

Η χρήση μεγάλων και πολύπλοκων μοντέλων γίνεται ολοένα και περισσότερο συχνή στις μέρες μας με την πληροφορία και τη διαθεσιμότητα των μεγάλων βάσεων δεδομένων να αυξάνει συνεχώς. Προτού σκεφτούμε τρόπους για να αντιμετωπίσουμε το πρόβλημα πολυπλοκότητας ενός μοντέλου, ένα σημαντικό ζήτημα είναι να δούμε αν υπάρχει κάποιος τρόπος να μετρηθεί η πολυπλοκότητα του μοντέλου ενσωματώνοντας την παράμετρο prior, η οποία θεωρείται αναπόσπαστο μέρος κάθε μοντέλου. Η πολυπλοκότητα όπως είδαμε έχει λάβει πολλές εννοιολογικές σημασίες, διαφορετικές από την έννοια της στατιστικής πολυπλοκότητας και εξυπηρετώντας διαφορετικούς σκοπούς. Παρακάτω περιγράφουμε τη στατιστική πολυπλοκότητα, ένα μέτρο πολυπλοκότητας γνωστό στη διεθνή βιβλιογραφία, το οποίο προέρχεται από την επιστήμη των υπολογιστών.

Θεωρούμε την ακόλουθη διαδοχική σειρά εξαγόμενων αποτελεσμάτων τα οποία προκύπτουν από τη ρίψη ενός "δίκαιου" νομίσματος

$$\vec{S} \equiv \ldots HHTHTTTHHHHHTHTHTTHHT \ldots.$$

Άπο αριστερά προς τα δεξιά, όλα τα σύμβολα κατανέμονται ανεξάρτητα και η πιθανότητα εμφάνισης για οποιοδήποτε από τα δύο σύμβολα (Η ή Τ) είναι $\frac{1}{2}$. Εδώ, η εντροπία της επόμενης ρίψης του νομίσματος ισούται με 1, αφού η πυκνότητα της εντροπίας της ρίψης του νομίσματος είναι 1. Στην επιστήμη των υπολογιστών το παραπάνω πείραμα καλείται ως ϵ-μηχανή για τη ρίψη ενός "δίκαιου" νομίσματος [28], [114]. Η ϵ-μηχανή αποτελεί ένα μοντέλο το οποίο περιγράφει την αρχική διαδικασία $\vec{S}$, υπό την έννοια ότι ένας τυχαίος περίπατος μέσω της ϵ-μηχανής θα αναπαράγει στατιστικά την αρχική ακολουθία συμβόλων τα οποία προκύπτουν από τη ρίψη ενός "δίκαιου" νομίσματος. Η ϵ-μηχανή λειτουργεί παρόμοια και στην περίπτωση ενός μεροληπτικού νομίσματος. Η στατιστική πολυπλοκότητα αποτελεί το μέτρο του μεγέθους της ϵ-μηχανής [57].

Ορισμός 7.1. *Η στατιστική πολυπλοκότητα ορίζεται ως το μέγεθος πληροφορίας της κατανομής μιας σειράς εξαγόμενων αποτελεσμάτων που παράγονται από ένα παρατηρούμενο σύστημα, όπως είναι και η ρίψη ενός νομίσματος.*

Από τώρα και στο εξής στις αποδείξεις που ακολουθούν, χρησιμοποιούμε τη λογαριθμική συνάρτηση, δεδομένου ότι το μέτρο του λογαρίθμου είναι πιο βολικό για διάφορους λόγους οι οποίοι παρουσιάστηκαν διεξοδικά από τον Shannon [208]. Η επιλογή μιας βάσης λογαρίθμου αντιστοιχεί στην επιλογή μιας μονάδας για τη μέτρηση πληροφοριών. Ο ίδιος ο Shannon περιέγραψε τις ιδιότητες των βάσεων λογαρίθμου 2, 10 και e, αλλά στην πράξη η εντροπία σχεδόν πάντα υπολογίζεται χρησιμοποιώντας βάση λογαρίθμου το 2. Ας σημειωθεί εδώ ότι οποιαδήποτε επιλογή μιας λογαριθμικής βάσης θα μπορούσε να χρησιμοποιηθεί για τις αποδείξεις μας με ισοδύναμα αποτελέσματα. Εδώ, θεωρούμε το λογάριθμο με βάση το 2, αφού αυτή είναι η πιο δημοφιλής επιλογή λογαριθμικής βάσης σε εφαρμογές που αφορούν την εντροπία του Shannon.

Θεώρημα 7.1. *Ας θεωρήσουμε μια ακολουθία καταστάσεων ενός παρατηρούμενου συστήματος, και ας συμβολίσουμε το σύνολο των καταστάσεων με $\{s_i, i = 1, \ldots, k\}$. Το σύνολο $\{s_i\}$ αποτελεί ένα μοντέλο k στοιχείων, και $p(s_i)$ είναι η πιθανότητα εύρεσης της i-οστής κατάστασης εφόσον το πείραμα έχει τρέξει πάνω στο σύνολο των καταστάσεων. Το μέγεθος πληροφορίας της κατανομής πάνω στο σύνολο των καταστάσεων, αποτελεί την ποσότητα της στατιστικής πολυπλοκότητας, υπολογίζεται μέσω της εντροπίας Shannon ως*

$$C = -\sum_{s_i} p(s_i) \log\, p(s_i), \tag{7.1}$$

και επηρεάζεται από την παράμετρο prior.

Απόδειξη. Ας θεωρήσουμε δύο μοντέλα με k καταστάσεις, και ας συμβολίσουμε το σύνολο των καταστάσεων με $\{s_i, i = 1, \ldots, k\}$. Ας συμβολίσουμε τώρα με $p(s_i)\epsilon[0,1]$ την πιθανότητα εύρεσης της i-οστής κατάστασης εφόσον το πείραμα έχει τρέξει πάνω στο σύνολο των καταστάσεων, δηλαδή την πιθανότητα του i-οστού εξαγόμενου αποτελέσματος, $i = 1, 2, \ldots, k$. Η διακριτή κατανομή πιθανότητας είναι ένα σύνολο ίσων μη αρνητικών πραγματικών αριθμών $p(s_1), p(s_2), \ldots, p(s_k)$, που αθροίζουν στη μονάδα.

Αρχικά θεωρούμε ένα μοντέλο M με k καταστάσεις οι οποίες συνοψίζονται στο διάνυσμα παραμέτρων $s = (s_1, s_2, \ldots, s_k)$. Η ισοπίθανη κατανομή του μοντέλου είναι $\{p(s_1), p(s_2), \ldots, p(s_k)\}$. Η στατιστική πολυπλοκότητα, όπως αυτή περιγράφεται από τη σχέση (7.1), είναι συμμετρική για τα $p(s_i)$'s, και για αυτόν το λόγο χωρίς βλάβη της γενικότητας, μπορούμε να θεωρήσουμε ένα άλλο μοντέλο με $p(s_2) - \varepsilon < p(s_1) + \varepsilon$ για οποιοδήποτε θετικό ε το οποίο παράγει μη αρνητικές πιθανότητες. Η μη ισοπίθανη κατανομή αυτού του μοντέλου είναι η $\{p(s_1) + \varepsilon, p(s_2) - \varepsilon, p(s_3), \ldots, p(s_k)\}$.

Η στατιστική πολυπλοκότητα της κατανομής $\{p(s_1) + \varepsilon, p(s_2) - \varepsilon, p(s_3), \ldots, p(s_k)\}$ μείον τη στατιστική πολυπλοκότητα της κατανομής $\{p(s_1), p(s_2), p(s_3), \ldots, p(s_k)\}$ ισούται με

$$-(p(s_1) + \varepsilon)\log(p(s_1) + \varepsilon) - (p(s_2) - \varepsilon)\log(p(s_2) - \varepsilon) + p(s_1)\log\, p(s_1) + p(s_2)\log\, p(s_2),$$

η οποία ισούται με

$$-(p(s_1)+\varepsilon)\log\left(p(s_1)\left(1+\tfrac{\varepsilon}{p(s_1)}\right)\right)-(p(s_2)-\varepsilon)\log\left(p(s_2)\left(1-\tfrac{\varepsilon}{p(s_2)}\right)\right)$$
$$+p(s_1)\log\, p(s_1)+p(s_2)\log\, p(s_2).$$

Η παραπάνω σχέση ισούται με

$$-\varepsilon\log p(s_1)+\varepsilon\log p(s_2)-(p(s_1)+\varepsilon)\log\left(1+\frac{\varepsilon}{p(s_1)}\right)-(p(s_2)-\varepsilon)\log\left(1-\frac{\varepsilon}{p(s_2)}\right),$$

που με τη σειρά της ισούται με

$$\varepsilon\log\frac{p(s_2)}{p(s_1)}-(p(s_1)+\varepsilon)\log\left(1+\frac{\varepsilon}{p(s_1)}\right)-(p(s_2)-\varepsilon)\log\left(1-\frac{\varepsilon}{p(s_2)}\right). \tag{7.2}$$

Κάτω από την υπόθεση του ισοπίθανου μοντέλου M, ισχύει ότι $\log\frac{p(s_2)}{p(s_1)}=0$. Αντικαθιστώντας τη σχέση $\log\frac{p(s_2)}{p(s_1)}=0$ στη σχέση (7.2), και παραγωγίζοντας τη σχέση (7.2) ως προς ε, παίρνουμε

$$-\log\left(1+\frac{\varepsilon}{p(s_1)}\right)+\log\left(1-\frac{\varepsilon}{p(s_2)}\right). \tag{7.3}$$

Από τη στιγμή που

$$\log(1+x)=\sum_{n=1}^{\infty}(-1)^{n-1}\frac{x^n}{n}=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots=x+\mathcal{O}\left(x^2\right),$$

όπου ο συμβολισμός $\mathcal{O}$ περιγράφει την οριακή συμπεριφορά της συνάρτησης για μικρό x, η σχέση (7.3) γίνεται

$$-\left(\frac{\varepsilon}{p(s_1)}+\mathcal{O}(\varepsilon^2)\right)+\left(-\frac{\varepsilon}{p(s_2)}-\mathcal{O}(\varepsilon^2)\right),$$

η οποία με τη σειρά της γίνεται

$$-2\left(\frac{\varepsilon}{p(s_1)}+\mathcal{O}(\varepsilon^2)\right),$$

η οποία είναι αρνητική για οποιοδήποτε θετικό ε εξαιτίας της υπόθεσης ότι τα $p(s_i)$ είναι ίσοι μη αρνητικοί πραγματικοί αριθμοί, έτσι ώστε $p(s_1)=p(s_2)$.

Αποδείξαμε ότι η προκύπτουσα παράγωγος έχει αρνητικό πρόσημο, το οποίο συνεπάγεται άμεσα ότι η στατιστική πολυπλοκότητα της κατανομής $\{p(s_1),p(s_2),p(s_3),\ldots,p(s_k)\}$ είναι μεγαλύτερη από εκείνη της κατανομής $\{p(s_1)+\varepsilon,p(s_2)-\varepsilon,p(s_3),\ldots,p(s_k)\}$. Εντοπίσαμε τη διαφορά μεταξύ της πολυπλοκότητας αυτών των μοντέλων που έχουν διαφορετικές εκ των προτέρων κατανομές πάνω στις παραμέτρους τους. Αποδείξαμε ότι λιγότερη εκ των προτέρων (prior) πληροφορία αντιστοιχεί σε περισσότερη αβεβαιότητα και μεγαλύτερη πολυπλοκότητα. Συμπερασματικά, αποδείξαμε ότι η παράμετρος prior επηρεάζει σημαντικά το μέτρο της στατιστικής πολυπλοκότητας ενός μοντέλου στη διακριτή περίπτωση, ή με άλλα λόγια η στατιστική πολυπλοκότητα είναι ευαίσθητη στην παράμετρο prior.

□

Ας σημειωθεί εδώ ότι χρησιμοποιούμε την έννοια της εντροπίας και ως εκ τούτου το παραπάνω αποτέλεσμα αποδεικνύεται εύκολα και είναι αναμενόμενο, δεδομένου ότι είναι ευρέως γνωστό ότι η εντροπία μεγιστοποιείται μοναδικά αν και μόνο αν η κατανομή πιθανότητας p είναι ομοιόμορφη, δηλαδή, $p(s_i)=\frac{1}{k}$ για όλα τα i [116]. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στις εργασίες [15], [64] και [41] για την έννοια της μέγιστης εντροπίας

στη μοντελοποίηση και τις εφαρμογές της στην επιστήμη. Στην παρούσα ενότητα, εμείς ενδιαφερόμαστε να εξετάσουμε την παραμέτρο prior και τη σύνδεσή της μέσω της εντροπίας με την έννοια της πολυπλοκότητας του μοντέλου, και αποδείξαμε ότι η παράμετρος prior επηρεάζει σημαντικά το μέτρο της στατιστικής πολυπλοκότητας ενός μοντέλου στη διακριτή περίπτωση. Η στατιστική πολυπλοκότητα (υπολογισμένη μέσω της εντροπίας Shannon), αν και ορίζεται για μια διακριτή τυχαία μεταβλητή, μπορεί να επεκταθεί σε καταστάσεις όπου η υπό εξέταση τυχαία μεταβλητή είναι συνεχής. Ο Wallis [223] αντιμετώπισε τη διακριτή κατανομή πιθανότητας ως ανάλογη του ιστογράμματος μιας συνεχούς τυχαίας μεταβλητής και παρακινούμενοι από αυτό εξετάζουμε τι συμβαίνει και στη συνεχή περίπτωση.

Λήμμα 7.1. *Έστω S μια συνεχής τυχαία μεταβλητή με συνάρτηση πυκνότητας πιθανότητας $p(s)$. Τότε η στατιστική πολυπλοκότητα δίνεται από τη σχέση*

$$C = -\int p(s) log\, p(s) ds,$$

και επηρεάζεται από την παράμετρο prior.

Απόδειξη. Θεωρούμε μια συνεχή τυχαία μεταβλητή S με συνάρτηση πυκνότητας πιθανότητας $p(s)$ και στατιστική πολυπλοκότητα ίση με

$$C = -\int p(s) \log\, p(s) ds.$$

Το εύρος της μεταβλητής S μπορεί να διαιρεθεί σε n διαστήματα (l_k, u_k), $k = 1, \ldots, n$, έτσι ώστε

$$C = -\sum_{k=1}^{n} \int_{l_k}^{u_k} p(s) \log\, p(s) ds. \tag{7.4}$$

Ο ορισμός αυτός σχετίζεται με την κατάσταση στην οποία η πυκνότητα μπορεί να αναπαρασταθεί από ένα ιστόγραμμα. Δεδομένου ότι η πυκνότητα αντιπροσωπεύεται από ένα ιστόγραμμα, θεωρούμε τη διακριτή κατανομή πιθανότητας ως ανάλογη με το ιστόγραμμα μιας συνεχούς τυχαίας μεταβλητής. Ο k-οστός όρος του ανωτέρου αθροίσματος σχετίζεται με το k-οστό bin ενός ιστογράμματος, με πλάτος $w_k = u_k - l_k$. Οι bin πιθανότητες p_k, $k = 1, \ldots, n$, ορίζονται ως $p_k = \int_{l_k}^{u_k} p(s) ds$, και έπειτα μπορούν να προσεγγιστούν από την ποσότητα $w_k\, p(s_k)$, όπου το $w_k\, p(s_k)$ αναπαριστά το εμβαδόν ενός ορθογωνίου ύψους $p(s_k)$ με το s_k να είναι μια αντιπροσωπευτική τιμή εντός του διαστήματος (l_k, u_k). Τότε,το k-οστό ολοκλήρωμα στο ανωτέρω άθροισμα μπορεί να προσεγγιστεί από την ποσότητα $w_k\, p(s_k)\, \log\, p(s_k)$. Η σχέση (7.4) μπορεί να εκφραστεί μέσω των πιθανοτήτων p_k, $k = 1, \ldots, n$, ως

$$C = -\sum_{k=1}^{n} p_k \log\left(\frac{p_k}{w_k}\right). \tag{7.5}$$

Έτσι καταλήγουμε στο συμπέρασμα ότι για μια συνεχή τυχαία μεταβλητή S, η στατιστική πολυπλοκότητα επηρεάζεται από την παράμετρο prior, δεδομένου ότι για $w_k = 1$, η σχέση (7.5) αντιστοιχεί στη στατιστική πολυπλοκότητα μιας διακριτής κατανομής, ενώ για w_k σταθερό αριθμό, ο όρος $\log\, w_k$ είναι ο ίδιος για κάθε $k = 1, \ldots, n$, και σε αυτή την περίπτωση πάλι το ανωτέρω άθροισμα είναι ευαίσθητο στην παράμετρο prior σύμφωνα με το Θεώρημα 7.1. Στην περίπτωση που το εύρος της συνεχούς τυχαίας μεταβλητής είναι $(l_1, u_n) = (-\infty, +\infty)$, ο πειραματιστής μπορεί να θεωρήσει για υπολογιστικούς σκοπούς τα l_1 και u_n να είναι πεπερασμένα, δηλαδή, το πρώτο και το τελευταίο διάστημα να έχουν πλάτος παρόμοιο με αυτό των εσωτερικών διαστημάτων. Ωστόσο, δεν είναι σαφές πόση πληροφορία χάνεται όταν θεωρούμε τα l_1 και u_n να είναι πεπερασμένα. Σε μια τέτοια περίπτωση, όταν το εύρος της μεταβλητής είναι μη φραγμένο, η τυπική επιλογή ενός σταθερού w_k μπορεί να μην είναι η κατάλληλη. Εάν το w_k δεν είναι σταθερός αριθμός, προτού να πραγματοποιηθούν

οι συγκρίσεις της εντροπίας μεταξύ των ιστογραμμάτων με διαφορετικά διαμορφωμένα bin, ο πειραματιστής μπορεί να κάνει bin by bin προσαρμογές, προκειμένου να οδηγήσει την περίπτωση των bins των οποίων το πλάτος αλλάζει, στην περίπτωση των bins σταθερού πλάτους ίσο με το w, καταλήγοντας έτσι στη σχέση

$$C = -\sum_{k=1}^{n} p_k \log\, p_k + \log\, w,$$

η οποία παραμένει ευαίσθητη στην παράμετρο prior σύμφωνα με το Θεώρημα 7.1. □

7.2.2 Ένα Νέο Μέτρο Στατιστικής Πολυπλοκότητας (EPC)

Δείξαμε παραπάνω ότι η παράμετρος prior επηρεάζει σημαντικά το μέτρο της στατιστικής πολυπλοκότητας ενός μοντέλου στη διακριτή καθώς και στη συνεχή περίπτωση. Το πιο κοινό χαρακτηριστικό που συνήθως αγνοείται από πολλά μέτρα πολυπλοκότητας είναι η παράμετρος prior, που όπως είδαμε αποτελεί αναπόσπαστο μέρος του μοντέλου. Προκειμένου λοιπόν να καθοριστεί ένα μοντέλο με ακρίβεια, ο πειραματιστής θα πρέπει να θεωρεί πέρα από την εξίσωση του μοντέλου και την παράμετρο prior. Έτσι, οποιοδήποτε μέτρο πολυπλοκότητας για να θεωρείται "καλό" και αξιόπιστο, θα πρέπει να λαμβάνει υπόψη του και τις δύο αυτές πτυχές για τον ακριβή καθορισμό του μοντέλου, και ένα τέτοιο μέτρο προτείνουμε παρακάτω.

Οι επιπτώσεις της πολυπλοκότητας στην προσαρμογή του μοντέλου μπορούν να απεικονιστούν θεωρώντας το εύρος των υποδειγμάτων δεδομένων (δηλαδή, τις κατανομές πιθανότητας των παρατηρήσεων που καθορίζονται από την εξίσωση του μοντέλου) που ένα μοντέλο μπορεί να καταλαμβάνει στο χώρο των δεδομένων [169]. Ο χώρος των δεδομένων ορίζεται ως το σύνολο όλων των πιθανών υποδειγμάτων δεδομένων που θα μπορούσαν να παρατηρηθούν σε έναν πειραματισμό. Το μέγεθος του χώρου των δεδομένων που καταλαμβάνεται από ένα μοντέλο σχετίζεται θετικά με την πολυπλοκότητα του μοντέλου [169]. Με πιο απλά λόγια, η πολυπλοκότητα ενός μοντέλου αναφέρεται στην ιδιότητα του μοντέλου να επιτρέπει την πρόβλεψη ενός μεγάλου εύρους υποδειγμάτων δεδομένων. Ο Vanpaemel [219] πρότεινε το κριτήριο Prior Predictive Complexity (PPC) για τη μέτρηση του κατά πόσο μεγάλο είναι αυτό το εύρος. Το κριτήριο PPC υπολογίζει το καθολικό διάστημα (Universal Interval-UI), δηλαδή το εύρος των πιθανών παρατηρούμενων αποτελεσμάτων, και το προβλεπόμενο διάστημα (Predicted Interval-PI), δηλαδή το εύρος των προβλεπόμενων αποτελεσμάτων του μοντέλου, και υπολογίζεται από τη σχέση

$$PPC = \frac{1}{m}\sum_{i=1}^{m}\frac{|PI_i|}{|UI_i|}.$$

Για την ακριβή μέτρηση του κατά πόσο μεγάλο είναι αυτό το εύρος των υποδειγμάτων δεδομένων και την ποσοτικοποίησή του προτείνουμε ένα νέο μέτρο πολυπλοκότητας, το Entropy Prior Complexity (EPC) το οποίο έχει παρόμοια μορφή με το PPC. Ας θεωρήσουμε το καθολικό διάστημα (UI) το οποίο εκφράζει το εύρος των πιθανών παρατηρούμενων αποτελεσμάτων, ανεξάρτητα από το μοντέλο, και το προβλεπόμενο διάστημα μοντέλου (Predicted Model Interval-PMI) το οποίο εκφράζει το εύρος του συνόλου των αποτελεσμάτων που το μοντέλο θα μπορούσε να προβλέψει ή με άλλα λόγια το εύρος των προβλεπόμενων αποτελεσμάτων του μοντέλου. Προτείνουμε το εύρος του PMI να υπολογίζεται χρησιμοποιώντας το μέτρο της εντροπίας ως προς την κατανομή των εξαγόμενων αποτελεσμάτων που καθορίζονται από το μοντέλο. Με αυτό τον τρόπο ο πειραματιστής δεν χρειάζεται να περιορίσει το PMI σε ένα μικρότερο διάστημα που περιλαμβάνει μία προκαθορισμένη ποσότητα εκ των προτέρων πληροφορίας (όπως συμβαίνει με το PPC), δεδομένου ότι ένα μοντέλο ορίζει μια

κατανομή πάνω σε όλες τις πιθανές εκβάσεις του. Ως εκ τούτου, το EPC ορίζεται ως

$$EPC = \frac{1}{m}\sum_{i=1}^{m}\frac{|PMI_i|}{|UI_i|},$$

δηλαδή, ο πληθάριθμος του PMI σε σχέση με τον πληθάριθμο του UI κατά μέσο όρο για όλες τις m συνθήκες.

Για παράδειγμα, ας θεωρήσουμε ένα πείραμα με n δοκιμές Bernoulli. Το εύρος του UI είναι $[0, n]$, απλά επειδή αυτό είναι το εκ των προτέρων εύρος όλων των πιθανών αποτελεσμάτων. Υποθέτουμε ότι το Y είναι η μεταβλητή απόκρισης με συνάρτηση μάζας πιθανότητας $p(Y = y) = p(y)$. Το PMI υπολογίζεται από την εντροπία της κατανομής πάνω στο Y ως

$$PMI = -p(y)\log p(y) - \{1 - p(y)\}\log\{1 - p(y)\}.$$

Εκτιμάμε την πολυπλοκότητα του μοντέλου χρησιμοποιώντας το μέτρο EPC το οποίο αντιπροσωπεύει ένα μέτρο που σχετίζεται με το μέγεθος πληροφορίας της κατανομής πάνω στην ακολουθία καταστάσεων που παράγονται από τη μεταβλητή απόκρισης Y. Το μέτρο EPC λαμβάνει υπόψη του την παράμετρο prior δεδομένου ότι ορίζεται βάση της στατιστικής πολυπλοκότητας, η οποία όπως δείξαμε παραπάνω είναι ευαίσθητη στην παράμετρο prior. Η κύρια εφαρμογή ενός μέτρου πολυπλοκότητας μοντέλου είναι στην επιλογή του καλύτερου μοντέλου. Το προτεινόμενο περιγραφικό μέτρο EPC μπορεί να χρησιμοποιηθεί για τη στατιστική εκτίμηση της πολυπλοκότητας ενός μοντέλου από τα παρατηρούμενα δεδομένα και μπορεί να χρησιμοποιηθεί σε ένα πείραμα Bernoulli με στόχο την επιλογή του βέλτιστου μοντέλου που αποτελείται από τους παράγοντες που έχουν σημαντική συμβολή στην απόκριση.

Η επιλογή του βέλτιστου μοντέλου (ένα από τα K υπό εξέταση μοντέλα) γίνεται σύμφωνα με τη μικρότερη τιμή του μέτρου EPC, επιτρέποντας στον πειραματιστή να επιλέξει εκείνο το μοντέλο που προσεγγίζει καλύτερα την πραγματική δομή βάση της οποίας τα παρατηρούμενα δεδομένα παράχθηκαν. Η παράμετρος K, η οποία προσδιορίζει τον αριθμό των μοντέλων που εξετάζονται και συγκρίνονται, καθορίζεται σύμφωνα με τις απαιτήσεις ενός πειραματισμού. Υπάρχουν διαφορές στην πολυπλοκότητα του μοντέλου κατά τη χρήση διαφορετικών σχεδιασμών, και προφανώς το μέτρο EPC εξαρτάται από την επιλογή σχεδιασμού. Ενσωματώσαμε επιτυχώς το μέτρο πολυπλοκότητας EPC στο πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών, θέτοντας για παράδειγμα το $K = \frac{m(m-1)}{2}$, όπου το m είναι ο αριθμός των στηλών του υπό εξέταση σχεδιασμού, με στόχο την αναζήτηση και εύρεση του βέλτιστου μοντέλου, μία μέθοδο που παρουσιάζουμε αναλυτικά σε επόμενη ενότητα αυτού του κεφαλαίου.

7.2.3 Υπολογισμός της Εκ των Προτέρων Περιθώριας Κατανομής (MPD)

Ας θεωρήσουμε τώρα την περίπτωση που ο πειραματιστής θέλει να γνωρίζει εκ των προτέρων την περιθώρια πιθανότητα των δεδομένων δοθέντος του μοντέλου. Παρακινούμενοι από την περιθώρια πιθανοφάνεια [42] [168], [120], θεωρούμε ένα μοντέλο M με k παραμέτρους, έστω $\boldsymbol{\theta}=(\theta_1, \theta_2, \ldots, \theta_k)$. Με x συμβολίζουμε τα δεδομένα και $p(x|M)$ τη συνάρτηση πιθανοφάνειας. Σε ένα τέτοιο μοντέλο βασισμένο στην πιθανοφάνεια, η πρόβλεψη παίρνει τη μορφή κατανομής (μία παραμετρική οικογένεια κατανομών πιθανότητας των παρατηρήσεων), και ως εκ τούτου ορίζουμε την εκ των προτέρων περιθώρια κατανομή (Marginal Prior Distribution-MPD) ως

$$MPD = p(x|M) = \int p(x|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}. \tag{7.6}$$

Αυτή η ποσότητα αποτελεί ένα k-διάστατο ολοκλήρωμα (όπου k ο αριθμός των παραμέτρων του μοντέλου) το οποίο στην πράξη είναι δύσκολο να υπολογιστεί αναλυτικά. Για το λόγο

αυτό, προτείνουμε έναν εναλλακτικό τρόπο υπολογισμού της MPD, ο οποίος βασίζεται στο γεγονός ότι ο λόγος των εκ των υστέρων πιθανοτήτων (posterior odds ratio) για ένα μοντέλο M περιλαμβάνει την MPD.

Ο λόγος των εκ των υστέρων πιθανοτήτων για ένα μοντέλο M συμβολίζεται με $p(M|x)$ και ο λόγος των εκ των προτέρων πιθανοτήτων για ένα μοντέλο M με $p(M)$. Τότε, σύμφωνα με το Θεώρημα του Bayes έχουμε ότι

$$p(M|x) = \frac{p(M)p(x|M)}{p(x)},$$

το οποίο μπορεί να γραφεί ως

$$p(M|x) \propto p(M)p(x|M),$$

δεδομένου ότι το $p(x)$ είναι σταθερό ανεξάρτητα από το M.

Χρησιμοποιώντας τη σχέση (7.6), η ποσότητα $p(M|x)$ μπορεί να υπολογιστεί ως

$$p(M|x) = p(M) \int p(x|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}.$$

Με άλλα λόγια,

$$\text{posterior odds ratio for } M = \text{prior odds ratio for } M \times \text{MPD}$$

δίνοντας

$$MPD = \frac{posterior\ odds\ ratio\ for\ M}{prior\ odds\ ratio\ for\ M}. \tag{7.7}$$

Θεώρημα 7.2. *Θεωρούμε ένα πείραμα Bernoulli στο οποίο η δίτιμη μεταβλητή H συμβολίζει την υποτιθέμενη κλάση H={1 για θετική κλάση; 0 για αρνητική κλάση}. Η δίτιμη μεταβλητή απόκρισης Y είναι η πραγματική κλάση. Κατά συνθήκη Y={1 για θετικό αποτέλεσμα; 0 για αρνητικό αποτέλεσμα}. Η εκ των προτέρων περιθώρια κατανομή (MPD) ενός πειράματος Bernoulli ορίζεται ως*

$$MPD = \frac{P(Y = y|H = 1)}{P(Y = y|H = 0)}.$$

Για $Y = 1$, έχουμε

$$MPD^{+} = \frac{P(Y = 1|H = 1)}{P(Y = 1|H = 0)},$$

ενώ για $Y = 0$, έχουμε

$$MPD^{-} = \frac{P(Y = 0|H = 1)}{P(Y = 0|H = 0)}.$$

Απόδειξη. Ας θεωρήσουμε το λόγο πιθανοτήτων μία παράμετρος να ανήκει στη θετική κλάση πριν ακόμα εκτελεστεί το πείραμα Bernoulli, δηλαδή, χωρίς τη γνώση του πειραματικού αποτελέσματος, τότε prior odds $= \frac{P(H=1)}{P(H=0)}$. Μετά την εκτέλεση του πειράματος, δηλαδή με την παρουσία του πειραματικού αποτελέσματος ο λόγος πιθανοτήτων θετικής κλάσης δίνεται από τη σχέση posterior odds $(Y) = \frac{P(H=1|Y)}{P(H=0|Y)}$. Αποδεικνύουμε το αποτέλεσμα, υπό την παρουσία ενός θετικού πειραματικού αποτελέσματος $Y = 1$, και βρίσκουμε

$$\begin{aligned}
\text{Posterior odds } (Y=1) &= \frac{P(H=1|Y=1)}{P(H=0|Y=1)} \\
&= \frac{\frac{P(H=1,Y=1)}{P(Y=1)}}{\frac{P(H=0,Y=1)}{P(Y=1)}} \\
&= \frac{P(Y=1|H=1)}{P(Y=1|H=0)}\frac{P(H=1)}{P(H=0)} \\
&= \frac{P(Y=1|H=1)}{P(Y=1|H=0)} \times \text{prior odds},
\end{aligned}$$

η οποία συνεπάγεται άμεσα ότι

$$\text{Posterior odds } (Y=1) \text{ / prior odds} = \tfrac{P(Y=1|H=1)}{P(Y=1|H=0)},$$

η οποία με τη σειρά της συνεπάγεται ότι

$$MPD^{+} = \frac{P(Y=1|H=1)}{P(Y=1|H=0)}$$

από τη σχέση (7.7).

Η απόδειξη υπό την παρουσία ενός αρνητικού πειραματικού αποτελέσματος γίνεται με παρόμοιο τρόπο. Όταν το $Y=0$, έχουμε ότι

$$MPD^{-} = \frac{P(Y=0|H=1)}{P(Y=0|H=0)}.$$

Ως εκ τούτου, ορίζουμε την εκ των προτέρων περιθώρια κατανομή (MPD) ενός πειράματος Bernoulli ως

$$MPD = \frac{P(Y=y|H=1)}{P(Y=y|H=0)}.$$

□

7.2.4 Εμπειρική Μελέτη

Προκειμένου να δείξουμε τον τρόπο με τον οποίο ο πειραματιστής μπορεί να υπολογίσει τις MPD παραμέτρους, πραγματοποιήσαμε Monte Carlo προσομοιώσεις. Για το πείραμα προσομοίωσης θεωρούμε μοντέλα λογιστικής παλινδρόμησης με τους συντελεστές να λαμβάνουν τυχαίες τιμές από το διάνυσμα $\boldsymbol{\beta}$ και θεωρούμε μοντέλα μόνο κύριων επιδράσεων. Η αρχική μας επιλογή στο πείραμα προσομοίωσης είναι $\boldsymbol{\beta}$ = [-0.8756, 3.1869, 2.3112, -1.6768, -3.0301, 2.4999, -2.6921, -4.0285, -3.7139, 3.1744, -3.7349, 4.4703, -2.2932, 2.8191, 2.4019, -3.8837, 4.4915, -2.4739, 3.7828, 4.1800, -2.6089] $'$. Θεωρούμε *n=100* πειραματικές εκτελέσεις με μία δίτιμη απόκριση y (0 ή 1) η οποία εξαρτάται από ένα σύνολο μεταβλητών παλινδρόμησης x_{i1},x_{i2},...,x_{i20}.

Οι πραγματικά ενεργές μεταβλητές επιλέγονται τυχαία από τις είκοσι ενδεχόμενες ενεργές μεταβλητές. Παράγουμε τις τιμές για τον πίνακα σχεδιασμού X= [x_0, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9}, x_{i10}, x_{i11}, x_{i12}, x_{i13}, x_{i14}, x_{i15}, x_{i16}, x_{i17}, x_{i18}, x_{i19}, x_{i20}] μέσω 10,000 Monte Carlo δοκιμών. Για τον πίνακα σχεδιασμού X, έχουμε ότι $x_0 = 1$ είναι η στήλη της σταθεράς, δηλαδή ένα (100×1) διάνυσμα αποτελούμενο από μονάδες, και οι μεταβλητές x_{i1},...,x_{i20} παράγονται από ένα συνδυασμό κανονικών και ομοιόμορφων κατανομών με διάφορες τιμές παραμέτρων. Συγκεκριμένα, 10,000 δείγματα παράχθηκαν από

κανονικές κατανομές $x_i \sim$ N(μ=60, σ=3), N(μ=80, σ=4), N(μ=100, σ=5) και 10,000 δείγματα παράχθηκαν από ομοιόμορφες κατανομές $x_i \sim$ U(a=3, b=10), U(a=5, b=15), U(a=10, b=20). Με αυτό τον τρόπο παράχθηκαν τα διανύσματα των τυχαίων μεταβλητών εισόδου $x_{i1},\ldots,x_{i20}$ του πίνακα σχεδιασμού.

Στο πλαίσιο του ελέγχου της μηδενικής υπόθεσης ($H = 0$) έναντι της εναλλακτικής υπόθεσης ($H = 1$), η από κοινού κατανομή των (H, Y) αναπαριστά τα δεδομένα $\{(H_i, Y_i); i = 1, \ldots, 100\}$. Υποθέτουμε ότι οι παρατηρήσεις (H_i, Y_i) είναι ανεξάρτητες από τώρα και στο εξής.

Πίνακας 7.1: Αποτελέσματα 10,000 Monte Carlo δοκιμών για τα λογιστικά μοντέλα παλινδρόμησης

	Αποτέλεσμα
$P(Y = 1\|H = 1)$	0.8889
$P(Y = 1\|H = 0)$	0.7805
MPD^+	1.1388
$P(Y = 0\|H = 1)$	0.1111
$P(Y = 0\|H = 0)$	0.2195
MPD^-	0.5061

Εφαρμόζοντας την προτεινόμενη μέθοδο, καθορίσαμε την εκ των προτέρων περιθώρια κατανομή μέσω των (MPD$^+$, MPD$^-$)=(1.1388, 0.5061) παραμέτρων, όπως φαίνεται στον Πίνακα 7.1. Οι (MPD$^+$, MPD$^-$) παράμετροι είναι οι ποσοδείκτες της σχέσης μεταξύ του λόγου των εκ των προτέρων πιθανοτήτων και του λόγου των εκ των υστέρων πιθανοτήτων. Εάν θεωρήσουμε ότι $P(H = 1)$ είναι η εκ των προτέρων πιθανότητα ένα υπόδειγμα να ανήκει στη θετική κλάση και $P(H = 1|Y)$ η εκ των υστέρων πιθανότητα, τότε η MPD είναι ο πολλαπλασιαστικός συντελεστής (multiplication factor) που σχετίζει τις εκ των προτέρων και εκ των υστέρων κατανομές. Παρουσιάζουμε παρακάτω ένα παράδειγμα για να δείξουμε την πρακτική χρησιμότητα των MPD παραμέτρων. Στη μελέτη προσομοίωσής μας, έχουμε να κάνουμε με μια ακολουθία n=100 ανεξάρτητων δοκιμών Bernoulli. Ας υποθέσουμε ότι η πιθανότητα επιτυχίας είναι $p = p(Y = 1) = 0.05$ και η πιθανότητα της αποτυχίας είναι $p(Y = 0) = 1 - p = 0.95$ μεταξύ όλων των ανεξάρτητων δοκιμών. Η εκ των προτέρων πληροφορία υπολογίζεται μέσω των παραμέτρων MPD$^+$=1.1388 και MPD$^-$=0.5061. Ως εκ τούτου, ο λόγος πιθανοτήτων της θετικής υποτιθέμενης κλάσης μεταξύ εκείνων των υποδειγμάτων που δίνουν θετικό πειραματικό αποτέλεσμα ισούται με

$$MPD^+ \tfrac{p}{1-p} = 1.1388 \tfrac{0.05}{0.95} = 0.0599.$$

Ομοίως, ο λόγος πιθανοτήτων της θετικής υποτιθέμενης κλάσης μεταξύ εκείνων των υποδειγμάτων που δίνουν αρνητικό πειραματικό αποτέλεσμα ισούται με

$$MPD^- \tfrac{p}{1-p} = 0.5061 \tfrac{0.05}{0.95} = 0.0266.$$

Αυτοί οι δύο λόγοι πιθανοτήτων είναι πολύ χρήσιμοι προκειμένου να συγκρίνουμε το κατά πόσο μεγαλύτερη είναι η μία πιθανότητα σε σχέση με την άλλη. Με άλλα λόγια, οι (MPD$^+$, MPD$^-$) λόγοι πιθανοτήτων μπορούν να δείξουν στον πειραματιστή πόσες φορές το θετικό αποτέλεσμα σε ένα πείραμα Bernoulli ($y = 1$) είναι πιο πιθανό να προκύψει από το να μην προκύψει ($y = 0$). Δηλαδή, οι πιθανότητες της θετικής υποτιθέμενης κλάσης είναι περίπου 5.9% και 2.6% μετά από θετικό αποτέλεσμα ($y = 1$) και μετά από αρνητικό αποτέλεσμα ($y = 0$), αντίστοιχα, για αυτό το πείραμα. Γενικά, οι (MPD$^+$, MPD$^-$) παράμετροι μπορούν να χρησιμοποιηθούν άμεσα μαζί με την πιθανότητα επιτυχίας/αποτυχίας ενός πειράματος Bernoulli, προκειμένου να υπολογιστεί η πιθανότητα μιας υποτιθέμενης κλάσης δοθέντος του πειραματικού αποτελέσματος. Οι MPD παράμετροι αποτελούν ουσιαστικά τις εμπειρικές εκτιμήτριες των λόγων πιθανοφάνειας και μπορούν να χρησιμοποιηθούν στο πλαίσιο ενός πειράματος Bernoulli ως μέτρα πληροφόρησης.

7.3 Πολυπλοκότητα Μοντέλου και Σχεδιασμοί Πειραμάτων

Υπάρχουν διαφορές στην πολυπλοκότητα του μοντέλου κατά τη χρήση διαφορετικών σχεσιασμών, και προφανώς το μέτρο EPC όπως αυτό παρουσιάστηκε παραπάνω, εξαρτάται από την επιλογή σχεδιασμού. Ενσωματώσαμε επιτυχώς το μέτρο πολυπλοκότητας EPC στο πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών, με στόχο την αναζήτηση και εύρεση του βέλτιστου μοντέλου μέσα από πολλά υποψήφια λογιστικά μοντέλα παλινδρόμησης, μία μέθοδο που παρουσιάζουμε βήμα προς βήμα παρακάτω.

7.3.1 Επιλογή Βέλτιστου Μοντέλου στους Υπερκορεσμένους Σχεδιασμούς Μέσω του Κριτηρίου EPC

Σε αυτή την ενότητα, παρουσιάζουμε μία μέθοδο στην οποία η παράμετρος της πολυπλοκότητας μοντέλου εμπλέκεται μέσα στη διαδικασία ανάλυσης των υπερκορεσμένων σχεδιασμών. Συγκεκριμένα, το μέτρο πολυπλοκότητας EPC ενσωματώνεται σε μία μέθοδο επιλογής βέλτιστου μοντέλου μέσα από πολλά υποψήφια λογιστικά μοντέλα παλινδρόμησης για την ανάλυση των υπερκορεσμένων σχεδιασμών. Η μέθοδος επιλογής βέλτιστου μοντέλου έχει ως στόχο την εύρεση των ενεργών κύριων επιδράσεων που φαίνεται να επηρεάζουν σημαντικά τη μεταβλητή απόκρισης Bernoulli και εκτελεί τα ακόλουθα βήματα:

Βήμα 1. Έστω $X = [x_1, x_2, \ldots, x_m]$ ο $n \times m$ πίνακας του υπερκορεσμένου σχεδιασμού, όπου x_j, $j = 1, 2, \ldots, m$, αντιστοιχεί στην j-οστή στήλη του πίνακα X, και $n \times 1$ το δίτιμο διάνυσμα απόκρισης Y.

Βήμα 2. Υπολογίζουμε το σύνολο των K μοντέλων, όπου $K = \frac{m(m-1)}{2}$.

Βήμα 3. Για κάθε μοντέλο i, $i = 1, 2, \ldots, K$, υπολογίζουμε το $UI = [0, n]$, όπου το n αναπαριστά τον αριθμό των δοκιμών Bernoulli και στη συνέχεια υπολογίζουμε το PMI ως $PMI = -p(y) \log p(y) - \{1 - p(y)\} \log\{1 - p(y)\}$.

Βήμα 4. Για κάθε μοντέλο i, $i = 1, 2, \ldots, K$, κανονικοποιούμε το PMI από το UI και έπειτα υπολογίζουμε το μέτρο EPC από τη σχέση

$$EPC = \frac{1}{m} \sum_{i=1}^{m} \frac{|PMI_i|}{|UI_i|}.$$

Βήμα 5. Τελικά, επιλέγουμε ως βέλτιστο μοντέλο εκείνο με τη μικρότερη τιμή του μέτρου EPC. Συγκεκριμένα, κρατάμε εκείνους τους παράγοντες που οδηγούν στο "στενότερο" εύρος του PMI σε σχέση με το UI, παρέχοντας με αυτό τον τρόπο ένα απλούστερο μοντέλο. Οι παράγοντες αυτοί τελικά δηλώνονται ως σημαντικοί.

Η επιλογή του βέλτιστου μοντέλου γίνεται σύμφωνα με τη μικρότερη τιμή του μέτρου EPC για τον εξής βασικό λόγο. Ένα στενό εύρος του PMI, σε σχέση με το UI, υποδεικνύει ένα απλούστερο μοντέλο, ενώ ένα πολύπλοκο μοντέλο χαρακτηρίζεται από ένα ευρύ PMI. Για να το θέσουμε διαφορετικά, όσο μεγαλύτερη είναι η ποσότητα της εκ των προτέρων πληροφορίας που υπάρχει σε ένα μοντέλο, τόσο στενότερο θα είναι το εύρος του PMI, και κατά συνέπεια, τόσο απλούστερο θα είναι και το μοντέλο. Αυτή η διαφορά στην ύπαρξη της εκ των προτέρων πληροφορίας μεταφράζεται σε διαφορά στην πολυπλοκότητα μέσω του μέτρου EPC.

7.3.2 Επιλογή Τιμών Κατωφλιού

Η επιλογή των τιμών κατωφλιού της προτεινόμενης μεθόδου διαδραματίζει πολύ σημαντικό ρόλο. Για έναν υπερκορεσμένο σχεδιασμό, δεν είναι συχνά υπολογιστικά εφικτό να περιλαμβάνει όλα τα πιθανά μοντέλα στη διαδικασία. Η αρχή της σποραδικότητας των επιδράσεων [23]

προτείνει τον περιορισμό σε ένα μόνο σύνολο μοντέλων, καθένα από τα οποία θα περιέχει μόνο λίγους παράγοντες. Είναι ευρέως αποδεκτό ότι η αποτελεσματικότητα των υπερκορεσμένων σχεδιασμών στην ανίχνευση ενεργών παραγόντων απαιτεί τον αριθμό αυτών των παραγόντων να είναι μικρός [23]. Η απόφαση σχετικά με την τιμή κατωφλιού s, η οποία στην πράξη καθορίζει το μέγιστο αριθμό των σημαντικών παραγόντων που μπορούν να αναγνωριστούν σε κάθε μοντέλο, έγινε βάση της αρχής της σποραδικότητας. Με αυτό τον τρόπο καθορίσαμε το $s = m/2$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου πίνακα σχεδιασμού X. Η παράμετρος K, η οποία προσδιορίζει τον αριθμό των μοντέλων που εξετάζονται και συγκρίνονται, καθορίζεται ως $K = \frac{m(m-1)}{2}$, όπου το m είναι ο αριθμός των στηλών του υπό εξέταση σχεδιασμού, και η επιλογή αυτής της τιμής κατωφλιού έγινε και πάλι βάση της αρχής της σποραδικότητας των επιδράσεων, και προτείνεται ως η καταλληλότερη επιλογή στην επιστημονική εργασία [160]. Η διαδικασία αυτή χαρακτηρίζεται ως many-models μέθοδος και παρουσιάστηκε πρώτη φορά στην εργασία [112].

7.3.3 Αξιολόγηση της Μεθόδου Επιλογής Βέλτιστου Μοντέλου

Για τη διερεύνηση και την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου πραγματοποιήσαμε μια εμπειρική μελέτη, θεωρώντας ένα ευρύ φάσμα προσομειωμένων μοντέλων και υπερκορεσμένων σχεδιασμών. Στη μελέτη προσομοίωσης χρησιμοποιήσαμε υπερκορεσμένους σχεδιασμούς οι οποίοι κατασκευάζονται από τα κλάσματα πινάκων Hadamard, σύμφωνα με τη μέθοδο που προτάθηκε από τον Lin [148]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 10$ παράγοντες και $n = 6$ πειραματικές εκτελέσεις. Επιπλέον, για τα πειράματα προσομοίωσης, χρησιμοποιήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς, όπως αυτοί κατασκευάστηκαν από τους Koukouvinos et al. [130]. Αυτοί οι υπερκορεσμένοι σχεδιασμοί έχουν n πειραματικές εκτελέσεις και $m = q \cdot (n-1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Στις προσομοιώσεις μας θεωρήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, m) τιμές: $(6, 10)$, $(8, 14)$, $(10, 18)$, $(12, 22)$, $(14, 26)$, $(16, 30)$, $(18, 34)$, $(20, 38)$ και $(22, 42)$. Εκτός από αυτούς τους σχεδιασμούς, θεωρήσαμε και τους s-block-ορθογώνιους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με n πειραματικές εκτελέσεις και $m = s(n-1)$ παράγοντες, όπως αυτοί κατασκευάστηκαν από τους Tang και Wu [215]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 22$ παράγοντες και $n = 12$ πειραματικές εκτελέσεις.

Για τα πειράματα προσομοίωσης, αναπτύξαμε λογιστικά μοντέλα παλινδρόμησης με τους συντελεστές παλινδρόμησης να λαμβάνουν τυχαίες τιμές από το διάνυσμα β. Να σημειώσουμε εδώ ότι μας ενδιαφέρει να μελετήσουμε μοντέλα που περιλαμβάνουν μόνο κύριες επιδράσεις, οπότε θεωρούμε ότι τα μοντέλα μας δεν συμπεριλαμβάνουν αλληλεπιδράσεις. Οι πραγματικά ενεργοί παράγοντες επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων και οι συντελεστές των μη ενεργών παραγόντων τέθηκαν ίσοι με μηδέν στο πραγματικό μοντέλο. Η κατανομή των αντιθέσεων δεν επηρεάζεται από το μέτρο (magnitude) των συντελεστών, αλλά εξαρτάται από το σχετικό μέγεθος (relative size) των συντελεστών. Δεδομένου ότι οι συνθήκες στην πράξη είναι συνήθως διαφορετικές από εκείνες των προσομοίωσεων και ότι ο πειραματιστής δεν ξέρει πόσοι και ποιοί παράγοντες μπορεί να είναι ενεργοί, προκειμένου να εξετάσουμε το πόσο ευαίσθητα είναι τα αποτελέσματα στην επιλογή και τον αριθμό των ενεργών στηλών-παραγόντων, μελετήσαμε ένα ευρύ φάσμα προσομειωμένων μοντέλων στα πειράματά μας στα οποία θεωρούμε για κάθε υπερκορεσμένο σχεδιασμό διαφορετικό αριθμό ενεργών παραγόντων, διαφορετική επιλογή ενεργών στηλών και διαφορετικά β. Στην εμπειρική μελέτη μας χρησιμοποιήσαμε όλα τα μοντέλα τα οποία απαριθμούνται στον Πίνακα 7.2.

Προκειμένου να αξιολογήσουμε την προτεινόμενη μέθοδο, εκτιμήσαμε τα σφάλματα Τύπου I και Τύπου II. Αυτό που μας ενδιαφέρει να ελέγξουμε στην εμπειρική μελέτη μας είναι το

Πίνακας 7.2: Μοντέλα μελέτης προσομοίωσης

Μοντέλο	ΥΣ από	m	n	β
1	[148]	10	6	[-1, 0, 8, 1, 22, 0, 19, 0, 0, 0]$'$
2	[148]	10	6	[-9, 0, 0, 0, 0, -8, 0, 0, 3, 0]$'$
3	[215]	22	12	[-1, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0, 9, 0, 0]$'$
4	[215]	22	12	[1, 0, 0, -13, 0, 0, 4, -3, 0, 0, -6, 0, 0, -7, 0, -24, 0, -5, 0, -21, 0, 0]$'$
5	[130]	22	12	[1, 0, 0, 0, -1, 0, 1, 0, 0, 0, 0, 0, -2, 0, 0, -4, 0, 0, 0, -1, 0, 0]$'$
6	[130]	22	12	[1, 2, 0, 0, 0, 6, 0, 0, 0, 0, 7, 0, 0, 0, -1, 0, 0, 5, 0, 18, 0, 0]$'$
7	[130]	18	10	[0, 0, -5, 0, -3, 4, 0, 0, 7, 0, 0, 1, 2, 0, 0, -7, 0, 3]$'$
8	[130]	18	10	[0, 0, -7, 0, 0, 6, 4, 0, 0, -3, 0, 0, 1, 1, 0, 0, 0, 1]$'$
9	[130]	18	10	[-1, 0, 0, 0, 14, 0, -2, 0, 0, 0, 0, 0, 0, 17, 0, 0, 0, 0]$'$
10	[130]	18	10	[0, -4, 1, 0, -2, 0, 3, -4, 0, 0, 0, -6, 0, 0, -2, 0, 2, 0]$'$
11	[130]	26	14	[-2, -1, 0, 0, 2, -3, 0, 0, -3, 0, 0, 2, 0, -4, 0, 0, 0, -6, 0, 0, -4, 0, 0, 0, -5, 0]$'$
12	[130]	26	14	[2, 0, 0, 5, 0, 0, -17, 0, 0, 0, 12, 0, 0, 3, 0, -8, 0, 0, -9, 0, 0, -1, 0, -4, 0, 0]$'$
13	[130]	30	16	[0, -1, 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]$'$
14	[130]	30	16	[-1, -2, -1, 0, 2, 0, 0, 0, 6, 0, 4, 0, 0, 0, 0, 0, -3, 2, 0, 0, 0, 0, 1, 0, 0, 7, 0, 0, 0, 0]$'$
15	[130]	34	18	[1, -13, 0, -2, 0, 0, 0, -3, 0, 0, 0, -9, 0, 0, 0, 14, 0, 0, 9, 0, 1, 1, 0, 0, 0, -12, 0, 0, 0, 0, 0, -6, 0, -2]$'$
16	[130]	34	18	[0, 0, 3, 4, 0, 0, 0, 0, -1, 0, 18, 0, 0, 0, 0, -7, 0, 0, 0, -11, 0, -19, 0, 0, 0, -5, 0, 0, 1, 1, 0, 0, 9, 0]$'$
17	[130]	38	20	[1, 0, 5, 0, -17, 0, 0 ,0, 0, 0, -23, 0, 0, 0, 0, 0, 0, 9, 0, 0, 0, 0, 26, 0, 0, 0, 0, 0, 12, 14, 0, 0, 0, 3, 0, 0, 0, 0]$'$
18	[130]	38	20	[0, 0, -5, 0, 0, 0, -7, 0, 0, 0, 12, 0, 0, 1, 0, -22, 0, 0, -1, 0, 0, 0, 0, 0, 0, -1, 0, 4, 0, 0, 0, 0, 0, 0, -2, 15, 0, 0, 0]$'$
19	[130]	42	22	[0, 0, 0, 0, -12, 0, 0, 0, 0, 1, 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, 1, 0, 0]$'$
20	[130]	42	22	[0, 0, 0, 0, -13, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, -4, 0, 0, 0, 5, 0, -14, 0, 0, 0, 0, -3, 2, 0, 0, 0, 0, 0, -4, 1, -1, 3, 0, 0]$'$
21	[130]	10	6	[-7, 0, 0, 0, 0, -7, 0, 0, 3, 0]$'$
22	[130]	10	6	[1, 0, -8, 0, -21, 0, -17, 0, 0, 0]$'$
23	[130]	14	8	[1, 0, -3 ,0 ,0, -1, 0, 0, 0, 0, 7, 0, 0, 0]$'$
24	[130]	14	8	[1, 4, -5, 2, 0, 0, 0, -1, 0, 0, -7, 0, 0, 0]$'$

κόστος του να δηλώσουμε μία ανενεργή επίδραση ως ενεργή (σφάλμα Τύπου I) καθώς και το κόστος του να δηλώσουμε μία ενεργή επίδραση ως ανενεργή (σφάλμα Τύπου II). Το ποσοστό σφάλματος Τύπου I και το ποσοστό σφάλματος Τύπου II αποτελούν κριτήρια τύπου "smaller the better". Επίσης, θεωρήσαμε ως κριτήριο αξιολόγησης την κάλυψη (coverage), που καλείται το μέσο ποσοστό των προσομοιώσεων στις οποίες το σύνολο των παραγόντων που δηλώθηκαν ως ενεργοί περιλαμβάνει όλους τους πραγματικά ενεργούς παράγοντες ("larger the better"). Στα πειράματά μας εκτελέσαμε τις προσομοιώσεις 1000 φορές και τα ληφθέντα αποτελέσματα συνοψίζονται στον Πίνακα 7.3 που ακολουθεί. Συγκεκριμένα, στον Πίνακα 7.3, η πρώτη στήλη αναφέρεται στον αριθμό που αντιστοιχεί σε κάθε μοντέλο που χρησιμοποιείται. Οι στήλες που ονομάζονται ως "Type I" και "Type II" αναφέρονται στις τιμές των σφαλμάτων Τύπου I και Τύπου II που αντιστοιχούν σε κάθε μοντέλο. Η τελευταία στήλη "Coverage" παρουσιάζει τις τιμές του κριτηρίου της κάλυψης για κάθε μοντέλο.

Γενικά, παρατηρούμε από τον Πίνακα 7.3, ότι η προτεινόμενη μέθοδος επιτυγχάνει ικανοποιητικά χαμηλές τιμές και για τα δύο υπό εκτίμηση σφάλματα, και σχεδόν για όλα τα μοντέλα. Εκτός από αυτό, σημειώνουμε ότι οι τιμές των σφαλμάτων παραμένουν ικανοποιητικά χαμηλές ακόμα και για τα μοντέλα στα οποία η αρχή της σποραδικότητας είναι ασθενέστερη, δηλαδή στα μοντέλα των οποίων ο αριθμός των ενεργών παραγόντων p είναι μεγαλύτερος από $\frac{n}{2}$. Συγκεκριμένα, οι τιμές των σφαλμάτων Τύπου II είναι σημαντικά χαμηλότερες από τις αντίστοιχες τιμές των σφαλμάτων Τύπου I. Σύμφωνα με τον Πίνακα 7.3, το μέσο ποσοστό σφάλματος Τύπου I σε 1000 προσομοιώσεις είναι της τάξης του 27% και το μέσο ποσοστό σφάλματος Τύπου II σε 1000 προσομοιώσεις είναι της τάξης του 13%. Το γεγονός αυτό μας δείχνει ότι η προτεινόμενη μέθοδος παρουσιάζει την τάση να δηλώνει εσφαλμένα περισσότερους μη ενεργούς παραγόντες ως ενεργούς, και λιγότερους ενεργούς παραγόντες ως μη ενεργούς.

Με άλλα λόγια, η προτεινόμενη μέθοδος τείνει να δηλώνει με υψηλότερο ποσοστό τις ανενεργές επιδράσεις ως ενεργές και με πολύ χαμηλότερο ποσοστό τις ενεργές επιδράσεις ως ανενεργές, και υπό αυτή την έννοια το κριτήριο EPC θα μπορούσε πράγματι να χαρακτηριστεί ως συντηρητικό (conservative). Επιπλέον, παρατηρούμε από τον Πίνακα 7.3, ότι η μέση τιμή για την ισχύ (power=1-Type II error rate) και την κάλυψη είναι 0.868 και 0.814, αντίστοιχα, γεγονός το οποίο μας υποδεικνύει ότι σχεδόν όλες οι ενεργές επιδράσεις ανιχνεύονται σωστά.

Λόγω του γεγονότος ότι τα σφάλματα Τύπου I είναι υψηλότερα σε σχέση με τα σφάλματα Τύπου II για όλα τα μοντέλα, εκτελέσαμε επιπλέον προσομοιώσεις για τον έλεγχο της απόδοσης της προτεινόμενης μεθόδου, αν θεωρήσουμε ότι δεν υπάρχουν ενεργοί παράγοντες. Σε μια τέτοια περίπτωση τα ποσοστά σφάλματος Τύπου II και η κάλυψη πλέον δεν είναι εφικτό να υπολογιστούν. Έτσι, υπολογίσαμε τα ποσοστά σφάλματος Τύπου I και το μέσο αριθμό των παραγόντων που δηλώθηκαν εσφαλμένα ως ενεργοί, και τα αποτελέσματα παρουσιάζονται στον Πίνακα 7.4. Ας σημειωθεί εδώ ότι η τελευταία στήλη του Πίνακα 7.4 που ονομάζεται ως "Avg. No." αναφέρεται στο μέσο αριθμό των παραγόντων που δηλώθηκαν εσφαλμένα ως ενεργοί.

Σύμφωνα με τον Πίνακα 7.4, τα μοντέλα με τους περισσότερους παράγοντες, όπως τα μοντέλα $(42, 22)$, $(38, 20)$ δηλώνουν το πολύ τέσσερις παράγοντες ως ενεργούς. Τα μοντέλα $(34, 18)$, $(30, 16)$, $(26, 14)$, δηλώνουν το πολύ τρεις παράγοντες ως ενεργούς. Αντίστοιχα, τα μοντέλα $(22, 12), (18, 10), (14, 8)$, δηλώνουν το πολύ δύο παράγοντες ως ενεργούς. Τα μοντέλα με τους λιγότερους παράγοντες, όπως τα μοντέλα $(10, 6)$ δηλώνουν το πολύ έναν παράγοντα ως ενεργό. Σχετικά με την απόδοση της μεθόδου, παρατηρούμε ότι καθώς μειώνεται ο αριθμός των παραγόντων του σχεδιασμού, μειώνεται και ο μέσος αριθμός των παραγόντων που δηλώθηκαν εσφαλμένα ως ενεργοί. Η προτεινόμενη μέθοδος φαίνεται να λειτουργεί καλά, δηλώνοντας μόνο έναν πολύ μικρό αριθμό παραγόντων ως ενεργούς, ενώ στην πραγματικότητα δεν υπάρχουν καθόλου ενεργοί παράγοντες.

Συμπερασματικά, τα δύο σημαντικότερα χαρακτηριστικά-πλεονεκτήματα της προτεινόμενης μεθόδου επιλογής βέλτιστου μοντέλου είναι πρώτον ότι η συμπερασματολογία της βασίζεται σε μια σειρά από μοντέλα και όχι σε ένα μοναδικό μοντέλο, δεύτερον η άμεση αντιστοιχία της με το βασικότερο στόχο του προβλήματος ανάλυσης των υπερκορεσμένων σχεδιασμών, δηλαδή, τον εντοπισμό όλων των σημαντικών παραγόντων καθώς επίσης και του βέλτιστου μοντέλου, και μάλιστα με τη μικρότερη δυνατή πολυπλοκότητα, γεγονός που επιτυγχάνεται με τη χρήση του κριτηρίου EPC.

Πίνακας 7.3: Απόδοση της προτεινόμενης μεθόδου για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων

Μοντέλο	Type I	Type II	Coverage
1	0.20	0.00	1.00
2	0.14	0.00	0.93
3	0.29	0.00	0.89
4	0.23	0.22	0.76
5	0.25	0.16	0.82
6	0.33	0.28	0.68
7	0.20	0.12	0.82
8	0.27	0.14	0.83
9	0.35	0.00	0.87
10	0.30	0.25	0.71
11	0.37	0.10	0.77
12	0.37	0.22	0.70
13	0.36	0.12	0.77
14	0.30	0.10	0.89
15	0.25	0.16	0.81
16	0.22	0.09	0.88
17	0.24	0.13	0.85
18	0.39	0.30	0.67
19	0.21	0.16	0.81
20	0.27	0.09	0.89
21	0.14	0.00	0.91
22	0.25	0.16	0.79
23	0.25	0.20	0.75
24	0.37	0.16	0.73

Πίνακας 7.4: Απόδοση της προτεινόμενης μεθόδου για τα μοντέλα 1-24 με χρήση 1000 προσομοιώσεων αν θεωρήσουμε ότι δεν υπάρχουν ενεργοί παράγοντες

Μοντέλο	(factors, runs)	Type I	Avg. No.
1	$(10, 6)$	0.11	0.54
2	$(10, 6)$	0.04	0.12
3	$(22, 12)$	0.20	2.22
4	$(22, 12)$	0.12	2.14
5	$(22, 12)$	0.15	2.23
6	$(22, 12)$	0.22	2.35
7	$(18, 10)$	0.09	1.34
8	$(18, 10)$	0.18	1.52
9	$(18, 10)$	0.24	1.93
10	$(18, 10)$	0.21	1.82
11	$(26, 14)$	0.26	3.11
12	$(26, 14)$	0.28	3.16
13	$(30, 16)$	0.24	3.36
14	$(30, 16)$	0.19	3.24
15	$(34, 18)$	0.14	3.19
16	$(34, 18)$	0.11	3.14
17	$(38, 20)$	0.12	3.56
18	$(38, 20)$	0.29	3.88
19	$(42, 22)$	0.12	3.94
20	$(42, 22)$	0.16	4.12
21	$(10, 6)$	0.03	0.57
22	$(10, 6)$	0.11	1.13
23	$(14, 8)$	0.16	1.67
24	$(14, 8)$	0.28	1.82

7.4 Συμπεράσματα

Στο κεφάλαιο αυτό μελετήσαμε την έννοια της πολυπλοκότητας ενός μοντέλου και δείξαμε ότι η παράμετρος prior επηρεάζει σημαντικά το μέτρο της στατιστικής πολυπλοκότητας ενός μοντέλου στη διακριτή καθώς και στη συνεχή περίπτωση. Προκειμένου λοιπόν να καθοριστεί ένα μοντέλο με ακρίβεια, ο πειραματιστής θα πρέπει να θεωρεί πέρα από την εξίσωση του μοντέλου και την παράμετρο prior. Ένα τέτοιο μέτρο πολυπλοκότητας προτείναμε το οποίο λαμβάνει υπόψη του και τις δύο αυτές πτυχές για τον ακριβή καθορισμό του μοντέλου.

Το προτεινόμενο μέτρο EPC αντιπροσωπεύει ένα μέτρο πολυπλοκότητας που σχετίζεται με το μέγεθος πληροφορίας της κατανομής πάνω στην ακολουθία καταστάσεων που παράγονται από τη μεταβλητή απόκρισης Y, και σημαντικό χαρακτηριστικό του αποτελεί το γεγονός ότι η διαφορά στην ύπαρξη της εκ των προτέρων πληροφορίας μεταφράζεται σε διαφορά στην πολυπλοκότητα μέσω του μέτρου EPC. Το προτεινόμενο περιγραφικό μέτρο EPC μπορεί να χρησιμοποιηθεί για τη στατιστική εκτίμηση της πολυπλοκότητας ενός μοντέλου από τα παρατηρούμενα δεδομένα και μπορεί να χρησιμοποιηθεί σε ένα πείραμα Bernoulli με στόχο την επιλογή του βέλτιστου μοντέλου που αποτελείται από τους παράγοντες που έχουν σημαντική συμβολή στην απόκριση. Η επιλογή του βέλτιστου μοντέλου γίνεται σύμφωνα με τη μικρότερη τιμή του μέτρου EPC, επιτρέποντας στον πειραματιστή να επιλέξει εκείνο το μοντέλο που προσεγγίζει καλύτερα την πραγματική δομή βάση της οποίας τα παρατηρούμενα δεδομένα παράχθηκαν. Επιπλέον, το μέτρο πολυπλοκότητας EPC ενσωματώνεται επιτυχώς στο πρόβλημα ανάλυσης των υπερκορεσμένων σχεδιασμών, με στόχο την αναζήτηση και εύρεση του βέλτιστου μοντέλου μέσα από πολλά υποψήφια λογιστικά μοντέλα παλινδρόμησης.

Επιπλέον, σε ένα μοντέλο βασισμένο στην πιθανοφάνεια, η πρόβλεψη παίρνει τη μορφή κατανομής γνωστή ως εκ των προτέρων περιθώρια κατανομή (MPD), η οποία στην πράξη είναι δύσκολο να υπολογιστεί αναλυτικά. Για το λόγο αυτό, προτείναμε έναν εναλλακτικό τρόπο υπολογισμού της, μέσω των MPD παραμέτρων. Οι MPD παράμετροι απαντούν πρα-

κτικά στο ερώτημα του κατά πόσο η παρουσία ενός πειραματικού αποτελέσματος αλλάζει την εκ των προτέρων γνώση της υποτιθέμενης κλάσης. Με άλλα λόγια, οι (MPD^+, MPD^-) παράμετροι ποσοτικοποιούν την αλλαγή στην πιθανότητα ένα υπόδειγμα να ανήκει σε μία κλάση που λαμβάνεται με τη γνώση του πειραματικού αποτελέσματος. Οι MPD παράμετροι αποτελούν ουσιαστικά τις εμπειρικές εκτιμήτριες των λόγων πιθανοφάνειας και μπορούν να χρησιμοποιηθούν στο πλαίσιο ενός πειράματος Bernoulli ως μέτρα πληροφόρησης.

ΚΕΦΑΛΑΙΟ **8**

Επιλογή του Καλύτερου Υποσυνόλου Μεταβλητών Μέσω του Κριτηρίου Πληροφορίας ϵBIC

Statisticians, like artists,
have the bad habit of falling
in love with their models.

—*George E.P. Box (1919–2013)*

Στο όγδοο αυτό κεφάλαιο προτείνουμε μία μέθοδο επιλογής του καλύτερου υποσυνόλου μεταβλητών, η οποία επεκτείνει την ιδέα επιλογής μεταβλητών σε προβλήματα ταξινόμησης ή πρόβλεψης, λαμβάνοντας ταυτόχρονα υπόψη τη μεμονωμένη προβλεπτική δύναμη της κάθε μεταβλητής. Για το σκοπό αυτό χρησιμοποιούμε ως κριτήριο την απόδοση ενός ταξινομητή που λαμβάνει υπόψη του μία και μόνο μεταβλητή κάθε φορά. Συγκεκριμένα, το κριτήριο που χρησιμοποιούμε είναι το εμβαδόν κάτω από την ROC καμπύλη (Area Under the ROC Curve-AUC). Η προτεινόμενη μέθοδος βασίζεται στη μετρική AUC η οποία χρησιμοποιείται εδώ από άλλη οπτική γωνία (να λάβουμε εκ των προτέρων πληροφορία για τους πιθανώς σημαντικούς παράγοντες) και σε μια τροποποίηση (ϵBIC) του κριτηρίου modified Bayesian Information Criterion (mBIC). Η σύγκριση της απόδοσης του νέου αυτού κριτηρίου ϵBIC με κλασικά κριτήρια πληροφορίας (όπως το AIC, BIC και mBIC) που χρησιμοποιούνται για την επιλογή του καλύτερου υποσυνόλου μεταβλήτων, γίνεται αρχικά εκτελώντας εκτενή πειράματα προσομοίωσης και έπειτα αναλύοντας μία πραγματική βάση δεδομένων, παρέχοντας και στις δύο περιπτώσεις πολύ ικανοποιητικά αποτελέσματα. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [134].

8.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Η επιλογή μεταβλητών αποτελεί το επίκεντρο πολλών ερευνών σε προβλήματα στατιστικής μοντελοποίησης μεγάλης διάστασης, σε διάφορους επιστημονικούς τομείς. Υπάρχουν δύο βασικοί τρόποι για την αντιμετώπιση του προβλήματος της επιλογής μεταβλητών, η κατάταξη των μεταβλητών (ranking) και η επιλογή του καλύτερου υποσυνόλου μεταβλητών (best subset). Πολλοί αλγόριθμοι επιλογής μεταβλητών περιλαμβάνουν ένα απλό ιεραρχικό κριτήριο για την κατάταξη των μεταβλητών, και χρησιμοποιούνται ευρέως σε ιατρικές μελέτες π.χ. για τη διάκριση μεταξύ υγιών και ασθενών. Τα κριτήρια κατάταξης περιλαμβάνουν κριτήρια συσχέτισης τα οποία επιβάλλουν μία κατάταξη σύμφωνα με τον έλεγχο καλής προσαρμογής των μεμονωμένων μεταβλητών, και κριτήρια της Θεωρίας Πληροφοριών τα οποία βασίζονται σε εμπειρικές εκτιμήσεις της αμοιβαίας πληροφορίας μεταξύ κάθε μεταβλητής και της μεταβλητής στόχου. Τα κριτήρια συσχέτισης ως συνήθως μπορούν να ανιχνεύσουν τη γραμμική εξάρτηση μεταξύ μιας μεμονωμένης επεξηγηματικής μεταβλητής και της μεταβλητής στόχου. Ένας απλός τρόπος για την άρση αυτού του περιορισμού είναι να γίνει μία μη γραμμική προσαρμογή της μεταβλητής στόχου με τις επιμέρους επεξηγηματικές μεταβλητές και έπειτα αυτές να κατατάσσονται σύμφωνα με τον έλεγχο καλής προσαρμογής, με κίνδυνο πάντα υπερπροσαρμογής.

Από την άλλη μεριά, πολλές επιστημονικές προσεγγίσεις απεικονίζουν τη χρησιμότητα των μεθόδων επιλογής του καλύτερου υποσυνόλου εκείνων των μεταβλητών που όλες "μαζί" έχουν καλή προβλεπτική ικανότητα, σε αντίθεση με τις μεθόδους κατάταξης των μεταβλητών σύμφωνα με τη μεμονωμένη προβλεπτική ικανότητά τους [101]. Οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών χρησιμοποιούνται με ορισμένα κριτήρια τα οποία συνδυάζουν στατιστικά μέτρα με συναρτήσεις ποινής για την αύξηση του αριθμού των μεταβλητών πρόβλεψης στο μοντέλο. Τα πιο δημοφιλή από τα εν λόγω κριτήρια είναι τα κριτήρια πληροφορίας, και συγκεκριμένα τα κριτήρια AIC (Akaike Information Criteria) και BIC (Bayesian Information Criteria), όπως αυτά περιγράφηκαν στο Κεφάλαιο 6.

Ας θεωρήσουμε τώρα την περίπτωση που ο πειραματιστής πρέπει να αναλύσει μία μεγάλη βάση δεδομένων και αναμένει ότι λίγες από τις πολλές υποψήφιες επεξηγηματικές μεταβλητές επηρεάζουν σημαντικά την απόκριση. Σε μία τέτοια περίπτωση κλασικά κριτήρια, όπως το AIC ή το BIC συνήθως υπερεκτιμούν τον αριθμό των μεταβλητών παλινδρόμησης. Οι Bogdan et al. [19], [20] πρότειναν μία τροποποιημένη μορφή του κριτηρίου BIC (modified BIC-mBIC) η οποία επιτρέπει την ενσωμάτωση εκ των προτέρων γνώσης για έναν αριθμό μεταβλητών παλινδρόμησης, και με αυτό τον τρόπο προλαμβάνει το προαναφερθέν φαινόμενο της υπερεκτίμησης. Ειδικότερα, το mBIC προσαρμόζεται χρησιμοποιώντας τη διωνυμική prior και συνιστά την επιλογή του μοντέλου που ελαχιστοποιεί την ποσότητα

$$mBIC = -2l + q\log(n) + 2q\log\left(\frac{1-p}{p}\right),$$

όπου p είναι η εκ των προτέρων πιθανότητα μία τυχαία επιλεγμένη μεταβλητή παλινδρόμησης να επηρεάζει τη μεταβλητή απόκρισης, l είναι η λογαριθμοπιθανοφάνεια του μοντέλου, q είναι ο αριθμός των παραμέτρων του μοντέλου και n είναι ο αριθμός των παρατηρήσεων.

Οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών επιλέγουν τα υποσύνολα εκείνων των μεταβλητών που όλες "μαζί" έχουν καλή προβλεπτική ικανότητα. Ωστόσο, μία μεταβλητή που δεν είναι χρήσιμη όταν λαμβάνεται μαζί με άλλες μεταβλητές, μπορεί να είναι χρήσιμη από μόνη της. Το γεγονός αυτό μας παρακίνησε να προτείνουμε μία μέθοδο η οποία επεκτείνει την ιδέα επιλογής του καλύτερου υποσυνόλου μεταβλητών σε προβλήματα δυαδικής ταξινόμησης ή πρόβλεψης, λαμβάνοντας υπόψη τη μεμονωμένη προβλεπτική δύναμη της κάθε μεταβλητής. Για το σκοπό αυτό χρησιμοποιούμε ως κριτήριο την απόδοση ενός ταξινομητή που λαμβάνει υπόψη του μία και μόνο μεταβλητή κάθε φορά, στα πλαίσια της ανάλυσης καμπύλων λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic-ROC). Συγκεκριμένα, το κριτήριο που χρησιμοποιούμε είναι το εμβαδόν κάτω από την ROC καμπύλη

(Area Under the ROC Curve-AUC), με το οποίο εκτιμάμε την πιθανότητα ένας ταξινομητής να κατατάξει ένα τυχαία επιλεγμένο θετικό υπόδειγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό υπόδειγμα. Αυτή η πιθανότητα εκτιμήθηκε από το σύνολο δεδομένων δοκιμής. Η καινοτομία της προτεινόμενης μεθόδου έγκειται στο γεγονός ότι συνδυάζουμε μία μη παραμετρική τεχνική (μετρική AUC) με μία μέθοδο βασισμένη στην πιθανοφάνεια (ϵBIC προτεινόμενο κριτήριο πληροφορίας). Σύντομη περιγραφή της στατιστικής συμπερασματολογίας γύρω από την ανάλυση ROC καμπύλων καθώς και τη μετρική AUC παρουσιάστηκε στο Κεφάλαιο 3 στην ενότητα 3.2.2 της παρούσας διατριβής. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει εναλλακτικά για περισσότερες λεπτομέρειες στις εργασίες [105], [25], [179], [180] και [181].

8.2 Η Προτεινόμενη Μέθοδος Επιλογής του Καλύτερου Υποσυνόλου Μεταβλητών

Σε αυτή την ενότητα παρουσιάζεται αναλυτικά η προτεινόμενη μέθοδος επιλογής του καλύτερου υποσυνόλου μεταβλητών, η οποία βασίζεται στη μετρική AUC καθώς επίσης και στο νέο κριτήριο πληροφορίας ϵBIC, και εκτελεί τα ακόλουθα βήματα:

Βήμα 1. Διαχωρίζουμε τα δεδομένα σε σύνολο εκπαίδευσης και δοκιμής. Δοθέντος ενός $n \times m$ πίνακα μοντέλου $X = [x_1, x_2, \ldots, x_m]$, όπου x_j, $j = 1, 2, \ldots, m$, είναι η j-οστή στήλη του πίνακα, καθώς και ενός $n \times 1$ διανύσματος Y, το οποίο είναι το δίτιμο διάνυσμα απόκρισης, υπολογίζουμε το μέτρο AUC κάθε μεταβλητής x_j σε σχέση με το Y.

Βήμα 2. Ταξινομούμε σε φθίνουσα σειρά τις τιμές του μέτρου AUC που προέκυψαν για κάθε μία μεταβλητή και τις τοποθετούμε σε ένα διάνυσμα. Το διάνυσμα των τιμών AUC είναι $AUC = (AUC_1, AUC_2, \ldots, AUC_m)$, όπου AUC_j, για $j = 1, 2, \ldots, m$, είναι η τιμή του μέτρου AUC που αντιστοιχεί στην j-οστή μεταβλητή.

Βήμα 3. Κρατάμε τις μεταβλητές του μοντέλου που βρέθηκαν να έχουν αντίστοιχες τιμές AUC_j μεγαλύτερες από την τιμή κατωφλιού θ=0.5. Αυτές οι μεταβλητές μαζί συνιστούν ένα μοντέλο καλύτερο από ένα μοντέλο τυχαίας πρόβλεψης (random guess model). Η τιμή κατωφλιού θ=0.5 καθορίζεται σε σχέση με τη μεταβλητή απόκρισης, δηλαδή, στο ενδιάμεσο σημείο μεταξύ του κέντρου βάρους των δύο κλάσεων (y=0 ή y=1).

Βήμα 4. Υπολογίζουμε την εκ των προτέρων πιθανότητα μία τυχαία επιλεγμένη μεταβλητή παλινδρόμησης να επηρεάζει τη μεταβλητή απόκρισης Y ως $p = \frac{r}{q}$, όπου r είναι ο αριθμός των σημαντικών μεταβλητών οι οποίες ανιχνεύτηκαν ακολουθώντας τα προηγούμενα βήματα και το q είναι ο αριθμός των παραμέτρων του μοντέλου, δηλαδή, x_1,...,x_m.

Βήμα 5. Τέλος, το ϵBIC συνιστά την επιλογή του μοντέλου που ελαχιστοποιεί την ποσότητα

$$\epsilon BIC = -2l + q\log(n) + 2\epsilon q \log\left(\frac{1-p}{p}\right),$$

για κάποιο προκαθορισμένο $0 < \epsilon < 1$, όπου l είναι η λογαριθμοπιθανοφάνεια του μοντέλου και n είναι ο αριθμός των παρατηρήσεων.

Υπάρχουν δύο πράγματα που πρέπει να σημειώσουμε σε αυτό το σημείο. Πρώτα από όλα, ο διαχωρισμός των δεδομένων σε σύνολο εκπαίδευσης και δοκιμής διαδραματίζει σημαντικό ρόλο για την κατασκευή και αξιολόγηση των μοντέλων. Το σύνολο δεδομένων εκπαίδευσης χρησιμοποιείται για την κατασκευή του προγνωστικού μοντέλου. Από την άλλη μεριά

το σύνολο δοκιμής είναι ένα σύνολο δεδομένων που είναι ανεξάρτητο από τα δεδομένα εκπαίδευσης, αλλά ακολουθεί την ίδια κατανομή πιθανότητας με τα δεδομένα εκπαίδευσης, και χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου. Εάν ένα μοντέλο προσαρμόζεται το ίδιο καλά και στα δύο αυτά σύνολα, ελάχιστη υπερπροσαρμογή (overfitting) έχει λάβει χώρα. Εάν το μοντέλο που προκύπτει από τα δεδομένα εκπαίδευσης είναι πολύ καλύτερο από αυτό του συνόλου δοκιμής, η υπερπροσαρμογή είναι η πιθανή αιτία. Συνήθως, όταν ένα σύνολο δεδομένων διαχωρίζεται σε σύνολο εκπαίδευσης και δοκιμής, τα περισσότερα από τα δεδομένα χρησιμοποιούνται για την εκπαίδευση (75%), και ένα μικρότερο τμήμα των δεδομένων χρησιμοποιείται για τη δοκιμή (25%). Αυτή την αναλογία θεωρήσαμε στη μελέτη μας δεδομένου ότι χρησιμοποιείται συχνά στην εξόρυξη δεδομένων, παρόλο που η αναλογία αυτή μπορεί να ποικίλει ανάλογα με τις απαιτήσεις μιας πειραματικής μελέτης.

Επίσης, ας σημειωθεί εδώ ότι η τιμή κατωφλιού θ, η οποία καθορίζει τον αριθμό των σημαντικών παραγόντων διαδραματίζει επίσης σημαντικό ρόλο. Η απόφαση σχετικά με το θ γίνεται σύμφωνα με το γεγονός ότι το θ ποικίλλει πάνω στις πιθανές τιμές μιας μεταβλητής, και για να προσφέρει χρήσιμη πληροφορία, ολόκληρη η καμπύλη ROC θα πρέπει να βρίσκεται πάνω από τη 45° γραμμή όπου ισχύει ότι *ευαισθησία*(θ) = 1- *ειδικότητα*(θ) [65]. Η προεπιλεγμένη τιμή κατωφλιού σε προβλήματα δυαδικής ταξινόμησης είναι το 0.5. Σε κάποιες πρόσφατες εργασίες, για παράδειγμα στις [139] και [199], συζητείται το γεγονός ότι το 0.5 μπορεί να μην είναι η βέλτιστη τιμή κατωφλιού σε προβλήματα ταξινόμησης, και το γεγονός ότι το AUC αγνοεί αυτό το κατώφλι απεικονίζει την ύπαρξη μεροληψίας στην αποκλειστική χρήση του μέτρου AUC για την επιλογή μοντέλου. Ωστόσο, το ενδιαφέρον μας εδώ είναι καθαρά η σύγκριση της απόδοσης των συνόλων εκπαίδευσης και δοκιμής. Το κατώφλι αυτό αποτελεί ένα σημείο που επιλέγεται βάση της καμπύλης ROC του συνόλου δεδομένων εκπαίδευσης. Εάν αυτό το κατώφλι παραμένει η καλύτερη δυνατή επιλογή στο σύνολο δεδομένων δοκιμής, η AUC παραμένει αμερόληπτη. Έτσι, ακόμη και αν αυτή η επιλογή του κατωφλιού μπορεί να μην είναι η βέλτιστη για την ταξινόμηση, δεν επηρεάζει την εγκυρότητα των συγκρίσεών μας, εφόσον η ίδια τιμή κατωφλιού χρησιμοποιείται για την αξιολόγηση των συνόλων εκπαίδευσης και δοκιμής, με την έννοια ότι κρατάμε τους παράγοντες του μοντέλου που έχουν αντίστοιχες τιμές για το AUC μεγαλύτερες από 0.5 τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο δοκιμής. Υπό αυτή την έννοια, το κατώφλι κατάταξης θ στη μελέτη μας αποτελεί μία από τις παραμέτρους καθορισμού του μοντέλου και όχι μία ξεχωριστή παράμετρο που πρέπει να βελτιστοποιηθεί. Έτσι, υποθέτουμε εδώ ίσες πιθανότητες για τις δύο κλάσεις (y=0 ή y=1) και ίσα κόστη εσφαλμένης κατάταξης. Ως εκ τούτου, η τιμή κατωφλιού θ ορίζεται ίση με 0.5.

8.3 Συγκριτική Μελέτη

Σε αυτή την ενότητα συγκρίνουμε την απόδοση τεσσάρων μεθόδων επιλογής του καλύτερου υποσυνόλου μεταβλητών με κριτήρια πληροφορίας τα AIC, BIC, mBIC και το προτεινόμενο ϵBIC, μέσω μιας εμπειρικής μελέτης προσομοίωσης και μιας πραγματικής ανάλυσης δεδομένων.

8.3.1 Κριτήρια Αξιολόγησης

Σε έναν ξεχωριστό έλεγχο υποθέσεων συνίσταται ο έλεγχος της μηδενικής υπόθεσης $H_0 : \beta_j = 0$ έναντι της εναλλακτικής $H_1 : \beta_j \neq 0$, για $j = 1, 2, \ldots, m$. Ένα σφάλμα Τύπου I προκύπτει όταν δηλώνουμε έναν ανενεργό παράγοντα ως ενεργό, αλλά ισχύει η H_0. Ένα σφάλμα Τύπου II προκύπτει όταν δηλώνουμε έναν ενεργό παράγοντα ως ανενεργό, αλλά ισχύει η H_1. Πρόσφατα, έχουμε δει μια σημαντική αύξηση του μεγέθους των διαθέσιμων δεδομένων. Είναι πλέον συχνή πρόκληση για ένα στατιστικολόγο, να βρει όσο το δυνατόν περισσότερα σημαντικά χαρακτηριστικά σε ένα σύνολο δεδομένων, αντί να διενεργήσει έναν ξεχωριστό έλεγχο υποθέσεων για ένα και μόνο χαρακτηριστικό. Στη μελέτη προσομοίωσης που

ακολουθεί, ασχολούμαστε με ένα πρόβλημα πολλαπλών δοκιμών και διενεργούμε πολλαπλούς ελέγχους υποθέσεων, δεδομένου ότι εκατοντάδες παραμέτρων εξετάζονται ταυτόχρονα (1000 επαναλήψεις). Έτσι, υπολογίζουμε το μέσο ποσοστό σφάλματος Τύπου I (απορρίπτουμε την πραγματική μηδενική υπόθεση) και το μέσο ποσοστό σφάλματος Τύπου II (αποδεχόμαστε την ψευδή μηδενική υπόθεση) σε πάνω από 1000 επαναλήψεις. Εδώ, διεξάγουμε τον έλεγχο της μηδενικής υπόθεσης $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$ (καμία από τις ανεξάρτητες μεταβλητές δεν έχουν προβλεπτική ικανότητα) έναντι της εναλλακτικής υπόθεσης $H_1 : \beta_1 \neq 0$ ή $\beta_2 \neq 0$ ή $\ldots \beta_m \neq 0$.

Όσον αφορά την πραγματική ανάλυση δεδομένων, χρησιμοποιούμε ένα μέτρο σφάλματος πολλαπλών δοκιμών, δηλαδή, το ποσοστό εσφαλμένων ευρημάτων (false discovery rate-FDR). Το FDR είναι ιδιαίτερα κατάλληλο για τις αναλύσεις που βασίζονται σε πραγματικά δεδομένα στα οποία ο πειραματιστής ενδιαφέρεται για την εξεύρεση αρκετών σημαντικών αποτελεσμάτων μεταξύ πολλών δοκιμών. Οι Benjamini και Hochberg [12] προτείνουν ότι το FDR είναι το καταλληλότερο μέτρο ελέγχου του ποσοστού σφάλματος σε πολλά εφαρμοσμένα προβλήματα πολλαπλών δοκιμών και πολλαπλού ελέγχου υποθέσεων, και ότι ο πειραματιστής θα πρέπει να στοχεύει στο να διατηρεί το FDR σε ένα επιθυμητό επίπεδο, διατηρώντας παράλληλα την ισχύ της κάθε δοκιμής όσο το δυνατόν μεγαλύτερη. Αρχικά, υπολογίζουμε τον αριθμό των εσφαλμένων θετικών (FP), δηλαδή, τον αριθμό των επιλεγμένων μεταβλητών που δεν εμφανίζονται στο πραγματικό μοντέλο, και τον αριθμό των λανθασμένων αρνητικών (FN), δηλαδή, τον αριθμό των πραγματικών μεταβλητών παλινδρόμησης που δεν εντοπίστηκαν. Αυτοί οι αριθμοί (TP, FP) στη συνέχεια χρησιμοποιούνται για τον υπολογισμό των τιμών στα ακόλουθα κριτήρια αξιολόγησης.

1. Η ισχύς (power=1-Type II error), η οποία ορίζεται ως $\frac{(k-\mathrm{FN})}{k}$, όπου k συμβολίζει τον αριθμό των επεξηγηματικών μεταβλητών με μη μηδενικούς συντελεστές παλινδρόμησης. Η ισχύς δεν προσδιορίζεται για $k = 0$, δεδομένου ότι οι περιπτώσεις για τις οποίες το $k = 0$ εξαιρούνται από την ανάλυση.

2. Το ποσοστό εσφαλμένων ευρημάτων FDR, το οποίο ορίζεται ως $\mathrm{FDR} = \frac{\mathrm{FP}}{(\mathrm{FP}+k-\mathrm{FN})}$. Εάν δεν υπάρχουν ευρήματα, τότε $\mathrm{FP} + k - \mathrm{FN} = 0$, αποδίδοντας $\mathrm{FDR} = 0$.

3. Ο αριθμός των εσφαλμένα ταξινομημένων μεταβλητών (misclassified regressors), που ορίζεται ως $\mathrm{MR} = \mathrm{FP} + \mathrm{FN}$.

4. Το μέσο σφάλμα πρόβλεψης d το οποίο ορίζεται ως $d = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|$, όπου n είναι ο αριθμός των παρατηρήσεων. Το σφάλμα πρόβλεψης υπολογίζει τη μέση απόλυτη απόκλιση μεταξύ των παρατηρήσεων και των προβλεπόμενων τιμών τους. Το μέσο σφάλμα πρόβλεψης d υπολογίζεται και για τα δύο σύνολα δεδομένων (εκπαίδευσης και δοκιμής).

8.3.2 Μοτίβο Προσομοιώσεων

Για τα πειράματα προσομοίωσής μας, αναπτύξαμε λογιστικά μοντέλα παλινδρόμησης που περιλαμβάνουν μόνο κύριες επιδράσεις αποτελούμενα από $n = 100$ πειραματικές εκτελέσεις και $m = 100$ επεξηγηματικές μεταβλητές με τους συντελεστές παλινδρόμησης να λαμβάνουν τυχαίες τιμές από το διάνυσμα β. Αξίζει να σημειωθεί, ότι ψάχνοντας το χώρο των πιθανών μοντέλων, στην περίπτωση που μία μεταβλητή δεν βελτιώνει την κατάλληλη ποινικοποιημένη πιθανοφάνεια διαγράφεται ως μη σημαντική, εκτιμώντας το συντελεστή παλινδρόμησής της β να ισούται με το 0. Από τώρα και στο εξής, υποθέτουμε ότι ο πίνακας σχεδιασμού $X = (x_{ij})$ κανονικοποιείται κατά τέτοιο τρόπο έτσι ώστε κάθε στήλη να έχει μέση τιμή 0 και διασπορά 1. Για κάθε μία από τις 1000 επαναλήψεις που εκτελέσαμε στα πειράματα προσομοίωσης, χρησιμοποιήσαμε το ακόλουθο πρωτόκολλο προσομοίωσης.

Για να καθορίσουμε τον αριθμό k των ενεργών παραγόντων-στηλών, χρησιμοποιήσαμε την ακόλουθη διαδικασία. Οι πραγματικά ενεργοί παράγοντες επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων και μόνο οι κύριες επιδράσεις ελήφθησαν υπόψη. Για κάθε πείραμα, εκτιμήσαμε όλες τις κύριες επιδράσεις, οι οποίες στη συνέχεια επισημάνθηκαν ως ενεργές ή μη ενεργές. Έπειτα υπολογίζοντας το ποσοστό των εκτιμώμενων ενεργών επιδράσεων που εμφανίζονται σε κάθε πείραμα, εκτιμήσαμε τον αριθμό k των ενεργών παραγόντων που μπορεί να αναγνωριστεί. Σημειώνουμε εδώ ότι για το Σενάριο I ($0 < p < 0.5$) εκτιμήσαμε ότι ο αριθμός k των ενεργών παραγόντων που μπορεί να αναγνωριστεί δεν υπερβαίνει το $\frac{m}{2}$, όπου m είναι ο αριθμός των στηλών του πίνακα σχεδιασμού. Για τους συντελεστές παλινδρόμησης β των ενεργών παραγόντων, ένα δείγμα μεγέθους k σχηματίστηκε από την κανονική κατανομή $N(4, 0.2)$, και τα πρόσημα ± 1 κατανεμήθηκαν τυχαία σε κάθε αριθμό. Για τους $(m-k)$ μη ενεργούς παράγοντες στο πραγματικό μοντέλο, οι συντελεστές παλινδρόμησής τους β ελήφθησαν τυχαία από την κανονική κατανομή $N(0, 0.2)$. Ο πίνακας σχεδιασμού είναι ο X=[x_1, x_2,..., x_{100}] όπου οι μεταβλητές x_1,...,x_{100} αποτελούν μία μίξη από διχοτομικές, κατηγορικές και συνεχείς μεταβλητές, οι οποίες λαμβάνουν τυχαία διαφορετικές τιμές όπως αυτές περιγράφονται στον Πίνακα 8.1.

Πίνακας 8.1: Η τυχαία επιλεγμένη δομή για τις x_j

x_j	Τιμές
x_j	~ N(10,5)
x_j	~ N(0,0.5)
x_j	~ U(a=5, b=15)
x_j	± 1 ίσα κατανεμημένα
x_j	± 1 τυχαία κατανεμημένα
x_j	παρήχθησαν τυχαία ώστε να έχουν 4 επίπεδα (0-3)

8.3.3 Η Παράμετρος ϵ

Τα αποτελέσματα της προτεινόμενης μεθόδου (ϵBIC) συγκρίνονται για αρκετά ϵ ($0 < \epsilon < 1$) με τις μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών με κριτήρια τα AIC, BIC και mBIC. Αξίζει να σημειωθεί ότι το προτεινόμενο κριτήριο ϵBIC είναι ισοδύναμο με το BIC για $\epsilon = 0$ και ισοδύναμο με το mBIC για $\epsilon = 1$.

Καθορίσαμε τη βέλτιστη τιμή της παραμέτρου ϵ εμπειρικά, μέσω προσομειώσεων. Η βέλτιστη τιμή για το ϵ επιλέγεται να είναι αυτή που επιτυγχάνει τις χαμηλότερες τιμές για τα σφάλματα Τύπου I και Τύπου II, ταυτόχρονα, για τον εξής βασικό λόγο. Τα ποσοστά σφάλματος Τύπου I και Τύπου II είναι εξίσου σημαντικά και πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα. Τα χαμηλά ποσοστά σφάλματος Τύπου I είναι σημαντικά, δεδομένου ότι η ικανότητα αποκλεισμού των περιττών παραγόντων μειώνει το κόστος των επιπλέον πειραμάτων και τα χαμηλά ποσοστά σφάλματος Τύπου II είναι ιδιαίτερα επιθυμητά δεδομένου ότι ο κύριος στόχος μας είναι να ανακαλύψουμε τους σημαντικούς παράγοντες που επηρεάζουν περισσότερο τη μεταβλητή απόκρισης.

Εξετάσαμε διάφορες τιμές για το ϵ (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95) με στόχο να παρέχουμε μια πιο γενική πρόταση σχετικά με το πως ο πειραματιστής θα επιλέξει τη βέλτιστη τιμή της παραμέτρου ϵ σύμφωνα με τις απαιτήσεις της μελέτης του, καθώς επίσης και για να καλύψουμε όλα τα πιθανά σενάρια, από τη στιγμή που το πρόβλημα επιλογής μεταβλητών είναι πολύπλευρο και εξυπηρετεί διαφορετικούς σκοπούς σε διαφορετικές περιπτώσεις. Έτσι, θεωρήσαμε δύο διαφορετικά σενάρια για την παράμετρο p, η οποία είναι η εκ των προτέρων πιθανότητα μία τυχαία επιλεγμένη μεταβλητή παλινδρόμησης να επηρεάζει τη μεταβλητή απόκρισης.

Σενάριο I : Εάν $0 < p < 0.5$, ο όρος ποινής $2\epsilon q \log\left(\frac{1-p}{p}\right)$ είναι θετικός. Σε αυτή την περίπτωση, όσο μεγαλύτερο είναι το ϵ τόσο αυξάνεται ο όρος ποινής.

Σενάριο ΙΙ : Εάν $0.5 < p < 1$, ο όρος ποινής $2\epsilon q \log\left(\frac{1-p}{p}\right)$ είναι αρνητικός. Σε αυτή την περίπτωση, όσο μεγαλύτερο είναι το ϵ τόσο μειώνεται ο όρος ποινής.

Αξίζει να σημειωθεί, ότι στην περίπτωση που το $p = 0.5$, το προτεινόμενο κριτήριο ϵBIC είναι ισοδύναμο με το BIC.

8.3.4 Αποτελέσματα Μελέτης Προσομοίωσης

Πραγματοποιήσαμε τις προσομοιώσεις 1000 φορές για κάθε ένα από τα δύο σενάρια, και τα ληφθέντα αποτελέσματα για το Σενάριο Ι και το Σενάριο ΙΙ συνοψίζονται στον Πίνακα 8.2 και στον Πίνακα 8.3, αντίστοιχα. Στους πίνακες αυτούς, η πρώτη στήλη αναφέρεται στο υπό εξέταση κριτήριο. Οι στήλες που ονομάζονται ως "Type I" και "Type II" αναφέρονται στις τιμές των σφαλμάτων Τύπου Ι και Τύπου ΙΙ, αντίστοιχα, κατά μέσο όρο σε 1000 επαναλήψεις για κάθε μέθοδο επιλογής του καλύτερου υποσυνόλου μεταβλητών.

Παρατηρούμε από τον Πίνακα 8.2 (Σενάριο Ι: $0 < p < 0.5$) ότι τα κριτήρια AIC και BIC έχουν μια ισχυρή τάση να υπερεκτιμούν τον αριθμό των μεταβλητών παλινδρόμησης, δεδομένου ότι το μέσο ποσοστό σφάλματος Τύπου Ι ισούται με 0.3275 για το AIC και 0.1650 για το BIC, αντίστοιχα. Το AIC εμφανίζει την υψηλότερη μέση πιθανότητα εμφάνισης σφάλματος Τύπου Ι, αφού το AIC τυπικά περιλαμβάνει περισσότερες μεταβλητές παλινδρόμησης (για μέγεθος δείγματος $n \geq 8$, η ποινή στη διάσταση του μοντέλου χρησιμοποιώντας το AIC είναι μικρότερη από την ποινή χρησιμοποιώντας το BIC). Το κριτήριο mBIC ελέγχει καλύτερα το συνολικό σφάλμα Τύπου Ι σε σύγκριση με το AIC και το BIC, μειώνοντας την τιμή του ποσοστού σφάλματος Τύπου Ι στο 0.14.

Παρατηρούμε επίσης από τον Πίνακα 8.2 (Σενάριο Ι: $0 < p < 0.5$) ότι το προτεινόμενο κριτήριο ϵBIC αποτρέπει την υπερεκτίμηση, και αποδίδει καλύτερα από το AIC και το BIC, δεδομένου ότι έχει μικρότερες τιμές για τα σφάλματα Τύπου Ι για όλα τα θεωρούμενα ϵ. Το προτεινόμενο κριτήριο ϵBIC επιτυγχάνει επίσης χαμηλότερες τιμές σφάλματος Τύπου ΙΙ για σχεδόν όλα τα θεωρούμενα ϵ σε σύγκριση με το mBIC, και κάτι τέτοιο αντιστοιχεί στις περιπτώσεις όπου η πλειοψηφία των ενεργών παραγόντων ανιχνεύονται σωστά. Παρατηρούμε επίσης ότι το ϵBIC επιτυγχάνει χαμηλότερες τιμές σφάλματος Τύπου Ι και Τύπου ΙΙ ταυτόχρονα σε σύγκριση με το mBIC για τιμή του $\epsilon = 0.85$.

Παρατηρούμε από τον Πίνακα 8.3 (Σενάριο ΙΙ: $0.5 < p < 1$) ότι η υψηλότερη μέση πιθανότητα εμφάνισης σφάλματος Τύπου Ι είναι για το AIC (0.3275), όπως αναμενόταν. Στην περίπτωση αυτή, το κριτήριο mBIC δεν προλαμβάνει το φαινόμενο της υπερεκτίμησης και το BIC (0.1625) αποδίδει καλύτερα από το mBIC (0.2000) παρουσιάζοντας χαμηλότερη τιμή σφάλματος Τύπου Ι. Το προτεινόμενο κριτήριο ϵBIC επιτυγχάνει χαμηλότερες τιμές σφάλματος Τύπου Ι για όλα τα θεωρούμενα ϵ σε σύγκριση με το AIC και το mBIC. Παρατηρούμε επίσης, ότι το προτεινόμενο κριτήριο ϵBIC επιτυγχάνει χαμηλότερες τιμές σφάλματος Τύπου ΙΙ για όλα τα θεωρούμενα ϵ σε σύγκριση με το BIC. Τέλος, το ϵBIC επιτυγχάνει επίσης χαμηλότερες τιμές σφάλματος Τύπου Ι και Τύπου ΙΙ ταυτόχρονα σε σύγκριση με το BIC για τιμή του $\epsilon = 0.1$.

Πίνακας 8.2: Απόδοση της μεθόδου με χρήση 1000 προσομοιώσεων για το Σενάριο I ($0 < p < 0.5$)

Κριτήριο	Type I	Type II
AIC	0.3275	0.0667
BIC	0.1650	0.1067
mBIC	0.1400	0.1150
ϵBIC (ϵ=0.1)	0.1550	0.1083
ϵBIC (ϵ=0.15)	0.1575	0.1000
ϵBIC (ϵ=0.2)	0.1500	0.1100
ϵBIC (ϵ=0.25)	0.1550	0.1000
ϵBIC (ϵ=0.3)	0.1500	0.1117
ϵBIC (ϵ=0.35)	0.1550	0.1017
ϵBIC (ϵ=0.4)	0.1450	0.1133
ϵBIC (ϵ=0.45)	0.1500	0.1050
ϵBIC (ϵ=0.5)	0.1450	0.1133
ϵBIC (ϵ=0.55)	0.1500	0.1067
ϵBIC (ϵ=0.6)	0.1425	0.1133
ϵBIC (ϵ=0.65)	0.1475	0.1067
ϵBIC (ϵ=0.7)	0.1425	0.1133
ϵBIC (ϵ=0.75)	0.1450	0.1134
ϵBIC (ϵ=0.8)	0.1400	0.1133
ϵBIC (ϵ=0.85)	0.1375	0.1134
ϵBIC (ϵ=0.9)	0.1400	0.1133
ϵBIC (ϵ=0.95)	0.1350	0.1150

Πίνακας 8.3: Απόδοση της μεθόδου με χρήση 1000 προσομοιώσεων για το Σενάριο II ($0.5 < p < 1$)

Κριτήριο	Type I	Type II
AIC	0.3275	0.0533
BIC	0.1625	0.0933
mBIC	0.2000	0.0817
ϵBIC (ϵ=0.1)	0.1625	0.0917
ϵBIC (ϵ=0.15)	0.1775	0.0917
ϵBIC (ϵ=0.2)	0.1800	0.0900
ϵBIC (ϵ=0.25)	0.1775	0.0833
ϵBIC (ϵ=0.3)	0.1850	0.0867
ϵBIC (ϵ=0.35)	0.1850	0.0833
ϵBIC (ϵ=0.4)	0.1850	0.0867
ϵBIC (ϵ=0.45)	0.1875	0.0917
ϵBIC (ϵ=0.5)	0.1850	0.0867
ϵBIC (ϵ=0.55)	0.1900	0.0917
ϵBIC (ϵ=0.6)	0.1850	0.0867
ϵBIC (ϵ=0.65)	0.1925	0.0900
ϵBIC (ϵ=0.7)	0.1850	0.0867
ϵBIC (ϵ=0.75)	0.1975	0.0900
ϵBIC (ϵ=0.8)	0.1925	0.0850
ϵBIC (ϵ=0.85)	0.1975	0.0850
ϵBIC (ϵ=0.9)	0.1950	0.0817
ϵBIC (ϵ=0.95)	0.1975	0.0817

Πίνακας 8.4: Χρόνοι εκτέλεσης (sec)

Μέθοδος	Best Subset (AIC)	Best Subset (BIC)	Best Subset (mBIC)
	$1.7784 * 10^4$	$1.8096 * 10^4$	$1.8174 * 10^4$

Πίνακας 8.5: Χρόνοι εκτέλεσης (sec)

ϵBIC	ϵ=0.1,0.15	ϵ=0.2,0.25	ϵ=0.3,0.35	ϵ=0.4,0.45	ϵ=0.5
	$3.6114 * 10^4$	$5.4366 * 10^4$	$7.2462 * 10^4$	$9.0871 * 10^4$	$10.8889 * 10^4$
ϵBIC	ϵ=0.55	ϵ=0.6,0.65	ϵ=0.7,0.75	ϵ=0.8,0.85	ϵ=0.9,0.95
	$10.8889 * 10^4$	$12.7063 * 10^4$	$14.5315 * 10^4$	$16.3489 * 10^4$	$18.1585 * 10^4$

Ο Πίνακας 8.4 εμφανίζει τους μέσους χρόνους εκτέλεσης (sec) για τις μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών με κριτήρια τα AIC, BIC και mBIC, και για τα δύο υπό εξέταση Σενάρια. Ο Πίνακας 8.5 παρουσιάζει τους μέσους χρόνους εκτέλεσης (sec) για την προτεινόμενη μέθοδο επιλογής του καλύτερου υποσυνόλου μεταβλητών με κριτήριο το ϵBIC (για πολλές τιμές του ϵ και για τα δύο υπό εξέταση Σενάρια).

Με λίγα λόγια, συμπεραίνουμε από την παραπάνω μελέτη προσομοίωσης, ότι η προτεινόμενη μέθοδος παρουσιάζει την τάση να δηλώνει σε ένα υψηλότερο ποσοστό ανενεργούς παράγοντες ως ενεργούς και σε πολύ χαμηλότερο ποσοστό ενεργούς παράγοντες ως ανενεργούς. Έτσι, η προτεινόμενη μέθοδος θα μπορούσε να χαρακτηριστεί ως συντηρητική υπό αυτή την έννοια. Δεδομένου ότι ο κύριος σκοπός της είναι να ανακαλύψει τους σημαντικούς παράγοντες που επηρεάζουν τη μεταβλητή απόκρισης και πρέπει να θεωρηθούν για περαιτέρω μελέτη, τα χαμηλά ποσοστά σφάλματος Τύπου ΙΙ είναι ιδιαίτερα επιθυμητά, παρότι και τα δύο ποσοστά σφάλματος (Τύπου Ι και Τύπου ΙΙ) είναι σημαντικά και πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα. Η προτεινόμενη μέθοδος κατορθώνει κάτι τέτοιο με επιτυχία. Το γεγονός ότι το ϵBIC καθυστερεί περισσότερο σε σύγκριση με τις άλλες μεθόδους που εκτελούν την επιλογή μεταβλητών σε παρόμοιους χρόνους, δεν επηρεάζει τις προαναφερθείσες επιθυμητές ιδιότητες πρόβλεψης της προτεινόμενης μεθόδου. Αξίζει να σημειωθεί εδώ ότι το κριτήριο mBIC ενσωματώνει την εκ των προτέρων κατανομή για τον αριθμό των μεταβλητών παλινδρόμησης, θεωρώντας τη διωνυμική prior, ενώ όταν δεν είναι διαθέσιμη εκ των προτέρων γνώση σχετικά με τον αριθμό των μεταβλητών παλινδρόμησης, οι Bogdan et al. [19] πρότειναν ως λύση την επιλογή μιας σταθεράς c κατά τέτοιο τρόπο ώστε το ποσοστό σφάλματος family-wise (FWER, η πιθανότητα ανίχνευσης τουλάχιστον ενός ψευδώς θετικού) για μέγεθος δείγματος $n \geq 200$ να διατηρείται σε επίπεδο κάτω από το 10%. Η τροποποιημένη αυτή εκδοχή του mBIC, δηλαδή το προτεινόμενο κριτήριο ϵBIC πλεονεκτεί στο γεγονός ότι μπορεί να χρησιμοποιηθεί σε οποιοδήποτε πείραμα οποιουδήποτε μεγέθους, επιτρέποντας την ενσωμάτωση της εκ των προτέρων γνώσης, υπολογίζοντας μέσω της μετρικής AUC την παράμετρο p, δηλαδή, την πιθανότητα μία τυχαία επιλεγμένη μεταβλητή παλινδρόμησης να επηρεάζει το Y.

Θα παρουσιάσουμε τώρα μια πιο γενική πρόταση σχετικά με ποιο από τα κριτήρια AIC, BIC, mBIC ή ϵBIC είναι καλύτερο, ανάλογα με την περίπτωση που εξετάζεται, από τη στιγμή που το πρόβλημα επιλογής μεταβλητών είναι πολύπλευρο και εξυπηρετεί διαφορετικούς σκοπούς σε διαφορετικές περιπτώσεις.

<u>Σενάριο Ι.</u> Εάν $0 < p < 0.5$, ο όρος ποινής $2\epsilon q \log\left(\frac{1-p}{p}\right)$ είναι θετικός. Σε αυτή την περίπτωση, όσο μεγαλύτερο είναι το ϵ τόσο αυξάνεται ο όρος ποινής. Ως εκ τούτου, σύμφωνα με τα αποτελέσματα του Πίνακα 8.2, προτείνουμε τη χρήση του κριτηρίου AIC, BIC ή ϵBIC για $0 < \epsilon < 0.5$ ως κριτήριο επιλογής μοντέλου, εάν η υπερπροσαρμογή είναι αποδεκτή ούτως ώστε να εξασφαλιστεί ότι μπορεί να συμπεριληφθούν όλες οι πιθανές σημαντικές μεταβλητές παλινδρόμησης. Ο πειραματιστής μπορεί να επιλέξει ποιο από τα κριτήρια AIC, BIC ή ϵBIC για $0 < \epsilon < 0.5$, είναι η καταλληλότερη επιλογή για τη μελέτη του, ανάλογα με το βαθμό της υπερπροσαρμογής που είναι αποδεκτός στα πλαίσια της μελέτης του. Από την άλλη μεριά, εάν το να συμπεριληφθεί μία επιπλέον μεταβλητή παλινδρόμησης στο μοντέλο δημιουργεί πρόβλημα, προτείνουμε τη χρήση του ϵBIC για $0.5 < \epsilon < 1$ ως κριτήριο επιλογής μοντέλου.

<u>Σενάριο ΙΙ.</u> Εάν $0.5 < p < 1$, ο όρος ποινής $2\epsilon q \log\left(\frac{1-p}{p}\right)$ είναι αρνητικός. Σε αυτή την περίπτωση, όσο μεγαλύτερο είναι το ϵ τόσο μειώνεται ο όρος ποινής. Ως εκ τούτου, σύμφωνα με τα αποτελέσματα του Πίνακα 8.3, προτείνουμε τη χρήση του κριτηρίου AIC, mBIC ή ϵBIC για $0.5 < \epsilon < 1$ ως κριτήριο επιλογής μοντέλου, εάν η υπερπροσαρμογή είναι αποδεκτή, έτσι ώστε να περιλαμβάνονται σίγουρα όλες οι σημαντικές μεταβλητές παλινδρόμησης. Ο πειραματιστής μπορεί να επιλέξει ποιο από τα κριτήρια AIC, mBIC ή ϵBIC για $0.5 < \epsilon < 1$, είναι η καταλληλότερη επιλογή για τη μελέτη του, ανάλογα με το βαθμό της υπερπροσαρμογής που είναι αποδεκτός στα πλαίσια της μελέτης του. Από την άλλη μεριά, εάν το να συμπεριληφθεί

μία επιπλέον μεταβλητή παλινδρόμησης στο μοντέλο δημιουργεί πρόβλημα, προτείνουμε τη χρήση του BIC ή του ϵBIC για $0 < \epsilon < 0.5$ ως κριτήριο επιλογής μοντέλου.

8.3.5 Ανάλυση Πραγματικών Δεδομένων

Η προτεινόμενη μέθοδος επιλογής του καλύτερου υποσυνόλου μεταβλητών (ϵBIC) εφαρμόζεται τώρα σε ένα πραγματικό ιατρικό σύνολο δεδομένων, και η απόδοσή της συγκρίνεται με κλασικές μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC, BIC και mBIC).

Τα πραγματικά δεδομένα προέρχονται από μία ετήσια έρευνα που διεξήχθη κατά την περίοδο 01/01/2005-31/12/2006 από την Ελληνική Εταιρεία Τραύματος και την Ομάδα Εντατικής Χειρουργικής, περιλαμβάνοντας 30 Γενικά Νοσοκομεία στην Ελλάδα, με βασικό σκοπό τη μελέτη των επιδράσεων διαφόρων προγνωστικών παραγόντων στη θεραπεία τραυματισμένων ατόμων. Η βάση δεδομένων αποτελείται από 92 επεξηγηματικές μεταβλητές που αποτελούν τους προγνωστικούς παράγοντες. Οι παράγοντες αυτοί που καταγράφηκαν, περιελάμβαναν δημοφραφικά στοιχεία των ασθενών, στοιχεία που αφορούν τις συνθήκες του ατυχήματος και πληροφορίες για την ενδονοσοκομειακή αντιμετώπιση των ασθενών. Επίσης, η βάση δεδομένων αποτελείται από 8862 ασθενείς, και για κάθε έναν από αυτούς καταγράφηκε η κατάστασή τους κατά την τελευταία παρακολούθηση, με τη δίτιμη μεταβλητή απόκρισης y να παίρνει μόνο δύο τιμές, με το 0 να συμβολίζει την "επιβίωση" (8413 ασθενείς) και το 1 την "αποβίωση" (449 ασθενείς), αντίστοιχα. Σύμφωνα με ιατρικές συμβουλές, όλοι οι προγνωστικοί παράγοντες θα πρέπει να τυγχάνουν ίσης μεταχείρισης κατά τη στατιστική ανάλυση και δεν υπάρχει κανένας παράγοντας για τον οποίο επιβάλλεται να διατηρείται πάντα στο μοντέλο. Το σύνολο δεδομένων Τραύματος που χρησιμοποιείται για την αξιολόγηση και τη σύγκριση των αποτελεσμάτων των τεσσάρων υπό εξέταση μεθόδων επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC, BIC, mBIC και το προτεινόμενο ϵBIC), παρουσιάζεται στον Πίνακα 8.6.

Το σύνολο των δεδομένων διαχωρίστηκε σε σύνολο εκπαίδευσης (75%) και σύνολο δοκιμής (25%), προκειμένου να αξιολογηθεί η απόδοση της πρόβλεψης πάνω σε νέα δεδομένα. Για τον παραπάνω διαχωρισμό, οι καταχωρήσεις της βάσης δεδομένων επιλέχθηκαν τυχαία για να δημιουργήσουν το σύνολο εκπαίδευσης και δοκιμής σύμφωνα με το προκαθορισμένο μέγεθός τους. Αρχικά, θεωρούμε έναν προγνωστικό παράγοντα τη φορά για να δούμε πόσο καλά κάθε προγνωστικός παράγοντας από μόνος του προβλέπει τη μεταβλητή-στόχο ($y = 0$ ή $y = 1$). Αυτό το πρώτο βήμα που προστίθεται στη διαδικασία ανάλυσης, είναι η κατάταξη των μεταβλητών σε σημαντικές και μη σημαντικές, το οποίο επιτρέπει να μειωθεί το αρχικό μέγεθος του συνόλου των μεταβλητών, δημιουργώντας ένα πιο εύχρηστο σύνολο χαρακτηριστικών για τη μοντελοποίηση. Οι μεταβλητές κατατάσσονται σύμφωνα με ένα προκαθορισμένο κριτήριο το οποίο εξαρτάται από τα επίπεδα μέτρησης των προγνωστικών παραγόντων. Μια κοινή τεχνική που χρησιμοποιείται στην εξόρυξη δεδομένων είναι η κατάταξη των μεταβλητών με βάση το μέτρο της σημαντικότητάς τους, το οποίο ορίζεται ως *(1-p)*, όπου p είναι το *p-value* ενός επιλεγμένου στατιστικού τεστ συσχέτισης μεταξύ της υπό εξέταση μεταβλητής και της μεταβλητής στόχο. Στη μελέτη μας, ορισμένοι προγνωστικοί παράγοντες είναι συνεχείς και μερικοί είναι κατηγορικοί. Το κριτήριο που χρησιμοποιείται για τις συνεχείς μεταβλητές είναι η τιμή *p-value* που προκύπτει από ένα one-way ANOVA F test, ενώ το κριτήριο για τις κατηγορικές μεταβλητές είναι η τιμή *p-value* που προκύπτει με βάση το Pearson chi-square test [177].

Τα *p-values* τα οποία προέκυψαν συγκρίνονται και ως εκ τούτου χρησιμοποιούνται για την κατάταξη των μεταβλητών. Στο αρχικό σύνολο των δεδομένων μας είχαμε 92 επεξηγηματικές μεταβλητές και εκτελώντας την παραπάνω διαδικασία επιλογής χαρακτηριστικών, εντοπίσαμε τις σημαντικότερες εξ'αυτών, για επίπεδο σημαντικότητας α=5%. Η διαδικασία επιλογής χαρακτηριστικών μας επιτρέπει να ελαχιστοποιηθεί το σύνολο των μεταβλητών παλινδρόμησης από τις 92 αρχικά διαθέσιμες σε 44 μεταβλητές. Ας σημειωθεί εδώ ότι σε διαφορετική περίπτωση, οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών δεν θα μπορούσαν

να εφαρμοστούν λόγω υπολογιστικής πολυπλοκότητας (n=8862). Ο Πίνακας 8.7 παρουσιάζει τις τιμές σημαντικότητας για αυτές τις 44 επιλεγμένες μεταβλητές.

Πίνακας 8.6: Μελέτη Τραύματος

Συνεχείς Μεταβλητές
$x1$: weight, kg
$x2$: age, years
$x3$: Glasgow Coma Score, score
$x4$: pulse, N/min
$x6$: systolic arterial blood pressure, mmHg
$x7$: diastolic arterial blood pressure, mmHg
$x8$: Hematocrit (Ht), %
$x9$: haemoglobin (Hb), g/dl
$x11$: white cell count, /ml
$x12$: platelet, /ml
$x14$: potassium, /ml
$x15$: glucose, mg %
$x16$: creatinine, mg %
$x17$: urea, mg %
$x18$: amylase, score
$x20$: Injury Severity Score, score
$x21$: Revised Trauma Score, score
Κατηγορικές Μεταβλητές
$x19$: evaluation of disability (0 = expected permanent big, 1 = expected permanent small, 2 = expected impermanent big, 3 = expected impermanent small, 4 = recovery)
$x23$: cause of injury (0 = fall, 1 = trochee accident, 2 = athletic, 3 = industrial, 4 = crime, 5 = other)
$x24$: means of transportation (0 = airplane, 1 = ambulance, 2 = car, 3 = on foot)
$x25$: Ambulance (0 = no, 1 = yes)
$x26$: hospital of records
$x27$: substructure of hospital (0 = orthopaedic, 1 = CT, 2 = vascular surgeon, 3 = neurosurgeon, 4 = Intensive Care Unit) $x28$: comorbidities (0 = no, 1 = yes)
$x31$: sex (0 = female, 1 = male)
$x35$: doctor's speciality (0 = angiochirurgeon, 1 = non specialist, 2 = general doctor, 3 = general surgeon, 4 = jawbonesurgeon, 5 = gynaecologist, 6 = thoraxsurgeon, 7 = neurosurgeon, 8 = orthopaedic, 9 = urologist, 10 = paediatrician, 11 = children surgeon, 12 = plastic surgeon)
$x36$: major doctor (0 = no, 1 = yes)
$x41$: dysphoria (0 = no, 1 = yes)
$x52$: collar (0 = no, 1 = yes)
$x55$: immobility of limbs (0 = no, 1 = yes)
$x56$: fluids (0 = no, 1 = yes)
$x64$: Radiograph E.R. (0 = no, 1 = yes)
$x66$: US (0 = no, 1 = yes)
$x67$: urea test (0 = no, 1 = yes)
$x71$: destination after the emergency room (0 = other hospital, 1 = clinic, 2 = unit of high care, 3 = intensive care unit I.C.U, 4 = operating room)
$x72$: surgical intervention (0 = no, 1 = yes)
$x86$: arrival at emergency room (0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 = 16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)
$x87$: exit from emergency room (0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 = 16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)
$x101$: head injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x102$: face injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x104$: breast injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x106$: spinal column injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x107$: upper limbs injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)
$x108$: lower limbs injury (0 = none, 1 =$AIS \leqslant 2$,2 =$AIS > 2$)

Πίνακας 8.7: Οι 44 σημαντικές επιλεγμένες μεταβλητές

Fields	$x71$	$x3$	$x21$	$x101$	$x20$	$x64$	$x56$	$x19$	$x8$	$x6$	$x26$
(*1-p*)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Fields	$x7$	$x35$	$x9$	$x36$	$x11$	$x66$	$x108$	$x72$	$x55$	$x27$	$x67$
(*1-p*)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Fields	$x41$	$x107$	$x52$	$x2$	$x16$	$x18$	$x104$	$x28$	$x12$	$x17$	$x31$
(*1-p*)	1.0	1.0	1.0	1.0	1.0	1.0	0.999	0.999	0.998	0.997	0.996
Fields	$x1$	$x4$	$x24$	$x15$	$x102$	$x106$	$x25$	$x86$	$x87$	$x14$	$x5$
(*1-p*)	0.994	0.988	0.988	0.988	0.985	0.978	0.965	0.964	0.962	0.961	0.952

Ακολουθώντας τα βήματα της προτεινόμενης μεθόδου, κατατάσσουμε τις μεταβλητές βάση της τιμής AUC που προέκυψε για κάθε μία. Κρατάμε εκείνες τις μεταβλητές του μοντέλου που έχουν τιμή AUC_j μεγαλύτερη από θ=0.5. Αυτές οι μεταβλητές δηλώνονται ως σημαντικές. Στη μελέτη μας, $r = 6$ μεταβλητές βρέθηκαν να έχουν $AUC_j > \theta = 0.5$, άρα το $p = \frac{r}{q} = \frac{6}{44}$, και θέτουμε το ϵ ίσο με 0.85 για το προτεινόμενο κριτήριο ϵBIC, δεδομένου ότι η τιμή αυτή βρέθηκε να είναι η βέλτιστη τιμή από τη μελέτη προσομοίωσης για το Σενάριο I ($0 < p < 0.5$). Τα αποτελέσματα της προτεινόμενης μεθόδου (ϵBIC) συγκρίνονται με αυτά των υπολοίπων μεθόδων επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC, BIC, mBIC). Οι εκτιμώμενοι συντελεστές παλινδρόμησης β και τα εκτιμώμενα τυπικά σφάλματα του επιλεγμένου βέλτιστου μοντέλου, για κάθε μία από τις τέσσερις υπό εξέταση μεθόδους, παρουσιάζονται στον Πίνακα 8.8. Οι μη σημαντικές μεταβλητές αποκλείονται από το μοντέλο, εκτιμώντας και συρρικνώνοντας τους συντελεστές τους ακριβώς στο μηδέν.

Πίνακας 8.8: Εκτιμώμενοι συντελεστές β και τυπικά σφάλματα (σε παρενθέσεις)

Μέθοδος	MLE	AIC	BIC	mBIC	ϵBIC
Intercept	-5.97 (0.21)	-5.97 (0.21)	-5.96 (0.21)	-5.96 (0.21)	-5.96 (0.21)
x_2	0.72 (0.09)	0.71 (0.09)	0.73 (0.09)	0.73 (0.09)	0.73 (0.09)
x_{11}	0.25 (0.07)	0.25 (0.07)	0.28 (0.07)	0.28 (0.07)	0.28 (0.07)
x_{16}	0.09 (0.05)	0.09 (0.05)	0 (-)	0 (-)	0 (-)
x_{20}	0.56 (0.05)	0.56 (0.05)	0.55 (0.05)	0.55 (0.05)	0.55 (0.05)
x_{23}	0.06 (0.10)	0 (-)	0 (-)	0 (-)	0 (-)
x_{25}	1.02 (0.15)	1.00 (0.14)	1.02 (0.14)	1.02 (0.14)	1.02 (0.14)
x_{27}	-0.16 (0.09)	-0.16 (0.09)	0 (-)	0 (-)	0 (-)
x_{71}	1.46 (0.07)	1.47 (0.07)	1.43 (0.07)	1.43 (0.07)	1.43 (0.07)
x_{101}	1.30 (0.08)	1.30 (0.08)	1.29 (0.08)	1.29 (0.08)	1.29 (0.08)

Ο Πίνακας 8.9 παρουσιάζει τους χρόνους εκτέλεσης (sec) για τις τέσσερις υπό εξέταση μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών.

Πίνακας 8.9: Χρόνοι εκτέλεσης (sec)

Μέθοδος	AIC	BIC	mBIC	ϵBIC
	$8.535 * 10^4$	$8.475 * 10^4$	$8.548 * 10^4$	$1.709 * 10^5$

Παρατηρούμε από τον Πίνακα 8.8 ότι τα κριτήρια BIC, mBIC και ϵBIC έχουν παρόμοια απόδοση και τις ίδιες ακριβώς τιμές για τα β και τα τυπικά σφάλματα. Το γεγονός αυτό δεν αποτελεί έκπληξη λαμβάνοντας υπόψη τις ιδιότητες του κριτηρίου BIC, δεδομένου ότι για καθορισμένο m ($m = 44$) και το n να αυξάνει ραγδαία ($n = 8862$) ή να τείνει στο άπειρο, τα κριτήρια mBIC και ϵBIC προσεγγίζουν ασυμπτωτικά το BIC. Το AIC αποτυγχάνει να αποκλείσει δύο περιττές μεταβλητές (x_{16}, x_{27}) σε σύγκριση με το BIC, mBIC και ϵBIC, όπως αναμενόταν, αφού το AIC σε πολλές πρακτικές εφαρμογές τυπικά περιλαμβάνει περισσότερες μεταβλητές παλινδρόμησης, όπως σημειώνεται και στην εργασία [20]. Η προτεινόμενη μέθοδος επιλογής του καλύτερου υποσυνόλου μεταβλητών βασισμένη στη μετρική AUC, επιτυγχάνει να προσδιορίσει το ίδιο υποσύνολο μεταβλητών (x_2, x_{11}, x_{25}, x_{20}, x_{71} και x_{101}) που επηρεάζουν σημαντικά την απόκριση σε σύγκριση με τα γνωστά κριτήρια BIC και mBIC.

Στη μελέτη τραύματος, το σύνολο εκπαίδευσης χρησιμοποιείται για να επιλέξει τις σημαντικές μεταβλητές παλινδρόμησης και να εκτιμήσει τους συντελεστές τους. Αφού προσδιοριστεί το μοντέλο με χρήση του συνόλου εκπαίδευσης, αξιολογούμε την απόδοση του μοντέλου κάνοντας προβλέψεις χρησιμοποιώντας το σύνολο δοκιμής. Επειδή τα δεδομένα στο σύνολο δοκιμής περιέχουν ήδη γνωστές τιμές για τη μεταβλητή απόκρισης που θέλουμε να προβλέψουμε, είναι εύκολο να εξετάσουμε αν οι προβλέψεις του μοντέλου είναι σωστές. Το σύνολο δοκιμής με τη σειρά του χρησιμοποιείται για να υπολογίσουμε τον αριθμό των FP και FN, και οι αριθμοί αυτοί χρησιμοποιούνται στη συνέχεια για να υπολογίσουμε τις τιμές των κριτηρίων απόδοσης (βλέπε ενότητα 8.3.1) κατά μέσο όρο για 1000 διαφορετικούς τυχαίους διαχωρισμούς. Τα ληφθέντα αποτελέσματα παρουσιάζονται στον Πίνακα 8.10.

Πίνακας 8.10: Αξιολόγηση της απόδοσης

Μέθοδος	AIC	BIC	mBIC	ϵBIC
FP	15.4	12.7	12.5	12.4
FN	1.92	3.45	2.95	2.48
Power	0.808	0.655	0.705	0.752
FDR	0.655	0.659	0.639	0.622
MR	17.32	16.15	15.45	14.88
d test	0.325	0.213	0.204	0.198
d training	0.432	0.397	0.381	0.376

Παρατηρούμε από τον Πίνακα 8.10, ότι τα περισσότερα από τα ευρήματα που ανιχνεύονται είναι ψευδώς θετικά για όλα τα κριτήρια πληροφορίας. Ο αριθμός των ψευδώς θετικών που ανιχνεύονται από το AIC είναι 15.4, και είναι σημαντικά υψηλότερος από τον αριθμό των ψευδώς θετικών που ανιχνεύονται από τα κριτήρια BIC, mBIC και ϵBIC τα οποία έχουν παρόμοια απόδοση. Η τάση του AIC να συμπεριλαμβάνει περισσότερα ψευδώς θετικά αντικατοπτρίζεται επίσης στην υψηλότερη τιμή του MR. Ας σημειωθεί εδώ ότι αν το κόστος ενός ψευδώς θετικού είναι το ίδιο με το κόστος ενός ψευδώς αρνητικού, τότε το MR είναι ανάλογο προς το συνολικό κόστος του πειράματος. Αυτή η ισχυρή τάση έχει σχετικά μικρή επιρροή στο σφάλμα πρόβλεψης χρησιμοποιώντας το κριτήριο AIC. Οι χαμηλές τιμές του σφάλματος πρόβλεψης d και για τα δύο σύνολα (εκπαίδευσης και δοκιμής) για το κριτήριο AIC, υποδεικνύουν ότι για μεγάλα μεγέθη δείγματος, η υπερεκτίμηση του αριθμού των μεταβλητών παλινδρόμησης δεν επιδεινώνει σημαντικά τις προβλεπτικές ιδιότητες του κριτηρίου AIC. Επιπλέον, το FDR της προτεινόμενης μεθόδου είναι ίσο με 0.622, μικρότερο από την αντίστοιχη τιμή των κριτηρίων AIC, BIC και mBIC. Η προτεινόμενη μέθοδος φαίνεται να αποδίδει καλύτερα από τις μεθόδους επιλογής μεταβλητών με τα κριτήρια AIC, BIC και mBIC, δεδομένου ότι έχει επίσης μικρότερο ποσοστό εσφαλμένης ταξινόμησης και μικρότερο σφάλμα πρόβλεψης τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο δοκιμής. Επιπλέον, η προτεινόμενη μέθοδος έχει την υψηλότερη τιμή ισχύος σε σύγκριση με τα κριτήρια BIC και mBIC.

Συμπερασματικά, η μελέτη τραύματος απεικονίζει τα πλεονεκτήματα της προτεινόμενης μεθόδου επιλογής του καλύτερου υποσυνόλου μεταβλητών. Παρά το γεγονός ότι έπρεπε να διαχειριστούμε ένα υψηλής διάστασης σύνολο δεδομένων, η προτεινόμενη μέθοδος επιτυγχάνει τον κύριο στόχο της, δηλαδή, να προσδιορίσει ένα μικρό υποσύνολο μεταβλητών που είναι επαρκές για τη μοντελοποίηση. Περαιτέρω, το προγνωστικό μοντέλο το οποίο προέκυψε περιλαμβάνει μόνο τα στοιχεία ύψιστης σημασίας για την πρόβλεψη της έκβασης των τραυματισμένων ασθενών. Ως αποτέλεσμα, το απλούστερο αυτό προγνωστικό μοντέλο το οποίο προέκυψε, είναι πλέον εύχρηστο και παρέχει συγκεκριμένες πληροφορίες που μπορούν να χρησιμοποιηθούν ως κατευθυντήριες γραμμές για τη διαχείριση των τραυματισμένων ασθενών, καθώς επίσης μπορεί να βοηθήσει την ιατρική κοινότητα να επικεντρωθεί ως επί το πλείστον σε αυτά τα χαρακτηριστικά που εντοπίστηκαν να σχετίζονται περισσότερο με την πρόβλεψη της έκβασης ενός ασθενή.

8.4 Συμπεράσματα

Η επιλογή του καλύτερου υποσυνόλου μεταβλητών παραμένει πρόκληση για έναν πειραματιστή και ένας πολλά υποσχόμενος τομέας της έρευνας. Στο κεφάλαιο αυτό, προτείναμε μία μέθοδο επιλογής του καλύτερου υποσυνόλου μεταβλητών, η οποία επιτυγχάνει να αποκλείσει τις περιττές μεταβλητές και να εντοπίσει μόνο τα στοιχεία ύψιστης σημαντικότητας. Ένα μειονέκτημα της προτεινόμενης μεθόδου είναι ο υπολογιστικός χρόνος που απαιτεί. Βέβαια, σε γενικές γραμμές, οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών είναι ιδιαίτερα χρονοβόρες, επειδή η εύρεση του καλύτερου υποσυνόλου μεταβλητών αποτελεί μία εξαντλητική αναζήτηση, δεδομένου ότι ερευνώνται όλα τα δυνατά υποσύνολα των μεταβλητών με στόχο να βρεθεί το βέλτιστο από αυτά. Αυτή η υπολογιστική δυσκολία αποθαρρύνει

την ευρεία χρήση των μεθόδων αυτών όταν υπάρχει ένας μεγάλος αριθμός προγνωστικών παραγόντων σε πρακτικά προβλήματα. Ωστόσο, η προτεινόμενη μέθοδος απέδωσε πολύ ικανοποιητικά αποτελέσματα στο πρόβλημα στατιστικής μοντελοποίησης υψηλής διάστασης, προσδιορίζοντας τους σημαντικούς προγνωστικούς παράγοντες που επηρεάζουν την έκβαση ενός τραυματισμένου ασθενή. Επιπλέον, η πραγματική ανάλυση των δεδομένων έδειξε ότι η προτεινόμενη μέθοδος έχει πολύ καλές ιδιότητες, δεδομένου ότι έχει μικρότερο ποσοστό εσφαλμένης ταξινόμησης, μικρότερο ποσοστό εσφαλμένων ευρημάτων, μικρότερο σφάλμα πρόβλεψης και υψηλότερη τιμή ισχύος σε σύγκριση με τις υπόλοιπες μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών που εξετάστηκαν. Η εκτενής μελέτη προσομοίωσης έδειξε ότι η προτεινόμενη μέθοδος επιτυγχάνει πολύ χαμηλές τιμές για το ποσοστό σφάλματος Τύπου ΙΙ, γεγονός το οποίο είναι ιδιαίτερα σημαντικό προκειμένου να μην παραλείπονται σημαντικοί παράγοντες του μοντέλου. Επιπλέον, το ποσοστό σφάλματος Τύπου Ι επίσης διατηρήθηκε σε χαμηλά επίπεδα για πολλές τιμές της παραμέτρου ϵ. Τέλος, η καινοτομία της προτεινόμενης μεθόδου επιλογής του καλύτερου υποσυνόλου μεταβλητών, έγκειται στο γεγονός ότι συνδυάζουμε μία μη παραμετρική τεχνική (μετρική AUC) με μία μέθοδο βασισμένη στην πιθανοφάνεια (ϵBIC), μία διαδικασία επιλογής μεταβλητών που δεν συνηθίζεται έως τώρα στη διεθνή βιβλιογραφία από όσο είμαστε σε θέση να γνωρίζουμε.

ΚΕΦΑΛΑΙΟ **9**

Επιλογή Μεταβλητών σε Προβλήματα Υψηλής Διάστασης: Μοντελοποίηση Μέσω Υπερκορεσμένων Σχεδιασμών

Observations always involve theory.

—*Edwin P. Hubble (1889–1953)*

Στο ένατο αυτό κεφάλαιο μελετάμε το πρόβλημα της επιλογής μεταβλητών σε προβλήματα παλινδρόμησης στην περίπτωση που οι παρατηρήσεις (εισόδου) και οι ετικέτες των κλάσεων (εξόδου) ενός πραγματικού μεγάλου συνόλου δεδομένων είναι διαθέσιμες. Αρχικά, εφαρμόζουμε και εξετάζουμε την απόδοση διάφορων στατιστικών μεθόδων που έχουν αναπτυχθεί κατά καιρούς στη βιβλιογραφία, με στόχο την ανάλυση ενός μεγάλου συνόλου δεδομένων (δεδομένα Τραύματος) που αποτελείται από 7000 παρατηρήσεις και 70 παράγοντες κινδύνου. Οι στατιστικές μέθοδοι οι οποίες εφαρμόζονται είναι οι διαδικασίες μη κοίλης ποινικοποιημένης πιθανοφάνειας (SCAD, LASSO και Hard), η λογιστική παλινδρόμηση, και τεχνικές επιλογής του καλύτερου υποσυνόλου μεταβλητών (με κριτήρια πληροφορίας τα AIC και BIC), οι οποίες χρησιμοποιούνται για την ανίχνευση πιθανών παραγόντων κινδύνου θανάτου. Σε αυτό το κεφάλαιο, παρουσιάζουμε επίσης μια νέα προσέγγιση επιλογής μεταβλητών εμπνευσμένη από τους υπερκορεσμένους σχεδιασμούς, δοθέντος ενός μεγάλου συνόλου δεδομένων παρατηρήσεων. Τα πλεονεκτήματα και η αποτελεσματικότητα αυτής της εναλλακτικής προσέγγισης για την αναγνώριση σημαντικών μεταβλητών σε μελέτες παρατήρησης, αξιολογείται θεωρώντας αρκετούς υπερκορεσμένους σχεδιασμούς δύο επιπέδων και μια ποικιλία από διαφορετικά στατιστικά μοντέλα σε σχέση με τους συνδυασμούς των παραγόντων και τον αριθμό των παρατηρήσεων. Τα ληφθέντα αποτελέσματα σε σύγκριση με τις υπόλοιπες μεθόδους ανάλυσης είναι ιδιαίτερα ενθαρρυντικά, δεδομένου ότι η μοντελοποίηση μέσω υπερκορεσμένων σχεδιασμών παρέχει συγκεκριμένες πληροφορίες που είναι λογικές και σύμφωνες με την ιατρική εμπειρία, και μπορεί επίσης να παρέχει κατευθυντήριες γραμμές για τη διαχείριση του τραύματος στην ιατρική κοινότητα. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [176].

9.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Η επιλογή μεταβλητών έχει γίνει το επίκεντρο πολλών ερευνών σε προβλήματα στατιστικής μοντελοποίησης μεγάλης διάστασης, σε διάφορους επιστημονικούς τομείς. Η χρήση των πειραματικών σχεδιασμών για την επιλογή μεταβλητών σε προβλήματα παλινδρόμησης με διαθέσιμα τα δεδομένα των παρατηρήσεων μιας μεγάλης βάσης δεδομένων εισήχθη από τους Pumplün et al. [185], [186]. Οι Schiffner και Weihs [204] επέκτειναν τη μελέτη προσομοίωσης στην εργασία [185] προκειμένου να επαληθεύσουν-επικυρώσουν τα αποτελέσματα, καθώς επίσης μελέτησαν και την καταλληλότητα των D-βέλτιστων σχεδιασμών για την εκμάθηση μεθόδων ταξινόμησης. Οι Rüping και Weihs [201] πρότειναν έναν αλγόριθμο εμπνευσμένο από τους στατιστικούς σχεδιασμούς πειραμάτων και τις μεθόδους πυρήνα για να αντιμετωπίσουν το πρόβλημα της επιλογής μεταβλητών δοθείσης μιας βάσης δεδομένων παρατηρήσεων. Αξίζει να σημειωθεί εδώ, ότι σε αντίθεση με τις παρατηρήσεις που προκύπτουν από τους πειραματικούς σχεδιασμούς, τεράστια σύνολα δεδομένων μερικές φορές γίνονται διαθέσιμα στις μέρες μας χωρίς προκαθορισμένους σκοπούς. Συνήθως, είναι προτιμότερο να βρεθούν μερικά ενδιαφέροντα χαρακτηριστικά από τα σύνολα δεδομένων τα οποία θα παρέχουν πολύτιμες πληροφορίες για την υποστήριξη της λήψης αποφάσεων [146].

Για τις πειραματικές καταστάσεις, όπου πραγματικά δεν υπάρχει καμία εκ των προτέρων γνώση των επιδράσεων των παραγόντων, αλλά μόνο μία ισχυρή παραδοχή της αρχής της σποραδικότητας των επιδράσεων των παραγόντων, και όπου ο στόχος είναι να εντοπίσουμε τους κυρίαρχους παράγοντες, εφόσον υπάρχουν, ο πειραματιστής θα πρέπει να εξετάσει σοβαρά το ενδεχόμενο να χρησιμοποιήσει τους υπερκορεσμένους σχεδιασμούς, όπως άλλωστε προτείνεται από τον Gilmour [92]. Στην πρόσφατη διεθνή βιβλιογραφία, υπάρχουν αρκετές ερευνητικές εργασίες, όπως για παράδειγμα οι [92] και [112], σχετικά με την πρακτική χρήση των υπερκορεσμένων σχεδιασμών σε πραγματικά προβλήματα της ζωής. Απαιτείται περισσότερη έρευνα για την καλύτερη πρακτική χρήση των υπερκορεσμένων σχεδιασμών, δεδομένου ότι οι καταστάσεις στις οποίες οι υπερκορεσμένοι σχεδιασμοί είναι πραγματικά υποσχόμενοι και ουσιαστικά έτοιμοι προς χρήση, στην πράξη, είναι περιορισμένες. Σχεδιασμοί με καλές ιδιότητες (αλλά όχι βέλτιστοι) είναι ήδη διαθέσιμοι για πολλά μεγέθη ενός πειράματος [92] αλλά η πρακτική χρήση τους εξακολουθεί να παραμένει ένα δύσκολο και απαιτητικό έργο. Οι παραπάνω λόγοι όπως είναι φυσικό, μας παρακίνησαν το ενδιαφέρον να μελετήσουμε το πρόβλημα επιλογής μεταβλητών με χρήση των υπερκορεσμένων σχεδιασμών, προκειμένου να αντιμετωπίσουμε γρήγορα και αποτελεσματικά ένα πρόβλημα στατιστικής μοντελοποίησης υψηλής διάστασης.

9.2 Περιγραφή Μεθόδων Επιλογής Μεταβλητών

Σε αυτή την ενότητα παρουσιάζονται συνοπτικά οι διάφορες στατιστικές μεθόδοι που χρησιμοποιήσαμε για την ανάλυση ενός μεγάλου συνόλου δεδομένων, και τελικά συγκρίναμε και αξιολογήσαμε την απόδοσή τους σε σχέση με αυτήν της προτεινόμενης μεθόδου επιλογής μεταβλητών μέσω των υπερκορεσμένων σχεδιασμών. Παρακάτω παρουσιάζουμε τις στατιστικές μεθόδους επιλογής μεταβλητών τις οποίες θεωρήσαμε στη μελέτη μας.

1. Το μοντέλο λογιστικής παλινδρόμησης χρησιμοποιείται συχνά για την ανάλυση δεδομένων που προκύπτουν από ιατρικές μελέτες, όπου συχνά εμφανίζεται η περίπτωση μίας δίτιμης μεταβλητής απόκρισης, λαμβάνοντας δύο πιθανές τιμές 1 ή 0, για την επιτυχία ή την αποτυχία, αντίστοιχα. Στο παρόν κεφάλαιο, υιοθετούμε το λογιστικό μοντέλο παλινδρόμησης (LR) (βλέπε Κεφάλαιο 5 και ενότητα 5.2.1 για περισσότερες λεπτομέρειες).

2. Οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών χρησιμοποιούνται με ορισμένα κριτήρια τα οποία συνδυάζουν στατιστικά μέτρα με συναρτήσεις ποινής. Στο παρόν κεφάλαιο θεωρήσαμε τα πιο δημοφιλή από τα εν λόγω κριτήρια, δηλαδή τα κριτήρια

πληροφορίας, και συγκεκριμένα τα κριτήρια AIC (Akaike Information Criteria) και BIC (Bayesian Information Criteria) (βλέπε Κεφάλαιο 6 και ενότητα 6.3.1 για περισσότερες λεπτομέρειες).

3. Οι Fan και Li [76] πρότειναν μία κλάση διαδικασιών επιλογής μεταβλητών μέσω της μη κοίλης ποινικοποιημένης πιθανοφάνειας (nonconcave penalized likelihood), σε μία προσπάθειά τους να επιλέξουν μεταβλητές και ταυτόχρονα να εκτιμήσουν τους συντελεστές παλινδρόμησης. Στο παρόν κεφάλαιο, υιοθετούμε τις μεθόδους αυτές και θεωρούμε ως συναρτήσεις ποινής τις SCAD, LASSO και Hard (βλέπε Κεφάλαιο 6 και ενότητα 6.3.3 για περισσότερες λεπτομέρειες).

9.3 Η Προτεινόμενη Μέθοδος Επιλογής Μεταβλητών

Σε αυτή την ενότητα, μελετάμε τη χρήση των υπερκορεσμένων σχεδιασμών για την επιλογή μεταβλητών δοθέντος ενός συνόλου δεδομένων παρατηρήσεων. Προσεγγίζουμε το πρόβλημα της επιλογής μεταβλητών θεωρώντας υπερκορεσμένους σχεδιασμούς, καθώς επίσης και διεξάγοντας διάφορες μεθόδους στατιστικής ανάλυσης. Το μέτρο που χρησιμοποιείται για τον προσδιορισμό των υπερκορεσμένων σχεδιασμών είναι το μέτρο της μη ορθογωνιότητας μεταξύ όλων των ζευγών των στηλών του σχεδιασμού. Κατά τη διάρκεια των πειραμάτων μας, μόνο μοντέλα κύριων επιδράσεων ελήφθησαν υπόψη. Παρουσιάζουμε αναλυτικά παρακάτω τα βήματα της προτεινόμενης μεθόδου επιλογής μεταβλητών και ταυτόχρονου προσδιορισμού όσο το δυνατόν καταλληλότερων υπερκορεσμένων σχεδιασμών για την ανάλυση ενός μεγάλου συνόλου πραγματικών δεδομένων.

Βήμα 1. Δοθέντος ενός αρχικού συνόλου δεδομένων, αποτελούμενο έστω από m προγνωστικές μεταβλητές εισόδου $\{x_{i1}, \ldots, x_{im}\}$ και N παρατηρήσεις, $i = 1, 2, \ldots, N$, καθορίζουμε τις ιδιότητες των υπερκορεσμένων σχεδιασμών.
- Επιθυμητός αριθμός γραμμών (n): $n = \frac{1}{100}$ των N διαθέσιμων παρατηρήσεων-πειραματικών εκτελέσεων.
- Επιθυμητός αριθμός στηλών (m): $m + 1 > n$.

Βήμα 2. Επιλέγουμε τυχαία από το αρχικό σύνολο δεδομένων 100 ($n \times m$) σχεδιασμούς σύμφωνα με την ομοιόμορφη κατανομή. Υποθέτουμε ότι κάθε ένας από τους αρχικούς m παράγοντες έχει δύο επίπεδα ± 1. Κατασκευάζουμε για κάθε σχεδιασμό τον υπερκορεσμένο πίνακα $X = [1, x_1, x_2, \ldots, x_m]$, ο οποίος είναι ο $n \times (m + 1)$ πίνακας του μοντέλου. Η πρώτη στήλη του X είναι η $1_n = [1, \ldots, 1]^T$, με τη στήλη j να αντιστοιχεί στα επίπεδα του $(j - 1)$-οστού παράγοντα για $j = 2, \ldots, m + 1$. Από τις $2, \ldots, m + 1$ στήλες του X, m στήλες ανατέθηκαν τυχαία σε κάθε σχεδιασμό. Από τις $1, \ldots, N$ γραμμές, $n = \frac{1}{100}N$ γραμμές ανατέθηκαν τυχαία σε κάθε σχεδιασμό.

Βήμα 3. Ψάχνουμε και αποκλείουμε τους σχεδιασμούς με τη χειρότερη ιδιότητα μη ορθογωνιότητας (η μέγιστη τιμή του βαθμού της μη ορθογωνιότητας μεταξύ των στηλών του σχεδιασμού είναι 1).

Βήμα 4. Ανιχνεύουμε τους υπερκορεσμένους σχεδιασμούς (ή σχεδιασμό) με την καλύτερη δυνατή ιδιότητα μη ορθογωνιότητας, δηλαδή τους υπερκορεσμένους σχεδιασμούς (ή σχεδιασμό) με το ελάχιστο επίπεδο συσχέτισης (συντελεστή συσχέτισης) μεταξύ των παραγόντων.

9.4 Ανάλυση Πραγματικών Δεδομένων

Η προτεινόμενη μέθοδος επιλογής μεταβλητών μέσω των υπερκορεσμένων σχεδιασμών εφαρμόζεται τώρα σε ένα σύνολο πραγματικών δεδομένων, και η απόδοσή της συγκρίνεται

με την απόδοση των προαναφερθέντων στατιστικών μεθόδων επιλογής μεταβλητών (SCAD, LASSO, Hard, AIC και BIC).

Τα πραγματικά δεδομένα προέρχονται από μία ετήσια έρευνα που διεξήχθη κατά την περίοδο 01/01/2005-31/12/2006 από την Ελληνική Εταιρεία Τραύματος και την Ομάδα Εντατικής Χειρουργικής, περιλαμβάνοντας 30 Γενικά Νοσοκομεία στην Ελλάδα, με βασικό σκοπό τη μελέτη των επιδράσεων διαφόρων προγνωστικών παραγόντων στη θεραπεία τραυματισμένων ατόμων. Η βάση δεδομένων αποτελείται από 70 επεξηγηματικές μεταβλητές που αποτελούν τους προγνωστικούς παράγοντες. Οι παράγοντες αυτοί που καταγράφηκαν περιελάμβαναν δημοφραφικά στοιχεία των ασθενών, στοιχεία που αφορούν τις συνθήκες του ατυχήματος και πληροφορίες για την ενδονοσοκομειακή αντιμετώπιση των ασθενών. Επίσης, η βάση δεδομένων αποτελείται από 7000 ασθενείς, και για κάθε έναν από αυτούς καταγράφηκε η κατάστασή τους κατά την τελευταία παρακολούθηση, με τη δίτιμη μεταβλητή απόκρισης y να παίρνει μόνο δύο τιμές, με το 0 να συμβολίζει την "επιβίωση" και το 1 την "αποβίωση, αντίστοιχα. Σύμφωνα με ιατρικές συμβουλές, όλοι οι προγνωστικοί παράγοντες θα πρέπει να τυγχάνουν ίσης μεταχείρισης κατά τη στατιστική ανάλυση και δεν υπάρχει κανένας παράγοντας για τον οποίο επιβάλλεται να διατηρείται πάντα στο μοντέλο.

9.4.1 Διαδοχική Επιλογή Μεταβλητών

Σε αυτή την ενότητα εκτελούμε διαδοχικά διάφορες μεθόδους επιλογής μεταβλητών, μέχρι την επίτευξη σύγκλισης του μοντέλου. Το *Model I* είναι το αρχικό μοντέλο, αποτελούμενο από όλο το διαθέσιμο σύνολο δεδομένων των $N = 7000$ ασθενών και 70 ενδεχόμενων παραγόντων κινδύνου. Οι ετικέτες-ονομασίες των 70 διαθέσιμων προγνωστικών παραγόντων του αρχικού μοντέλου *Model I* παρουσιάζονται στον Πίνακα 9.2. Λόγω του γεγονότος ότι οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC και BIC) είναι εξαιρετικά χρονοβόρες έως ανέφικτο να υλοποιηθούν από άποψη υπολογιστικού κόστους και χρόνου, η εφαρμογή τους δεν ήταν δυνατή σε αυτό το στάδιο της μελέτης.

Το *Model II* ελήφθη από την πρώτη εκτέλεση καθεμιάς από τις LR, SCAD, LASSO και Hard μεθόδους, και 29 μεταβλητές ανιχνεύθηκαν ως στατιστικά σημαντικές. Αυτές οι στατιστικά σημαντικές 29 μεταβλητές (x_9, x_{11}, x_{12}, x_{13}, x_{14}, x_{16}, x_{17}, x_{18}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}, x_{27}, x_{29}, x_{35}, x_{36}, x_{37}, x_{38}, x_{39}, x_{40}, x_{41}, x_{42}, x_{43}, x_{44}, x_{45}, x_{46}) επιλέχθηκαν από τουλάχιστον μία από αυτές τις μεθόδους, κρατήθηκαν και χρησιμοποιήθηκαν για περαιτέρω ανάλυση.

Το *Model III* ελήφθη από τη δεύτερη εκτέλεση καθεμιάς από τις LR, SCAD, LASSO και Hard μεθόδους, και 14 μεταβλητές ανιχνεύθηκαν ως στατιστικά σημαντικές. Αυτές οι στατιστικά σημαντικές 14 μεταβλητές επιλέχθηκαν από τουλάχιστον μία από αυτές τις μεθόδους, κρατήθηκαν και χρησιμοποιήθηκαν για περαιτέρω ανάλυση. Αξίζει να σημειωθεί εδώ, ότι όταν τέθηκαν σε εφαρμογή οι τέσσερις παραπάνω μέθοδοι για τρίτη φορά, παρατηρήσαμε ότι οι τιμές των εκτιμώμενων συντελεστών διατηρήθηκαν διαφορετικές από το μηδέν, με αποτέλεσμα τη σύγκλιση του μοντέλου.

Ως εκ τούτου, ακολουθώντας τις διαδικασίες επιλογής μεταβλητών LR, SCAD, LASSO και Hard, μετά από δύο βήματα, βρήκαμε το *Model III* το οποίο αποτελείται από 14 στατιστικά σημαντικές μεταβλητές, οι οποίες παρουσιάζονται στον Πίνακα 9.1. Οι εκτιμώμενοι συντελεστές και τα τυπικά σφάλματα του *Model III* παρουσιάζονται επίσης στον Πίνακα 9.1. Η άγνωστη παράμετρος λ εκτιμήθηκε ως 0.0008, 0.0001 και 0.0008 για την SCAD, LASSO και Hard συνάρτηση ποινής, αντίστοιχα. Με την εκτιμηθείσα τιμή του λ, ο εκτιμητής της ποινικοποιημένης πιθανοφάνειας αποκτήθηκε στο δεύτερο, τρίτο και δεύτερο βήμα επαναλήψεων, για την ποινικοποιημένη πιθανοφάνεια με την SCAD, LASSO και Hard συνάρτηση ποινής, αντίστοιχα.

Η εφαρμογή των μεθόδων επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC και BIC) δεν ήταν δυνατή μέχρι αυτό το στάδιο της πειραματικής μελέτης εξαιτίας της υπολογιστικής πολυπλοκότητας. Για αυτόν το λόγο εφαρμόσαμε τις δύο αυτές μεθόδους επιλογής

του καλύτερου υποσυνόλου μεταβλητών (AIC και BIC) μόνο στις 14 στατιστικά σημαντικές μεταβλητές του τελικού μοντέλου *Model III*. Παρατηρούμε από τον Πίνακα 9.1 ότι οι μέθοδοι SCAD, Hard, LR και AIC καταλήγουν στο ίδιο μοντέλο, ενώ οι εκτιμήσεις της LASSO είναι συστηματικά χαμηλότερες. Η μέθοδος επιλογής μεταβλητών με κριτήριο πληροφορίας το BIC, απορρίπτει παράγοντες που μπορεί να είναι στατιστικά σημαντικοί. Για παράδειγμα, ο εκτιμώμενος συντελεστής για τη μεταβλητή "ακτίνες X στη σπονδυλική στήλη" (x_{42}) τίθεται ίσος με μηδέν, ενώ η μεταβλητή αυτή έχει επιλεγεί σε όλα τα άλλα θεωρούμενα μοντέλα. Σημειώνουμε εδώ ότι τα εκτιμώμενα τυπικά σφάλματα και οι συντελεστές της LASSO είναι σταθερά μικρότερα σε σύγκριση με τις μεθόδους LR, SCAD και Hard. Αυτό σημαίνει ότι η μεροληψία των LASSO εκτιμητών ποινικοποιημένης πιθανοφάνειας είναι μεγαλύτερη.

Όσον αφορά τους χρόνους εκτέλεσης, παρατηρούμε ότι η μέθοδος Hard ήταν πολύ πιο γρήγορη με 67 sec έναντι της SCAD που χρειάστηκε σχεδόν 147 sec, και της LASSO η οποία χρειάστηκε 490 sec. Οι μέθοδοι επιλογής του καλύτερου υποσυνόλου μεταβλητών με τα κριτήρια BIC και AIC χρειάστηκαν 275331 sec και 276422 sec, αντίστοιχα. Σε γενικές γραμμές, η μέθοδος Hard ήταν η πιο γρήγορη και αποτελεσματική σε σύγκριση με τις υπόλοιπες υπό εξέταση μεθόδους.

Πίνακας 9.1: Εκτιμώμενοι συντελεστές και τυπικά σφάλματα (σε παρανθέσεις) για το *Model III*

Μέθοδος	LR	SCAD	LASSO	Hard	Best Subset (BIC)	Best Subset (AIC)
Intercept	-4.08 (0.11)	-4.08 (0.11)	-4.06 (0.11)	-4.08 (0.11)	-4.06 (0.11)	-4.08 (0.11)
x_9	0.33 (0.07)	0.33 (0.07)	0.33 (0.06)	0.33 (0.07)	0.32 (0.06)	0.33 (0.07)
x_{11}	-0.47 (0.10)	-0.47 (0.10)	-0.46 (0.10)	-0.47 (0.10)	-0.47 (0.09)	-0.47 (0.10)
x_{13}	0.24 (0.06)	0.24 (0.06)	0.24 (0.06)	0.24 (0.06)	0.23(0.06)	0.24 (0.06)
x_{16}	-0.20 (0.06)	-0.20 (0.06)	-0.20 (0.06)	-0.20 (0.06)	-0.26 (0.06)	-0.20 (0.06)
x_{20}	-0.24 (0.10)	-0.24 (0.10)	-0.23 (0.10)	-0.24 (0.10)	0 (-)	-0.24 (0.10)
x_{21}	-0.67 (0.12)	-0.67 (0.12)	-0.66 (0.12)	-0.67 (0.12)	-0.69 (0.12)	-0.67 (0.12)
x_{24}	0.65 (0.07)	0.65 (0.07)	0.65 (0.07)	0.65 (0.07)	0.63 (0.07)	0.65 (0.07)
x_{26}	-0.18 (0.08)	-0.18 (0.07)	-0.17 (0.07)	-0.18 (0.07)	0 (-)	-0.18 (0.07)
x_{29}	-0.80 (0.12)	-0.80 (0.12)	-0.79 (0.12)	-0.80 (0.12)	-0.82 (0.12)	-0.80 (0.12)
x_{37}	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)
x_{38}	-0.71 (0.06)	-0.71 (0.06)	-0.71 (0.06)	-0.71 (0.06)	-0.74 (0.06)	-0.71 (0.06)
x_{39}	-0.34 (0.09)	-0.34 (0.09)	-0.34 (0.08)	-0.34 (0.09)	-0.42 (0.08)	-0.34 (0.09)
x_{42}	-0.21 (0.09)	-0.21 (0.09)	-0.20 (0.09)	-0.21 (0.09)	0 (-)	-0.21 (0.09)
x_{43}	-0.46 (0.09)	-0.46 (0.09)	-0.45 (0.09)	-0.46 (0.09)	-0.45 (0.09)	-0.46 (0.09)

Πίνακας 9.2: Μελέτη Τραύματος

Model I (levels ±1):
x_1: gender
x_2: arterial hypertension
x_3: coronary disease
x_4: heart failure
x_5: arrhythmia
x_6: asthma
x_7: chronic obstructive pulmonary disease (COPD)
x_8: chronic kidney disease
x_9: car accident
x_{10}: none safety measures at the side of the accident
x_{11}: collar soft at the side of the accident
x_{12}: RL at the side of the accident
x_{13}: unconscious at the side of the accident
x_{14}: life belt
x_{15}: airbags
x_{16}: transfer to a clinic, of the hospital
x_{17}: transfer to surgery
x_{18}: transfer to general surgery
x_{19}: transfer to orthopedic clinic
x_{20}: expected big temporary handicap
x_{21}: expected small temporary handicap
x_{22}: recovery
x_{23}: transport by ambulance (in other hospital)
x_{24}: transport by ambulance in emergency room
x_{25}: transport by car in emergency room
x_{26}: checking in the emergency room by general surgeon
x_{27}: checking in the emergency room by neurosurgeon
x_{28}: ICP Monitoring
x_{29}: DW
x_{30}: US Triplex
x_{31}: evacuation
x_{32}: A.T.L.S
x_{33}: capillary refill
x_{34}: peritoneum points
x_{35}: sweating
x_{36}: peripatetic
x_{37}: central cyanosis
x_{38}: peripheral vein
x_{39}: head x-ray
x_{40}: upper part of spinal column x-ray
x_{41}: pelvis x-ray
x_{42}: spinal column x-ray
x_{43}: extremities x-ray
x_{44}: CT abdomen
x_{45}: CT extremities
x_{46}: resident on charge
x_{47}: CT thorax
x_{48}: thorax x-ray
x_{49}: nasogastric tube
x_{50}: fluids
x_{51}: chest drainage
x_{52}: catheter
x_{53}: pericardiocentesis
x_{54}: blood
x_{55}: thoracotomy
x_{56}: angiography
x_{57}: diagnostic peritoneal lavage (DPL)
x_{58}: embolism
x_{59}: toxicology testing
x_{60}: ultrasound (US)
x_{61}: urea testing
x_{62}: mild trauma
x_{63}: Radiograph E.R.
x_{64}: immobility of limbs
x_{65}: face injury
x_{66}: head injury
x_{67}: breast injury
x_{68}: spinal column injury
x_{69}: upper limbs injury
x_{70}: lower limbs injury

9.4.2 Επιλογή Μεταβλητών Μέσω Υπερκορεσμένων Σχεδιασμών

Εφαρμόσαμε την προτεινόμενη μέθοδο επιλογής μεταβλητών στο αρχικό σύνολο δεδομένων και 100 ($n = 70 \times m = 70$) σχεδιασμοί επιλέχθηκαν τυχαία σύμφωνα με την ομοιόμορφη κατανομή. Με άλλα λόγια, 100 υπερκορεσμένοι σχεδιασμοί κατασκευάστηκαν με βάση τυχαία δείγματα των παρατηρήσεων του συνόλου δεδομένων τραύματος, δηλαδή, με τυχαία επιλογή κάθε φορά $n = 70$ εκτελέσεων από τις $N = 7000$ διαθέσιμες πειραματικές εκτελέσεις. Εντοπίσαμε 5 υπερκορεσμένους σχεδιασμούς $m = 70$ παραγόντων και $n = 70$ εκτελέσεων με τον ελάχιστο συντελεστή συσχέτισης. Το μειονέκτημα αυτών των υπερκορεσμένων σχεδιασμών είναι ότι λόγω της φύσης του προβλήματος δεν ήταν δυνατόν να είναι ισορροπημένοι, δηλαδή κάθε στήλη του σχεδιασμού να έχει επίπεδα 1 και -1 να εμφανίζονται εξίσου συχνά. Δεδομένου ότι οι παράγοντες αυτοί που επιλέχθηκαν εξαρτώνται άμεσα από τα χαρακτηριστικά των ασθενών που μελετήθηκαν και δεν μπορεί να προκαθοριστούν, ήταν αδύνατο για εμάς να λάβουμε υπόψη μας κριτήρια βελτιστοποίησης ισορροπημένων υπερκορεσμένων σχεδιασμών δύο επιπέδων, γεγονός που θα μας προστάτευε περισσότερο από την αναπόφευκτη και πιθανή ύπαρξη ταυτόσημων σχεδιασμών εμπλεκόμενων στη στατιστική ανάλυση.

Εκτελέσαμε διαδοχικά διάφορες μεθόδους επιλογής μεταβλητών στα κατασκευασμένα 5 σύνολα δεδομένων, με καθένα από αυτά να αποτελείται από $n = 70$ εκτελέσεις και $m = 70$ παράγοντες μέχρι να επιτευχθεί η σύγκλιση του μοντέλου. Το $Model_{SSDs}$ ελήφθη από την πρώτη εκτέλεση καθεμιάς από τις LR, SCAD, LASSO και Hard μεθόδους, και 13 μεταβλητές ανιχνεύθηκαν ως στατιστικά σημαντικές και από τους πέντε θεωρούμενους υπερκορεσμένους σχεδιασμούς. Αυτές οι στατιστικά σημαντικές 13 μεταβλητές (x_1, x_2, x_3, x_{10}, x_{11}, x_{16}, x_{19}, x_{38}, x_{42}, x_{43}, x_{47}, x_{48}, x_{62}) επιλέχθηκαν από τουλάχιστον μία από τις προαναφερθείσες μεθόδους, κρατήθηκαν και χρησιμοποιήθηκαν για περαιτέρω ανάλυση.

Η εφαρμογή των μέθοδων επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC και BIC) ήταν δυνατή σε αυτό το στάδιο της μελέτης. Για το λόγο αυτό εφαρμόσαμε τις μεθόδους LR, SCAD, LASSO, Hard καθώς επίσης τις AIC και BIC στις 13 στατιστικά σημαντικές μεταβλητές του $Model_{SSDs}$. Με αυτό τον τρόπο βρήκαμε ένα νέο μοντέλο αποτελούμενο από 5 στατιστικά σημαντικές μεταβλητές. Αξίζει να σημειωθεί εδώ, ότι όταν τέθηκαν σε εφαρμογή οι παραπάνω μέθοδοι για άλλη μία φορά, παρατηρήσαμε ότι οι τιμές των εκτιμώμενων συντελεστών διατηρήθηκαν διαφορετικές από το μηδέν, με αποτέλεσμα τη σύγκλιση του μοντέλου. Έτσι, εντοπίσαμε το "καλύτερο" δυνατό μοντέλο (*Final* $Model_{SSDs}$) αποτελούμενο από 5 στατιστικά σημαντικές μεταβλητές, όπως αυτές παρουσιάζονται στον Πίνακα 9.3. Οι εκτιμώμενοι συντελεστές αυτού του μοντέλου αποκτήθηκαν θεωρώντας και τις 7000 παρατηρήσεις, γεγονός που καθιστά εφικτή την αξιολόγηση και τη σύγκριση των αποτελεσμάτων με το *Model III*. Η άγνωστη παράμετρος λ εκτιμήθηκε ίση με 0.0007, 0.0001 και 0.0007 για την SCAD, LASSO και Hard συνάρτηση ποινής, αντίστοιχα. Ο απαιτούμενος αριθμός βημάτων έτσι ώστε να αποκτηθούν οι εκτιμητές ποινικοποιημένης πιθανοφάνειας ήταν δύο, τρία και δύο για την SCAD, LASSO και Hard συνάρτηση ποινής, αντίστοιχα.

Αξίζει να σημειωθεί εδώ, ότι ακόμα και αν οι κατασκευασμένοι υπερκορεσμένοι σχεδιασμοί μπορεί να μην είναι οι καλύτεροι δυνατοί, ωστόσο, αυτοί οι υπερκορεσμένοι σχεδιασμοί κατόρθωσαν να αναγνωρίσουν τους 5 στατιστικά σημαντικούς παράγοντες οι οποίοι συμπεριλαμβάνονται και στο *Model III*, χρησιμοποιώντας αρχικά μόνο το 1/100 των διαθέσιμων πειραματικών εκτελέσεων. Παρατηρούμε επίσης μια σταθερότητα στα αποτελέσματα που προέκυψαν από τη στατιστική ανάλυση και των πέντε κατασκευασμένων υπερκορεσμένων σχεδιασμών. Σε ιατρικούς όρους, ένα μοντέλο με όσο το δυνατόν λιγότερους σημαντικούς παράγοντες ήταν επιθυμητό στην περίπτωσή μας. Παρατηρούμε από τον Πίνακα 9.3 ότι και οι έξι μέθοδοι έδωσαν τα ίδια αποτελέσματα, και ως εκ τούτου η ιατρική κοινότητα καλό θα ήταν να λάβει σοβαρά υπόψη της αυτούς τους πέντε παράγοντες, προκειμένου να μειωθεί το ποσοστό θανάτων από τραύμα στην Ελλάδα. Πράγματι, οι 5 αυτοί κοινοί παράγοντες θα πρέπει να θεωρούνται ως στατιστικά σημαντικοί, και όλοι οι άλλοι θα μπορούσαν να διατηρηθούν για περαιτέρω έρευνα. Ο πρώτος και τρίτος παράγοντας αναφέρονται σε δράσεις

στο μέρος του ατυχήματος. Ο δεύτερος παράγοντας αφορά την απόφαση που ένας γιατρός θα πρέπει να λάβει αφού ο ασθενής εξεταστεί στο δωμάτιο έκτακτης ανάγκης, και οι δύο τελευταίοι παράγοντες αφορούν δράσεις που αναλαμβάνονται στο δωμάτιο έκτακτης ανάγκης.

Πίνακας 9.3: Εκτιμώμενοι συντελεστές και τυπικά σφάλματα (σε παρανθέσεις) για το *Final* $Model_{SSDs}$

Μέθοδος	LR	SCAD	LASSO	Hard	Best Subset (BIC)	Best Subset (AIC)
Intercept	-3.93 (0.10)	-3.93 (0.10)	-3.92 (0.10)	-3.93 (0.10)	-3.93 (0.10)	-3.93 (0.10)
x_{11}	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)
x_{16}	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)
x_{38}	-0.65 (0.05)	-0.65 (0.05)	-0.65 (0.05)	-0.65 (0.05)	-0.65(0.05)	-0.65 (0.05)
x_{42}	-0.49 (0.09)	-0.49 (0.09)	-0.49 (0.08)	-0.49 (0.09)	-0.49 (0.09)	-0.49 (0.09)
x_{43}	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)

9.4.3 Συγκριτικά Αποτελέσματα

Αρχικά, εκτελέσαμε διαδοχικά διάφορες μεθόδους επιλογής μεταβλητών σε ολόκληρο το σύνολο δεδομένων μέχρι να πετύχουμε σύγκλιση του μοντέλου. Ως εκ τούτου, εντοπίσαμε το τελικό μοντέλο *Model III* αποτελούμενο από 14 στατιστικά σημαντικές μεταβλητές (βλέπε Πίνακα 9.1). Οι 14 αυτές στατιστικά σημαντικές μεταβλητές που επιλέχθηκαν είναι οι x_9, x_{11}, x_{13}, x_{16}, x_{20}, x_{21}, x_{24}, x_{26}, x_{29}, x_{37}, x_{38}, x_{39}, x_{42}, x_{43}.

Στη συνέχεια προσεγγίσαμε το πρόβλημα επιλογής μεταβλητών θεωρώντας αρκετούς υπερκορεσμένους σχεδιασμούς δύο επιπέδων. Μετά την ενσωμάτωση των πέντε κατασκευασμένων υπερκορεσμένων σχεδιασμών, εφαρμόσαμε διάφορες μεθόδους στατιστικής ανάλυσης, μέχρι την επίτευξη σύγκλισης του μοντέλου. Ως εκ τούτου, εντοπίσαμε το "καλύτερο" δυνατό μοντέλο *Final* $Model_{SSDs}$ που αποτελείται από 5 στατιστικά σημαντικές μεταβλητές (βλέπε Πίνακα 9.3). Οι 5 αυτές στατιστικά σημαντικές μεταβλητές που επιλέχθηκαν είναι οι x_{11}, x_{16}, x_{38}, x_{42}, x_{43}.

Υπάρχουν τέσσερις σημαντικές παρατηρήσεις που αξίζει να αναφερθούν εδώ.

1. Το *Final* $Model_{SSDs}$ αποτελείται από 5 μεταβλητές οι οποίες ανιχνεύθηκαν ως στατιστικά σημαντικές και από τους πέντε κατασκευασμένους υπερκορεσμένους σχεδιασμούς.

2. Οι 5 στατιστικά σημαντικές μεταβλητές του *Final* $Model_{SSDs}$ (βλέπε Πίνακα 9.3) βρέθηκαν επίσης ως στατιστικά σημαντικές και συμπεριλαμβάνονται στο *Model III* (βλέπε Πίνακα 9.1). Εάν συγκρίνουμε τα αποτελέσματα των στατιστικά σημαντικών μεταβλητών που δίνονται στον Πίνακα 9.1 και στον Πίνακα 9.3, παρατηρούμε ότι υπάρχουν μόνο πέντε σύμφωνες μεταβλητές (x_{11}, x_{16}, x_{38}, x_{42}, x_{43}) οι οποίες αποκτήθηκαν μέσω της διαδοχικής επιλογής μεταβλητών και μέσω της προτεινόμενης μεθόδου. Αξίζει να σημειωθεί εδώ ότι αυτές οι σύμφωνες πέντε μεταβλητές συνιστούν και το τελικό μοντέλο (*Final* $Model_{SSDs}$) που προσδιορίζεται από την εφαρμογή της προτεινόμενης μεθόδου και τη βοήθεια των μεθόδων ποινικοποιημένης πιθανοφάνειας.

3. Η προτεινόμενη μέθοδος επιτυγχάνει να αναγνωρίσει αυτές τις 5 στατιστικά σημαντικές μεταβλητές (*Final* $Model_{SSDs}$) χρησιμοποιώντας αρχικά μόνο το 1/100 των διαθέσιμων πειραματικών εκτελέσεων.

4. Με την εκτέλεση της διαδοχικής επιλογής μεταβλητών, η μέθοδος επιλογής του καλύτερου υποσυνόλου μεταβλητών ελαχιστοποιώντας το κριτήριο BIC βρέθηκε να απορρίπτει παράγοντες που μπορεί να είναι στατιστικά σημαντικοί. Η μεταβλητή "ακτίνες X στη σπονδυλική στήλη" (x_{42}) αναγνωρίστηκε ως μη σημαντική από το BIC, ενώ έχει επιλεγεί ως σημαντική σε όλα τα άλλα θεωρούμενα μοντέλα (βλέπε Πίνακα 9.1). Η προτεινόμενη μέθοδος κατόρθωσε να αναγνωρίσει τη μεταβλητή "ακτίνες X στη

σπονδυλική στήλη" (x_{42}) ως στατιστικά σημαντική, ακόμη και ελαχιστοποιώντας το κριτήριο BIC (βλέπε Πίνακα 9.3). Το γεγονός αυτό υποδηλώνει ότι η προτεινόμενη μεθοδολογία μέσω των υπερκορεσμένων σχεδιασμών τείνει να μειώσει τη μεροληψία για τους εκτιμητές των συντελεστών και των τυπικών σφαλμάτων.

9.5 Συμπεράσματα

Η πρόσφατη εξάπλωση των βάσεων δεδομένων μεγάλης διάστασης καθιστά την επιλογή μεταβλητών ή τη μείωση της διάστασης ενός προβλήματος, ένα ερευνητικό πρόβλημα ζωτικής σημασίας για την κατασκευή ενός μοντέλου, καθώς επίσης αποτελεί και πρόκληση για έναν πειραματιστή λόγω της περίπλοκης δομής που προσπαθεί να απλοποιήσει. Η συνετή και αποτελεσματική επιλογή μεταβλητών όχι μόνο απλά βελτιώνει την προβλεπτική ικανότητα ενός μοντέλου, αλλά γενικά εξασφαλίζει μία καλύτερη κατανόηση της υποκείμενης έννοιας που δημιουργεί τα δεδομένα. Το κεφάλαιο αυτό έρχεται να παρουσιάσει μία εκτενή εφαρμογή των μεθόδων ποινικοποιημένης πιθανοφάνειας (SCAD, LASSO και Hard), καθώς επίσης και των μεθόδων επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC και BIC), σε συνδυασμό με τη χρήση των υπερκορεσμένων σχεδιασμών για την επιλογή μεταβλητών δοθέντος ενός μεγάλου συνόλου δεδομένων παρατηρήσεων.

Η μελέτη μας έδειξε ότι η λογιστική παλινδρόμηση σε σύγκριση με τις μεθόδους AIC και BIC, αποτυγχάνει να εξαλείψει τις πιθανές μη σημαντικές μεταβλητές, και οι μέθοδοι AIC ή BIC είναι ιδιαίτερα χρονοβόρες, σε ορισμένες περιπτώσεις ακόμα και αδύνατον να εφαρμοστούν. Οι SCAD, LASSO και Hard μέθοδοι υπερέχουν ξεκάθαρα ως προς την απόδοσή τους σε σύγκριση με τις LR, AIC και BIC, δεδομένου ότι είναι πολύ πιο γρήγορες και πιο αποτελεσματικές επειδή επιλέγουν τις σημαντικές μεταβλητές βελτιστοποιώντας ταυτόχρονα την ποινικοποιημένη πιθανοφάνεια, και ως εκ τούτου τα τυπικά σφάλματα των παραμέτρων του μοντέλου μπορούν να εκτιμηθούν με περισσότερη ακρίβεια.

Δεδομένου ότι οι μέθοδοι SCAD, LASSO και Hard είναι γρήγορες καθώς επίσης και εύκολα υλοποιήσιμες, ακόμη και σε ένα πρόβλημα μεγάλης διάστασης, καλό θα ήταν να εφαρμόζονται όλες κατά τη διάρκεια μίας στατιστικής ανάλυσης. Με τον τρόπο αυτό, οι παράγοντες που επιλέγονται από αυτές τις τρεις μεθόδους ποινικοποιημένης πιθανοφάνειας θα πρέπει να θεωρούνται στατιστικά σημαντικοί, και οι υπόλοιποι παράγοντες που έχουν επιλεγεί από τουλάχιστον μία από αυτές τις μεθόδους θα μπορούσαν να διατηρηθούν για περαιτέρω έρευνα. Η ενσωμάτωση των υπερκορεσμένων σχεδιασμών στη στατιστική ανάλυση μιας βάσης δεδομένων κρίθηκε αποτελεσματική και χρήσιμη, δεδομένου ότι μας επέτρεψε με τη χρήση μόνο λίγων παρατηρήσεων (μόνο το 1/100 των διαθέσιμων πειραματικών εκτελέσεων) να αποκτήσουμε ένα φειδωλό μοντέλο που προσδιορίζει τους σημαντικούς προγνωστικούς παράγοντες οι οποίοι επηρεάζουν την έκβαση ενός ασθενή.

ΚΕΦΑΛΑΙΟ **10**

Επιλογή Σχεδιασμού σε Βάσεις Δεδομένων

We are drowning in information
and starving for knowledge.

—*Rutherford D. Roger (1985)*

Στο δέκατο αυτό κεφάλαιο, μελετάμε την εφαρμογή κάποιων τεχνικών εξόρυξης δεδομένων, ενός γενετικού αλγορίθμου καθώς επίσης και τη χρήση των πειραματικών σχεδιασμών σε βάσεις δεδομένων, προκειμένου να καθορίσουμε τις πιο σημαντικές μεταβλητές σε προβλήματα παλινδρόμησης σε περιπτώσεις που οι παρατηρήσεις ειδόδου και οι ετικέτες κλάσης εξόδου μιας μεγάλης βάσης δεδομένων είναι διαθέσιμες. Προτείνουμε ένα κρυπτογραφικό σχήμα καθοδηγούμενο από την ίδια τη βάση των δεδομένων, με στόχο την κρυπτογράφηση συγκεκριμένων πεδίων της βάσης δεδομένων και την επιλογή ενός βέλτιστου σχεδιασμού αποτελούμενου από τις μεταβλητές της βάσης δεδομένων που ανιχνεύθηκαν να επηρεάζουν σημαντικά τη μεταβλητή απόκρισης. Η προτεινόμενη προσέγγιση επιλογής σχεδιασμού είναι αρκετά υποσχόμενη, δεδομένου ότι μας επιτρέπει να ανακτήσουμε βέλτιστους σχεδιασμούς που ανήκουν στην κλάση των υπερκορεσμένων σχεδιασμών, χρησιμοποιώντας μόνο ένα πολύ μικρό ποσοστό των διαθέσιμων πειραματικών εκτελέσεων, γεγονός που καθιστά τη στατιστική ανάλυση μιας μεγάλης βάσης δεδομένων υπολογιστικά εφικτή και οικονομικά προσιτή. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [136].

10.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Στο προηγούμενο Κεφάλαιο 9 αναφερθήκαμε διεξοδικά στη χρήση των πειραματικών σχεδιασμών για την επιλογή μεταβλητών σε προβλήματα παλινδρόμησης με διαθέσιμα τα δεδομένα των παρατηρήσεων μιας μεγάλης βάσης δεδομένων. Επίσης, αναφερθήκαμε και στις ερευνητικές εργασίες που συναναντάμε στην πρόσφατη βιβλιογραφία, σχετικά με την πρακτική χρήση των υπερκορεσμένων σχεδιασμών σε πραγματικά προβλήματα της ζωής, όπως για παράδειγμα είναι οι εργασίες [185], [186], [204], [201] και [112]. Παρ' όλα αυτά, απαιτείται περισσότερη έρευνα για την καλύτερη πρακτική χρήση των υπερκορεσμένων σχεδιασμών, δεδομένου ότι οι καταστάσεις στις οποίες οι υπερκορεσμένοι σχεδιασμοί είναι πραγματικά υποσχόμενοι και ουσιαστικά έτοιμοι προς χρήση, στην πράξη, είναι περιορισμένες [92]. Σχεδιασμοί με καλές ιδιότητες (αλλά όχι βέλτιστοι) είναι ήδη διαθέσιμοι για πολλά μεγέθη ενός πειράματος, αλλά η πρακτική χρήση τους εξακολουθεί να παραμένει ένα δύσκολο και απαιτητικό έργο.

Στο προηγούμενο Κεφάλαιο 9, προτείναμε μία μέθοδο στην οποία θεωρήσαμε τους υπερκορεσμένους σχεδιασμούς για την επιλογή μεταβλητών σε μία βάση δεδομένων, χρησιμοποιώντας ταυτόχρονα μεθόδους ποινικοποιημένης πιθανοφάνειας (SCAD, LASSO και Hard), καθώς επίσης και μεθόδους επιλογής του καλύτερου υποσυνόλου μεταβλητών (AIC και BIC). Η μέθοδος αυτή παρουσιάζει όμως τρία βασικά μειονεκτήματα. Πρώτον, η μέθοδος ναι μεν εφαρμόζεται στο αρχικό σύνολο δεδομένων αλλά μόνο ένα προκαθορισμένος από τον πειραματιστή αριθμός σχεδιασμών επιλέγεται προς εξέταση, με άλλα λόγια η μέθοδος δεν μπορεί να σαρώσει όλη τη βάση δεδομένων και να ελέγξει όλους τους πιθανούς συνδυασμούς που μπορεί να προκύψουν από τους m παράγοντες και τις n πειραματικές εκτελέσεις. Δεύτερον, ο πειραματιστής είναι αυτός που καθορίζει τον επιθυμητό αριθμό γραμμών και τον επιθυμητό αριθμό στηλών των σχεδιασμών που θα σχηματιστούν. Τρίτον, οι σχεδιασμοί επιλέγονται τυχαία σύμφωνα με την ομοιόμορφη κατανομή, με άλλα λόγια κατασκευάζονται με βάση τυχαία δείγματα των παρατηρήσεων του συνόλου δεδομένων, δηλαδή, με τυχαία επιλογή κάθε φορά n προκαθορισμένων εκτελέσεων από τις N αρχικές διαθέσιμες πειραματικές εκτελέσεις. Λόγω των παραπάνω παρατηρήσεων και της φύσης του προβλήματος ήταν αδύνατο να προστατευτούμε από την αναπόφευκτη και πιθανή ύπαρξη ταυτόσημων σχεδιασμών εμπλεκόμενων στη στατιστική ανάλυση. Οι παραπάνω λόγοι όπως είναι φυσικό μας παρακίνησαν το ενδιαφέρον να μελετήσουμε εκ νέου το πρόβλημα επιλογής μεταβλητών με χρήση των υπερκορεσμένων σχεδιασμών προτείνοντας ένα κρυπτογραφικό σχήμα καθοδηγούμενο από την ίδια τη βάση των δεδομένων, γεγονός που μας επιτρέπει να αποφύγουμε και ταυτόχρονα να αντιμετωπίσουμε τα παραπάνω μειονεκτήματα.

Συγκεκριμένα, προτείνουμε μία μέθοδο επιλογής βέλτιστου σχεδιασμού, καθοδηγούμενη από τη βάση των δεδομένων, προκειμένου να κρυπτογραφήσουμε συγκεκριμένα πεδία της βάσης δεδομένων που αντιστοιχούν στις σημαντικές μεταβλητές ενός προβλήματος παλινδρόμησης, σε περιπτώσεις που οι παρατηρήσεις και οι ετικέτες κλάσης μιας βάσης δεδομένων είναι διαθέσιμες. Το προτεινόμενο κρυπτογραφικό σχήμα το καθοδηγούμενο από τα δεδομένα, είναι ένας συνδυασμός γενετικού αλγορίθμου, metaheuristics και τεχνικών εξόρυξης δεδομένων και επιτρέπει στον πειραματιστή να προσδιορίσει τον καταλληλότερο υπερκορεσμένο σχεδιασμό, ο οποίος ανακτάται από μια βάση δεδομένων για τους σκοπούς της επιλογής μεταβλητών. Η στενή αλληλεπίδραση μεταξύ της εξόρυξης δεδομένων και της στατιστικής ανάλυσης επέτρεψε την επιτυχή μετάβαση από τη συλλογή των δεδομένων, μέσω της μοντελοποίησης των υποκείμενων δομών, σε κατανοητά και κερδοφόρα αποτελέσματα.

10.2 Τεχνικές Εξόρυξης Δεδομένων και Γενετικοί Αλγόριθμοι

Σε αυτή την ενότητα παρουσιάζουμε συνοπτικά τις τεχνικές που ενσωματώνουμε και υλοποιούμε στην προτεινόμενη μέθοδο με στόχο την ανάκτηση βέλτιστων σχεδιασμών μέσα από μία πραγματική βάση δεδομένων.

10.2.1 Εξόρυξη Κανόνων Συσχέτισης

Η εξόρυξη κανόνων συσχέτισης βρίσκει ενδιαφέρουσες συσχετίσεις μεταξύ ενός μεγάλου συνόλου στοιχείων των διαθέσιμων δεδομένων. Οι κανόνες συσχέτισης υποδεικνύουν συνθήκες για τα στοιχεία που εμφανίζονται συχνά μαζί σε ένα συγκεκριμένο σύνολο δεδομένων. Οι κανόνες συσχέτισης παρέχουν πληροφορίες αυτού του τύπου, με τη μορφή "εάν-τότε" ("if-then"). Εκτός από το προηγηθέν μέρος (το "αν" μέρος) και το συνακόλουθο μέρος (το "τότε" μέρος) ενός κανόνα, ένας κανόνας συσχέτισης έχει δύο κριτήρια που εκφράζουν το βαθμό αβεβαιότητας για τον κανόνα [228]. Το πρώτο κριτήριο καλείται υποστήριξη (support) και αναφέρεται στο ποσοστό των καταχωρήσεων στα δεδομένα εκπαίδευσης των οποίων το προηγηθέν μέρος είναι σωστό. Το δεύτερο κριτήριο είναι γνωστό ως εμπιστοσύνη (confidence), βασίζεται στις καταχωρήσεις για τις οποίες το προηγηθέν μέρος είναι σωστό, και είναι το ποσοστό αυτών των καταχωρήσεων για τις οποίες το συνακόλουθο μέρος είναι επίσης σωστό. Με άλλα λόγια, αποτελεί το ποσοστό των σωστών προβλέψεων που βασίζονται στον κανόνα συσχέτισης. Κανόνες με μικρότερη εμπιστοσύνη ή υποστήριξη από την προκαθορισμένη τιμή κριτηρίου αποκλείονται από την ανάλυση συσχετίσεων.

Ο αλγόριθμος Generalized Rule Induction (GRI) αποτελεί την καταλληλότερη μέθοδο για τη μελέτη μας, δεδομένου ότι μπορεί να διαχειριστεί κατηγορικές ή αριθμητικές μεταβλητές εισόδου, και απαιτεί κατηγορικές μεταβλητές εξόδου. Ο αλγόριθμος GRI εφαρμόζει μια προσέγγιση της Θεωρίας Πληροφοριών [210] προκειμένου να προσδιορίσει το ενδιαφέρον ενός υποψήφιου κανόνα συσχέτισης χρησιμοποιώντας το ποσοτικό μέτρο J. Ο αλγόριθμος GRI χρησιμοποιεί αυτό το μέτρο J για να υπολογίσει το κατά πόσο ενδιαφέρον ένας κανόνας μπορεί να είναι και χρησιμοποιεί συγκεκριμένα όρια-φράγματα σχετικά με τις πιθανές τιμές που μπορεί να λάβει αυτό το μέτρο, προκειμένου να περιορίσει το χώρο αναζήτησης των κανόνων. Οι κανόνες συσχέτισης που παράγονται από τον αλγόριθμο GRI έχουν τη μορφή " Εάν X=x τότε Y=y" όπου X και Y είναι δύο χαρακτηριστικά (πεδία) και x και y είναι οι τιμές αυτών των χαρακτηριστικών (πεδίων). Ο αλγόριθμος GRI εξάγει κανόνες συσχέτισης με την υψηλότερη περιεκτικότητα σε πληροφορίες βάση του δείκτη J, ο οποίος λαμβάνει υπόψη του τόσο την υποστήριξη όσο και την εμπιστοσύνη των κανόνων. Ο GRI στοχεύει στην ιεράρχηση των ανταγωνιστικών κανόνων, περιορισμό του χώρου έρευνας κανόνων και εξακρίβωση των καλύτερων κανόνων συσχέτισης που περιγράφουν ικανοποιητικά τη βάση δεδομένων. Αυτό το προκαταρκτικό στάδιο της στατιστικής ανάλυσης είναι πολύ σημαντικό, διότι επιτρέπει στον πειραματιστή να εντοπίσει τα πεδία και τις εγγραφές που είναι πιο πιθανό να έχουν ενδιαφέρον στη μοντελοποίηση, και να δημιουργήσει ένα σχήμα καθοδηγούμενο από τα δεδομένα για την κρυπτογράφηση συγκεκριμένων πεδίων της βάσης δεδομένων.

10.2.2 Απλός Γενετικός Αλγόριθμος

Στη μελέτη αυτή υποθέτουμε κάποια βασική εξοικείωση με κλασικές έννοιες των γενετικών αλγορίθμων, όπως είναι η αναπαραγωγή (reproduction), η μετάλλαξη (mutation) και η διασταύρωση (crossover), και θα περιγράψουμε μόνο ότι απαιτείται για την προσέγγισή μας. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στις εργασίες [94] και [62] για περισσότερες λεπτομέρειες σχετικά με τις απαραίτητες έννοιες για την πλήρη περιγραφή ενός απλού γενετικού αλγορίθμου (Simple Genetic Algorithm-SGA).

Στην προτεινόμενη μέθοδο, ενσωματώνουμε έναν απλό γενετικό αλγόριθμο, συγκεκριμένα αντιστοιχούμε καταχωρήσεις της βάσης δεδομένων σε χρωμοσώματα του απλού γενετικού αλγορίθμου, και μας ενδιαφέρει η περίπτωση που το μήκος του χρωμοσώματος είναι πολύ μικρό σε σύγκριση με το συνολικό αριθμό των χαρακτηριστικών της βάσης δεδομένων. Η βάση δεδομένων κωδικοποιείται σε τιμές $\{-1, 1\}$ και χρησιμοποιούμε την ίδια κωδικοποίηση για τον απλό γενετικό αλγόριθμο. Κάθε αλγόριθμος βελτιστοποίησης εξαρτάται από μία συγκεκριμένη αντικειμενική συνάρτηση η οποία χρειάζεται να βελτιστοποιηθεί. Για τους υπερκορεσμένους σχεδιασμούς, ένα ευρέως χρησιμοποιούμενο κριτήριο βελτιστοποίησης είναι το r_{max} κριτήριο. Ένα εύλογο κριτήριο για να συγκρίνουμε υπερκορεσμένους σχεδιασμούς είναι η ελαχιστοποίηση του $max_{i<j}|r_{ij}|$, όπου r_{ij} είναι η συσχέτιση δύο στηλών, c_i και c_j. Η μεγαλύτερη απόλυτη τιμή της συσχέτισης r_{ij} μεταξύ όλων των ζευγών των στηλών συμβολίζεται με το r_{max}.

10.2.3 L_1-norm Μηχανές Διανυσματικής Υποστήριξης

Οι Μηχανές Διανυσματικής Υποστήριξης (SVMs) αποτελούν μία μέθοδο εκμάθησης με πλήρη επίβλεψη, η οποία βασίζεται σε ιδέες που προέρχονται από τη στατιστική θεωρία εκμάθησης [220], [33], και έχουν εφαρμογές σε πολλούς επιστημονικούς κλάδους. Οι SVMs μοντελοποιούν μια μεταβλητή απόκρισης βασισμένες σε μια ή περισσότερες επεξηγηματικές μεταβλητές με στόχο την εύρεση του βέλτιστου υπερεπιπέδου. Δεδομένου ότι στη μελέτη μας έχουμε να αντιμετωπίσουμε ένα δυαδικό πρόβλημα ταξινόμησης, το βέλτιστο υπερεπίπεδο είναι αυτό με το μέγιστο περιθώριο διαχωρισμού μεταξύ δύο κλάσεων. Οι SVMs μπορούν επίσης να εφαρμοστούν σε προβλήματα παλινδρόμησης με την εισαγωγή μιας συνάρτησης απώλειας και μιας συνάρτησης ποινής. Αντί πλέον να προσπαθούμε να κατατάξουμε μια νέα άγνωστη επεξηγηματική μεταβλητή σε μία από τις δύο κλάσεις $y_i = \pm 1$ τώρα επιθυμούμε να προβλέψουμε την πραγματική τιμή εξόδου, με τα δεδομένα εκπαίδευσης να είναι της μορφής $\{x_i, y_i\}$.

Οι SVMs εναλλακτικά μπορούν επίσης να διατυπωθούν ως ένα πρόβλημα εκτίμησης μιας κανονικοποιημένης συνάρτησης, που αντιστοιχεί σε μια συνάρτηση απώλειας συν έναν όρο κανονικοποίησης για τους προσαρμοσμένους συντελεστές [233]. Η Least Absolute Shrinkage and Selection Operator (LASSO) μέθοδος [216] είναι μία από τις πιο διαδεδομένες τεχνικές εκτίμησης παραμέτρων σε προβλήματα παλινδρόμησης. Η L_1-norm ποινή (LASSO) προσαρμόστηκε στη μεθοδολογία των SVMs προκειμένου να εκτελείται αυτόματη επιλογή μεταβλητών σε προβλήματα κατηγοριοποίησης (όπως για παράδειγμα στις εργασίες [26], [236] και [84]). Πρόσφατα, ο όρος ποινής ελαστικού δικτύου (elastic net [238]) προσαρμόστηκε στη μεθοδολογία των SVMs. Η μέθοδος elastic net SVM [224], [225] είναι ιδιαίτερα χρήσιμη για τις περιπτώσεις στις οποίες ο αριθμός των μεταβλητών υπερβαίνει το μέγεθος του δείγματος. Η L_1-norm SVM είναι η καταλληλότερη μέθοδος για την ανάλυση των πραγματικών δεδομένων μας, αφού η διάσταση των δεδομένων (m=44 μεταβλητές εισόδου) δεν είναι μεγαλύτερη από τον αριθμό των δειγμάτων εκπαίδευσης (n=8862 παρατηρήσεις).

10.3 Η Προτεινόμενη Μέθοδος Επιλογής Σχεδιασμού

Σε αυτή την ενότητα παρουσιάζουμε αναλυτικά την προτεινόμενη μέθοδο ανάκτησης ενός βέλτιστου υπερκορεσμένου σχεδιασμού από τις καταχωρήσεις καθώς επίσης και από συγκεκριμένα πεδία της βάσης δεδομένων. Η προτεινόμενη μέθοδος προχωρά βήμα προς βήμα ως ακολούθως.

Βήμα 1. Έστω ότι το y_i συμβολίζει την i-οστή απόκριση στο σύνολο δεδομένων και το x_i συμβολίζει το m-διάστασης διάνυσμα των επεξηγηματικών μεταβλητών του συνόλου των δεδομένων.

Βήμα 2. Διαχωρίζουμε τα δεδομένα σε σύνολο εκπαίδευσης (90%) και σύνολο εξέτασης (10%). Για το διαχωρισμό, οι καταχωρήσεις επιλέχθηκαν τυχαία για να δημιουργήσουν το σύνολο εκπαίδευσης και εξέτασης σύμφωνα με το προκαθορισμένο μέγεθός τους.

Βήμα 3. Εκτελούμε τον αλγόριθμο GRI στα $(x_i, y_i)_{i=1,\ldots,n}$ και παράγουμε τους κανόνες συσχέτισης που χρησιμοποιούνται για την αρχικοποίηση των τυχαίων χρωμοσωμάτων, που με τη σειρά τους χρησιμοποιούνται στη δεξαμενή ζευγαρώματος (mating pool) του απλού γενετικού αλγορίθμου. Συγκεκριμένα, εφαρμόζουμε το GRI έργο εξόρυξης επαναληπτικά, δηλαδή, εκτελούμε μια "άνω" αναζήτηση σε ολόκληρο το σύνολο εκπαίδευσης και στη συνέχεια εκτελούμε μια "κάτω" αναζήτηση στο υπόλοιπο, για να απομακρύνουμε τα τμήματα με χαμηλές αποδόσεις.

Βήμα 4. Επιλέγουμε τις αρχικές πειραματικές εκτελέσεις του βέλτιστου σχεδιασμού ταιριάζοντας τους παραγόμενους GRI κανόνες συσχέτισης με τα αντίστοιχα πεδία της βάσης δεδομένων. Εάν n_d είναι ο αριθμός των πειραματικών εκτελέσεων που ανακτώνται χρησιμοποιώντας τους κανόνες GRI, τότε η αλγοριθμική διαδικασία προχωρά καθορίζοντας τον συνολικό αριθμό των εκτελέσεων n του υπερκορεσμένου σχεδιασμού ως $n = n_d + n_h$, όπου n_h είναι ο αριθμός των πειραματικών εκτελέσεων που επιλέγονται από τον απλό γενετικό αλγόριθμο.

Βήμα 5. Χρησιμοποιούμε ως σημείο διασταύρωσης το σημείο 1 (1-point crossover) και κρατάμε ενήμερα τα επιλεγμένα χαρακτηριστικά από τους κανόνες GRI χωρίς όμως να αλλάζουμε τις τιμές τους.

Βήμα 6. Ξεκινάμε μια ευρετική αναζήτηση, χρησιμοποιώντας ως αρχική κατάσταση τους επιλεγμένους παράγοντες του βέλτιστου σχεδιασμού. Ειδικότερα, καθοδηγούμε την ευρετική μέθοδο να αναζητήσει για k παράγοντες όπου ένα υποσύνολο από αυτούς διατηρείται πάντα σταθερό και οι υπόλοιποι είναι τυχαιοποιημένοι.

Βήμα 7. Τέλος, ανακτούμε τους υπόλοιπους παράγοντες του βέλτιστου σχεδιασμού σύμφωνα με το κριτήριο βελτιστοποίησης r_{max}. Μετά την υλοποίηση του απλού γενετικού αλγορίθμου παράγεται ένας βέλτιστος σχεδιασμός, ο οποίος ανήκει στην κλάση των υπερκορεσμένων σχεδιασμών, όταν ανιχνεύεται τιμή για το κριτήριο $r_{max} < 1$.

10.4 Εφαρμογή και Πειραματικά Αποτελέσματα

Σε αυτή την ενότητα εξετάζουμε την απόδοση της προτεινόμενης μεθόδου επιλογής βέλτιστου σχεδιασμού, υπό την παρουσία της βάσης των πραγματικών δεδομένων.

10.4.1 Περιγραφή της Βάσης Δεδομένων

Τα πραγματικά δεδομένα που θεωρούμε στη μελέτη μας προέρχονται από μία ετήσια έρευνα που διεξήχθη κατά την περίοδο 01/01/2005-31/12/2006 από την Ελληνική Εταιρεία Τραύματος και την Ομάδα Εντατικής Χειρουργικής, περιλαμβάνοντας 30 Γενικά Νοσοκομεία στην Ελλάδα, με βασικό σκοπό τη μελέτη των επιδράσεων διαφόρων προγνωστικών παραγόντων στη θεραπεία τραυματισμένων ατόμων. Η βάση δεδομένων αποτελείται από 44 δίτιμες επεξηγηματικές μεταβλητές που αποτελούν τους προγνωστικούς παράγοντες. Οι παράγοντες αυτοί που καταγράφηκαν περιελάμβαναν δημοφραφικά στοιχεία των ασθενών, στοιχεία που αφορούν τις συνθήκες του ατυχήματος και πληροφορίες για την ενδονοσοκομειακή αντιμετώπιση των ασθενών. Επίσης, η βάση δεδομένων αποτελείται από 8862 ασθενείς, και για κάθε έναν από αυτούς καταγράφηκε η κατάστασή τους κατά την τελευταία παρακολούθηση, με τη δίτιμη μεταβλητή απόκρισης y να παίρνει μόνο δύο τιμές, με το -1 να συμβολίζει την

"επιβίωση" (8413 ασθενείς) και το 1 την "αποβίωση" (449 ασθενείς), αντίστοιχα. Σύμφωνα με ιατρικές συμβουλές, όλοι οι προγνωστικοί παράγοντες θα πρέπει να τυγχάνουν ίσης μεταχείρισης κατά τη στατιστική ανάλυση και δεν υπάρχει κανένας παράγοντας για τον οποίο επιβάλλεται να διατηρείται πάντα στο μοντέλο. Το σύνολο των δεδομένων διαχωρίστηκε σε σύνολο εκπαίδευσης (90%) ($n = 7975$) και εξέτασης (10%) ($n = 887$), προκειμένου να αξιολογηθεί η απόδοση της πρόβλεψης πάνω σε νέα δεδομένα. Το σύνολο δεδομένων Τραύματος που χρησιμοποιήθηκε παρουσιάζεται στον Πίνακα 10.1.

Πίνακας 10.1: Μελέτη Τραύματος

Μεταβλητές
x_1: ambulance (-1 = no, 1 = yes)
x_2: comorbidities (-1 = no, 1 = yes)
x_3: sex (-1 = female, 1 = male)
x_4: evacuation (-1 = no, 1 = yes)
x_5: major doctor (-1 = no, 1 = yes)
x_6: A.T.L.S (-1 = no, 1 = yes)
x_7: capillary refill (-1 = no, 1 = yes)
x_8: pale (-1 = no, 1 = yes)
x_9: intrahospital transport (-1 = no, 1 = yes)
x_{10}: dysphoria (-1 = no, 1 = yes)
x_{11}: central cyanosis (-1 = no, 1 = yes)
x_{12}: peritoneum points (-1 = no, 1 = yes)
x_{13}: oxygen (-1 = no, 1 = yes)
x_{14}: intubation (-1 = no, 1 = yes)
x_{15}: mechanical ventilation (-1 = no, 1 = yes)
x_{16}: cardiopulmonary resuscitation (-1 = no, 1 = yes)
x_{17}: chest drainage (-1 = no, 1 = yes)
x_{18}: pericardiocentesis (-1 = no, 1 = yes)
x_{19}: catheter (-1 = no, 1 = yes)
x_{20}: nasogastric tube (-1 = no, 1 = yes)
x_{21}: collar (-1 = no, 1 = yes)
x_{22}: spinal immobilisation (-1 = no, 1 = yes)
x_{23}: pelvic immobilisation (-1 = no, 1 = yes)
x_{24}: limb immobilisation (-1 = no, 1 = yes)
x_{25}: fluids (-1 = no, 1 = yes)
x_{26}: blood (-1 = no, 1 = yes)
x_{27}: ICP monitoring (-1 = no, 1 = yes)
x_{28}: thoracotomy (-1 = no, 1 = yes)
x_{29}: angiography (-1 = no, 1 = yes)
x_{30}: embolism (-1 = no, 1 = yes)
x_{31}: diagnostic peritoneal lavage (DPL) (-1 = no, 1 = yes)
x_{32}: gases (vacuum phenomenon) (-1 = no, 1 = yes)
x_{33}: Radiograph E.R. (-1 = no, 1 = yes)
x_{34}: computed tomography (CT) (-1 = no, 1 = yes)
x_{35}: ultrasound (US) (-1 = no, 1 = yes)
x_{36}: urea testing (-1 = no, 1 = yes)
x_{37}: toxicology testing (-1 = no, 1 = yes)
x_{38}: surgical intervention (-1 = no, 1 = yes)
x_{39}: intrahospital CT (-1 = no, 1 = yes)
x_{40}: intrahospital US (-1 = no, 1 = yes)
x_{41}: intrahospital M.R.I (-1 = no, 1 = yes)
x_{42}: intrahospital angiography (-1 = no, 1 = yes)
x_{43}: complications (-1 = no, 1 = yes)
x_{44}: Intensive Care Unit (ICU) stay (-1 = no, 1 = yes)

10.4.2 Κριτήρια Αξιολόγησης της Απόδοσης

Η αξιολόγηση της αξιοπιστίας ενός ταξινομητή είναι απαραίτητη για τη διασφάλιση της ποιότητας των δεδομένων. Το πλέον συνηθισμένο κριτήριο για την αξιολόγηση της ποιότητας ενός μοντέλου ταξινόμησης είναι το κριτήριο της διάκρισης, το οποίο μετρά το κατά πόσο καλά διαχωρίζονται οι δύο κλάσεις στο σύνολο δεδομένων. Στη μελέτη μας θεωρούμε τα πιο συχνά χρησιμοποιούμενα μέτρα διάκρισης για την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου, και υιοθετούμε τους κλασικούς ορισμούς που χρησιμοποιούνται στην δυαδική ταξινόμηση [228].

Δοθέντος ενός ταξινομητή και μίας καταχώρησης, υπάρχουν τέσσερις πιθανές εκβάσεις. Οι θετικές καταχωρήσεις οι οποίες προβλέπονται σωστά ως θετικές (True Positive-TP), οι θετικές καταχωρήσεις οι οποίες προβλέπονται εσφαλμένα ως αρνητικές (False Negative-FN), οι αρνητικές καταχωρήσεις οι οποίες προβλέπονται εσφαλμένα ως θετικές (False Positive-FP), και τέλος οι αρνητικές καταχωρήσεις οι οποίες προβλέπονται σωστά ως αρνητικές (True Negative-TN).

Η ακρίβεια ταξινόμησης (accuracy) χρησιμοποιείται ως πρώτο κριτήριο. Η ακρίβεια ορίζεται ως το ποσοστό των σωστά ταξινομημένων καταχωρήσεων στο σύνολο εξέτασης. Τα άλλα δύο κριτήρια που χρησιμοποιούνται είναι η ευαισθησία και η ειδικότητα που είναι δύο στατιστικά μέτρα της απόδοσης ενός τεστ δυαδικής ταξινόμησης, και συνδέονται στενά με τις έννοιες των σφαλμάτων Τύπου I και Τύπου II. Η ευαισθησία (sensitivity) μετρά το ποσοστό των πραγματικά θετικών που έχουν αναγνωριστεί σωστά ως θετικά, ενώ η ειδικότητα (specificity) μετρά το ποσοστό των πραγματικά αρνητικών που έχουν αναγνωριστεί σωστά ως αρνητικά. Τα άλλα κριτήρια που χρησιμοποιούνται στην ανάκτηση πληροφοριών είναι η ανάκληση (recall) η οποία αντιστοιχεί στην ευαισθησία και η ακρίβεια (precision) που είναι το ποσοστό των αληθώς θετικών μεταξύ όλων των προβλεπόμενων θετικών. Στα προβλήματα ταξινόμησης, ένα άλλο ευρέως διαδεδομένο μέτρο απόδοσης είναι ο γεωμετρικός μέσος των τιμών ακρίβειας των κλάσεων (G-mean), ο οποίος θέτει όλες τις κατηγορίες επί ίσοις όροις, και δεν δίνει μεγαλύτερη προτεραιότητα στις σπάνιες θετικές κλάσεις. Ένα μέτρο απόδοσης που επιτρέπει κάτι τέτοιο είναι το F-μέτρο (F-measure) το οποίο δεν λαμβάνει υπόψη του την απόδοση των αρνητικών κλάσεων.

Σε προβλήματα δυαδικής ταξινόμησης, τα προαναφερθέντα μέτρα απόδοσης ορίζονται ως ακολούθως:

- Accuracy= $\frac{TP+TN}{TP+FP+TN+FN}$,
- Sensitivity (Recall) = $\frac{TP}{TP+FN}$,
- Specificity = $\frac{TN}{TN+FP}$,
- Precision = $\frac{TP}{TP+FP}$,
- G-mean = $\sqrt{\text{sensitivity} * \text{specificity}}$,
- F-measure= $\frac{\text{precision}*\text{recall}}{\beta\text{precision}+(1-\beta)\text{recall}}$,

όπου η παράμετρος β με $0 < \beta < 1$ επιτρέπει στον πειραματιστή να εκχωρήσει τα σχετικά βάρη στην ακρίβεια και στην ανάκληση, με το 0.5 να τους αποδίδει την ίδια σημασία.

Ένα άλλο δημοφιλές στατιστικό εργαλείο είναι οι καμπύλες λειτουργικού χαρακτηριστικού δέκτη (ROC) [179] οι οποίες εξ'ορισμού χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός συστήματος με διχοτομικά εξαγόμενα αποτελέσματα. Μια καμπύλη ROC παρουσιάζεται ως το γράφημα της ευαισθησίας έναντι του 1-ειδικότητα για όλες τις πιθανές τιμές αποκοπής. Παραδοσιακά, το εμβαδόν κάτω από την καμπύλη ROC (AUC) χρησιμοποιείται ως συνοπτικός δείκτης ακρίβειας ενός τεστ [105] και είναι χρήσιμο ως περιγραφικό μέτρο της συνολικής απόδοσης ενός τεστ. Στατιστικά μιλώντας, το εμβαδόν κάτω από την καμπύλη ROC (AUC) ενός ταξινομητή είναι η πιθανότητα ένας ταξινομητής να κατατάξει ένα τυχαία επιλεγμένο θετικό υπόδειγμα υψηλότερα από ένα τυχαία επιλεγμένο αρνητικό υπόδειγμα.

Σε προβλήματα λήψης αποφάσεων, τα λάθη διαφόρων τύπων πρέπει να σταθμιστούν σε σχέση με το κόστος. Στους σχεδιασμούς πειραμάτων, υπάρχει ένα κόστος του να δηλώσουμε έναν ανενεργό παράγοντα ως ενεργό (σφάλμα Τύπου I), καθώς επίσης και ένα κόστος του να δηλώσουμε ένα ενεργό παράγοντα ως ανενεργό (σφάλμα Τύπου II). Η ευαισθησία και η ειδικότητα μπορούν εναλλακτικά να περιγραφούν ως εξής:

- Ευαισθησία=$P(\hat{y} = 1|y = 1)$= 1-σφάλμα Τύπου ΙΙ,
- Ειδικότητα=$P(\hat{y} = -1|y = -1)$=1-σφάλμα Τύπου Ι.

10.4.3 Επιλογή Υπερκορεσμένων Σχεδιασμών

Για την αρχικοποίηση των τυχαίων χρωμοσωμάτων που χρησιμοποιήθηκαν στη δεξαμενή ζευγαρώματος του απλού γενετικού αλγορίθμου, χρησιμοποιήσαμε τους ακόλουθους GRI κανόνες συσχέτισης που παρουσιάζονται στον Πίνακα 10.2.

Πίνακας 10.2: Κανόνες συσχέτισης

GRI	Rules Antecedents	Consequent	Support	Confidence
If	x1 = 1.0 and x3 = 1.0 and x14 = 1.0 and x15 = 1.0 and x16 = 1.0	then y=1.0	0.21	94.12
If	x14 = 1.0 and x16 = 1.0	then y=1.0	0.49	94.87
If	x5 = 1.0 and x15 = 1.0 and x16 = 1.0	then y=1.0	0.21	94.12

Με αυτό τον τρόπο επιλέξαμε τις τρεις αρχικές πειραματικές εκτελέσεις για το βέλτιστο σχεδιασμό, ταιριάζοντας τους τρεις παραγόμενους GRI κανόνες συσχέτισης με τα αντίστοιχα πεδία της βάσης δεδομένων. Ο βέλτιστος υπερκορεσμένος σχεδιασμός που ανακτήθηκε από τη βάση δεδομένων μέσω των κανόνων GRI και του απλού γενετικού αλγορίθμου, περιλαμβάνει 6 πειραματικές εκτελέσεις και 8 παράγοντες, οι οποίοι είναι οι $x8$: χλωμότητα, $x5$: βασικός γιατρός, $x7$: επαναπλήρωση των τριχοειδών, $x13$: παροχή οξυγόνου, $x6$: ATLS, $x2$: συνοδά νοσήματα, $x14$: διασωλήνωση, $x11$: κεντρική κυάνωση, και λαμβάνουν ως τιμές τις -1 = όχι, 1 = ναι. Ο Πίνακας 10.3 παρουσιάζει το βέλτιστο υπερκορεσμένο σχεδιασμό που ανακτήθηκε από τη βάση δεδομένων.

Πίνακας 10.3: Ο 1ος βέλτιστος υπερκορεσμένος σχεδιασμός

record ID	x_8	x_5	x_7	x_{13}	x_6	x_2	x_{14}	x_{11}	y
968	1	-1	-1	1	-1	-1	1	-1	1
936	-1	1	-1	1	-1	-1	1	-1	1
5587	1	1	-1	1	-1	-1	1	1	1
7344	-1	1	-1	1	-1	-1	-1	-1	-1
3806	-1	1	-1	-1	-1	1	-1	-1	-1
2530	-1	-1	-1	-1	1	-1	-1	-1	-1

Ο σχεδιασμός που ανακτήθηκε από τις [x_8,x_5,x_7,x_{13},x_6,x_2,x_{14},x_{11}] έχει τιμή για το κριτήριο $r_{max} = 0.7071$.

Σημειώνουμε εδώ, ότι ο απλός γενετικός αλγόριθμος επίσης εντόπισε έναν ακόμα βέλτιστο σχεδιασμό, ο οποίος ανήκει στην κλάση των υπερκορεσμένων σχεδιασμών, με τιμή για το $r_{max} < 1$. Παρακάτω παρουσιάζουμε στον Πίνακα 10.4 το δεύτερο βέλτιστο σχεδιασμό που επιλέχθηκε μέσω των κανόνων GRI και του απλού γενετικού αλγορίθμου, αποτελούμενο επίσης από 6 πειραματικές εκτελέσεις και 8 παράγοντες.

Πίνακας 10.4: Ο 2ος βέλτιστος υπερκορεσμένος σχεδιασμός

record ID	x_8	x_5	x_7	x_{13}	x_6	x_2	x_{16}	x_{11}	y
968	1	-1	-1	1	-1	-1	1	-1	1
936	-1	1	-1	1	-1	-1	1	-1	1
5587	1	1	-1	1	-1	-1	1	1	1
7344	-1	1	-1	1	-1	-1	-1	-1	-1
3806	-1	1	-1	-1	-1	1	-1	-1	-1
2530	-1	-1	-1	-1	1	-1	-1	-1	-1

Ο σχεδιασμός που ανακτήθηκε από τις [x_8,x_5,x_7,x_{13},x_6,x_2,x_{16},x_{11}] έχει τιμή για το κριτήριο $r_{max} = 0.7071$.

Αξίζει να σημειωθεί εδώ ότι η στήλη x_7, και στους δύο πίνακες σχεδιασμού, είναι πλήρως ταυτόσημη με το μέσο αφού όλες οι τιμές της είναι ίσες με -1. Ως εκ τούτου, η επίδραση

της είναι αδύνατον να εκτιμηθεί. Το γεγονός αυτό δεν αποτελεί πρόβλημα για τη μελέτη μας, η οποία είναι σχεδιασμένη έτσι ώστε να εξυπηρετεί μόνο σκοπούς κρησαρίσματος. Επίσης παρατηρούμε ότι η μόνη διαφορά μεταξύ του πρώτου και δεύτερου βέλτιστου σχεδιασμού που ανακτώνται από τη βάση δεδομένων, είναι η διαφορετική επιλογή της 7-ης στήλης, δηλαδή της x_{14} αντί της x_{16}, αντίστοιχα. Η μεταβλητή x_{16} συμβολίζει την καρδιοπνευμονική ανάνηψη ($-1 =$ όχι, $1 =$ ναι). Δεδομένου ότι τα πρόσημα ± 1 κατανέμονται πανομοιότυπα τόσο στη στήλη x_{14} όσο και στη στήλη x_{16} των αντίστοιχων πινάκων σχεδιασμού, τα ληφθέντα αποτελέσματα θα είναι εντελώς όμοια είτε θεωρήσουμε προς χρήση τον πρώτο ή το δεύτερο βέλτιστο σχεδιασμό για περαιτέρω ανάλυση. Το γεγονός ότι και οι δύο σχεδιασμοί που ανακτώνται από τη βάση δεδομένων, έχουν ακριβώς τα ίδια πρόσημα ± 1 να κατανέμονται σε κάθε στήλη του πίνακα σχεδιασμού με τον ίδιο ακριβώς τρόπο είναι ιδιαίτερα σημαντικό, δεδομένου ότι παρέχει ισχυρά αποδεικτικά στοιχεία ότι ο επιθυμητός βέλτιστος σχεδιασμός πράγματι ανιχνεύεται σωστά (είτε ο σχεδιασμός σχηματίζεται από τις $[x_8,x_5,x_7,x_{13},x_6,x_2,x_{14},x_{11}]$ είτε από τις $[x_8,x_5,x_7,x_{13},x_6,x_2,x_{16},x_{11}]$).

10.4.4 Συγκριτικά Αποτελέσματα

Η L_1-norm SVM μεθοδολογία χρησιμοποιείται για την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου. Αρχικά εκτελούμε την L_1-norm SVM μέθοδο σε όλο το σύνολο δεδομένων τραύματος, αποτελούμενο από τους 8862 ασθενείς και τους 44 πιθανούς παράγοντες κινδύνου (*Model I*). Το γεγονός ότι η L_1-norm SVM είναι γρήγορη στην εκπαίδευση, την ταξινόμηση και δεν είναι υπολογιστικά ακριβή και χρονοβόρα, μας επέτρεψε να εφαρμόσουμε αυτή τη μέθοδο σε ολόκληρο το διαθέσιμο μεγάλης διάστασης σύνολο δεδομένων. Έπειτα εκτελούμε την L_1-norm SVM χρησιμοποιώντας ως πίνακα σχεδιασμού το σχεδιασμό $SSD_{(6,8)}$ που ανιχνεύθηκε μέσω της προτεινόμενης μεθόδου, και ως απόκριση τις αντίστοιχες ετικέτες της κλάσης εξόδου (*Model II*). Τα ληφθέντα αποτελέσματα παρουσιάζονται στον Πίνακα 10.5.

Πίνακας 10.5: Συγκριτική απόδοση μοντέλων

Κριτήριο	*Model I*	*Model II*
Training error	0.04%	0.01%
Accuracy	96%	99%
Sensitivity	20%	100%
Recall	20%	100%
Specificity	99%	100%
Precision	68%	100%
G-mean	45%	100%
F-measure	31%	100%
AUC	0.57	0.88

Ένα τέλειος προγνωστικός παράγοντας θα μπορούσε να περιγραφεί ως 100% "ευαίσθητος" και 100% "ειδικός". Η ευαισθησία και η ειδικότητα αφορούν την ικανότητα ενός τεστ να εντοπίσει σωστά τα θετικά και αρνητικά αποτελέσματα, αντιστοίχως. Ένα τεστ με υψηλή ευαισθησία και υψηλή ειδικότητα μπορεί να θεωρηθεί αξιόπιστο δεδομένου ότι έχει ένα πολύ χαμηλό ποσοστό σφάλματος Τύπου ΙΙ και ένα πολύ χαμηλό ποσοστό σφάλματος Τύπου Ι, αντίστοιχα.

Η L_1-norm SVM (*Model I*) προσεγγίζει το ποσοστό του 99% για την ειδικότητα, πράγμα που σημαίνει ότι ο ταξινομητής αναγνωρίζει σχεδόν όλα τα πραγματικά αρνητικά. Με άλλα λόγια αυτό σημαίνει ότι έχει ένα πολύ χαμηλό ποσοστό σφάλματος Τύπου Ι. Αυτό το μέτρο από μόνο του δεν απεικονίζει το κατά πόσο καλά ο ταξινομητής αναγνωρίζει τις θετικές περιπτώσεις και γι'αυτό είναι αναγκαίο να ληφθεί υπόψη και η ευαισθησία του χρησιμοποιούμενου ταξινομητή. Όταν η L_1-norm SVM (*Model I*) αξιολογείται με βάση την ευαισθησία, έχει ένα σαφές μειονέκτημα παρουσιάζοντας χαμηλότερα ποσοστά, πράγμα που σημαίνει ότι τα ποσοστά σφάλματος Τύπου ΙΙ είναι υψηλότερα. Γενικά, ο Πίνακας 10.5 υποδεικνύει

ότι η L_1-norm SVM (*Model I*) τείνει να δηλώνει σε ένα χαμηλότερο ποσοστό ανενεργούς παράγοντες ως ενεργούς, και σε υψηλότερο ποσοστό ενεργούς παράγοντες ως ανενεργούς. Το εμβαδόν κάτω από την καμπύλη λειτουργικού χαρακτηριστικού δέκτη (AUC) για το *Model I* έχει σημαντικά χαμηλότερη τιμή ($\text{AUC}_{Model\ I}$=0.57) σε σύγκριση με το *Model II* ($\text{AUC}_{Model\ II}$=0.88).

Όταν η L_1-norm SVM εφαρμόζεται στην προτεινόμενη μέθοδο (*Model II*) έχει σαφώς καλύτερη ακρίβεια ταξινόμησης, ευαισθησία και ειδικότητα που προσεγγίζει σχεδόν το απόλυτο ποσοστό του 100%. Σε γενικές γραμμές, ο Πίνακας 10.5 δείχνει ότι η προτεινόμενη μέθοδος (*Model II*) έχει πολύ χαμηλά σφάλματα Τύπου I (το ποσοστό σφάλματος Τύπου I είναι σχεδόν μηδενικό, και αυτό αντιστοιχεί σε περιπτώσεις όπου σχεδόν κανένας από τους ανενεργούς παράγοντες δεν δηλώνεται ως ενεργός) και πολύ χαμηλά σφάλματα Τύπου II (το ποσοστό σφάλματος Τύπου II είναι σχεδόν μηδενικό, και αυτό αντιστοιχεί σε περιπτώσεις όπου σχεδόν κανένας από τους ενεργούς παράγοντες δεν δηλώνεται ως ανενεργός). Ως εκ τούτου, η προτεινόμενη μέθοδος παρουσιάζει μια σταθερή συμπεριφορά υπό αυτή την έννοια.

Οι τιμές ευαισθησίας και ειδικότητας από μόνες τους δεν μπορούν να χρησιμοποιηθούν για να προσδιορίσουν αν μια δοκιμασία είναι χρήσιμη στην πράξη, και μπορεί να είναι παραπλανητικές. Στις ιατρικές διαγνώσεις, η ευαισθησία είναι η ικανότητα του τεστ να εντοπίσει σωστά τα άτομα με τη νόσο (ποσοστό αληθινών θετικών), ενώ η ειδικότητα είναι η ικανότητα του τεστ να εντοπίσει σωστά τα άτομα χωρίς τη νόσο (ποσοστό αληθινών αρνητικών). Το χειρότερο σενάριο για την ευαισθησία και την ειδικότητα πρέπει να υπολογίζεται, προκειμένου να αποφευχθεί η εξάρτηση από πειράματα με λίγα και αμφίβολα αποτελέσματα. Ένας συνήθης τρόπος για να γίνει αυτό είναι να υπολογίσουμε τα διωνυμικά διαστήματα εμπιστοσύνης ποσοστού για την ευαισθησία και ειδικότητα, δίνοντας το εύρος των τιμών εντός του οποίου η σωστή τιμή βρίσκεται σε ένα συγκεκριμένο επίπεδο εμπιστοσύνης (95%). Ο Πίνακας 10.6 δείχνει ότι οι τιμές της ευαισθησίας και της ειδικότητας τόσο για το *Model I* όσο και για το *Model II* είναι ικανοποιητικές, δεδομένου ότι οι εκτιμώμενες τιμές βρίσκονται μέσα στο αντίστοιχο εκτιμώμενο διάστημα εμπιστοσύνης. Πρέπει να σημειωθεί εδώ ότι τα τρία βασικά πράγματα που συνήθως επηρεάζουν το εύρος του διαστήματος εμπιστοσύνης είναι το επίπεδο εμπιστοσύνης, η μεταβλητότητα και το μέγεθος του δείγματος. Το γεγονός ότι τα διαστήματα εμπιστοσύνης στον Πίνακα 10.6, για το *Model II*, είναι πολύ ευρύτερα απ'ότι για το *Model I*, δεν αποτελεί έκπληξη δεδομένου ότι τα μικρότερα μεγέθη δείγματος παράγουν ευρύτερα διαστήματα.

Πίνακας 10.6: Διαστήματα εμπιστοσύνης (95%)

	Sensitivity			Specificity		
Μοντέλο	Εκτιμώμενη Τιμή	Lower Limit	Upper Limit	Εκτιμώμενη Τιμή	Lower Limit	Upper Limit
Model I	0.994891	0.993057	0.996255	0.204036	0.168201	0.245076
Model II	0.999999	0.309989	1.000000	0.999999	0.309989	1.000000

Η L_1-norm SVM όταν εφαρμόζεται σε όλο το σύνολο δεδομένων (*Model I*) ανίχνευσε ένα σύνολο 33 μεταβλητών από τις 44 αρχικά διαθέσιμες μεταβλητές, ως στατιστικά σημαντικές. Το παραγόμενο *Model I* απέκλεισε 11 μεταβλητές ως μη σημαντικές, τις x_{19}, x_{20}, x_{23}, x_{24}, x_{29}, x_{30}, x_{31}, x_{38}, x_{40}, x_{42}, x_{43}. Η L_1-norm SVM όταν εφαρμόζεται στην προτεινόμενη μέθοδο (*Model II*) επίσης κατόρθωσε να αποκλείσει αυτές τις 11 μεταβλητές ως μη σημαντικές. Επιπλέον, η προτεινόμενη μέθοδος (*Model II*) κατόρθωσε να αποκλείσει περισσότερες μη σημαντικές μεταβλητές, οδηγώντας έτσι σε ένα πιο φειδωλό μοντέλο σε σύγκριση με το *Model I*. Η προτεινόμενη μέθοδος (*Model II*) ανίχνευσε ένα σύνολο 8 μεταβλητών από τις 44 αρχικά διαθέσιμες μεταβλητές, ως στατιστικά σημαντικές, τις x_2, x_5, x_6, x_7, x_8, x_{11}, x_{13}, x_{14} (ή x_{16}). Δεδομένου ότι η απόφαση μεταξύ των μεταβλητών x_{14} και x_{16} φαίνεται να είναι αυθαίρετη και προκειμένου να αποφευχθεί ο αποκλεισμός μιας πιθανής σημαντικής μεταβλητής, πραγματοποιήσαμε περαιτέρω ανάλυση εξετάζοντας τρία νέα μοντέλα, συμπεριλαμβάνοντας

αρχικά μόνο την x_{14}, έπειτα μόνο την x_{16}, και τέλος ένα μοντέλο, συμπεριλαμβανομένων και των δύο μεταβλητών x_{14} και x_{16}. Τα ληφθέντα αποτελέσματα παρουσιάζονται στην επόμενη ενότητα.

10.4.5 Περαιτέρω Ανάλυση

Η L_1-norm SVM μεθοδολογία χρησιμοποιείται επίσης για την αξιολόγηση της απόδοσης των νέων μοντέλων που θεωρούνται για μεταγενέστερη ανάλυση. Αρχικά εκτελούμε την L_1-norm SVM μέθοδο στους $m = 8$ πιθανούς παράγοντες κινδύνου που ανιχνεύθηκαν από τον 1ο σχεδιασμό, δηλαδή στους $[x_8, x_5, x_7, x_{13}, x_6, x_2, x_{14}, x_{11}]$ και στους $n = 8862$ ασθενείς (*Model A*). Έπειτα εκτελούμε την L_1-norm SVM μέθοδο στους $m = 8$ πιθανούς παράγοντες κινδύνου που ανιχνεύθηκαν από τον 2ο σχεδιασμό, δηλαδή στους $[x_8, x_5, x_7, x_{13}, x_6, x_2, x_{16}, x_{11}]$ και στους $n = 8862$ ασθενείς (*Model B*). Τέλος, εκτελούμε την L_1-norm SVM μέθοδο στους $m = 9$ πιθανούς παράγοντες κινδύνου που ανιχνεύθηκαν και από τους δύο σχεδιασμούς, δηλαδή στους $[x_8, x_5, x_7, x_{13}, x_6, x_2, x_{14}, x_{16}, x_{11}]$ και στους $n = 8862$ ασθενείς (*Model C*). Τα ληφθέντα αποτελέσματα παρουσιάζονται στους Πίνακες 10.7 και 10.8.

Πίνακας 10.7: Συγκριτική απόδοση μοντέλων

Κριτήριο	*Model A*	*Model B*	*Model C*
Training error	0.04976 %	0.04513 %	0.04513 %
Accuracy	0.95024 %	0.95486 %	0.95486 %
Sensitivity	0.02017 %	0.11211 %	0.11435 %
Recall	0.02017 %	0.11211 %	0.11435 %
Specificity	0.99952 %	0.99952 %	0.99941 %
Precision	0.69231 %	0.92593 %	0.91071 %
G-mean	0.14202 %	0.33475 %	0.33806 %
F-measure	0.03921 %	0.20000 %	0.20319 %
AUC	0.70096	0.69024	0.68281

Πίνακας 10.8: Διαστήματα εμπιστοσύνης (95%)

	Sensitivity			Specificity		
Μοντέλο	Εκτιμώμενη Τιμή	Lower Limit	Upper Limit	Εκτιμώμενη Τιμή	Lower Limit	Upper Limit
Model A	0.02017	0.009879	0.039343	0.99952	0.998694	0.999848
Model B	0.11211	0.085092	0.145999	0.99952	0.998694	0.999848
Model C	0.11435	0.087064	0.148470	0.99941	0.998527	0.999781

Παρατηρούμε από τον Πίνακα 10.7 ότι το *Model B* ή το *Model C* υπερέχουν ξεκάθαρα σε σύγκριση με το *Model A* όσον αφορά την απόδοσή τους, σύμφωνα με τις τιμές του σφάλματος εκπαίδευσης, της ευαισθησίας, της ανάκλησης, της ακρίβειας, του γεωμετρικού μέσου και του F-μέτρου. Τα *Model B* και *Model C* έχουν παρόμοια απόδοση με μικρές διαφορές στις τιμές των εξεταζόμενων κριτηρίων. Επιπλέον, παρατηρούμε από τον Πίνακα 10.8 ότι οι τιμές της ευαισθησίας και της ειδικότητας και για τα τρία μοντέλα είναι ικανοποιητικές, δεδομένου ότι οι εκτιμώμενες τιμές βρίσκονται μέσα στο αντίστοιχο εκτιμώμενο διάστημα εμπιστοσύνης. Δεδομένου ότι η απόφαση της επιλογής μεταξύ του *Model B* (που περιλαμβάνει μόνο την x_{16}) ή του *Model C* (που περιλαμβάνει τις x_{14} και x_{16}) δεν φαίνεται να είναι σαφής και μπορεί να είναι αυθαίρετη λόγω της παρόμοιας απόδοσης και των δύο αυτών μοντέλων, προτείνουμε την επιλογή του *Model C* ως το "καλύτερο" παραγόμενο μοντέλο προκειμένου να αποφευχθεί ο αποκλεισμός μιας πιθανής σημαντικής μεταβλητής.

10.5 Συμπεράσματα

Ο καινοτόμος χαρακτήρας της παραπάνω μελέτης έγκειται στη χρήση της κλάσης των υπερκορεσμένων σχεδιασμών, σε συνδυασμό με μεθόδους εξόρυξης δεδομένων και γενετικών

αλγορίθμων, που μας επέτρεψε να αντιμετωπίσουμε το πρόβλημα της επιλογής μεταβλητών σε μια μεγάλης διάστασης βάση δεδομένων με μια εφικτή υπολογιστική προσπάθεια. Η προτεινόμενη μέθοδος κατόρθωσε να δηλώνει σε ένα πολύ χαμηλό ποσοστό, ανενεργούς παράγοντες ως ενεργούς, και ενεργούς παράγοντες ως ανενεργούς, αντίστοιχα. Η προτεινόμενη μέθοδος χρησιμοποιώντας τους υπερκορεσμένους σχεδιασμούς μπορεί να αποδειχθεί πολύ χρήσιμη για τη στατιστική ανάλυση των μεγάλων δεδομένων, δεδομένου ότι μας επέτρεψε να προσδιορίσουμε αποτελεσματικά και φειδωλά τους σημαντικούς προγνωστικούς παράγοντες (9 στατιστικά σημαντικές από τις αρχικά 44 διαθέσιμες μεταβλητές πρόβλεψης) χρησιμοποιώντας μόνο λίγες πειραματικές εκτελέσεις (6 εκτελέσεις από τις αρχικά 8862 διαθέσιμες εκτελέσεις). Προτείνουμε αυτή την εναλλακτική προσέγγιση με σκοπό την επιλογή μεταβλητών δοθείσης μιας βάσης δεδομένων παρατηρήσεων, διότι επιτρέπει στον πειραματιστή να χρησιμοποιεί μόνο ένα πολύ μικρό ποσοστό των διαθέσιμων εκτελέσεων, γεγονός που κάνει τη στατιστική ανάλυση μιας μεγάλης βάσης δεδομένων υπολογιστικά εφικτή και οικονομικά προσιτή. Η προτεινόμενη προσέγγιση μπορεί να προτιμηθεί από πειραματιστές που θεωρούν υπολογιστικά ακριβό το να εξετάσουν όλα τα διαθέσιμα επίπεδα των παραγόντων που πιθανώς μπορεί να συνδέονται με υψηλό κόστος και θα ήθελαν να ανιχνεύσουν και να εξετάσουν μόνο συγκεκριμένα επίπεδα ορισμένων παραγόντων, προκειμένου να ελαχιστοποιηθεί το πειραματικό κόστος.

Μέρος **III**

Στατιστική Ανάλυση Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις

ΚΕΦΑΛΑΙΟ **11**

Εισαγωγικά Στοιχεία και Βασικές Έννοιες

Logic merely sanctions
the conquests of the intuition.

—*Jacques S. Hadamard (1865–1963)*

Στο ενδέκατο αυτό κεφάλαιο θα αναφερθούμε στην ύπαρξη των αλληλεπιδράσεων στους υπερκορεσμένους σχεδιασμούς. Οι περισσότερες μέθοδοι για την κατασκευή και την ανάλυση των υπερκορεσμένων σχεδιασμών θεωρούν μόνο τις κύριες επιδράσεις και αγνοούν ή θεωρούν αμελητέα την ύπαρξη αλληλεπιδράσεων. Θα παρουσιάσουμε μια σύντομη ιστορική αναδρομή σχετικά με το τι έχει προταθεί βιβλιογραφικά έως τώρα για την κατασκευή και την ανάλυση των υπερκορεσμένων σχεδιασμών συμπεριλαμβανομένων των αλληλεπιδράσεων. Επίσης, θα παρουσιάσουμε τις βασικές αρχές οι οποίες διέπουν τις μεθόδους κρησαρίσματος των υπερκορεσμένων σχεδιασμών με αλληλεπιδράσεις.

11.1 Ιστορική Αναδρομή

Υπερκορεσμένοι σχεδιασμοί δύο επιπέδων ονομάζονται οι κλασματικοί παραγοντικοί σχεδιασμοί των οποίων ο αριθμός των παραγόντων m είναι μεγαλύτερος ή ίσος του αριθμού των πειραματικών εκτελέσεων n, δηλαδή $m \geq n$ ή $m > n - 1$. Η πρόσφατη έρευνα έχει επικεντρωθεί στην κλάση των υπερκορεσμένων σχεδιασμών, γεγονός που οφείλεται στις νέες μαθηματικές τους ιδιότητες και στο οικονομικό μέγεθος των πειραματικών εκτελέσεών τους. Οι περισσότερες από τις μεθόδους κατασκευής και ανάλυσης των υπερκορεσμένων σχεδιασμών θεωρούν μόνο τις κύριες επιδράσεις και αγνοούν την ύπαρξη αλληλεπιδράσεων. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στο Κεφάλαιο 1, στις ενότητες 1.3.1 και 1.3.2 όπου παρουσιάζεται αναλυτικά η προηγούμενη συνεισφορά για την κατασκευή και την ανάλυση των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων, αντίστοιχα. Επίσης, σημαντικές πρόσφατες ιστορικές ανασκοπήσεις αποτελούν οι εργασίες [92] και [123] σχετικά με τις μεθόδους κατασκευής των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων, καθώς και οι εργασίες [145], [98] και [91] σχετικά με τις μεθόδους ανάλυσης των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων. Από όσο καλά γνωρίζουμε, πολύ λίγη μελέτη έχει γίνει για την κατασκευή ή την ανάλυση των υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις.

Συγκεκριμένα, ο Wu [229] πρότεινε μία μέθοδο κατασκευής των υπερκορεσμένων σχεδιασμών μέσω μερικώς ταυτόσημων αλληλεπιδράσεων. Ο Lin [150] πρότεινε μία μέθοδο ανίχνευσης των σημαντικών αλληλεπιδράσεων, σε υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων, μέσω μιας κατάλληλης κατασκευής ενός υπερκορεσμένου σχεδιασμού. Οι Allen και Bernshteyn [4] πρότειναν κριτήρια για τη δημιουργία και την αξιολόγηση υπερκορεσμένων σχεδιασμών με υψηλή πιθανότητα εντοπισμού όλων των σημαντικών παραγόντων με χρήση γραφημάτων κύριων επιδράσεων και αλληλεπιδράσεων. Οι Liu, Ruan και Dean [154] μελέτησαν την κατασκευή και την ανάλυση των $E(s^2)$ αποτελεσματικών υπερκορεσμένων σχεδιασμών συμπεριλαμβανομένων των αλληλεπιδράσεων. Οι Jones, Lin και Nachtsheim [118] μελέτησαν τους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς στους οποίους κάποιοι από τους ενδεχόμενους όρους μπορεί να αντιστοιχούν σε αλληλεπιδράσεις δύο παραγόντων. Οι Draguljic et al. [68] μελέτησαν τεχνικές κρησαρίσματος υπερκορεσμένων σχεδιασμών υπό την παρουσία αλληλεπιδράσεων. Ο Cossari [52] προσάρμοσε τη μέθοδο των Box-Meyer [24] για την ανάλυση των υπερκορεσμένων σχεδιασμών, διερεύνησε την απόδοσή τους υπό την παρουσία των αλληλεπιδράσεων δύο παραγόντων, και κατέληξε στο συμπέρασμα ότι η μέθοδος Box-Meyer μπορεί να είναι αποτελεσματική όταν οι αλληλεπιδράσεις θεωρούνται στην ανάλυση των υπερκορεσμένων σχεδιασμών.

11.2 Τεχνικές Κρησαρίσματος με Αλληλεπιδράσεις

Σε πολλούς επιστημονικούς πειραματισμούς, ο κύριος στόχος είναι η ταυτόχρονη μελέτη ενός μεγάλου αριθμού επιδράσεων πολλών παραγόντων και ο εντοπισμός μόνο ενός μικρού υποσυνόλου αυτών που έχουν πραγματικά μια σημαντική επιρροή στη μεταβλητή απόκρισης. Ο πειραματιστής μπορεί να διεξάγει ένα πείραμα ελέγχου-κρησαρίσματος, προκειμένου να προσδιοριστούν οι σημαντικοί παράγοντες κατά τρόπο αποτελεσματικό, από άποψη κόστους και εργασίας. Οι Hamada και Wu [104], και πρόσφατα οι Phoa, Wong και Xu [183] επεσήμαναν την ανάγκη του να λαμβάνονται υπόψη οι σημαντικές αλληλεπιδράσεις στην ανάλυση των πειραματικών σχεδιασμών, δεδομένου ότι η αγνόησή τους μπορεί να επηρεάσει τη στατιστική συμπερασματολογία, να οδηγήσει σε απώλεια σημαντικών αποτελεσμάτων, να οδηγήσει σε μεροληπτικές εκτιμήσεις και να αυξήσει τη μεταβλητότητα στην απόκριση. Έτσι, η υπόθεση ότι όλες οι αλληλεπιδράσεις είναι υπαρκτές σε μια διαδικασία ανάλυσης ή κρησαρίσματος ενός σχεδιασμού διαδραματίζει πολύ σημαντικό ρόλο.

Διαφορετικές στρατηγικές για την ανίχνευση των αλληλεπιδράσεων έχουν αναπτυχθεί

εκτενώς στη βιβλιογραφία, όπως είναι η Classical Screening (CS), Classical Group Screening (CGS) και η Interaction Group Screening (IGS), όλες με δύο στάδια πειραματισμού. Η μέθοδος του κρησαρίσματος ομάδας (group screening) εισήχθη από τον Dorfman [67] στα πλαίσια του ελέγχου δειγμάτων αίματος. Ο Watson [226] επέκτεινε τη μέθοδο του κρησαρίσματος ομάδας για το κρησάρισμα παραγόντων πειραματικών σχεδιασμών. Ο Morris [166] στην εργασία του έκανε μια ιστορική αναδρομή της μεθόδου κρησαρίσματος ομάδας και της χρήσης της για την αναζήτηση των ενεργών παραγόντων.

Η κλασική στρατηγική κρησαρίσματος (CS) ανιχνεύει τις σημαντικές κύριες επιδράσεις στο πρώτο στάδιο του πειράματος και στο δεύτερο στάδιο, η στρατηγική CS προχωρά με ανίχνευση των σημαντικών αλληλεπιδράσεων μεταξύ των παραγόντων των οποίων οι κύριες επιδράσεις προσδιορίστηκαν ως σημαντικές κατά το πρώτο στάδιο. Η μέθοδος CS μπορεί να λειτουργήσει ικανοποιητικά όταν το πείραμα έχει ένα μέτριο αριθμό παραγόντων, και όταν οι μοναδικές υπαρκτές αλληλεπιδράσεις είναι μεταξύ των παραγόντων με μεγάλες κύριες επιδράσεις [141]. Οι μέθοδοι CGS και IGS αποτελούν μεθόδους κρησαρίσματος ομάδας με δύο βήματα (two-stage group screening) και έχουν προταθεί για την επιτυχία ενός πειράματος κρησαρίσματος στο οποίο το σύνολο των παραγόντων είναι μεγάλο, κάποιες από τις αλληλεπιδράσεις αναμένονται σημαντικές, και η μέθοδος CS μπορεί να απαιτεί πλέον πάρα πολλές παρατηρήσεις για να είναι εφαρμόσιμη [141], [221], [222].

11.3 Βασικές Αρχές Τεχνικών Κρησαρίσματος

Σε αυτή την ενότητα παρουσιάζουμε τις τέσσερις βασικές αρχές στις οποίες βασίζονται οι περισσότερες από τις στρατηγικές κρησαρίσματος καθώς επίσης και οι προτεινόμενες μέθοδοι κατασκευής και ανάλυσης υπερκορεσμένων σχεδιασμών με αλληλεπιδράσεις, που θα παρουσιάσουμε σε αυτό το τρίτο μέρος της παρούσας διατριβής.

Οι βασικές αρχές που διέπουν γενικά τις στρατηγικές κρησαρίσματος περιγράφονται ως ακολούθως.

1. Αρχή της σποραδικότητας των επιδράσεων (effect sparsity ή principle of parsimony), βάση της οποίας μόνο λίγες από τις ενδεχόμενες επιδράσεις είναι ενεργές και επηρεάζουν σημαντικά την απόκριση [23].

2. Αρχή της κληρονομικότητας (effect heredity), βάση της οποίας η πιθανότητα μιας αλληλεπίδρασης να είναι ενεργή εξαρτάται από τις αντίστοιχες "γονεϊκές" κύριες επιδράσεις. Υπό ισχυρή κληρονομικότητα (strong heredity), μία αλληλεπίδραση δύο παραγόντων είναι ενεργή αν και οι δύο γονεϊκές κύριες επιδράσεις της είναι ενεργές, ενώ υπό ασθενή κληρονομικότητα (weak heredity) αρκεί μόνο μία από τις γονεϊκές επιδράσεις να είναι ενεργή [104], [49].

3. Αρχή της ιεραρχίας (effect hierarchy), βάση της οποίας οι επιδράσεις χαμηλότερης τάξης είναι πιο σημαντικές από τις επιδράσεις υψηλότερης τάξης [230].

4. Αρχή της υπό συνθήκη ανεξαρτησίας (conditional independence), βάση της οποίας οι όροι μιας συγκεκριμένης τάξης θεωρούνται ανεξάρτητοι, και εξαρτώνται από όλους τους όρους χαμηλότερης τάξης. Ας σημειωθεί εδώ, ότι μόνον επιδράσεις πρώτης και δεύτερης τάξης αναμένεται να βρεθούν ενεργές [49].

Όλες αυτές οι υποθέσεις-αρχές μπορεί να είναι περιοριστικές για μία πιο γενική πειραματική μελέτη. Από τώρα και στο εξής, στις μεθόδους κατασκευής και ανάλυσης υπερκορεσμένων σχεδιασμών με αλληλεπιδράσεις, που θα παρουσιάσουμε σε επόμενα κεφάλαια, θεωρούμε όχι μόνο την αρχή της σποραδικότητας των επιδράσεων, δεδομένου ότι στην πράξη συχνά απαιτείται η μελέτη ενός μέτριου ή μεγαλύτερου αριθμού παραγόντων. Για το λόγο

αυτό θεωρούμε κάθε φορά τρία διαφορετικά σενάρια και υποθέτουμε ένα μικρό, μεσαίο και μεγαλύτερο αριθμό παραγόντων, αντίστοιχα, να είναι ενεργοί στην έναρξη του πειράματος.

Εάν πάλι τώρα, όλοι οι υποψήφιοι ενεργοί παράγοντες και οι αλληλεπιδράσεις τους (όροι πρώτης-, δεύτερης-, τρίτης-, τέταρτης-, ...-τάξης) θεωρούνται στη μελέτη, τότε το κύριο εμπόδιο για την αποτελεσματική εφαρμογή οποιασδήποτε στρατηγικής κρησαρίσματος από άποψη υπολογιστικού χρόνου και κόστους, είναι ο πολύ μεγάλος αριθμός των παραγόντων που εμπλέκονται στους υπερκορεσμένους σχεδιασμούς. Αυτό συνεπάγεται ένα πολύ μεγάλο αριθμό μοντέλων που χρειάζεται να διερευνηθούν, και οδηγεί σε μια ανέφικτη υπολογιστική προσπάθεια. Από τώρα και στο εξής, στις μεθόδους κατασκευής και ανάλυσης υπερκορεσμένων σχεδιασμών που θα παρουσιάσουμε σε επόμενα κεφάλαια, υποθέτουμε ότι μόνο πρώτης και δεύτερης τάξης επιδράσεις (κύριες επιδράσεις και αλληλεπιδράσεις δεύτερης τάξης) αναμένεται να βρεθούν ενεργές, γεγονός που υποδηλώνει ότι ο αριθμός των αντιπροσωπευτικών μοντέλων μειώνεται σημαντικά.

Ωστόσο, ακόμη και κάτω από αυτή την υπόθεση, ο κύριος στόχος ενός πειράματος κρησαρίσματος να εντοπίσει το υποσύνολο των πραγματικά σημαντικών παραγόντων και των αλληλεπιδράσεών τους εξακολουθεί να παραμένει ένα δύσκολο έργο. Στις μεθόδους που θα παρουσιάσουμε παρακάτω συμπεριλαμβάνουμε τις αλληλεπιδράσεις δύο παραγόντων στην αρχή του πειραματισμού, προκειμένου να αποκλειστούν οι αλληλεπιδράσεις ελάχιστης σημαντικότητας όσο το δυνατόν γρηγορότερα. Ας σημειωθεί σε αυτό το σημείο, ότι για να είναι ένας οποιοσδήποτε πειραματισμός επιτυχής, οι ενδεχόμενοι ενεργοί παράγοντες θα έπρεπε να είναι γνωστοί ή τουλάχιστον να υπάρχει εκ των προτέρων πληροφορία για τη σημαντικότητα κάποιων, δεδομένου ότι καλό θα ήταν οι εμπειρογνώμονες να έχουν μια αρχική κατανόηση των δεδομένων που ενδέχεται να προκύψουν από διαφορετικά σενάρια. Σε πειράματα κρησαρίσματος, όπου οι σημαντικοί παράγοντες δεν είναι απαραίτητα γνωστοί εκ των προτέρων, η επίτευξη αυτού του στόχου φαίνεται δύσκολη [222]. Για το λόγο αυτό, μια επιπλέον παραδοχή των προτεινόμενων μεθόδων που θα παρουσιάσουμε διεξοδικά παρακάτω είναι ότι ο αριθμός των ενδεχόμενων σημαντικών παραγόντων υποτίθεται σωστά εκ των προτέρων.

Επιπλέον, όπως προαναφέραμε οι παραδοσιακές τεχνικές πειραματισμού για τη βελτίωση ενός προϊόντος ή μιας διαδικασίας ξεκινούν στο πρώτο στάδιο με την εξέταση μόνο των κύριων επιδράσεων των παραγόντων, και χρησιμοποιούν περαιτέρω πειραματισμό (δεύτερο στάδιο) για να εξετάσουν τις αλληλεπιδράσεις μεταξύ των παραγόντων των οποίων οι κύριες επιδράσεις ανιχνεύτηκαν ως σημαντικές σε σχέση με την απόκριση, από τα αποτελέσματα τα οποία προέκυψαν στο πρώτο στάδιο του πειραματισμού. Αυτή η προσέγγιση απαιτεί την παραδοχή ότι σταθερά ισχύει η αρχή της ισχυρής κληρονομικότητας που αναφέρει ότι οι αλληλεπιδράσεις προκύπτουν μόνο μεταξύ αυτών των παραγόντων με ενεργές κύριες επιδράσεις, όπως επισημαίνεται και στην εργασία [68]. Ωστόσο, τα αποτελέσματα που προκύπτουν από πολλές πρακτικές εφαρμογές παρέχουν ενδείξεις ότι η αρχή της ισχυρής κληρονομικότητας αποτυγχάνει αρκετά συχνά, όπως για παράδειγμα στις εργασίες [165], [222], [206] και [68]. Κατά συνέπεια, στις μεθόδους κατασκευής και ανάλυσης υπερκορεσμένων σχεδιασμών με αλληλεπιδράσεις που θα παρουσιάσουμε σε επόμενα κεφάλαια, επιτρέπουμε τη χαλάρωση της αρχής της ισχυρής κληρονομικότητας, και συμπεριλαμβάνουμε τις αλληλεπιδράσεις δύο παραγόντων στο αρχικό στάδιο του πειραματισμού, προκειμένου να αποκλείσουμε τις αλληλεπιδράσεις ελάχιστης σημαντικότητας και πληροφόρησης, όσο το δυνατόν γρηγορότερα.

ΚΕΦΑΛΑΙΟ **12**

Ανάλυση Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις Μέσω Ποινικοποιημένων Μηχανών Διανυσματικής Υποστήριξης

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

—*Donald E. Knuth (1996)*

Στο δωδέκατο αυτό κεφάλαιο μελετάμε τη στατιστική ανάλυση υπερκορεσμένων σχεδιασμών όχι μόνο κύριων επιδράσεων, αλλά συμπεριλαμβανομένων και των αλληλεπιδράσεών τους. Το πρόβλημα εύρεσης και σχηματισμού του καταλληλότερου μοντέλου αντιμετωπίζεται ως ένα πρόβλημα βελτιστοποίησης, του οποίου η επίλυση οδηγεί στο βέλτιστο υπόδειγμα. Η προτεινόμενη ποινικοποιημένη μέθοδος βελτιστοποίησης διατυπωμένη μέσω συναρτήσεων απώλειας και συναρτήσεων ποινής, οδηγεί σε μια πιο άμεση και συνεκτική τεχνική ανάλυσης, η οποία φαίνεται αρκετά υποσχόμενη ακόμα και για την εύρεση των σημαντικών αλληλεπιδράσεων δεύτερης τάξης πέρα από τις κύριες επιδράσεις. Αναπτύσσεται μία μέθοδος κρησαρίσματος, η συμπεριφορά της οποίας μελετάται χρησιμοποιώντας το λογιστικό μοντέλο παλινδρόμησης, με στόχο την εύρεση και εκτίμηση των σημαντικών παραγόντων των υπερκορεσμένων σχεδιασμών. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου δημοσιεύθηκαν στην επιστημονική εργασία [135].

12.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Οι περισσότερες μέθοδοι για την κατασκευή και την ανάλυση των υπερκορεσμένων σχεδιασμών θεωρούν μόνο τις κύριες επιδράσεις, και αγνοούν ή θεωρούν αμελητέα την ύπαρξη αλληλεπιδράσεων. Στο προηγούμενο Κεφάλαιο 11 αναφερθήκαμε διεξοδικά σε κάποιες μεθόδους κατασκευής και ανάλυσης υπερκορεσμένων σχεδιασμών με αλληλεπιδράσεις που έχουν προταθεί στη βιβλιογραφία. Ακόμη όμως και αν η ύπαρξη των αλληλεπιδράσεων δεν θεωρείται αμελητέα στην ανάλυση των υπερκορεσμένων σχεδιασμών, το πλαίσιο ανάλυσης μέχρι τώρα περιορίζεται στην κεντρική παραδοχή ότι τα πειραματικά δεδομένα προέρχονται από ένα γραμμικό μοντέλο. Παρακινούμενοι από αυτό το γεγονός, μελετάμε εδώ μία μη συνηθισμένη προσέγγιση, συγκεκριμένα υιοθετούμε τις ποινικοποιημένες μηχανές διανυσματικής υποστήριξης, για τον έλεγχο των κύριων επιδράσεων και αλληλεπιδράσεων δύο παράγοντων σε υπερκορεσμένους σχεδιασμούς δύο επιπέδων, κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων με δίτιμη απόκριση.

Επιπλέον, στο προηγούμενο Κεφάλαιο 11 αναφερθήκαμε σε διαφορετικές στρατηγικές που έχουν αναπτυχθεί έως σήμερα στη διεθνή βιβλιογραφία για την ανίχνευση των αλληλεπιδράσεων, βασισμένες όλες σε δύο στάδια πειραματισμού. Συγκεκριμένα, οι τεχνικές αυτές ξεκινούν στο πρώτο στάδιο με την εξέταση μόνο των κύριων επιδράσεων των παραγόντων, και χρησιμοποιούν περαιτέρω πειραματισμό (δεύτερο στάδιο) για να εξετάσουν τις αλληλεπιδράσεις μεταξύ των παραγόντων των οποίων οι κύριες επιδράσεις ανιχνεύτηκαν ως σημαντικές σε σχέση με την απόκριση, από τα αποτελέσματα τα οποία προέκυψαν στο πρώτο στάδιο του πειραματισμού. Αυτή η προσέγγιση απαιτεί την παραδοχή ότι σταθερά ισχύει η αρχή της ισχυρής κληρονομικότητας που αναφέρει ότι οι αλληλεπιδράσεις προκύπτουν μόνο μεταξύ αυτών των παραγόντων με ενεργές κύριες επιδράσεις. Ωστόσο, όπως προαναφέραμε στο Κεφάλαιο 11 τα αποτελέσματα που προκύπτουν από πολλές πρακτικές εφαρμογές παρέχουν ενδείξεις ότι η αρχή της ισχυρής κληρονομικότητας αποτυγχάνει αρκετά συχνά. Παρακινούμενοι από αυτήν τη διαπίστωση, προτείνουμε μία μέθοδο των ποινικοποιημένων μηχανών διανυσματικής υποστήριξης η οποία είναι πλεονεκτική και διαφέρει από τις υπόλοιπες σε σχέση με το ότι επιτρέπει τη χαλάρωση της αρχής της ισχυρής κληρονομικότητας, και συμπεριλαμβάνει τις αλληλεπιδράσεις δύο παραγόντων στο αρχικό στάδιο του πειραματισμού, προκειμένου να αποκλειστούν οι αλληλεπιδράσεις ελάχιστης σημαντικότητας και πληροφόρησης, όσο το δυνατόν γρηγορότερα. Οι βασικές παραδοχές στις οποίες βασίζεται η προτεινόμενη μέθοδος παρουσιάζονται στο Κεφάλαιο 11 στην ενότητα 11.3.

12.2 Ποινικοποιημένες Μηχανές Διανυσματικής Υποστήριξης

Σε αυτή την ενότητα περιγράφουμε αναλυτικά τη μέθοδο των ποινικοποιημένων μηχανών διανυσματικής υποστήριξης, η οποία είναι ικανή να προσδιορίζει ποιες μεταβλητές παλινδρόμησης επηρεάζουν σημαντικά την απόκριση, και εκτελεί ταυτόχρονα αυτόματη επιλογή μεταβλητών για το σκοπό αυτό. Υιοθετούμε αυτή τη μέθοδο ποινικοποίησης, δεδομένου ότι είναι ικανή να προσδιορίσει τις σημαντικές κύριες επιδράσεις αλλά και τις αλληλεπιδράσεις δύο παραγόντων ενός μοντέλου. Θεωρούμε πολλά διαφορετικά σενάρια και τα αποτελέσματα που προκύπτουν παρέχουν καθοδήγηση στον πειραματιστή όσον αφορά την αποτελεσματική χρήση και την επιλογή της καταλληλότερης τεχνικής των ποινικοποιημένων μηχανών διανυσματικής υποστήριξης για το κρησάρισμα των υπερκορεσμένων σχεδιασμών.

12.2.1 Μαθηματική Περιγραφή των Μηχανών Διανυσματικής Υποστήριξης

Οι μέθοδοι κατηγοριοποίησης και πρόβλεψης διαδραματίζουν σημαντικό ρόλο σε ένα ευρύ φάσμα επιστημονικών πειραμάτων. Οι Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines-SVMs) αποτελούν μία μέθοδο εκμάθησης με πλήρη επίβλεψη, η οποία βασίζεται σε ιδέες που προέρχονται από τη στατιστική θεωρία εκμάθησης [220], [33], και έχουν εφαρμογές σε πολλούς επιστημονικούς κλάδους. Οι SVMs μοντελοποιούν μια μεταβλητή απόκρισης βασισμένες σε μία ή περισσότερες επεξηγηματικές μεταβλητές με στόχο την εύρεση του βέλτιστου υπερεπιπέδου. Δεδομένου ότι στη μελέτη μας έχουμε να αντιμετωπίσουμε ένα δυαδικό πρόβλημα ταξινόμησης, το βέλτιστο υπερεπίπεδο είναι αυτό με το μέγιστο περιθώριο διαχωρισμού μεταξύ δύο κλάσεων.

Οι SVMs μπορούν επίσης να εφαρμοστούν σε προβλήματα παλινδρόμησης με την εισαγωγή μιας συνάρτησης απώλειας και μιας συνάρτησης ποινής. Αντί να προσπαθούμε πλέον να κατατάξουμε μια νέα άγνωστη επεξηγηματική μεταβλητή σε μία από τις δύο κλάσεις $y_i = \pm 1$ τώρα επιθυμούμε να προβλέψουμε την πραγματική τιμή εξόδου, με τα δεδομένα εκπαίδευσης να είναι της μορφής $\{x_i, y_i\}$. Η μεθοδολογία των SVMs μπορεί να επεκταθεί και να διαχειριστεί δεδομένα που δεν είναι πλήρως γραμμικά διαχωρίσιμα. Ένα σύνολο δεδομένων το οποίο δεν είναι γραμμικά διαχωρίσιμο στο διδιάστατο χώρο, μπορεί να διαχωριστεί στο μη γραμμικό χώρο με χρήση συναρτήσεων πυρήνα που χρησιμοποιούνται ευρέως στην ταξινόμηση και στην παλινδρόμηση. Όπως προαναφέραμε οι SVMs συνήθως αναπτύσσονται ως μια μέθοδος εύρεσης του μέγιστου περιθωρίου του υπερεπιπέδου. Όμως, οι SVMs εναλλακτικά μπορούν επίσης να διατυπωθούν ως ένα πρόβλημα εκτίμησης μιας κανονικοποιημένης συνάρτησης, που αντιστοιχεί σε μια συνάρτηση απώλειας συν έναν όρο κανονικοποίησης για τους προσαρμοσμένους συντελεστές [233]. Παρουσιάζουμε παρακάτω το μαθηματικό πλαίσιο πάνω στο οποίο βασίζονται οι SVMs, το οποίο χρησιμοποιείται στη μελέτη προσομοίωσης καθώς επίσης και στα πειράματα κρησαρίσματος που εκτελούμε αργότερα.

Έστω ότι τα $\{x_1, x_2, \ldots, x_n\}$ συμβολίζουν τα n διανύσματα εισόδου, όπου το $x_i \in \Re^m$ είναι ένα διάνυσμα με m μεταβλητές πρόβλεψης, και έστω ότι τα $\{y_1, y_2 \ldots, y_n\}$ αναπαριστούν τις αντίστοιχες ετικέτες κλάσης εξόδου, όπου $y_i \in \{1, -1\}, i = 1, 2, \ldots n$. Οι SVMs βρίσκουν το υπερεπίπεδο το οποίο διαχωρίζει τις δύο κλάσεις των σημείων των δεδομένων μέσω ενός γραμμικού ορίου

$$f(x) = \beta_0 + x^T \beta,$$

όπου $x = (x_1, \ldots, x_m)^T$, το $\beta = (\beta_1, \ldots, \beta_m)^T$ συμβολίζει τους συντελεστές του υπερεπιπέδου για $j = 1, 2, \ldots m$, και το β_0 συμβολίζει το συντελεστή του σταθερού όρου του υπερεπιπέδου.

Οι SVMs επιτυγχάνουν να βρουν το βέλτιστο υπερεπίπεδο που είναι αυτό με το μέγιστο περιθώριο διαχωρισμού μεταξύ δύο κλάσεων (δηλ. το υπερεπίπεδο το οποίο διαχωρίζει τις δύο κλάσεις μέσω της μεγαλύτερης απόστασης), λύνοντας ένα κυρτό πρόβλημα βελτιστοποίησης το οποίο ισοδύναμα μπορεί να μετατραπεί σε "loss and penalty" μορφή, δηλαδή,

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + pen_\lambda(\beta), \tag{12.1}$$

όπου η συνάρτηση απώλειας $[1 - \cdot]_+ = \max[1 - \cdot, 0]$ καλείται ως hinge συνάρτηση απώλειας, και το $pen_\lambda(\beta)$ συμβολίζει τον όρο ποινής των προσαρμοσμένων συντελεστών.

12.2.2 Η L_1-norm SVM

Ο όρος ποινής στην απλή κανονική μορφή των SVMs είναι η L_2-norm ποινή (ποινή παλινδρόμησης διάσελου) και διαμορφώνεται ως

$$pen_\lambda(\beta) = \lambda\|\beta\|_2^2.$$

Αξίζει να σημειωθεί εδώ ότι δεν συμπεριλαμβάνουμε την L_2-norm SVM μέθοδο ποινικοποίησης στην εμπειρική μας μελέτη, δεδομένου ότι η L_2-norm ποινή έχει το βασικό περιορισμό ότι μπορεί και συρρικνώνει τους συντελεστές παλινδρόμησης, αλλά δεν μπορεί να τους καθορίσει ακριβώς στο μηδέν και συνεπώς δεν μπορεί να εκτελεστεί αυτόματη επιλογή μεταβλητών. Αρκετές ποινικοποιημένες μέθοδοι επιλογής μεταβλητών για τις SVMs έχουν προταθεί για το σκοπό αυτό, τις οποίες παρουσιάζουμε εν συντομία παρακάτω.

Η Least Absolute Shrinkage and Selection Operator (LASSO) μέθοδος [216] είναι μία από τις πιο διαδεδομένες τεχνικές εκτίμησης παραμέτρων σε προβλήματα παλινδρόμησης. Οι Bradley και Mangasarian [26] προσάρμοσαν στη μεθοδολογία των SVMs την L_1-norm ποινή (LASSO) αντικαθιστώντας την L_2-norm ποινή προκειμένου να εκτελείται αυτόματη επιλογή μεταβλητών. Οι Zhu et al. [236] επίσης υιοθέτησαν την ιδέα της χρήσης της L_1-norm ποινής και των περιορισμών της, για την αυτόματη επιλογή ματαβλητών σε προβλήματα κατηγοριοποίησης, χρησιμοποιώντας την ποινικοποιημένη μέθοδο L_1-norm SVM. Οι Fung και Mangasarian [84] πρότειναν μία εναλλακτική διατύπωση της L_1-norm SVM χρησιμοποιώντας newton γραμμικό προγραμματισμό.

Η L_1-norm ποινή διαμορφώνεται ως

$$pen_\lambda(\beta) = \lambda\|\beta\|_1 = \lambda\sum_{j=1}^{m}|\beta_j|.$$

Άρα, το πρόβλημα βελτιστοποίησης των SVMs όπως αυτό περιγράφεται από τη σχέση (12.1) αναδιατυπώνεται πλέον ως

$$\min_{\beta_0,\beta}\sum_{i=1}^{n}[1-y_if(x_i)]_+ + \lambda\sum_{j=1}^{m}|\beta_j|,$$

όπου το λ είναι η παράμετρος κανονικοποίησης που ελέγχει την ισορροπία μεταξύ της απώλειας και της ποινής. Η L_1-norm ποινή μειώνει τη διακύμανση των εκτιμήσεων, βελτιώνει την ακρίβεια πρόβλεψης και επιτρέπει ένα είδος συνεχούς αυτόματης επιλογής μεταβλητών.

12.2.3 Η L_1-norm Huberized SVM

Σε αυτή την ενότητα προτείνουμε τη χρήση της L_1-norm Huberized SVM η οποία συνδυάζει την huberized hinge συνάρτηση απώλειας και την L_1-norm ποινή. Η L_1-norm Huberized SVM μέθοδος προσαρμόζει την L_1-norm ποινή στις SVMs και χρησιμοποιεί την huberized hinge συνάρτηση απώλειας $\phi(y_if(x_i))$ αντί να χρησιμοποιεί την καθιερωμένη hinge συνάρτηση απώλειας $[1-\cdot]_+$ των SVMs. Με αυτό τον τρόπο, η L_1-norm Huberized SVM έχει τη μορφή "loss and penalty". Συγκεκριμένα, χρησιμοποιεί την huberized hinge συνάρτηση απώλειας για να μετρήσει την απώλεια και εκτελεί αυτόματη επιλογή μεταβλητών με παρόμοιο τρόπο με αυτόν της L_1-norm SVM εξαλείφοντας τις μη σημαντικές μεταβλητές, δηλαδή, συρρικνώνοντας τους συντελεστές παλινδρόμησης των μη σημαντικών μεταβλητών ακριβώς στο μηδέν.

Άρα, το πρόβλημα βελτιστοποίησης των SVMs όπως αυτό περιγράφεται από τη σχέση

(12.1) αναδιατυπώνεται πλέον ως

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \phi(y_i f(x_i)) + \lambda \sum_{j=1}^{m} |\beta_j|,$$

όπου το λ είναι η παράμετρος κανονικοποίησης που ελέγχει την ισορροπία μεταξύ της απώλειας και της ποινής, και

$$\phi(y_i f(x_i)) = \begin{cases} 0, & \text{για} \quad y_i f(x_i) > 1, \\ \frac{(1-y_i f(x_i))^2}{2\delta}, & \text{για} \quad 1-\delta < y_i f(x_i) \leq 1, \\ 1 - y_i f(x_i) - \frac{\delta}{2}, & \text{για} \quad y_i f(x_i) \leq 1-\delta, \end{cases}$$

όπου $\delta \geq 0$ μία προκαθορισμένη σταθερά. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στην εργασία [200] για περισσότερες λεπτομέρειες σε σχέση με την huberized hinge συνάρτηση απώλειας.

12.3 Εμπειρική Μελέτη Προσομοίωσης

Για τη διερεύνηση και την αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου πραγματοποιήσαμε μια εμπειρική μελέτη, θεωρώντας ένα ευρύ φάσμα προσομειωμένων μοντέλων και υπερκορεσμένων σχεδιασμών. Στη μελέτη προσομοίωσης χρησιμοποιήσαμε υπερκορεσμένους σχεδιασμούς οι οποίοι κατασκευάζονται από τα κλάσματα πινάκων Hadamard, σύμφωνα με τη μέθοδο που προτάθηκε από τον Lin [148]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 10$ παράγοντες και $n = 6$ πειραματικές εκτελέσεις. Επιπλέον, για τα πειράματα προσομοίωσης, χρησιμοποιήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς, όπως αυτοί κατασκευάστηκαν από τους Koukouvinos et al. [130]. Αυτοί οι υπερκορεσμένοι σχεδιασμοί έχουν n πειραματικές εκτελέσεις και $m = q \cdot (n-1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Στις προσομοιώσεις μας θεωρήσαμε τους $E(s^2)$-βέλτιστους και minimax-βέλτιστους κυκλικούς υπερκορεσμένους σχεδιασμούς με τις ακόλουθες (n, m) τιμές: $(6, 10)$, $(8, 14)$ $(10, 18)$, $(12, 22)$. Εκτός από αυτούς τους σχεδιασμούς, θεωρήσαμε και τους s-block-ορθογώνιους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με n πειραματικές εκτελέσεις και $m = s(n-1)$ παράγοντες, όπως αυτοί κατασκευάστηκαν από τους Tang και Wu [215]. Συγκεκριμένα, χρησιμοποιήσαμε τον υπερκορεσμένο σχεδιασμό με $m = 22$ παράγοντες και $n = 12$ πειραματικές εκτελέσεις.

Για την εφαρμογή των προαναφερθέντων μεθόδων θεωρήσαμε γενικευμένα γραμμικά μοντέλα στα οποία η μεταβλητή απόκρισης παίρνει μόνο δύο πιθανές τιμές που συμβολίζονται με -1 και 1. Στα πειράματα προσομοίωσης για κάθε υπερκορεσμένο σχεδιασμό που θεωρούμε, αναπτύξαμε λογιστικά μοντέλα παλινδρόμησης προκειμένου να καθορίσουμε τις ετικέτες εξόδου $y_i \in \{1, -1\}, i = 1, 2, \ldots n$, δοθέντων των m μεταβλητών πρόβλεψης εισόδου $x_i \in \Re^m$.

Ας υποθέσουμε ότι έχουμε n πειραματικές εκτελέσεις με μια δίτιμη απόκριση y (-1 ή 1) η οποία εξαρτάται από ένα σύνολο μεταβλητών παλινδρόμησης $x_1, x_2, \ldots, x_m$. Συμβολίζουμε με $y = 1$ την επιτυχία και με $y = -1$ την αποτυχία, και μοντελοποιούμε τη μέση απόκριση με $P(x_i)$, όπου το $P(x_i)$ είναι η πιθανότητα επιτυχίας και τα x_i συμβολίζουν τις μεταβλητές παλινδρόμησης στο i-οστό σημείο των δεδομένων. Το λογιστικό μοντέλο για το $P(x_i)$ δίνεται από τη σχέση

$$P(x_i) = \frac{1}{1 + e^{-x_i^T \beta}} \tag{12.2}$$

όπου ο όρος $x_i^T \beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{im}$ είναι η γραμμική προβλέπουσα με $0 \leq P(x_i) \leq 1$.

12.3.1 Κριτήρια Αξιολόγησης της Απόδοσης

Για τη σύγκριση και την αξιολόγηση των προτεινόμενων μεθόδων κρησαρίσματος, θεωρήσαμε τρία ευρέως χρησιμοποιούμενα μέτρα αξιολόγησης. Αρχικά, η ακρίβεια χρησιμοποιείται ως πρώτο κριτήριο. Η ακρίβεια (larger-the-better) ορίζεται ως το ποσοστό των επιδράσεων που ανιχνεύτηκαν σωστά. Τα άλλα δύο χρησιμοποιούμενα κριτήρια είναι η ευαισθησία και η ειδικότητα. Η ευαισθησία (larger-the-better) μετρά το ποσοστό των ενεργών επιδράσεων που ανιχνεύθηκαν σωστά ως ενεργές, ενώ η ειδικότητα (larger-the-better) μετρά το ποσοστό των ανενεργών επιδράσεων οι οποίες σωστά δηλώθηκαν ως ανενεργές. Τα τελευταία δύο στατιστικά μέτρα απόδοσης συνδέονται στενά με τις έννοιες των σφαλμάτων Τύπου I και Τύπου II. Στους πειραματικούς σχεδιασμούς, το σφάλμα Τύπου I (1-ειδικότητα) είναι το κόστος του να δηλώσεις μία ανενεργή επίδραση ως ενεργή, και το σφάλμα Τύπου II (1-ευαισθησία) το κόστος του να δηλώσεις μία ενεργή επίδραση ως ανενεργή.

Επιπρόσθετα, θεωρούμε δύο άλλα ευρέως χρησιμοποιούμενα μέτρα αξιολόγησης μοντέλου. Αρχικά, υπολογίζουμε το σφάλμα εκπαίδευσης (smaller-the-better) το οποίο αναπαριστά τη μέση απώλεια πληροφορίας πάνω στα δεδομένα εκπαίδευσης, και έπειτα εκτιμάμε το σφάλμα πρόβλεψης (smaller-the-better) πάνω σε νέα δεδομένα, το οποίο αναπαριστά το αναμενόμενο σφάλμα πάνω σε ένα ανεξάρτητο δείγμα εξέτασης το οποίο σχηματίζεται από την ίδια κατανομή με αυτήν των δεδομένων εκπαίδευσης.

12.3.2 Κρησάρισμα Υπερκορεσμένων Σχεδιασμών Κύριων Επιδράσεων

Σε αυτή την ενότητα κατά τη διάρκεια των πειραμάτων προσομοίωσης, μόνο μοντέλα κύριων επιδράσεων ελήφθησαν υπόψη. Παρακάτω περιγράφουμε το πλαίσιο ανάλυσης που θεωρήσαμε για τους υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων.

1. Τα δεδομένα παράχθηκαν από το λογιστικό μοντέλο παλινδρόμησης όπως αυτό περιγράφεται στη σχέση (12.2). $X = [1_n, x_1, x_2, \ldots, x_m]$ είναι ο $n \times (m+1)$ πίνακας του μοντέλου. Υποθέτουμε ότι κάθε ένας από τους m κύριους παράγοντες έχει δύο επίπεδα, ± 1. Η πρώτη στήλη του X είναι η $1_n = [1, \ldots, 1]^T$, με τη στήλη j να αντιστοιχεί στα επίπεδα του $(j-1)$-οστού παράγοντα για $j = 2, \ldots, m+1$.

2. Οι πραγματικά ενεργές μεταβλητές επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ πιθανών ενεργών παραγόντων. Από τις $2, \ldots, m+1$ στήλες του X, p στήλες ανατέθηκαν σε ενεργούς παράγοντες τυχαία για κάθε ένα από τα σενάρια που ακολουθούν:

- Σενάριο I: μικρή πολυπλοκότητα με $p < \frac{n}{2}$ (μικρός αριθμός ενεργών παραγόντων),
- Σενάριο II: μεσαία πολυπλοκότητα με $p = \frac{n}{2}$ (μεσαίος αριθμός ενεργών παραγόντων),
- Σενάριο III: μεγάλη πολυπλοκότητα με $p > \frac{n}{2}$ (μεγάλος αριθμός ενεργών παραγόντων).

3. Για να ληφθούν οι συντελεστές για τους ενεργούς παράγοντες, ένα δείγμα μεγέθους p σχηματίστηκε από την κανονική κατανομή $N(\mu = 4, 0.2)$, και τα πρόσημα ± 1 τυχαία κατανεμημένα σε κάθε αριθμό.

4. Για τους μη ενεργούς παράγοντες, στο πραγματικό μοντέλο, οι συντελεστές τους ελήφθησαν τυχαία από την κανονική κατανομή $N(0, 0.2)$.

Η κατανομή των αντιθέσεων δεν επηρεάζεται από το μέτρο (magnitude) των συντελεστών, αλλά εξαρτάται από το σχετικό μέγεθος (relative size) των συντελεστών. Σημειώνουμε εδώ ότι σε οποιαδήποτε μελέτη υπερκορεσμένων σχεδιασμών, εάν ο αναλυτής θεωρεί τους υπό μελέτη παράγοντες να έχουν είτε τεράστιες επιδράσεις ή πολύ μικρές επιδράσεις, τότε

δοθέντων αυτών των περιορισμών, πιθανόν να είναι η μόνη περίπτωση στην οποία ο αναλυτής ελπίζει να πάρει ξεκάθαρα αποτελέσματα, αλλά ένα τέτοιο σενάριο είναι μη ρεαλιστικό. Στην εμπειρική μας μελέτη προσομοίωσης, εξετάζουμε ένα πιο ρεαλιστικό σενάριο, θέτοντας το $\mu = 4$ που αντιστοιχεί σε ένα ενδιάμεσο μέγεθος των επιδράσεων. Δεδομένου ότι οι συνθήκες στην πράξη είναι συνήθως διαφορετικές από εκείνες των προσομοιώσεων και ότι ο πειραματιστής δεν ξέρει πόσοι και ποιοί παράγοντες μπορεί να είναι ενεργοί, προκειμένου να εξετάσουμε το πόσο ευαίσθητα είναι τα αποτελέσματα στην επιλογή και τον αριθμό των ενεργών στηλών-παραγόντων, μελετήσαμε ένα ευρύ φάσμα προσομειωμένων μοντέλων στα πειράματά μας, στα οποία θεωρούμε για κάθε υπερκορεσμένο σχεδιασμό διαφορετικό αριθμό ενεργών παραγόντων, διαφορετική επιλογή ενεργών στηλών και διαφορετικά β. Θεωρήσαμε την τυχαία ανάθεση των ενεργών παραγόντων σε στήλες, διαδικασία που είναι σημαντική για να αφαιρέσουμε τη μεροληψία επιλογής. Επίσης, η επιλογή των κατανομών στα προηγούμενα βήματα 3 και 4, γενικά εξασφαλίζει το διαχωρισμό μεταξύ των συντελεστών των ενεργών και ανενεργών παραγόντων. Για το λόγο αυτό, θεωρήσαμε αρκετά μοντέλα που ήταν διαφορετικά από την άποψη αυτή.

Για κάθε μοντέλο, 1000 σύνολα δεδομένων παρήχθησαν και τα ληφθέντα αποτελέσματα σύμφωνα με την L_1-norm SVM και την L_1-norm Huberized SVM, συνοψίζονται στον Πίνακα 12.1 και Πίνακα 12.2, αντιστοίχως. Σε αυτούς τους πίνακες, η πρώτη στήλη αναφέρει τον αριθμό του μοντέλου. Η δεύτερη στήλη με την ονομασία "$SSD_{(n,m)}$" αντιστοιχεί στο χρησιμοποιούμενο υπερκορεσμένο σχεδιασμό. Η τρίτη στήλη με την ονομασία "Σενάριο" παρουσιάζει τον επιλεγμένο αριθμό των p ενεργών παραγόντων για καθένα από τα τρία εξεταζόμενα Σενάρια I, II και III. Όλες οι άλλες στήλες δίνουν τις τιμές των κριτηρίων απόδοσης για κάθε μοντέλο μετά την εκτέλεση της L_1-norm SVM και της L_1-norm Huberized SVM, λαμβάνοντας υπόψη μόνο υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων. Η τελευταία στήλη με τον τίτλο "Ενεργοί Παράγοντες" παρουσιάζει το μέσο αριθμό των ενεργών παραγόντων που ανιχνεύθηκαν, δηλαδή, τους παράγοντες με $\beta_j \neq 0$, όπου β_j αντιστοιχεί στο συντελεστή του υπερεπιπέδου του j-οστού παράγοντα.

Παρατηρούμε από τους Πίνακες 12.1 και 12.2 ότι η L_1-norm SVM υπερέχει ξεκάθαρα της L_1-norm Huberized SVM για το κρησάρισμα κύριων επιδράσεων στους υπερκορεσμένους σχεδιασμούς. Παρατηρούμε ότι η L_1-norm Huberized SVM φαίνεται να έχει μεγαλύτερα ποσοστά σφαλμάτων πρόβλεψης (για 11 από τα 18 μοντέλα) σε σύγκριση με την L_1-norm SVM, και σημαντικά υψηλότερα ποσοστά σφαλμάτων εκπαίδευσης. Για τα 18 θεωρούμενα μοντέλα, οι τιμές του σφάλματος εκπαίδευσης για την L_1-norm Huberized SVM κυμαίνονται στο εύρος [0.40, 0.80], ενώ το αντίστοιχο εύρος για την L_1-norm SVM είναι μόνο [0.10, 0.45] όπως φαίνεται στο Σχήμα 12.1. Παρατηρούμε επίσης ότι η L_1-norm SVM κατέχει σημαντικά υψηλότερες τιμές ακρίβειας σε σύγκριση με την L_1-norm Huberized SVM μέθοδο. Οι τιμές ακρίβειας για την L_1-norm Huberized SVM κυμαίνονται στο εύρος [0.20, 0.60], ενώ το αντίστοιχο εύρος για την L_1-norm SVM είναι το [0.55, 0.90] όπως φαίνεται στο Σχήμα 12.2.

Οι προκύπτουσες τιμές της ευαισθησίας και της ειδικότητας από την L_1-norm Huberized SVM είναι πολύ χαμηλές για όλα τα θεωρούμενα μοντέλα, ενώ η L_1-norm SVM κατέχει σημαντικά υψηλότερες τιμές για την ευαισθησία και την ειδικότητα. Οι τιμές της ευαισθησίας και της ειδικότητας για την L_1-norm Huberized SVM κυμαίνονται στο εύρος [0.15, 0.55], ενώ οι αντίστοιχες τιμές για την L_1-norm SVM κυμαίνονται στο εύρος [0.45, 0.90] όπως φαίνεται στο Σχήμα 12.3. Τα αποτελέσματα της μελέτης προσομοίωσης αποκαλύπτουν ότι η L_1-norm Huberized SVM δεν κατάφερε να παρέχει ούτε καν μοντέλα τυχαίας πρόβλεψης, όπως εύκολα φαίνεται από τις τιμές της ευαισθησίας και της ειδικότητας που είναι σχεδόν χαμηλότερες από $\frac{1}{2}$. Η L_1-norm Huberized SVM απέτυχε να προσδιορίσει τους σημαντικούς παράγοντες του κάθε μοντέλου, όπως θα μπορούσε κανείς να δει εύκολα από την τελευταία στήλη του Πίνακα 12.2 με τίτλο "Ενεργοί Παράγοντες", όπου παρατηρούμε ότι ο μέσος αριθμός των ενεργών παραγόντων που ανιχνεύθηκαν είναι μηδέν για όλες σχεδόν τις υπό εξέταση περιπτώσεις.

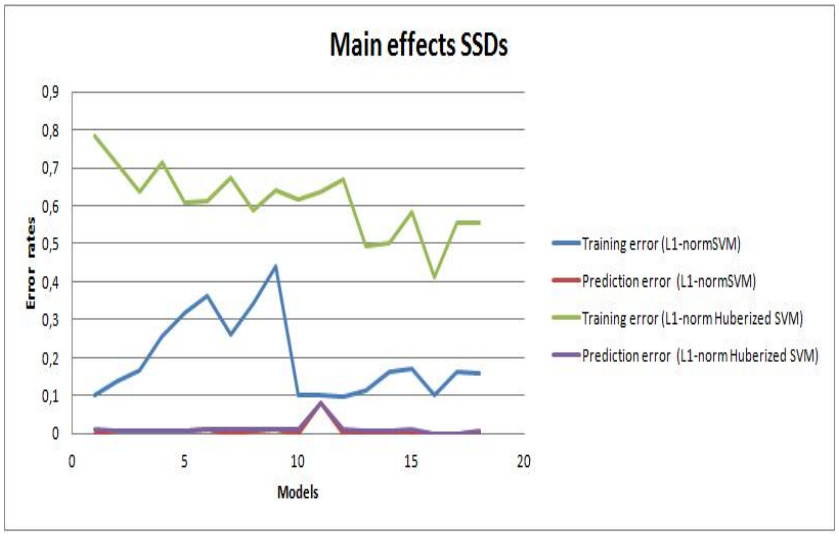
Συμπερασματικά, όταν θεωρούμε υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων ($m > n$), η L_1-norm SVM έχει πολύ καλύτερη απόδοση ανιχνεύοντας του σημαντικούς παράγοντες για μία δίτιμη απόκριση, και η L_1-norm Huberized SVM δεν φαίνεται να είναι μια λογική επιλογή ως μέθοδος κρησαρίσματος. Το αποτέλεσμα αυτό δεν αποτελεί έκπληξη δεδομένου ότι η L_1-norm Huberized SVM αναμένεται να είναι ιδιαίτερα χρήσιμη και αποδοτική σε περιπτώσεις στις οποίες ο αριθμός των μεταβλητών υπερέχει κατά πολύ το μέγεθος του δείγματος ($m \gg n$). Καταλήγουμε λοιπόν στο συμπέρασμα ότι όταν θεωρούμε υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων ($m > n$) η L_1-norm SVM φαίνεται να είναι μια λογική επιλογή ως μέθοδος κρησαρίσματος. Συγκεκριμένα, ο Πίνακας 12.1 δείχνει ξεκάθαρα ότι οι βέλτιστες τιμές των κριτηρίων απόδοσης για την L_1-norm SVM (υψηλότερες τιμές ακρίβειας, ευαισθησίας και ειδικότητας, και χαμηλότερες τιμές σφάλματος) προκύπτουν κάτω από το Σενάριο I (μικρός αριθμός ενεργών παραγόντων). Οι τιμές αυτές επισημαίνονται στον Πίνακα 12.1 με έντονα γράμματα.

Πίνακας 12.1: Αξιολόγηση της απόδοσης της L_1-norm SVM για ΥΣ κύριων επιδράσεων

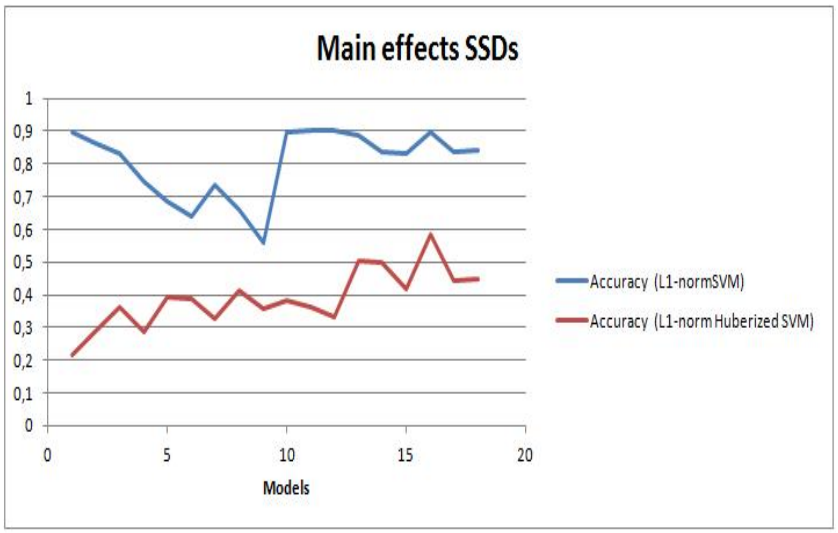
Μοντέλο	$SSD_{(n,m)}$	Σενάριο	Σφάλμα Εκπαίδευσης	Σφάλμα Πρόβλεψης	Ακρίβεια	Ευαισθησία	Ειδικότητα	Ενεργοί Παράγοντες
1	$SSD_{(6,10)}$	**I (p=2)**	**0.1033**	**0.0000**	**0.8967**	**0.8500**	**0.8300**	**0.68**
2	$SSD_{(6,10)}$	II (p=3)	0.1400	0.0067	08600	0.7600	0.8100	0.57
3	$SSD_{(6,10)}$	III (p=4)	0.1683	0.0067	0.8317	0.7600	0.7100	0.52
4	$Lin_{(6,10)}$	**I (p=2)**	**0.2550**	**0.0067**	**0.7450**	**0.7100**	**0.6200**	**0.69**
5	$Lin_{(6,10)}$	II (p=3)	0.3167	0.0067	0.6833	0.5300	0.6200	0.60
6	$Lin_{(6,10)}$	III (p=4)	0.3617	0.0000	0.6383	0.5300	0.5200	0.41
7	$SSD_{(8,14)}$	**I (p=2)**	**0.2625**	**0.0000**	**0.7375**	**0.6500**	**0.6900**	**1.15**
8	$SSD_{(8,14)}$	II (p=4)	0.3425	0.0075	0.6575	0.5200	0.6300	1.33
9	$SSD_{(8,14)}$	III (p=6)	0.4412	0.000	0.5587	0.5200	0.4500	1.30
10	$SSD_{(10,18)}$	**I (p=2)**	**0.0970**	**0.0000**	**0.9030**	**0.8567**	**0.8917**	**1.94**
11	$SSD_{(10,18)}$	II (p=5)	0.1000	0.0800	0.9000	0.8517	0.8883	2.21
12	$SSD_{(10,18)}$	III (p=8)	0.1010	0.0000	0.8990	0.8483	0.8717	2.49
13	$SSD_{(12,22)}$	**I (p=3)**	**0.1150**	**0.0017**	**0.8850**	**0.8782**	**0.8857**	**2.36**
14	$SSD_{(12,22)}$	II (p=6)	0.1617	0.0017	0.8383	0.7976	0.8156	2.94
15	$SSD_{(12,22)}$	III (p=10)	0.1708	0.0000	0.8292	0.7832	0.8227	3.33
16	$TangWu_{(12,22)}$	**I (p=3)**	**0.1033**	**0.0000**	**0.8967**	**0.8734**	**0.9031**	**2.22**
17	$TangWu_{(12,22)}$	II (p=6)	0.1567	0.0000	0.8433	0.7944	0.8461	3.02
18	$TangWu_{(12,22)}$	III (p=10)	0.1642	0.0000	0.8358	0.7928	0.8319	3.18

Πίνακας 12.2: Αξιολόγηση της απόδοσης της L_1-norm Huberized SVM για ΥΣ κύριων επιδράσεων

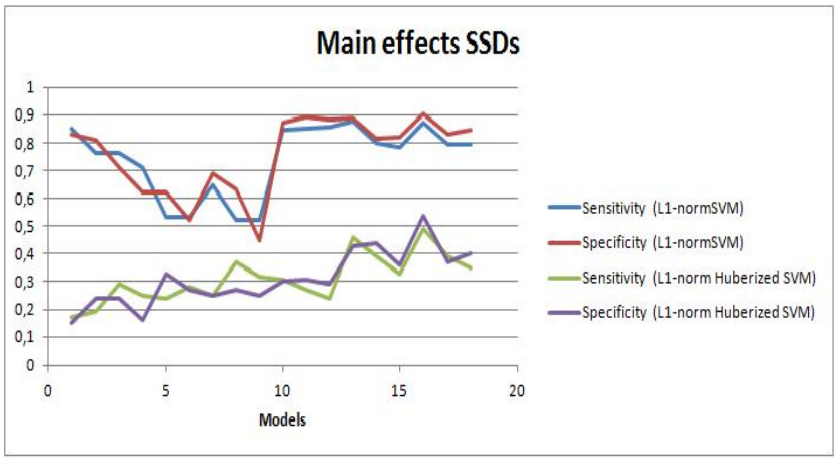
Μοντέλο	$SSD_{(n,m)}$	Σενάριο	Σφάλμα Εκπαίδευσης	Σφάλμα Πρόβλεψης	Ακρίβεια	Ευαισθησία	Ειδικότητα	Ενεργοί Παράγοντες
1	$SSD_{(6,10)}$	I (p=2)	0.7833	0.0100	0.2167	0.1700	0.1500	0.00
2	$SSD_{(6,10)}$	II (p=3)	0.7100	0.0067	0.2900	0.1900	0.2400	0.00
3	$SSD_{(6,10)}$	III (p=4)	0.6383	0.0067	0.3617	0.2900	0.2400	0.00
4	$Lin_{(6,10)}$	I (p=3)	0.7150	0.0067	0.2850	0.2500	0.1600	0.00
5	$Lin_{(6,10)}$	II (p=3)	0.6067	0.0067	0.3933	0.2400	0.3300	0.00
6	$Lin_{(6,10)}$	III (p=4)	0.6117	0.0100	0.3883	0.2800	0.2700	0.00
7	$SSD_{(8,14)}$	I (p=2)	0.6737	0.0100	0.3262	0.2500	0.2500	0.00
8	$SSD_{(8,14)}$	II (p=4)	0.5888	0.0100	0.4113	0.3700	0.2700	0.00
9	$SSD_{(8,14)}$	III (p=6)	0.6412	0.0100	0.3588	0.3200	0.2500	0.00
10	$SSD_{(10,18)}$	I (p=2)	0.6150	0.0100	0.3850	0.3100	0.3000	0.00
11	$SSD_{(10,18)}$	II (p=5)	0.6380	0.0800	0.3620	0.2700	0.3100	0.00
12	$SSD_{(10,18)}$	III (p=8)	0.6680	0.0100	0.3320	0.2400	0.2900	0.00
13	$SSD_{(12,22)}$	I (p=3)	0.4950	0.0083	0.5050	0.4600	0.4300	0.14
14	$SSD_{(12,22)}$	II (p=6)	0.5017	0.0083	0.4983	0.3900	0.4400	0.06
15	$SSD_{(12,22)}$	III (p=10)	0.5817	0.0100	0.4183	0.3300	0.3600	0.01
16	$TangWu_{(12,22)}$	I (p=3)	0.4133	0.0000	0.5867	0.4900	0.5400	0.18
17	$TangWu_{(12,22)}$	II (p=6)	0.5558	0.0000	0.4442	0.3900	0.3700	0.09
18	$TangWu_{(12,22)}$	III (p=10)	0.5533	0.0083	0.4467	0.3500	0.4000	0.03



Σχήμα 12.1: Σύγκριση σφαλμάτων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων



Σχήμα 12.2: Ακρίβεια μοντέλων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων



Σχήμα 12.3: Ανάλυση ευαισθησίας και ειδικότητας υπερκορεσμένων σχεδιασμών κύριων επιδράσεων

Δύο βασικοί περιορισμοί της L_1-norm ποινής συζητήθηκαν εκτενώς στην εργασία [238]. Θα σχολιάσουμε σε αυτό το σημείο το πως αυτοί οι δύο βασικοί περιορισμοί της L_1-norm SVM μεθόδου είναι πλεονεκτικοί για το πρόβλημα κρησαρίσματος των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων. Ένας περιορισμός της L_1-norm SVM είναι το γεγονός ότι ο αριθμός των μεταβλητών που επιλέγεται από την L_1-norm ποινή είναι άνω φραγμένος από το μέγεθος του δείγματος n (δηλαδή, η L_1-norm ποινή μπορεί να κρατήσει το πολύ n μεταβλητές εισόδου ως σημαντικές). Εξ'ορισμού ένας υπερκορεσμένος σχεδιασμός είναι ένας σχεδιασμός για τον οποίο το μέγεθος των εκτελέσεων δεν είναι αρκετά μεγάλο για να εκτιμηθούν όλες οι κύριες επιδράσεις. Άρα, σε μία ανάλυση υπερκορεσμένων σχεδιασμών επιθυμούμε ο αριθμός των μεταβλητών που επιλέγονται να είναι άνω φραγμένος από το μέγεθος του δείγματος. Ο Πίνακας 12.1 δείχνει ότι η L_1-norm SVM έχει πολύ καλύτερη απόδοση κάτω από το Σενάριο I με το $p < \frac{n}{2}$. Καταλήγουμε λοιπόν στο συμπέρασμα ότι η L_1-norm SVM υπερέχει ξεκάθαρα ως προς την απόδοσή της για το κρησάρισμα υπερκορεσμένων σχεδιασμών κύριων επιδράσεων, όταν ο αριθμός των μεταβλητών που επιλέγεται από την L_1-norm ποινή είναι άνω φραγμένος από το $\frac{n}{2}$ (το οποίο είναι μικρότερο από το μέγεθος του δείγματος n). Ένας άλλος περιορισμός της L_1-norm SVM μεθόδου είναι το γεγονός ότι στην περίπτωση που υπάρχουν αρκετές υψηλά συσχετισμένες μεταβλητές εισόδου στο σύνολο δεδομένων που επηρεάζουν τη μεταβλητή εξόδου, η L_1-norm ποινή τείνει να επιλέξει μόνο μία ή λίγες από αυτές τις μεταβλητές, και συρρικνώνει τις υπόλοιπες στο μηδέν. Το γεγονός αυτό δεν αποτελεί πρόβλημα για το κρησάρισμα των κύριων επιδράσεων στους υπερκορεσμένους σχεδιασμούς για δύο βασικούς λόγους. Πρώτον, ακόμα και αν ένας μεγάλος αριθμός παραγόντων μελετάται σε μία ανάλυση υπερκορεσμένων σχεδιασμών, μόνο λίγοι από αυτούς αναμένεται να είναι σημαντικοί (αρχή της σποραδικότητας των επιδράσεων) και δεύτερον, η L_1-norm SVM μέθοδος είχε σταθερά την καλύτερη απόδοση κάτω από το Σενάριο I, το οποίο υποθέτει ένα μικρό αριθμό παραγόντων να είναι ενεργοί. Οι παραπάνω λόγοι μας οδηγούν στο συμπέρασμα ότι οι περιορισμοί της L_1-norm ποινής είναι πλεονεκτικοί για το πρόβλημα κρησαρίσματος των υπερκορεσμένων σχεδιασμών.

12.3.3 Κρησάρισμα Υπερκορεσμένων Σχεδιασμών Κύριων Επιδράσεων και Αλληλεπιδράσεων σε Ένα Στάδιο

Σε αυτή την ενότητα προτείνουμε μία τεχνική κρησαρίσματος στην οποία εξετάζουμε τις κύριες επιδράσεις των παραγόντων και τις αλληλεπιδράσεις δύο παραγόντων των υπερκορεσμένων σχεδιασμών σε ένα και μόνο στάδιο. Το πείραμα τρέχει ταυτόχρονα πάνω στους αρχικούς κύριους παράγοντες και στις αλληλεπιδράσεις δύο παραγόντων αυτών, προκειμένου να ανιχνεύσει την ομάδα των ενεργών κύριων επιδράσεων και αλληλεπιδράσεων. Παρακάτω περιγράφουμε το πλαίσιο ανάλυσης που θεωρήσαμε για το κρησάρισμα των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων συμπεριλαμβανομένων και των αλληλεπιδράσεων.

1. $X = [1_n, x_1, x_2, \ldots, x_m]$ είναι ο $n \times (m+1)$ πίνακας του μοντέλου. Υποθέτουμε ότι κάθε ένας από τους m κύριους παράγοντες έχει δύο επίπεδα, ± 1. Η πρώτη στήλη του X είναι η $1_n = [1, \ldots, 1]^T$, με τη στήλη j να αντιστοιχεί στα επίπεδα του $(j-1)$-οστού παράγοντα για $j = 2, \ldots, m+1$.

2. Αναμειγνύουμε τις κύριες επιδράσεις και τις αλληλεπιδράσεις δύο παραγόντων. Κατασκευάζουμε όλες τις $\binom{m}{2}$ στήλες αλληλεπιδράσεων, και τις συνδυάζουμε με τις m στήλες κύριων επιδράσεων. Τώρα έχουμε $m + \binom{m}{2}$ στήλες σχεδιασμού. Ο $n \times (m+1)$ πίνακας X, τώρα αυξάνεται σε έναν $n \times \left(\frac{m(m+1)}{2} + 1\right)$ πίνακα $X^+ = [1, x_1, x_2, \ldots, x_{\frac{m(m+1)}{2}}]$. Η πρώτη στήλη του X^+ είναι η $1_n = [1, \ldots, 1]^T$, με τη στήλη j να αντιστοιχεί στα επίπεδα του $(j-1)$-οστού παράγοντα για $j = 2, \ldots, \frac{m(m+1)}{2} + 1$. Τα δεδομένα παράχθηκαν από το λογιστικό μοντέλο παλινδρόμησης, όπως αυτό περιγράφεται στη σχέση (12.2).

3. Οι πραγματικά ενεργές μεταβλητές επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, \frac{m(m+1)}{2}\}$ πιθανών ενεργών παραγόντων. Από τις $2, \ldots, \frac{m(m+1)}{2} + 1$ στήλες του X^+, p στήλες ανατέθηκαν σε ενεργούς παράγοντες τυχαία για κάθε ένα από τα σενάρια που ακολουθούν:

- Σενάριο I: μικρή πολυπλοκότητα με $p < \frac{n}{2}$ (μικρός αριθμός ενεργών παραγόντων),
- Σενάριο II: μεσαία πολυπλοκότητα με $p = \frac{n}{2}$ (μεσαίος αριθμός ενεργών παραγόντων),
- Σενάριο III: μεγάλη πολυπλοκότητα με $p > \frac{n}{2}$ (μεγάλος αριθμός ενεργών παραγόντων).

4. Για να ληφθούν οι συντελεστές για τους ενεργούς παράγοντες, ένα δείγμα μεγέθους p σχηματίστηκε από την κανονική κατανομή $N(\mu = 4, 0.2)$, κα τα πρόσημα ± 1 τυχαία κατανεμημένα σε κάθε αριθμό.

5. Για τους μη ενεργούς παράγοντες, στο πραγματικό μοντέλο, οι συντελεστές τους ελήφθησαν τυχαία από την κανονική κατανομή $N(0, 0.2)$.

Για κάθε μοντέλο συμπεριλαμβανομένων των αλληλεπιδράσεων δύο παραγόντων, 100 σύνολα δεδομένων παρήχθησαν εξαιτίας της υπολογιστικής πολυπλοκότητας, και τα ληφθέντα αποτελέσματα σύμφωνα με την L_1-norm SVM και την L_1-norm Huberized SVM, συνοψίζονται στον Πίνακα 12.3 και Πίνακα 12.4, αντιστοίχως. Σε αυτούς τους πίνακες, η πρώτη στήλη αναφέρει τον αριθμό του μοντέλου. Η δεύτερη στήλη με την ονομασία "$SSD_{(n,m)}$" αντιστοιχεί στο χρησιμοποιούμενο υπερκορεσμένο σχεδιασμό. Η τρίτη στήλη με την ονομασία "Σενάριο" παρουσιάζει τον επιλεγμένο αριθμό των p ενεργών παραγόντων για καθένα από τα τρία εξεταζόμενα Σενάρια I, II και III. Όλες οι άλλες στήλες δίνουν τις τιμές των κριτηρίων απόδοσης για κάθε μοντέλο μετά την εκτέλεση της L_1-norm SVM και της L_1-norm Huberized SVM, λαμβάνοντας υπόψη υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων και αλληλεπιδράσεων. Η τελευταία στήλη με τον τίτλο "Ενεργοί Παράγοντες" παρουσιάζει το μέσο αριθμό των ενεργών παραγόντων που ανιχνεύθηκαν, δηλαδή, τους παράγοντες με $\beta_j \neq 0$, όπου β_j αντιστοιχεί στο συντελεστή του υπερεπιπέδου του j-οστού παράγοντα.

Παρατηρούμε από τους Πίνακες 12.3 και 12.4 ότι η L_1-norm Huberized SVM υπερέχει ξεκάθαρα της L_1-norm SVM για το κρησάρισμα κύριων επειδράσεων και αλληλεπιδράσεων δύο παραγόντων σε υπερκορεσμένους σχεδιασμούς. Οι Πίνακες 12.3 και 12.4 δείχνουν ότι η L_1-norm Huberized SVM έχει χαμηλότερα ποσοστά σφάλματος πρόβλεψης και εκπαίδευσης σε σύγκριση με την L_1-norm SVM όπως φαίνεται στο Σχήμα 12.4. Η L_1-norm Huberized SVM έχει τιμές σφάλματος πρόβλεψης ίσες με το μηδέν για 14 από τα 18 θεωρούμενα μοντέλα, ενώ η L_1-norm SVM για 8 από τα 18 μοντέλα, αντίστοιχα. Για 10 από τα 18 θεωρούμενα μοντέλα, οι τιμές του σφάλματος εκπαίδευσης της L_1-norm Huberized SVM είναι χαμηλότερες σε σύγκριση με τις αντίστοιχες τιμές της L_1-norm SVM. Παρατηρούμε ότι η L_1-norm Huberized SVM έχει σχεδόν παρόμοιες τιμές ακρίβειας σε σύγκριση με την L_1-norm SVM. Οι τιμές ακρίβειας και για τις δύο αυτές μεθόδους κυμαίνονται στο εύρος [0.75, 1] όπως φαίνεται στο Σχήμα 12.5.

Οι τιμές για την ευαισθησία και την ειδικότητα που προέκυψαν από τις δύο μεθόδους L_1-norm Huberized SVM και L_1-norm SVM είναι υψηλές για όλα τα θεωρούμενα μοντέλα, δεδομένου ότι οι τιμές αυτές κυμαίνονται στο εύρος [0.60, 1] όπως φαίνεται στο Σχήμα 12.6. Η L_1-norm Huberized SVM έχει υψηλότερες τιμές ειδικότητας για 10 από τα 18 θεωρούμενα μοντέλα σε σύγκριση με την L_1-norm SVM. Οι προκύπτουσες τιμές ευαισθησίας από την L_1-norm Huberized SVM είναι επίσης υψηλότερες για 10 από τα 18 θεωρούμενα μοντέλα σε σύγκριση με την L_1-norm SVM. Αυτές αντιστοιχούν σε περιπτώσεις όπου η πλειοψηφία των ενεργών επιδράσεων ανιχνεύονται σωστά. Οι Πίνακες 12.3 και 12.4 δείχνουν ότι η L_1-norm Huberized SVM τείνει να δηλώνει σε ένα χαμηλότερο ποσοστό ανενεργές επιδράσεις ως

ενεργές σε σύγκριση με την L_1-norm SVM, καθώς επίσης η L_1-norm Huberized SVM τείνει να δηλώνει σε ένα χαμηλότερο ποσοστό ενεργές επιδράσεις να είναι ανενεργές σε σύγκριση με την L_1-norm SVM.

Η εμπειρική απόδοση βασισμένη στην παραπάνω μελέτη προσομοίωσης (Πίνακας 12.3 και Πίνακας 12.4) αποκαλύπτει ότι η L_1-norm Huberized SVM φαίνεται να υπερέχει της L_1-norm SVM για το κρησάρισμα κύριων επιδράσεων και αλληλεπιδράσεων σε υπερκορεσμένους σχεδιασμούς (χαμηλότερες τιμές σφάλματος εκπαίδευσης και πρόβλεψης, υψηλότερες τιμές ευαισθησίας και ειδικότητας) για περιπτώσεις στις οποίες ο αριθμός των μεταβλητών υπερβαίνει κατά πολύ το μέγεθος του δείγματος ($m \gg n$). Το γεγονός ότι η L_1-norm Huberized SVM τείνει να δηλώνει σε ένα χαμηλότερο ποσοστό ενεργές επιδράσεις ως ανενεργές είναι πολύ σημαντικό, δεδομένου ότι η κύρια χρήση των υπερκορεσμένων σχεδιασμών είναι ο εντοπισμός των σημαντικών παραγόντων που πρέπει να διερευνηθούν περαιτέρω, και τα χαμηλά σφάλματα Τύπου ΙΙ είναι ιδιαίτερα επιθυμητά, μολονότι τόσο τα σφάλματα Τύπου Ι όσο και τα σφάλματα Τύπου ΙΙ είναι σημαντικά και πρέπει να διατηρούνται όσο το δυνατόν χαμηλότερα. Η L_1-norm Huberized SVM επιτυγχάνει κάτι τέτοιο πολύ ικανοποιητικά.

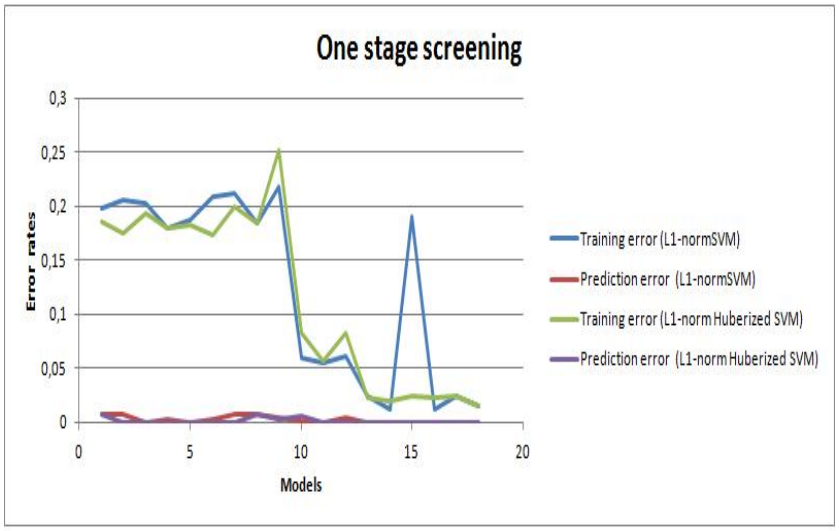
Συμπερασματικά, καταλήγουμε στο ότι όταν ο πειραματιστής θεωρεί υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων συμπεριλαμβανομένων των αλληλεπιδράσεων, η L_1-norm Huberized SVM φαίνεται να είναι μία λογική επιλογή ως μέθοδος κρησαρίσματος. Συγκεκριμένα, ο Πίνακας 12.4 δείχνει ξεκάθαρα ότι οι βέλτιστες τιμές των κριτηρίων απόδοσης για την L_1-norm Huberized SVM (υψηλότερες τιμές ακρίβειας, ευαισθησίας και ειδικότητας, και χαμηλότερες τιμές σφάλματος) προκύπτουν κάτω από το Σενάριο ΙΙ (μεσαίος αριθμός ενεργών παραγόντων). Οι τιμές αυτές επισημαίνονται στον Πίνακα 12.4 με έντονα γράμματα.

Πίνακας 12.3: Αξιολόγηση της απόδοσης της L_1-norm SVM για ΥΣ κύριων επιδράσεων και αλληλεπιδράσεων

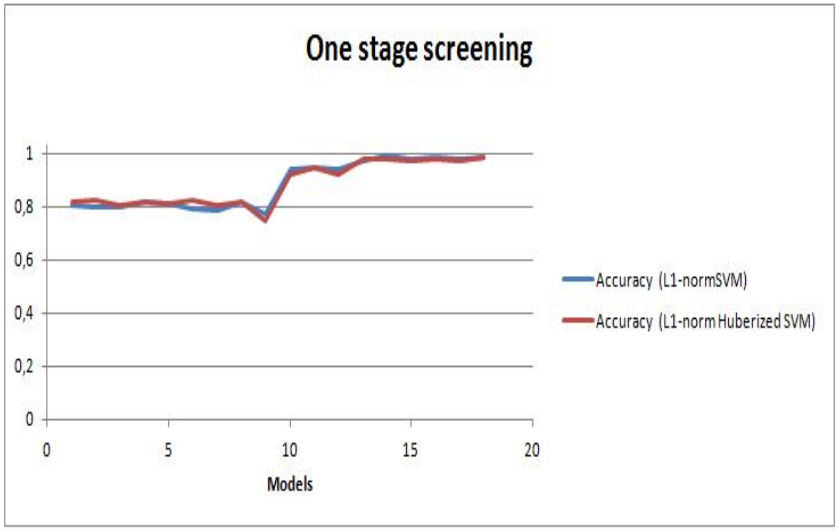
Μοντέλο	$SSD_{(n,m)}$	Σενάριο	Σφάλμα Εκπαίδευσης	Σφάλμα Πρόβλεψης	Ακρίβεια	Ευαισθησία	Ειδικότητα	Ενεργοί Παράγοντες
1	$SSD_{(6,55)}$	Ι (p=2)	0.1983	0.0067	0.8017	0.7300	0.6000	1.89
2	$SSD_{(6,55)}$	ΙΙ (p=3)	0.2050	0.0067	0.7950	0.7100	0.6100	1.97
3	$SSD_{(6,55)}$	ΙΙΙ (p=4)	0.2033	0.0003	0.7967	0.6900	0.6600	2.06
4	$Lin_{(6,55)}$	Ι (p=2)	0.1900	0.0033	0.8100	0.6500	0.7300	1.75
5	$Lin_{(6,55)}$	ΙΙ (p=3)	0.1866	0.0000	0.8133	0.7300	0.6500	1.88
6	$Lin_{(6,55)}$	ΙΙΙ (p=4)	0.2083	0.0033	0.7916	0.6000	0.7100	1.75
7	$SSD_{(8,105)}$	Ι (p=2)	0.2125	0.0075	0.7875	0.7200	0.6500	6.45
8	$SSD_{(8,105)}$	ΙΙ (p=4)	0.1839	0.0075	0.8161	0.6900	0.7200	5.28
9	$SSD_{(8,105)}$	ΙΙΙ (p=6)	0.2187	0.0050	0.7725	0.7300	0.6400	5.96
10	$SSD_{(10,171)}$	Ι (p=2)	0.0590	0.0020	0.9410	0.8950	0.9050	3.83
11	$SSD_{(10,171)}$	ΙΙ (p=5)	0.0550	0.0000	0.9450	0.9350	0.8833	4.44
12	$SSD_{(10,171)}$	ΙΙΙ (p=8)	0.0610	0.0040	0.9390	0.8783	0.9133	4.62
13	$SSD_{(12,253)}$	Ι (p=3)	0.0250	0.0000	0.9750	0.9687	0.9347	5.48
14	$SSD_{(12,253)}$	ΙΙ (p=6)	0.0116	0.0000	0.9883	0.9817	0.9800	6.16
15	$SSD_{(12,253)}$	ΙΙΙ (p=10)	0.0191	0.0000	0.9808	0.9805	0.9457	5.57
16	$TangWu_{(12,253)}$	Ι (p=3)	0.0125	0.0000	0.9875	0.9745	0.9840	5.71
17	$TangWu_{(12,253)}$	ΙΙ (p=6)	0.0241	0.0000	0.9758	0.9731	0.9325	6.46
18	$TangWu_{(12,253)}$	ΙΙΙ (p=10)	0.0158	0.0000	0.9847	0.9505	0.9880	7.16

Πίνακας 12.4: Αξιολόγηση της απόδοσης της L_1-norm Huberized SVM για ΥΣ κύριων επιδράσεων και αλληλεπιδράσεων

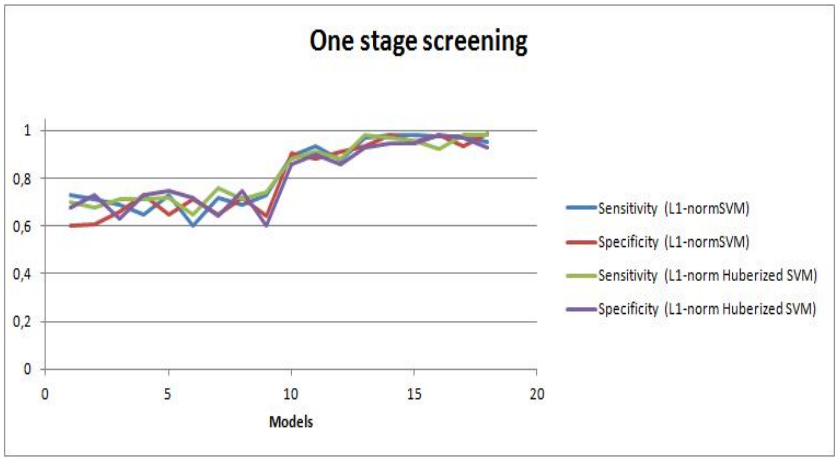
Μοντέλο	$SSD_{(n,m)}$	Σενάριο	Σφάλμα Εκπαίδευσης	Σφάλμα Πρόβλεψης	Ακρίβεια	Ευαισθησία	Ειδικότητα	Ενεργοί Παράγοντες
1	$SSD_{(6,55)}$	I (p=2)	0.1850	0.0067	0.8150	0.7000	0.6800	2.01
2	$SSD_{(6,55)}$	**II (p=3)**	**0.1750**	**0.0000**	**0.8250**	**0.6800**	**0.7300**	**2.05**
3	$SSD_{(6,55)}$	III (p=4)	0.1933	0.0000	0.8067	0.7100	0.6300	2.34
4	$Lin_{(6,55)}$	I (p=2)	0.1733	0.0000	0.8267	0.7100	0.7300	1.85
5	$Lin_{(6,55)}$	**II (p=3)**	**0.1800**	**0.0000**	**0.8200**	**0.7200**	**0.7500**	**1.98**
6	$Lin_{(6,55)}$	III (p=4)	0.1883	0.0000	0.8117	0.6500	0.7200	1.76
7	$SSD_{(8,105)}$	I (p=2)	0.1988	0.0000	0.8013	0.7600	0.6400	6.57
8	$SSD_{(8,105)}$	**II (p=4)**	**0.1838**	**0.0075**	**0.8163**	**0.7100**	**0.7500**	**5.34**
9	$SSD_{(8,105)}$	III (p=6)	0.2512	0.0025	0.7488	0.7400	0.6000	6.05
10	$SSD_{(10,171)}$	I (p=2)	0.0820	0.0060	0.9180	0.8817	0.8580	4.17
11	$SSD_{(10,171)}$	**II (p=5)**	**0.0560**	**0.0000**	**0.9440**	**0.9113**	**0.8983**	**5.39**
12	$SSD_{(10,171)}$	III (p=8)	0.0820	0.0000	0.9180	0.8790	0.8500	4.97
13	$SSD_{(12,253)}$	I (p=3)	0.0233	0.0000	0.9767	0.9840	0.9286	7.96
14	$SSD_{(12,253)}$	**II (p=6)**	**0.0200**	**0.0000**	**0.9800**	**0.9711**	**0.9480**	**6.86**
15	$SSD_{(12,253)}$	III (p=10)	0.0250	0.0000	0.9750	0.9551	0.9480	7.50
16	$TangWu_{(12,253)}$	I (p=3)	0.0233	0.0000	0.9767	0.9240	0.9840	6.72
17	$TangWu_{(12,253)}$	**II (p=6)**	**0.0150**	**0.0000**	**0.9850**	**0.9823**	**0.9680**	**6.75**
18	$TangWu_{(12,253)}$	III (p=10)	0.0250	0.0000	0.9750	0.9826	0.9271	7.86



Σχήμα 12.4: Σύγκριση σφαλμάτων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων



Σχήμα 12.5: Ακρίβεια μοντέλων υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων



Σχήμα 12.6: Ανάλυση ευαισθησίας και ειδικότητας υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων

12.3.4 Ιδιότητες Ποινικοποιημένων Μηχανών Διανυσματικής Υποστήριξης

Υπάρχουν έξι σηματικές ιδιότητες-κλειδιά που αξίζει να αναφερθούμε σε αυτό το σημείο.

1. Η μέθοδος των ποινικοποιημένων SVMs (L_1-norm SVM και L_1-norm Huberized SVM) φαίνεται να είναι εύρωστη σε σχέση με τον αριθμό των ενεργών παραγόντων που ανιχνεύονται, υπό την έννοια ότι ο συνολικός μέσος αριθμός των παραγόντων που βρέθηκαν να είναι ενεργές για κάθε υπερκορεσμένο σχεδιασμό είναι περίπου ο ίδιος και για τα τρία διαφορετικά θεωρούμενα σενάρια.

2. Η μέθοδος των ποινικοποιημένων SVMs είναι ευαίσθητη στις παραμέτρους που εμπλέκονται στο μοντέλο, υπό την έννοια ότι η L_1-norm SVM έχει καλύτερη απόδοση από την L_1-norm Huberized SVM για το κρησάρισμα υπερκορεσμένων σχεδιασμών μόνο κύριων επιδράσεων, ενώ η L_1-norm Huberized SVM υπερέχει της L_1-norm SVM όταν οι αλληλεπιδράσεις συμπεριλαμβάνονται στο μοντέλο.

3. Η L_1-norm SVM είναι ευαίσθητη στο θεωρούμενο σενάριο. Η καλύτερη απόδοση για το κρησάρισμα υπερκορεσμένων σχεδιασμών μόνο κύριων επιδράσεων από την L_1-norm SVM προέρχεται κάτω από το Σενάριο I (μικρός αριθμός ενεργών παραγόντων). Η δεύτερη καλύτερη απόδοση προέρχεται κάτω από το Σενάριο II (μεσαίος αριθμός ενεργών παραγόντων), και τελευταία ακολουθεί η απόδοση κάτω από το Σενάριο III (μεγάλος αριθμός ενεργών παραγόντων).

4. Η L_1-norm Huberized SVM είναι ευαίσθητη στο θεωρούμενο σενάριο. Η καλύτερη απόδοση για το κρησάρισμα υπερκορεσμένων σχεδιασμών κύριων επιδράσεων και αλληλεπιδράσεων από την L_1-norm Huberized SVM προέρχεται κάτω από το Σενάριο II (μεσαίος αριθμός ενεργών παραγόντων). Η δεύτερη καλύτερη απόδοση προέρχεται κάτω από το Σενάριο I (μικρός αριθμός ενεργών παραγόντων), και τελευταία ακολουθεί η απόδοση κάτω από το Σενάριο III (μεγάλος αριθμός ενεργών παραγόντων).

5. Η L_1-norm SVM παρουσιάζει σταθερότητα στις προβλέψεις σε περιπτώσεις στις οποίες ο αριθμός των μεταβλητών υπερβαίνει ελάχιστα το μέγεθος του δείγματος ($m > n$).

6. Η L_1-norm Huberized SVM παρουσιάζει σταθερότητα στις προβλέψεις σε περιπτώσεις στις οποίες ο αριθμός των μεταβλητών υπερβαίνει κατά πολύ το μέγεθος του δείγματος ($m \gg n$).

12.4 Ανάλυση Πραγματικών Δεδομένων

Σε αυτή την ενότητα εξετάζουμε την απόδοση της προτεινόμενης μεθόδου κρησαρίσματος με κίνητρο την παρουσία πραγματικών δεδομένων. Η αρχική βάση δεδομένων αποτελείται από 44 επεξηγηματικές μεταβλητές που αποτελούν τους προγνωστικούς παράγοντες, και από 8862 ασθενείς. Το σύνολο δεδομένων Τραύματος που χρησιμοποιήθηκε παρουσιάζεται στον Πίνακα 10.1 στο Κεφάλαιο 10.

Ο σκοπός αυτής της ανάλυσης πραγματικών δεδομένων είναι να αξιολογήσει την πρακτική χρήση της προτεινόμενης προσέγγισης, αλλά και να δώσει κάποια εικόνα για το πώς η προτεινόμενη μέθοδος κρησαρίσματος συμβάλλει προς την επικείμενη χρήση των υπερκορεσμένων σχεδιασμών σε πραγματικά προβλήματα της ζωής. Στο Κεφάλαιο 10 ασχοληθήκαμε με ένα μεγάλης διάστασης πρόβλημα στατιστικής μοντελοποίησης δοθέντων των παρατηρήσεων μιας βάσης δεδομένων, προτείνοντας μία εναλλακτική προσέγγιση με στόχο την κρυπτογράφηση συγκεκριμένων πεδίων της βάσης δεδομένων θεωρώντας υπερκορεσμένους σχεδιασμούς. Για περισσότερες λεπτομέρειες σχετικά με τον αλγόριθμο επιλογής βέλτιστου υπερκορεσμένου σχεδιασμού από τη βάση δεδομένων, ο ενδιαφερόμενος αναγνώστης μπορεί να

ανατρέξει στο Κεφάλαιο 10 ή στην εργασία [136]. Στη μελέτη του παρόντος κεφαλαίου, χρησιμοποιούμε μόνο το βέλτιστο υπερκορεσμένο σχεδιασμό που ανακτήθηκε από την προαναφερθείσα βάση δεδομένων, ο οποίος περιλαμβάνει 6 πειραματικές εκτελέσεις και 8 παράγοντες, και παρουσιάζεται στον Πίνακα 12.5.

Πίνακας 12.5: Ο βέλτιστος υπερκορεσμένος σχεδιασμός

record ID	x_8	x_5	x_7	x_{13}	x_6	x_2	x_{14}	x_{11}	y
968	1	-1	-1	1	-1	-1	1	-1	1
936	-1	1	-1	1	-1	-1	1	-1	1
5587	1	1	-1	1	-1	-1	1	1	1
7344	-1	1	-1	1	-1	-1	-1	-1	-1
3806	-1	1	-1	-1	-1	1	-1	-1	-1
2530	-1	-1	-1	-1	1	-1	-1	-1	-1

Αναμειγνύουμε τώρα τις κύριες αυτές επιδράσεις που ανιχνεύθηκαν με τις αλληλεπιδράσεις δύο παραγόντων. Ο υπερκορεσμένος σχεδιασμός του Πίνακα 12.5 ο οποίος περιλαμβάνει 6 πειραματικές εκτελέσεις και 8 παράγοντες, αυξάνεται τώρα σε έναν υπερκορεσμένο σχεδιασμό με 6 πειραματικές εκτελέσεις και 36 παράγοντες.

Αρχικά εκτελούμε την L_1-norm SVM και την L_1-norm Huberized SVM μέθοδο σε όλη τη βάση των δεδομένων η οποία αποτελείται από 8862 ασθενείς και 44 πιθανούς παράγοντες κινδύνου ($Model_{database}$). Έπειτα εκτελούμε την L_1-norm SVM και την L_1-norm Huberized SVM χρησιμοποιώντας ως πίνακα σχεδιασμού τον υπερκορεσμένο σχεδιασμό μόνο κύριων επιδράσεων, δηλαδή τον $SSD_{(6_{runs},8_{factors})}$, και ως απόκριση y τις αντίστοιχες ετικέτες κλάσης εξόδου (τελευταία στήλη του Πίνακα 12.5). Το μοντέλο αυτό ονομάζεται ως $Model_{SSD}$. Στο τέλος, εκτελούμε την L_1-norm SVM και την L_1-norm Huberized SVM χρησιμοποιώντας ως πίνακα σχεδιασμού τον υπερκορεσμένο σχεδιασμό κύριων επιδράσεων συμπεριλαμβανομένων των αλληλεπιδράσεων δύο παραγόντων, δηλαδή τον $SSD_{(6_{runs},36_{factors})}$, και ως απόκριση y τις αντίστοιχες ετικέτες κλάσης εξόδου (τελευταία στήλη του Πίνακα 12.5). Το μοντέλο αυτό ονομάζεται ως $Model_{SSDinter}$.

Για κάθε ένα από τα παραπάνω μοντέλα, 1000 σύνολα δεδομένων παράχθηκαν και τα ληφθέντα αποτελέσματα σύμφωνα με την L_1-norm SVM και την L_1-norm Huberized SVM συνοψίζονται στον Πίνακα 12.6. Στον Πίνακα 12.6, η πρώτη στήλη και η δεύτερη στήλη αναφέρεται στη χρησιμοποιούμενη μέθοδο και στο χρησιμοποιούμενο μοντέλο, αντίστοιχα. Όλες οι άλλες στήλες δίνουν τις τιμές των κριτηρίων απόδοσης για κάθε μοντέλο μετά την εκτέλεση της L_1-norm SVM και της L_1-norm Huberized SVM. Ο Πίνακας 12.7 παρουσιάζει τις επιδράσεις που εντοπίστηκαν ως ενεργές για κάθε μοντέλο μετά την εκτέλεση της L_1-norm SVM και της L_1-norm Huberized SVM.

Μετά την εκτέλεση των μεθόδων L_1-norm SVM και L_1-norm Huberized SVM σε όλη τη βάση δεδομένων ($Model_{database}$) παρατηρούμε ότι και οι δύο αυτές μέθοδοι έχουν παρόμοια απόδοδη (ταυτόσημες τιμές για τα σφάλματα εκπαίδευσης και πρόβλεψης, για την ακρίβεια, ευαισθησία και ειδικότητα).

Μετά την εκτέλεση των μεθόδων L_1-norm SVM και L_1-norm Huberized SVM στο σχεδιασμό μόνο κύριων επιδράσεων $SSD_{(6_{runs},8_{factors})}$ παρατηρούμε ότι και οι δύο αυτές μέθοδοι έχουν ταυτόσημες τιμές για το σφάλμα πρόβλεψης και την ειδικότητα. Παρ'όλα αυτά, η L_1-norm SVM υπερέχει της L_1-norm Huberized SVM για το $Model_{SSD}$ δεδομένου ότι έχει χαμηλότερη τιμή σφάλματος εκπαίδευσης και υψηλότερες τιμές ακρίβειας και ευαισθησίας. Η L_1-norm SVM ανίχνευσε 4 κύριες επιδράσεις ως ενεργές (x_2, x_7, x_{13}, x_{14}). Η L_1-norm Huberized SVM ανίχνευσε 2 κύριες επιδράσεις ως ενεργές (x_2, x_{14}).

Μετά την εκτέλεση των μεθόδων L_1-norm SVM και L_1-norm Huberized SVM στο σχεδιασμό κύριων επιδράσεων συμπεριλαμβανομένων των αλληλεπιδράσεων $SSD_{(6_{runs},36_{factors})}$, παρατηρούμε ότι και οι δύο αυτές μέθοδοι έχουν ταυτόσημες τιμές για την ειδικότητα. Η L_1-norm Huberized SVM υπερέχει της L_1-norm SVM για το $Model_{SSDinter}$ δεδομένου ότι έχει χαμηλότερες τιμές σφάλματος εκπαίδευσης και πρόβλεψης, υψηλότερες τιμές ευαισθησίας και

ακρίβειας. Η L_1-norm SVM ανίχνευσε 2 ενεργές κύριες επιδράσεις (x_2, x_{14}) και 2 ενεργές αλληλεπιδράσεις δύο παραγόντων ($x_{2,11}$, $x_{7,14}$). Η L_1-norm Huberized SVM ανίχνευσε 2 ενεργές κύριες επιδράσεις (x_2, x_{14}) και 4 ενεργές αλληλεπιδράσεις δύο παραγόντων ($x_{2,8}$, $x_{2,11}$, $x_{6,14}$, $x_{7,14}$).

Συμπεραίνουμε λοιπόν από τα αποτελέσματα της πραγματικής ανάλυσης δεδομένων ότι η L_1-norm SVM δίνει πολύ καλύτερα αποτελέσματα από την L_1-norm Huberized SVM για το κρησάρισμα κύριων επιδράσεων σε έναν υπερκορεσμένο σχεδιασμό, ενώ η L_1-norm Huberized SVM υπερέχει της L_1-norm SVM όταν οι όροι των αλληλεπιδράσεων συμπεριλαμβάνονται στο μοντέλο.

Πίνακας 12.6: Συγκριτική απόδοση μοντέλων

Μέθοδος	Μοντέλο	Σφάλμα Εκπαίδευσης	Σφάλμα Πρόβλεψης	Ακρίβεια	Ευαισθησία	Ειδικότητα
L_1-norm SVM	$Model_{database}$	0.04	0.08	0.96	0.20	0.99
L_1-norm Huberized SVM	$Model_{database}$	0.04	0.08	0.96	0.20	0.99
L_1-norm SVM	$Model_{SSD}$	0.01	0.00	0.99	0.99	0.99
L_1-norm Huberized SVM	$Model_{SSD}$	0.11	0.00	0.89	0.86	0.99
L_1-norm SVM	$Model_{SSDinter}$	0.03	0.08	0.97	0.91	0.99
L_1-norm Huberized SVM	$Model_{SSDinter}$	0.01	0.00	0.99	0.99	0.99

Πίνακας 12.7: Οι ενεργές επιδράσεις

Μέθοδος	x_2	x_5	x_6	x_7	x_8	x_{11}	x_{13}	x_{14}	$x_{2,8}$	$x_{2,11}$	$x_{6,14}$	$x_{7,14}$
L_1-norm SVM ($Model_{SSD}$)	x			x			x	x				
L_1-norm Huberized SVM ($Model_{SSD}$)	x							x				
L_1-norm SVM ($Model_{SSDinter}$)	x							x		x		x
L_1-norm Huberized SVM ($Model_{SSDinter}$)	x							x	x	x	x	x

12.5 Συμπεράσματα

Στο κεφάλαιο αυτό προτείναμε μία μέθοδο ανάλυσης δεδομένων από υπερκορεσμένους σχεδιασμούς δύο επιπέδων, με την παραδοχή των γενικευμένων γραμμικών μοντέλων για δίτιμη απόκριση, υιοθετώντας τη μεθοδολογία των ποινικοποιημένων μηχανών διανυσματικής υποστήριξης, η οποία προσδιορίζει ποιες μεταβλητές επηρεάζουν σημαντικά την απόκριση και εκτελεί αυτόματη επιλογή μεταβλητών για αυτό το σκοπό. Διεξήγαμε διαφορετικές μεθόδους στατιστικής ανάλυσης, θεωρώντας διάφορα σενάρια, και τα αποτελέσματα τα οποία προέκυψαν είναι πολύ ικανοποιητικά, δεδομένου ότι παρέχουν οδηγίες για τη χρήση και την επιλογή της κατάλληλης τεχνικής των ποινικοποιημένων μηχανών διανυσματικής υποστήριξης σε δύο διαφορετικές περιπτώσεις. Η L_1-norm SVM είναι κατάλληλη για το κρησάρισμα υπερκορεσμένων σχεδιασμών μόνο κύριων επιδράσεων, ενώ η L_1-norm Huberized SVM είναι κατάλληλη όταν οι όροι αλληλεπίδρασης συμπεριλαμβάνονται στο μοντέλο. Τα παραπάνω συμπεράσματα τα οποία προέκυψαν από την εκτενή εμπειρική μελέτη προσομοίωσης συμπίπτουν και για την ανάλυση πραγματικών δεδομένων την οποία διεξήγαμε μετέπειτα.

ΚΕΦΑΛΑΙΟ **13**

Κατασκευή Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις

Man's longing for perfection
finds expression in
the theory of optimization.

—*Beightler, Phillips and Wilde (1979)*

Στο δέκατο τρίτο αυτό κεφάλαιο μελετάμε τη δομή ταυτόσημων επιδράσεων στους υπερκορεσμένους σχεδιασμούς συμπεριλαμβανομένων των αλληλεπιδράσεων. Παρουσιάζουμε μία μέθοδο βελτιστοποίησης, η οποία αποτελεί μία τεχνική προς τα πίσω απαλοιφής καθοδηγούμενη από ένα σχεδιασμό (backward elimination design-driven optimization-BEDDO), με ένα κύριο στόχο στο μυαλό, να μελετήσουμε τη δομή ταυτόσημων επιδράσεων και να εξαλείψουμε τους παράγοντες που ανιχνεύονται να είναι ταυτόσημοι μεταξύ τους μέσα στον υπερκορεσμένο σχεδιασμό. Για την υλοποίηση της προτεινόμενης μεθόδου BEDDO, ενσωματώνουμε και συνδυάζουμε στατιστικά μέτρα βασισμένα στη συσχέτιση, τα οποία έχουν ληφθεί από δύο επιστημονικά πεδία, την κλασική θεωρία ελέγχου και τους σχεδιασμούς πειραμάτων, και παρουσιάζουμε επίσης ένα νέο κριτήριο βελτιστότητας το οποίο αποτελεί μία τροποποιημένη μορφή του Cronbach άλφα συντελεστή. Με αυτό τον τρόπο και με τη βοήθεια των υπολογιστών παρέχουμε μία νέα κλάση μη ισορροπημένων υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις, οι οποίοι απορρέουν άμεσα από τη μέθοδο βελτιστοποίησης BEDDO. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου παρουσιάζονται στην επιστημονική εργασία [43].

13.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Στο Κεφάλαιο 11 αναφερθήκαμε στο γεγονός ότι οι περισσότερες από τις μεθόδους κατασκευής και ανάλυσης των υπερκορεσμένων σχεδιασμών θεωρούν μόνο τις κύριες επιδράσεις, και αγνοούν ή θεωρούν αμελητέα την ύπαρξη των αλληλεπιδράσεων. Οι Hamada και Wu [104], και πρόσφατα οι Phoa, Wong και Xu [183] επεσήμαναν την ανάγκη του να λαμβάνονται υπόψη οι σημαντικές αλληλεπιδράσεις στο κρησάρισμα των πειραματικών σχεδιασμών, δεδομένου ότι η αγνόησή τους μπορεί να επηρεάσει σημαντικά τη στατιστική συμπερασματολογία. Σε αυτό το πνεύμα, προτείνουμε μία μέθοδο βελτιστοποίησης (BEDDO), με στόχο να εξαλείψουμε τους παράγοντες που ανιχνεύονται να είναι ταυτόσημοι μεταξύ τους μέσα στο σχεδιασμό και με αυτό τον τρόπο παράγουμε μία νέα κλάση μη ισορροπημένων υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις.

Από καιρό στη βιβλιογραφία, μία βασική παραδοχή-υπόθεση για τους υπερκορεσμένους σχεδιασμούς είναι ότι πρέπει να είναι ισορροπημένοι. Επειδή στη βιβλιογραφία, υφίστανται πολυάριθμες χρήσεις της έννοιας της "ισορροπίας", στην παρούσα διατριβή, αναφερόμαστε στους σχεδιασμούς των οποίων κάθε παράγοντας παρατηρείται τον ίδιο αριθμό φορών στην υψηλή και χαμηλή στάθμη ως "ισορροπημένους", και σε διαφορετική περίπτωση ως "μη-ισορροπημένους". Στις περισσότερες εφαρμοσμένες μελέτες κρησαρίσματος υποτίθεται ότι ένας παράγοντας πρέπει να είναι ισορροπημένος, ώστε να μεγιστοποιεί την πληροφορία με ελάχιστο κόστος. Στην προτεινόμενη μέθοδο BEDDO, το γεγονός ότι χαλαρώνουμε αυτό τον περιορισμό της ισορροπίας για τη δημιουργία μη ισορροπημένων υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις, παρέχει σχεδιασμούς με καλές στατιστικές ιδιότητες που συγκρίνονται ευνοϊκά με ισορροπημένους ομόλογούς τους για πολλά κριτήρια βελτιστοποίησης, συμπεριλαμβανομένου του $E(s^2)$. Παράγουμε μη ισορροπημένους υπερκορεσμένους σχεδιασμούς που στην ουσία μπορούν να χαρακτηριστούν ως μερικώς ισορροπημένοι, υπό την έννοια ότι m_1 από τις m στήλες είναι ισορροπημένες (έχουν ίσο αριθμό "-1's" και "1's") και οι υπόλοιπες $(m - m_1)$ $(= m_2)$ στήλες είναι μη ισορροπημένες. Τέτοιου τύπου σχεδιασμοί με την ιδιότητα της μερικής ισορροπίας μπορεί να προκύψουν συχνά στην πράξη εφόσον η ύπαρξη των αλληλεπιδράσεων θεωρηθεί είτε στην κατασκευή είτε στην ανάλυση των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων. Οι μερικώς ισορροπημένοι σχεδιασμοί μπορεί να ενδιαφέρουν έναν πειραματιστή ο οποίος χρειάζεται σχεδιασμούς όσο το δυνατόν πιο ισορροπημένους και μπορεί να είναι προτιμητέοι σε περιπτώσεις στις οποίες τα επίπεδα συγκεκριμένων παραγόντων σχετίζονται με υψηλά κόστη. Οι Allen και Bernshteyn [4] επίσης επισήμαναν ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί μπορεί να είναι προτιμητέοι στους πειραματιστές που θα ήθελαν να ελαχιστοποιήσουν το πειραματικό κόστος σε περιπτώσεις που το να επιτευχθούν συγκεκριμένα επίπεδα κάποιων παραγόντων κοστίζει πολύ.

Επιπρόσθετα, όπως προαναφέραμε στο Κεφάλαιο 11 οι παραδοσιακές τεχνικές πειραματισμού για τη βελτίωση ενός προϊόντος ή μιας διαδικασίας ξεκινούν με την εξέταση μόνο των κύριων επιδράσεων των παραγόντων στο πρώτο στάδιο, και χρησιμοποιούν περαιτέρω πειραματισμό (δεύτερο στάδιο) για να εξετάσουν τις αλληλεπιδράσεις μεταξύ των παραγόντων των οποίων οι κύριες επιδράσεις ανιχνεύτηκαν ως σημαντικές σε σχέση με την απόκριση, από τα αποτελέσματα τα οποία προέκυψαν στο πρώτο στάδιο του πειραματισμού. Αυτή η προσέγγιση απαιτεί την παραδοχή ότι σταθερά ισχύει η αρχή της ισχυρής κληρονομικότητας που αναφέρει ότι οι αλληλεπιδράσεις προκύπτουν μόνο μεταξύ αυτών των παραγόντων με ενεργές κύριες επιδράσεις. Ωστόσο, τα αποτελέσματα που προκύπτουν από πολλές πρακτικές εφαρμογές παρέχουν ενδείξεις ότι η αρχή της ισχυρής κληρονομικότητας αποτυγχάνει αρκετά συχνά [165], [222], [206], [68]. Κατά συνέπεια, η προτεινόμενη μέθοδος BEDDO πλεονεκτεί δεδομένου ότι επιτρέπει τη χαλάρωση της αρχής της ισχυρής κληρονομικότητας, και συμπεριλαμβάνει τις αλληλεπιδράσεις δύο παραγόντων στο αρχικό στάδιο του πειραματισμού, προκειμένου να ανιχνεύσει και να αποκλείσει τις αλληλεπιδράσεις ελάχιστης σημα-

ντικότητας και πληροφόρησης, όσο το δυνατόν γρηγορότερα.

Παρακινούμενοι από τα παραπάνω, προτείνουμε μία μέθοδο βελτιστοποίησης η οποία επιτυγχάνει να κατανοήσει τα υποδείγματα μεταξύ του συνόλου των στηλών-παραγόντων (χωρίς να θεωρεί μεταβλητή απόκρισης και να απαιτεί τη γνώση της), αξιολογεί το βαθμό της πλησιέστερης ορθογωνιότητας (near-orthogonality), και τελικά επιλέγει τον πιο πληροφοριακό σχεδιασμό συμπεριλαμβανομένων των αλληλεπιδράσεων. Η χρήση του BEDDO έχει ως υποπροϊόν το de-aliasing των παραγόντων και επιτυγχάνει να ανακτήσει ένα σχεδιασμό όσο το δυνατόν πλησιέστερα στον ορθογώνιο σε σχέση με κάποια στατιστικά μέτρα βασισμένα στη συσχέτιση. Με αυτό τον τρόπο, η προτεινόμενη μέθοδος επιτρέπει στον πειραματιστή να αποκτήσει ένα είδος εκ των προτέρων γνώσης των πιθανών σημαντικών παραγόντων, και το γεγονός αυτό οδηγεί σε λιγότερο πολύπλοκα και ακριβά πειράματα για περαιτέρω εκτέλεση.

13.2 Τροποποίηση του Cronbach Άλφα Συντελεστή

Σε αυτή την ενότητα, παρουσιάζουμε το γεγονός το οποίο μας παρακίνησε να τροποποιήσουμε τον Cronbach άλφα συντελεστή, και να τον υιοθετήσουμε ως ένα κριτήριο βελτιστοποίησης στους υπερκορεσμένους σχεδιασμούς. Επίσης, παρουσιάζουμε κάποιες από τις ιδιότητές του, κάτω φράγματα για αυτό το τροποποιημένο κριτήριο και αξιολογούμε την απόδοσή του μέσω μιας συγκριτικής μελέτης.

13.2.1 Κίνητρο Χρήσης και Τροποποίησης του Cronbach Άλφα Συντελεστή

Ο σωστός σχεδιασμός μιας πειραματικής μελέτης διαδραματίζει πολύ σημαντικό ρόλο για τρεις πολύ βασικούς λόγους. Πρώτον, μπορεί να επηρεάσει σημαντικά τη στατιστική συμπερασματολογία, δεύτερον μία φτωχά σχεδιασμένη μελέτη μπορεί να μην είναι ικανή να απαντήσει στις επιστημονικές υποθέσεις που τίθενται και τρίτον η πολυπλοκότητα και το υπολογιστικό κόστος των πειραμάτων αυξάνεται συνεχώς στις μέρες μας. Συμπεραίνουμε λοιπόν ότι υπάρχει μία εμφανής ανάγκη για το σχεδιασμό αποτελεσματικών μελετών με στόχο να μειώσουμε τις απαιτούμενες πηγές υπολογιστικού κόστους και χρόνου της έρευνας που θα διεξάγουμε. Μόνο μία προσεκτική θεώρηση ενός σχεδιασμού μπορεί να οδηγήσει σε μία αποτελεσματική μελέτη με ακριβείς εκτιμήσεις περισσότερων επιδράσεων με το ίδιο κόστος.

Η διαδικασία της επιλογής των καλύτερων αντικειμένων (items) από ένα αρχικό σύνολο διαθέσιμων αντικειμένων σε μία μελέτη ονομάζεται ανάλυση αντικειμένων ή ανάλυση αξιοπιστίας. Ο όρος αντικείμενο συνήθως χρησιμοποιείται σε μία ανάλυση αξιοπιστίας, αλλά με τον όρο αντικείμενα μπορεί να εννοούμε οτιδήποτε, όπως ερωτήσεις, ποσοδείκτες, μετρήσεις των επιπέδων των παραγόντων σε ένα βιομηχανικό πείραμα. Τα αντικείμενα πολύ συχνά αναφέρονται και ως μεταβλητές. Η ανάλυση αξιοπιστίας έρχεται στο προσκήνιο όταν μεταβλητές σε κλιμακωτή μορφή χρησιμοποιούνται ως προγνωστικές μεταβλητές σε αντικειμενικά προβλήματα στατιστικής μοντελοποίησης. Η ανάλυση αξιοπιστίας είναι ιδιαίτερα χρήσιμη όταν οι μεταβλητές προτίθεται να χρησιμοποιηθούν για περαιτέρω προγνωστικές αναλύσεις. Ένα από τα πιο ευρέως χρησιμοποιούμενα εργαλεία για την εκτίμηση της αξιοπιστίας είναι ο συντελεστής άλφα (γνωστός ως "Cronbach's α"). Προτάθηκε από τον Cronbach [56] και αποτελεί γενίκευση της φόρμουλας που δόθηκε από τους Kuder-Richardson για τα δυαδικά δεδομένα (KR-20) [138]. Η βιβλιογραφία προσφέρει πολλές διαφορετικές περιγραφές του συντελεστή άλφα. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στην εργασία [51] όπου αναφέρονται και ενσωματώνονται ορισμένες από τις πολλές περιγραφές που έχουν γίνει σχετικά με το συντελεστή άλφα. Σε αυτό το σημείο θεωρούμε σημαντικό να εξηγήσουμε το τι ακριβώς μετράει ο συντελεστής άλφα και το σκοπό χρήσης του, προκειμένου να αναδείξουμε την αποτελεσματική χρήση του για το κρησάρισμα των υπερκορεσμένων σχεδιασμών. Ο συντελεστής άλφα είναι μία συνάρτηση του βαθμού ως προς το ποιοί παράγοντες σε ένα πείραμα

έχουν υψηλές ομοιότητες και συνεπώς χαμηλή μοναδικότητα. Με άλλα λόγια, ο συντελεστής άλφα αποτελεί συνάρτηση της αλληλοσύνδεσης και της αλληλεξάρτησης μεταξύ των παραγόντων. Από τεχνικής απόψεως, ο συντελεστής άλφα είναι μία συνάρτηση του αριθμού των παραγόντων σε ένα πείραμα καθώς επίσης και μία συνάρτηση της μέσης αλληλοσυσχέτισης των παραγόντων σε ένα πείραμα.

Ο συντελεστής άλφα χρησιμοποιείται ευρέως στις κοινωνικές επιστήμες, στις επιχειρήσεις και σε άλλους επιστημονικούς κλάδους για την αξιολόγηση της κλίμακας αξιοπιστίας. Στους σχεδιασμούς πειραμάτων, η κλιμάκωση των επιπέδων του παράγοντα αντιστοιχεί σε μετατροπή των επιπέδων του παράγοντα έτσι ώστε η υψηλή τιμή να γίνεται "+1" και η χαμηλή τιμή να γίνεται "-1". Το γεγονός ότι ο συντελεστής άλφα αποτελεί ένα μέτρο της αλληλοσυσχέτισης των παραγόντων σε ένα πείραμα, αποτέλεσε το κίνητρο ώστε να υιοθετήσουμε μία τροποποιημένη μορφή του (σε συνδυασμό με άλλα στατιστικά μέτρα συσχέτισης), και να το χρησιμοποιήσουμε από μία άλλη οπτική γωνία, δηλαδή, να διερευνήσουμε τη δομή ταυτόσημων επιδράσεων σε υπερκορεσμένους σχεδιασμούς δύο επιπέδων που περιλαμβάνουν αλληλεπιδράσεις, προκειμένου να ανιχνεύσουμε και να εξαλείψουμε τους παράγοντες του σχεδιασμού που είναι αλληλένδετοι μεταξύ τους.

13.2.2 Cronbach Άλφα Συντελεστής

Ας υποθέσουμε ότι μετράμε την ποσότητα X που είναι το άθροισμα των m συνιστωσών $x_1, x_2, \ldots, x_m$. Η κλασική φόρμουλα του Cronbach άλφα συντελεστή δίνεται από τη σχέση

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} \sigma_{x_j}^2}{\sigma_X^2}\right),$$

όπου σ_X^2 είναι η διακύμανση του X, και $\sigma_{x_j}^2$ η διακύμανση της j-οστής συνιστώσας του X.

Ο συντελεστής άλφα υπολογίζεται από έναν πίνακα συνδιακυμάνσεων ως

$$\alpha_{raw} = \frac{m\bar{c}}{\bar{v} + (m-1)\bar{c}},$$

όπου $\bar{v}$ είναι ο μέσος των m στοιχείων στην κύρια διαγώνιο και $\bar{c}$ είναι ο μέσος των στοιχείων εκτός της διαγωνίου στον πίνακα συνδιακυμάνσεων.

Η τυποποιημένη μορφή του συντελεστή άλφα υπολογίζεται από έναν πίνακα συσχετίσεων ως

$$\alpha_{std} = \frac{m\bar{r}}{1 + (m-1)\bar{r}},$$

όπου το $\bar{r}$ είναι ο μέσος των $m(m-1)/2$ συντελεστών συσχέτισης ενός άνω τριγωνικού ή κάτω τριγωνικού πίνακα συσχετίσεων. Σημειώνουμε εδώ ότι αν όλες οι μεταβλητές έχουν μηδενική μέση τιμή και διακύμανση ίση με τη μονάδα, η κλασική μορφή του συντελεστή άλφα και η τυποποιημένη μορφή του είναι ισοδύναμες. Ο δείκτης *std* (για την τυποποιημένη μορφή) δείχνει ότι ο Cronbach άλφα συντελεστής υπάρχει και σε μη-τυποποιημένη εκδοχή *raw*, η οποία βασίζεται σε συσχετίσεις και όχι σε συνδιακυμάνσεις. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στην εργασία [50] για περισσότερες λεπτομέρειες σχετικά με την εκτίμηση του συντελεστή άλφα.

13.2.3 Το Τροποποιημένο Κριτήριο

Καθ'όλη την προτεινόμενη μέθοδο, θεωρούμε υπερκορεσμένους σχεδιασμούς που επιτρέπουν σε κάθε παράγοντα να έχει μόνο δύο επίπεδα στο πείραμα. Αναφερόμαστε στο επίπεδο "+1" ως το υψηλό επίπεδο, και στο επίπεδο "-1" ως το χαμηλό επίπεδο. Όταν θεωρείται ένας

υπερκορεσμένος σχεδιασμός, η εγκατάλειψη της ορθογωνιότητας είναι αναπόφευκτη. Επειδή η έλλειψη ορθογωνιότητας συνήθως οδηγεί και σε χαμηλότερη απόδοση, ο πειραματιστής επιδιώκει σχεδιασμούς όσο το δυνατόν πιο κοντά στον ορθογώνιο [149].

Έστω X ένας υπερκορεσμένος σχεδιασμός για τον οποίο m είναι ο αριθμός των παραγόντων, n είναι ο αριθμός των πειραματικών εκτελέσεων ($m \geq n$), και οι τιμές των επιπέδων για τις στήλες των παραγόντων c_i και c_j κωδικοποιούνται με ± 1. Ένας απλός τρόπος για να μετρηθεί ο βαθμός της μη ορθογωνιότητας μεταξύ δύο στηλών c_i και c_j είναι να θεωρήσουμε τη συσχέτισή τους (r_{ij}). Η συσχέτιση μεταξή δύο στηλών c_i και c_j ενός υπερκορεσμένου σχεδιασμού υπολογίζεται από την κλασική φόρμουλα του Pearson. Επιθυμούμε την ελάχιστη δυνατή τιμή για τη συσχέτιση r_{ij} μεταξύ όλων των ζευγών των στηλών. Όταν $r_{ij} = 0$ σημαίνει ότι οι c_i και c_j είναι ασυσχέτιστες και όταν $r_{ij} = \pm 1$ σημαίνει ότι οι c_i και c_j είναι πλήρως ταυτόσημες. Αξίζει να σημειωθεί εδώ ότι στην περίπτωση ενός ισορροπημένου υπερκορεσμένου σχεδιασμού η συσχέτιση ορίζεται ως $r_{ij} = \frac{c_i' c_j}{n}$. Η μεγαλύτερη απόλυτη τιμή της συσχέτισης r_{ij} μεταξύ όλων των ζευγών των στηλών συμβολίζεται με r_{max}.

Παρακάτω παρουσιάζουμε την τροποποιημένη μορφή του Cronbach άλφα συντελεστή. Ο τροποποιημένος συντελεστής άλφα βασίζεται σε ένα ευρέως χρησιμοποιούμενο κριτήριο στους υπερκορεσμένους σχεδιασμούς, το οποίο εισήγαγε ο Lin [149] και καλείται ως μέση τετραγωνική συσχέτιση. Η μέση τετραγωνική συσχέτιση ρ^2 είναι ο μέσος των r_{ij}^2 και ορίζεται ως

$$\rho^2 = \frac{\sum\limits_{i<j} r_{ij}^2}{\binom{m}{2}}. \tag{13.1}$$

Η εναλλακτική μορφή του Cronbach άλφα συντελεστή υπολογίζεται από έναν πίνακα συσχετίσεων. Ο τροποποιημένος συντελεστής άλφα που προτείνουμε συμβολίζεται με $\alpha_{rhosquared}$ και ορίζεται ως

$$\alpha_{rhosquared} = \frac{m\rho^2}{1 + (m-1)\rho^2}, \tag{13.2}$$

όπου το ρ^2 είναι η μέση τετραγωνική συσχέτιση ή με άλλα λόγια ο μέσος των $m(m-1)/2$ τετραγώνων των συντελεστών συσχέτισης που βρίσκονται πάνω από τη διαγώνιο του πίνακα συσχετίσεων. Το $\alpha_{rhosquared}$ υποδεικνύει τον ολικό συντελεστή άλφα ενός υπερκορεσμένου σχεδιασμού αποτελούμενου από m στήλες παραγόντων.

Αξίζει να σημειωθεί εδώ ότι η συσχέτιση μεταξύ δύο στηλών ενός ισορροπημένου υπερκορεσμένου σχεδιασμού μπορεί να οριστεί εναλλακτικά ως $r_{ij} = \frac{s_{ij}}{n}$, όπου το s_{ij} είναι το *i,j*-οστό στοιχείο του $X^{\mathrm{T}}X$ πίνακα πληροφορίας. Ο όρος s_{ij} μετρά το βαθμό μη ορθογωνιότητας μεταξύ των παραγόντων c_i και c_j. Εάν το $s_{ij} = 0$, οι παράγοντες c_i και c_j είναι ορθογώνιοι.

Αντικαθιστώντας το $r_{ij} = \frac{s_{ij}}{n}$ στη σχέση (13.1) λαμβάνουμε τη σχέση

$$\rho^2 = \frac{\frac{1}{n^2}\sum\limits_{i<j} s_{ij}^2}{\binom{m}{2}}, \tag{13.3}$$

όπου το $\sum\limits_{i<j} s_{ij}^2$ είναι το άθροισμα των τετραγώνων των στοιχείων πάνω από τη διαγώνιο του $X^{\mathrm{T}}X$. Αντικαθιστώντας τη σχέση (13.3) στη σχέση (13.2) μπορούμε εναλλακτικά να υπολογίσουμε το συντελεστή $\alpha_{rhosquared}$ από τον πίνακα πληροφορίας ενός ισορροπημένου υπερκορεσμένου σχεδιασμού.

Ας σημειωθεί εδώ ότι το $\rho^2 = \frac{E(s^2)}{n^2}$, όπως αναφέρεται στην εργασία [149], όπου το $E(s^2) = \frac{\sum\limits_{i<j} s_{ij}^2}{\binom{m}{2}}$ είναι το αποκαλούμενο "μέσο s^2" το οποίο εισήχθη από τους Booth και Cox [21]. Για έναν ισορροπημένο σχεδιασμό X με m παράγοντες σε δύο επίπεδα ο κάθε

ένας, και n εκτελέσεις ($m \geq n$), οι Nguyen [172] και οι Tang και Wu [215], ανεξάρτητα, παρείχαν ένα κάτω φράγμα

$$E(s^2) \geq \frac{n^2(m-n+1)}{(n-1)(m-1)} \tag{13.4}$$

για το $E(s^2)$.

Ο συντελεστής $\alpha_{rhosquared}$ μπορεί να ερμηνευθεί ως το κατώτερο όριο εκτίμησης της αλληλοσυσχέτισης μεταξύ των παραγόντων ενός πειράματος, και κυμαίνεται από 0 (χαμηλότερος βαθμός αλληλοσυσχέτισης) έως 1 (υψηλότερος βαθμός αλληλοσυσχέτισης). Η έννοια της ορθογωνιότητας σχετίζεται με τη δομή συσχετίσεων μεταξύ των παραγόντων ενός πειραματικού σχεδιασμού. Ένας σχεδιασμός που δεν εμφανίζει συσχετίσεις καλείται ορθογώνιος σχεδιασμός. Εάν οι στήλες των παραγόντων ενός σχεδιασμού εμφανίζουν μηδενικές συσχετίσεις, τότε κάθε στήλη του σχεδιασμού θα ενεργεί ανεξάρτητα από όλες τις άλλες στήλες. Έτσι, το $\alpha_{rhosquared}$ αναμένεται να ισούται με το μηδέν μόνο στην κλάση των ορθογώνιων σχεδιασμών. Στην κλάση των υπερκορεσμένων σχεδιασμών που μελετάμε στο παρόν κεφάλαιο, επιθυμούμε το χαμηλότερο βαθμό αλληλοσυσχέτισης μεταξύ των παραγόντων, όμως δεν είναι δυνατόν να έχουμε ένα συντελεστή $\alpha_{rhosquared}$ ίσο με το μηδέν, λόγω της δομής του σχεδιασμού. Επιπρόσθετα, σημειώνουμε εδώ ότι από τη στιγμή που ο συντελεστής $\alpha_{rhosquared}$ έχει ανάλογη μορφή με το $E(s^2)$, βασισμένοι στη σχέση (13.2) μπορούμε εύκολα να δείξουμε ότι ο συντελεστής $\alpha_{rhosquared}$ είναι μία αυστηρά αύξουσα συνάρτηση του $E(s^2)$, και ως αποτέλεσμα, κάτω φράγματα για το προτεινόμενο κριτήριο μπορούν εύκολα να αποκτηθούν βάση των κάτω φραγμάτων που ισχύουν για το $E(s^2)$. Οι παραπάνω λόγοι όπως ήταν φυσικό μας παρακίνησαν να ψάξουμε για κάτω φράγματα για το συντελεστή $\alpha_{rhosquared}$. Παρουσιάζουμε παρακάτω ένα κάτω φράγμα για το συντελεστή $\alpha_{rhosquared}$, για το οποίο απαραίτητη προϋπόθεση αποτελεί ο σχεδιασμός να είναι ισορροπημένος.

Θεώρημα 13.1. *Για έναν ισορροπημένο σχεδιασμό X με m παράγοντες σε δύο επίπεδα ο κάθε ένας και n πειραματικές εκτελέσεις ($m \geq n$), το κάτω φράγμα για το $\alpha_{rhosquared}$ είναι $(m-n+1)/(m-1)$.*

Απόδειξη. Δεδομένου ότι $\rho^2 = \frac{E(s^2)}{n^2}$, η σχέση

$$\alpha_{rhosquared} = \frac{m\rho^2}{1+(m-1)\rho^2},$$

μπορεί εναλλακτικά να γραφεί ως

$$\alpha_{rhosquared} = \frac{m\frac{E(s^2)}{n^2}}{1+(m-1)\frac{E(s^2)}{n^2}},$$

η οποία δίνει ως αποτέλεσμα

$$\alpha_{rhosquared} = \frac{mE(s^2)}{n^2+(m-1)E(s^2)}.$$

Χρησιμοποιώντας τη σχέση (13.4), έχουμε ότι

$$\alpha_{rhosquared} \geq \frac{m\frac{n^2(m-n+1)}{(n-1)(m-1)}}{n^2+(m-1)\frac{n^2(m-n+1)}{(n-1)(m-1)}},$$

η οποία δίνει ως αποτέλεσμα

$$\alpha_{rhosquared} \geq \frac{mn^2(m-n+1)}{n^2(m-1)[(n-1)+(m-n+1)]},$$

η οποία με τη σειρά της δίνει ως αποτέλεσμα

$$\alpha_{rhosquared} \geq \frac{(m-n+1)}{(m-1)}. \tag{13.5}$$

□

13.2.4 Συγκριτική Απόδοση του Τροποποιημένου Κριτηρίου

Προκειμένου να αξιολογηθεί η απόδοση του τροποποιημένου κριτηρίου, υπολογίσαμε και συγκρίναμε διάφορες τιμές κριτηρίων (ρ^2, r_{max}, $E(s^2)$, $\alpha_{rhosquared}$) πάνω σε μία σειρά από $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς (για τους οποίους $m = 2(n-1)$ όπως αυτοί παρουσιάζονται στην εργασία [130]). Τα ληφθέντα αποτελέσματα σύμφωνα με κάθε υπό εξέταση κριτήριο παρουσιάζονται στον Πίνακα 13.1. Σημειώστε εδώ ότι όλα τα κριτήρια που εξετάζονται είναι τύπου smaller-the-better.

Πίνακας 13.1: Τιμές κριτηρίων για ισορροπημένους $E(s^2)$-βέλτιστους ΥΣ

Εκτελέσεις	Παράγοντες	ρ^2	r_{max}	$E(s^2)$	$\alpha_{rhosquared}$
6	10	0.1111	0.3333	4.0000	0.5556
8	14	0.0769	0.5000	4.9231	0.5385
10	18	0.0588	0.6000	5.8824	0.5294
12	22	0.0476	0.3333	6.8571	0.5238
14	26	0.0400	0.4286	7.8400	0.5200
16	30	0.0345	0.2500	8.8276	0.5172
18	34	0.0303	0.3333	9.8182	0.5152
20	38	0.0270	0.2000	10.8108	0.5135
22	42	0.0244	0.2727	11.8049	0.5122

Για την ισορροπημένη περίπτωση, οι Tang και Wu [215] έδειξαν ότι το κάτω φράγμα (13.4) για το $E(s^2)$ μπορεί να επιτευχθεί εάν και μόνο εάν $m = q(n-1)$ και $n \equiv 0 \bmod 4$ ή εάν $m = 2q(n-1)$ και $n \equiv 2 \bmod 4$ για κάποιο θετικό ακέραιο q. Άρα, το n είναι άρτιος. Ένας σχεδιασμός καλείται $E(s^2)$-βέλτιστος όταν αυτό το φράγμα επιτυγχάνεται. Όταν το $n \equiv 0 \bmod 4$, οι πιθανές τιμές της συσχέτισης είναι $0, 4/n, 8/n, \ldots, 1$, και όταν $n \equiv 2 \bmod 4$, οι πιθανές τιμές της συσχέτισης είναι $2/n, 6/n, 10/n, \ldots, 1$.

Παρατηρούμε από τον Πίνακα 13.1 ότι οι θεωρούμενοι ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν $m = 2(n-1)$ παράγοντες και $n \equiv 0 \bmod 4$ ή $n \equiv 2 \bmod 4$ πειραματικές εκτελέσεις. Αυτοί οι υπερκορεσμένοι σχεδιασμοί επιτυγχάνουν το κάτω φράγμα (13.4) για το $E(s^2)$, και άρα είναι βέλτιστοι σε σχέση με το κριτήριο $E(s^2)$. Επιπλέον, παρατηρούμε από τον Πίνακα 13.1 ότι αυτοί οι υπερκορεσμένοι σχεδιασμοί επίσης επιτυγχάνουν το κάτω φράγμα (13.5) για το κριτήριο $\alpha_{rhosquared}$, και άρα είναι βέλτιστοι σε σχέση με το κριτήριο $\alpha_{rhosquared}$. Συμπεραίνουμε λοιπόν ότι οι ισορροπημένοι υπερκορεσμένοι σχεδιασμοί οι οποίοι επιτυγχάνουν το κάτω φράγμα για το $E(s^2)$, επίσης επιτυγχάνουν το κάτω φράγμα για το $\alpha_{rhosquared}$. Οι δύο τελευταίες στήλες στον Πίνακα 13.1 δείχνουν ότι οι ισορροπημένοι $E(s^2)$-βέλτιστοι υπερκορεσμένοι σχεδιασμοί αποδίδουν καλά με το κριτήριο $\alpha_{rhosquared}$.

Τα παραδείγματα στον Πίνακα 13.1 υποδεικνύουν ότι οι ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν παρόμοια απόδοση για τα κριτήρια $E(s^2)$ και $\alpha_{rhosquared}$. Το γεγονός ότι το τροποποιημένο κριτήριο βελτιστοποίησης έχει παρόμοια απόδοση σε σύγκριση με το $E(s^2)$ αποτελεί μία στατιστική ιδιότητα η οποία αναμένεται δεδομένου ότι ο συντελεστής $\alpha_{rhosquared}$ έχει ανάλογη μορφή με το $E(s^2)$. Ωστόσο, το κύριο πλεονέκτημα του τροποποιημένου κριτηρίου είναι η εύκολη ενσωμάτωσή του στη μέθοδο BEDDO και η άμεση αντιστοιχία του με τον κύριο στόχο του κρησαρίσματος και της εξάλειψης των αλληλεπιδράσεων μικρής πληροφορήσεως το ταχύτερο δυνατόν. Με τον τρόπο αυτό, η BEDDO μέθοδος βελτιστοποίησης κάτω από το τροποποιημένο κριτήριο και με τη βοήθεια των υπολογιστών παράγει μία νέα κλάση μη ισορροπημένων υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις.

13.3 Η Προτεινόμενη Μέθοδος Βελτιστοποίησης

Η προτεινόμενη μέθοδος βελτιστοποίησης αποτελεί μία τεχνική προς τα πίσω απαλοιφής καθοδηγούμενη από ένα σχεδιασμό, και αποτελεί μία προσέγγιση βασισμένη στη συσχέτιση με στόχο το κρησάρισμα ενός σχεδιασμού. Πριν παρουσιάσουμε την προτεινόμενη μέθοδο, περιγράφουμε παρακάτω τα στατιστικά μέτρα στα οποία βασίζεται η μέθοδός μας.

13.3.1 Στατιστικά Μέτρα Συσχέτισης

Το πρώτο μέτρο το οποίο υπολογίζουμε είναι ο συνολικός συντελεστής άλφα ενός υπερκορεσμένου σχεδιασμού, δηλαδή, το $\alpha_{rhosquared}$ όπως αυτό ορίζεται στη σχέση (13.2). Η τιμή του συντελεστή άλφα επηρεάζεται από τον αριθμό των παραγόντων καθώς επίσης και από το βαθμό της αλληλοσύνδεσης μεταξύ των παραγόντων. Στους υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις, ο συνολικός συντελεστής άλφα έχει την τάση να είναι αρκετά υψηλός, δείχνοντας ότι κάποιοι παράγοντες είναι περιττοί και θα πρέπει να εξαλειφθούν. Για αυτόν το λόγο, η τιμή του συνολικού συντελεστή άλφα σε συνδυασμό με άλλα στατιστικά μέτρα χρησιμοποιούνται για να αναδείξουν την αλληλοσύνδεση μεταξύ των παραγόντων του υπό εξέταση σχεδιασμού. Τα άλλα δύο μέτρα που θεωρούμε στον αλγόριθμο BEDDO είναι ο r_{corr} συντελεστής και ο $\hat{\alpha}_{rhosquared}$ συντελεστής εάν ένας συγκεκριμένος παράγοντας αφαιρεθεί από το σχεδιασμό. Τα δύο μέτρα αυτά (r_{corr} και $\hat{\alpha}_{rhosquared}$) διαδραματίζουν σημαντικό ρόλο στην προτεινόμενη μέθοδο δεδομένου ότι και τα δύο αξιολογούν τις συνέπειες του να μην θεωρήσουμε ένα συγκεκριμένο παράγοντα τη φορά ως μέρος του σχεδιασμού.

Ο συντελεστής με το συμβολισμό r_{corr} αποτελεί το διορθωμένο συντελεστή συσχέτισης μεταξύ ενός παράγοντα και του αθροίσματος των υπολοίπων παραγόντων του σχεδιασμού. Η συσχέτιση διορθώνεται με την έννοια ότι αποτελεί τη συσχέτιση μεταξύ ενός συγκεκριμένου παράγοντα και των υπολοίπων παραγόντων του σχεδιασμού, χωρίς αυτός ο συγκεκριμένος παράγοντας να θεωρείται μέρος του σχεδιασμού. Ας υποθέσουμε ότι έχουμε m παράγοντες σε έναν υπερκορεσμένο σχεδιασμό, τότε ο συντελεστής r_{corr} του j-οστού παράγοντα είναι η συσχέτιση μεταξύ του j-οστού παράγοντα και του αθροίσματος των υπολοίπων $m-1$ παραγόντων. Ο υπολογισμός του συντελεστή r_{corr} αποτελεί ένα τρόπο να δούμε το κατά πόσο ένας παράγοντας αλληλοσυσχετίζεται με όλους τους άλλους παράγοντες που παραμένουν στο σχεδιασμό. Στην κλασική θεωρία ελέγχου, συχνά το φράγμα του 0.40 χρησιμοποιείται ως κανόνας κατωφλιού [18], δείχνοντας ότι ο παράγοντας συσχετίζεται ασθενώς με τους άλλους παράγοντες. Στη μελέτη μας, υιοθετούμε αυτό τον κανόνα κατωφλιού, και όσο χαμηλότερη είναι η τιμή του r_{corr} ενός παράγοντα, τόσο πιο απίθανο ο παράγοντας να αλληλοσυσχετίζεται με τους υπόλοιπους παράγοντες του σχεδιασμού.

Ο συντελεστής $\hat{\alpha}_{rhosquared}$ του j-οστού παράγοντα υποδεικνύει τον αναμενόμενο προκύπτον συντελεστή άλφα εάν ο j-οστός παράγοντας αφαιρεθεί από το σχεδιασμό. Με άλλα λόγια, ο συντελεστής $\hat{\alpha}_{rhosquared}$ αναπαριστά το ποιος θα ήταν ο συντελεστής $\alpha_{rhosquared}$ εάν ο j-οστός παράγοντας αφαιρεθεί από το σχεδιασμό και δεν συμπεριληφθεί στον υπολογισμό του συντελεστή άλφα.

Κάθε αλγόριθμος βελτιστοποίησης εξαρτάται από μία συγκεκριμένη αντικειμενική συνάρτηση η οποία χρειάζεται να βελτιστοποιηθεί. Για τους υπερκορεσμένους σχεδιασμούς, όπως προαναφέραμε, ένα ευρέως χρησιμοποιούμενο κριτήριο βελετιστοποίησης είναι το r_{max} κριτήριο. Ένα εύλογο κριτήριο για να συγκρίνουμε υπερκορεσμένους σχεδιασμούς είναι η ελαχιστοποίηση του $max_{i<j}|r_{ij}|$, όπου r_{ij} είναι η συσχέτιση δύο στηλών c_i και c_j. Η μεγαλύτερη απόλυτη τιμή της συσχέτισης r_{ij} μεταξύ όλων των ζευγών των στηλών συμβολίζεται με r_{max}.

13.3.2 Ο Προτεινόμενος Αλγόριθμος Βελτιστοποίησης (BEDDO)

Σε έναν υπερκορεσμένο σχεδιασμό, εξ'ορισμού υπάρχουν λιγότερες πειραματικές εκτελέσεις σε σχέση με τις επιδράσεις που χρειάζεται να εκτιμηθούν. Όταν ο πειραματιστής θεωρεί υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις, το φαινόμενο της ταύτισης μεταξύ των κύριων επιδράσεων και των αλληλεπιδράσεων δύο παραγόντων λαμβάνει χώρα. Σε μία τέτοια περίπτωση, οι ταυτόσημες επιδράσεις των παραγόντων μπορεί να αλληλοεξουδετερώνονται, και οι ταυτόσημες μικρές επιδράσεις των παραγόντων μπορεί να συγχωνεύονται. Τα δύο αυτά φαινόμενα μπορούν να συμβούν σε οποιοδήποτε κλασματικό παραγοντικό σχεδιασμό, όπως επίσης και σε έναν υπερκορεσμένο σχεδιασμό που περιλαμβάνει αλληλεπιδράσεις. Ένα από τα κυριότερα προβλήματα που έχει να αντιμετωπίσει ο πειραματιστής όταν θεωρεί στη μελέτη του υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις, είναι ότι οι κύριες επιδράσεις ή οι αλληλεπιδράσεις δεν είναι εντελώς ταυτόσημες, αλλά οι εκτιμήσεις ορισμένων κύριων επιδράσεων και αλληλεπιδράσεων συσχετίζονται.

Στόχος μας σε αυτή τη μελέτη είναι να πάρουμε την περισσότερη πληροφορία που μπορεί να εξαχθεί από το σχεδιασμό καθώς επίσης και κάθε παράγοντας να συνεισφέρει μοναδική πληροφορία. Η προτεινόμενη μέθοδος επιτυγχάνει να κατανοήσει τα υποδείγματα μεταξύ του συνόλου των στηλών-παραγόντων, να αξιολογήσει το βαθμό της πλησιέστερης ορθογωνιότητας και τέλος, να επιλέξει τον υπερκορεσμένο σχεδιασμό μεγαλύτερης πληροφορήσεως συμπεριλαμβανομένων των αλληλεπιδράσεων.

Προτείνουμε τη μέθοδο BEDDO προκειμένου να εξαλειφθούν οι παράγοντες που εντοπίζονται να είναι ταυτόσημοι μεταξύ τους στο σχεδιασμό. Με τον τρόπο αυτό, καθορίζουμε άμεσα τους παράγοντες που περιέχουν όλες τις απαραίτητες σημαντικές πληροφορίες. Η μέθοδος BEDDO περιλαμβάνει τις αλληλεπιδράσεις δύο παραγόντων στις αρχές του πειραματισμού, και έτσι επιτυγχάνει να αποκλείσει τις αλληλεπιδράσεις μικρής πληροφορήσεως το ταχύτερο δυνατόν. Ο αλγόριθμος BEDDO είναι εφαρμόσιμος μόνο στις στήλες των παραγόντων, χωρίς να θεωρεί μεταβλητή απόκρισης και να απαιτεί τη γνώση της. Ο BEDDO αλγόριθμος μετά την υλοποίησή του έχει ως υποπροϊόν το de-aliasing των παραγόντων και επιτυγχάνει να ανακτήσει ένα σχεδιασμό όσο το δυνατόν πλησιέστερα στον ορθογώνιο, υποκείμενος σε κάποια στατιστικά μέτρα βασισμένα στη συσχέτιση (το συνολικό $\alpha_{rhosquared}$, r_{corr}, $\hat{\alpha}_{rhosquared}$ και το r_{max}). Τα βήματα της προτεινόμενης μεθόδου βελτιστοποίησης (backward elimination design-driven optimization-BEDDO) παρουσιάζονται παρακάτω σε μορφή ψευδοκώδικα.

Αλγόριθμος 13.1.

Require: *$X = [x_1, x_2, \ldots, x_m]$ is the initial $(n \times m)$ main effects SSD matrix {the j-th column of X corresponds to the levels of the j-th factor for $j = 1, \ldots, m$}*

1: ***for*** *$j = 1$ to m* ***do***
2: *generate all $\binom{m}{2}$ two-factor interactions columns*
3: ***end for***
4: *entertain the m factors and their $\binom{m}{2}$ two-factor interactions {the $(n \times m)$ SSD matrix X, now increases to an $n \times \left(\frac{m(m+1)}{2}\right)$ SSD matrix $X^+ = [x_1, x_2, \ldots, x_{\frac{m(m+1)}{2}}]$; the j-th column of X^+ corresponds to the levels of the j-th factor for $j = 1, \ldots, \frac{m(m+1)}{2}$}*
5: ***BEGIN BEDDO***
6: *compute the overall alpha coefficient of X^+ $\{\alpha_{rhosquared}\}$*
7: ***for*** *$j = 1$ to $\frac{m(m+1)}{2}$* ***do***
8: *compute the r_{corr} and the $\hat{\alpha}_{rhosquared}$ if the j-th column is deleted from X^+*
9: ***end for***
10: *sort r_{corr} and $\hat{\alpha}_{rhosquared}$ values in descending order*
11: ***for all*** *j, $j = 1, \ldots, \frac{m(m+1)}{2}$* ***do***
12: ***if*** *$|r_{corr}| > 0.40$* ***then***
13: ***EXCLUDE*** *x_j*
14: ***else if*** *$|r_{corr}| \leq 0.40$* ***then***
15: ***keep searching***
16: ***end if***

17: ***for all*** *j, $j = m+1, \ldots, \frac{m(m+1)}{2}$ such that* ***do***
18: *have a value of r_{corr} higher than the lowest observed r_{corr} value among factors* ***and*** ***if*** *$\hat{\alpha}_{rhosquared} \ll \alpha_{rhosquared}$* ***then***
19: ***EXCLUDE*** *x_j*
end if
20: ***end for***
21: ***end for***
22: *generate the new $(n \times k)$ SSD matrix (X_{BEDDO}) after excluding the factors identified as interrelated following the previous steps*
23: *compute r_{max} of the generated X_{BEDDO}*
24: ***if*** *$r_{max} = 1$* ***then***
25: ***repeat***
26: *the previous processing*
27: ***until*** *the generated SSD has a value of $r_{max} < 1$*
28: ***else if*** *$r_{max} < 1$* ***then***
29: ***STOP BEDDO***
30: ***end if***

13.4 Κατασκευή Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών

Από καιρό στη βιβλιογραφία, μία βασική παραδοχή-υπόθεση για τους υπερκορεσμένους σχεδιασμούς είναι ότι πρέπει να είναι ισορροπημένοι. Πρόσφατες μελέτες συζητούν τα πιθανά πλεονεκτήματα του να χαλαρώσουμε την απαίτηση της ισορροπίας. Για παράδειγμα, οι Allen και Bernshteyn [4] παράγουν μία νέα κλάση μη ισορροπημένων υπερκορεσμένων σχεδιασμών οι οποίοι έχουν πλεονεκτήματα όσον αφορά την απόδοσή τους. Ο Gilmour [92] επισημαίνει ότι οι υπερκορεσμένοι σχεδιασμοί συνήθως κατασκευάζονται έτσι ώστε να βελτιστοποιούν το κριτήριο $E(s^2)$ κάτω από τον περιορισμό και την παραδοχή της ισορροπίας, αλλά αυτό φαίνεται να μην σχετίζεται με τον τρόπο με τον οποίο τα δεδομένα των υπερκορεσμένων σχεδιασμών αναλύονται. Οι Jones, Lin και Nachtsheim [118] έδειξαν ότι το να χαλαρώσεις τον περιορισμό της ισορροπίας για την κατασκευή Bayesian D-βέλτιστων υπερκορεσμένων σχεδιασμών, παράγει σχεδιασμούς οι οποίοι συγκρίνονται ευνοϊκά με ισορροπημένους ομόλογούς τους για πολλά κριτήρια, συμπεριλαμβανομένου του $E(s^2)$.

Οι Booth και Cox το 1962 πρότειναν το κριτήριο $E(s^2)$ το οποίο επιλέγει ένα σχεδιασμό ελαχιστοποιώντας το άθροισμα των τετραγώνων των εσωτερικών γινομένων μεταξύ των στηλών c_i και c_j του X $(i, j = 2, \ldots m+1,\ i \neq j)$. Η πρώτη στήλη του X είναι η $1_n = [1, \ldots, 1]^T$. Οι Marley και Woods το 2010 επέκτειναν αυτό τον ορισμό για να συμπεριλάβουν το εσωτερικό γινόμενο της πρώτης στήλης με κάθε άλλη στήλη του X, δηλαδή,

$$E(s^2) = \frac{\sum\limits_{i<j} s_{ij}^2}{\binom{m+1}{2}}, \tag{13.6}$$

όπου το s_{ij} είναι το i,j-οστό στοιχείο του πίνακα $X^{\mathrm{T}}X$ $(i, j = 1, \ldots m+1)$. Οι δύο ορισμοί είναι ισοδύναμοι για τους ισορροπημένους σχεδιασμούς. Η ορθογωνιότητα αναφέρεται στη σχέση μεταξύ των παραγόντων με άλλους παράγοντες και η ισορροπία αναφέρεται στη σχέση μεταξύ των παραγόντων και του σταθερού όρου. Η ορθογωνιότητα ή η πλησιέστερη ορθογωνιότητα είναι μια σημαντική ιδιότητα στους πειραματικούς σχεδιασμούς, και οι υπερκορεσμένοι σχεδιασμοί εξ' ορισμού δεν μπορούν να ικανοποιήσουν την κατά ζεύγη ορθογωνιότητα όλων των στηλών των παραγόντων στον πίνακα σχεδιασμού, όπως επισημάνθηκε στην εργασία [115]. Όταν ένας σχεδιασμός είναι ισορροπημένος, ο σταθερός όρος είναι ορθογώνιος σε κάθε παράγοντα. Μια τέτοια ιδιότητα δεν ισχύει στην περίπτωση ενός μη ισορροπημένου σχεδιασμού και ο ορισμός (13.6) των Marley και Woods επιτρέπει στον πειραματιστή να συγκρίνει μη ισορροπημένους σχεδιασμούς όσον αφορά την ορθογωνιότητα μεταξύ του σταθερού όρου και των παραγόντων. Σημειώνουμε εδώ ότι η κλασική μορφή του $E(s^2)$ μπορεί

να υπολογιστεί για μη ισορροπημένους σχεδιασμούς, ωστόσο, η αξιολόγησή τους με το κάτω φράγμα (13.4) δεν επιτρέπεται.

Για την αξιολόγηση της απόδοσης και της πρακτικής εφαρμογής της μεθόδου BEDDO, εξετάζουμε τρεις υπερκορεσμένους σχεδιασμούς με διαφορετικά μεγέθη και επίπεδα υπερκορεσμού. Οι $E(s^2)$-βέλτιστοι και minimax-βέλτιστοι κυκλικοί υπερκορεσμένοι σχεδιασμοί από τους Koukouvinos et al. [130] θεωρούνται στη μελέτη μας. Σύμφωνα με αυτήν τη μέθοδο κατασκευής, οι $E(s^2)$-βέλτιστοι και minimax-βέλτιστοι κυκλικοί υπερκορεσμένοι σχεδιασμοί έχουν n πειραματικές εκτελέσεις και $m = q \cdot (n-1)$ παράγοντες, όπου το q είναι άρτιος αριθμός. Στη μελέτη προσομοίωσής μας, χρησιμοποιούμε ως αρχικούς υπερκορεσμένους σχεδιασμούς κύριων επιδράσεων, τους υπερκορεσμένους σχεδιασμούς με (n, m) τιμές: $(6, 10)$, $(8, 14)$ και $(10, 18)$ για το Παράδειγμα 1, 2 και 3, αντίστοιχα.

Αξίζει να σημειωθεί εδώ ότι ο αλγόριθμος BEDDO συμπεριλαμβανομένων των αλληλεπιδράσεων μπορεί να παράγει μεγαλύτερους σχεδιασμούς ξεκινώντας από μικρότερους σχεδιασμούς, μικρότερους σχεδιασμούς ξεκινώντας από μεγαλύτερους σχεδιασμούς ή σχεδιασμούς του ίδιου μεγέθους. Σχετικά με αυτή την παρατήρηση, από πρακτική σκοπιά, η μέθοδος BEDDO είναι ιδιαίτερα χρήσιμη για τις περιπτώσεις στις οποίες ο πειραματιστής υποθέτει ότι οι αλληλεπιδράσεις δεν είναι αμελητέες, και επιθυμεί να γνωρίζει εκ των προτέρων το σχεδιασμό που θα επιλέξει για περαιτέρω πειραματισμό καθώς επίσης και τον επιθυμητό αριθμό των παραγόντων που θα χρησιμοποιηθούν στο πείραμα.

13.4.1 Παράδειγμα 1: $\text{SSD}_{(6,10)}$

Θεωρούμε ένα πείραμα με $m = 10$ παράγοντες και $n = 6$ πειραματικές εκτελέσεις του $\text{SSD}_{(6,10)}$ όπως παρουσιάζεται στον Πίνακα 13.2. Έχουμε ότι ο $X = [x_a, x_b, \ldots, x_j]$ είναι ο αρχικός (6×10) πίνακας του υπερκορεσμένου σχεδιασμού. Μετά την ανάμειξη των $\binom{10}{2}$ στηλών των αλληλεπιδράσεων δύο παραγόντων με τις 10 γονεϊκές στήλες των παραγόντων, αποκτούμε έναν (6×55) πίνακα $X^+ = [x_a, x_b, \ldots, x_{hj}, x_{ij}]$. Εκτελούμε επαναληπτικά τον αλγόριθμο BEDDO, και μετά από ακριβώς τρεις εκτελέσεις κατασκευάζουμε ένα νέο μη ισορροπημένο υπερκορεσμένο σχεδιασμό αποτελούμενο από 6 εκτελέσεις και 7 παράγοντες με τιμή για το κριτήριο $r_{max} = 0.7071 < 1$. Ο κατασκευασμένος μη ισορροπημένος υπερκορεσμένος σχεδιασμός $\text{SSD}_{(6,7)}$ παρουσιάζεται στον Πίνακα 13.3. Σε αυτή την περίπτωση, ξεκινήσαμε από ένα μεγαλύτερο σχεδιασμό ($m = 10$) και κατασκευάσαμε ένα μικρότερο σχεδιασμό ($m(= 6)$+interactions(=1)=7).

Πίνακας 13.2: Παράδειγμα 1: Ο αρχικός ισορροπημένος $\text{SSD}_{(6,10)}$

Εκτέλεση	x_a	x_b	x_c	x_d	x_e	x_f	x_g	x_h	x_i	x_j
1	-1	1	-1	-1	1	1	1	-1	-1	-1
2	1	-1	1	-1	-1	-1	1	1	-1	-1
3	-1	1	-1	1	-1	-1	-1	1	1	-1
4	-1	-1	1	-1	1	-1	-1	-1	1	1
5	1	-1	-1	1	-1	1	-1	-1	-1	1
6	1	1	1	1	1	1	1	1	1	1

Πίνακας 13.3: Παράδειγμα 1: Ο νέος μη ισορροπημένος $\text{SSD}_{(6,7)}$

Εκτέλεση	x_b	x_c	x_f	x_g	x_i	x_j	x_{fg}
1	1	-1	1	1	-1	-1	1
2	-1	1	-1	1	-1	-1	-1
3	1	-1	-1	-1	1	-1	1
4	-1	1	-1	-1	1	1	1
5	-1	-1	1	-1	-1	1	-1
6	1	1	1	1	1	1	1

13.4.2 Παράδειγμα 2: SSD$_{(8,14)}$

Θεωρούμε ένα πείραμα με $m = 14$ παράγοντες και $n = 8$ πειραματικές εκτελέσεις του SSD$_{(8,14)}$ όπως παρουσιάζεται στον Πίνακα 13.4. Έχουμε ότι ο $X = [x_a, x_b, \ldots, x_n]$ είναι ο αρχικός (8×14) πίνακας του υπερκορεσμένου σχεδιασμού. Μετά την ανάμειξη των $\binom{14}{2}$ στηλών των αλληλεπιδράσεων δύο παραγόντων με τις 14 γονεϊκές στήλες των παραγόντων, αποκτούμε έναν (8×105) πίνακα $X^+ = [x_a, x_b, \ldots, x_{ln}, x_{mn}]$. Εκτελούμε επαναληπτικά τον αλγόριθμο BEDDO, και μετά από ακριβώς μία εκτελέση κατασκευάζουμε ένα νέο μη ισορροπημένο υπερκορεσμένο σχεδιασμό αποτελούμενο από 8 εκτελέσεις και 35 παράγοντες με τιμή για το κριτήριο $r_{max} = 0.5774 < 1$. Ο κατασκευασμένος μη ισορροπημένος υπερκορεσμένος σχεδιασμός SSD$_{(8,35)}$ παρουσιάζεται στον Πίνακα 13.5. Σε αυτή την περίπτωση, ξεκινήσαμε από ένα μικρότερο σχεδιασμό ($m = 14$) και κατασκευάσαμε ένα μεγαλύτερο σχεδιασμό ($m(= 14)$+interactions(=21)=35).

Πίνακας 13.4: Παράδειγμα 2: Ο αρχικός ισορροπημένος SSD$_{(8,14)}$

Εκτέλεση	x_a	x_b	x_c	x_d	x_e	x_f	x_g	x_h	x_i	x_j	x_k	x_l	x_m	x_n
1	1	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1
2	-1	1	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	1
3	-1	-1	1	1	-1	1	-1	1	-1	1	1	-1	-1	-1
4	-1	-1	-1	1	1	-1	1	-1	1	-1	1	1	-1	-1
5	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	1	1	-1
6	-1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	1	1
7	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Πίνακας 13.5: Παράδειγμα 2: Ο νέος μη ισορροπημένος SSD$_{(8,35)}$

Εκτέλεση	x_a	x_b	x_c	x_d	x_e	x_f	x_g	x_h	x_i	x_j	x_k	x_l	x_m	x_n
1	1	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1
2	-1	1	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	1
3	-1	-1	1	1	-1	1	-1	1	-1	1	1	-1	-1	-1
4	-1	-1	-1	1	1	-1	1	-1	1	-1	1	1	-1	-1
5	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	1	1	-1
6	-1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	1	1
7	1	-1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Εκτέλεση	x_{ah}	x_{al}	x_{am}	x_{bi}	x_{bm}	x_{bn}	x_{ch}	x_{cj}	x_{cn}	x_{dh}	x_{di}	x_{dk}	x_{ei}	x_{ej}
1	1	-1	1	1	1	-1	-1	1	1	1	1	-1	-1	1
2	1	1	1	1	-1	1	-1	1	1	1	-1	1	1	1
3	-1	1	1	1	1	1	1	1	-1	1	-1	1	1	-1
4	1	-1	1	-1	1	1	1	1	1	-1	1	1	1	-1
5	-1	1	1	1	-1	1	1	-1	1	1	1	1	-1	1
6	1	1	-1	-1	1	1	1	1	-1	1	1	-1	1	1
7	1	1	-1	1	1	-1	1	-1	1	-1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Εκτέλεση	x_{el}	x_{fj}	x_{fk}	x_{fm}	x_{gk}	x_{gl}	x_{gn}
1	1	1	1	-1	1	1	1
2	-1	-1	1	1	1	1	-1
3	1	1	1	-1	-1	1	1
4	1	1	-1	1	1	1	-1
5	1	1	-1	1	1	-1	1
6	1	-1	1	1	1	-1	1
7	-1	1	1	1	-1	1	1
8	1	1	1	1	1	1	1

13.4.3 Παράδειγμα 3: SSD$_{(10,18)}$

Θεωρούμε ένα πείραμα με $m = 18$ παράγοντες και $n = 10$ πειραματικές εκτελέσεις του SSD$_{(10,18)}$ όπως παρουσιάζεται στον Πίνακα 13.6. Έχουμε ότι ο $X = [x_a, x_b, \ldots, x_r]$ είναι ο αρχικός (10×18) πίνακας του υπερκορεσμένου σχεδιασμού. Μετά την ανάμειξη των $\binom{18}{2}$ στηλών των αλληλεπιδράσεων δύο παραγόντων με τις 18 γονεϊκές στήλες των παραγόντων,

αποκτούμε έναν (10×171) πίνακα $X^+ = [x_a, x_b, \ldots, x_{pr}, x_{qr}]$. Εκτελούμε επαναληπτικά τον αλγόριθμο BEDDO, και μετά από ακριβώς μία εκτελέση κατασκευάζουμε ένα νέο μη ισορροπημένο υπερκορεσμένο σχεδιασμό αποτελούμενο από 10 εκτελέσεις και 27 παράγοντες με τιμή για το κριτήριο $r_{max} = 0.600 < 1$. Ο κατασκευασμένος μη ισορροπημένος υπερκορεσμένος σχεδιασμός $\text{SSD}_{(10,27)}$ παρουσιάζεται στον Πίνακα 13.7. Σε αυτή την περίπτωση, ξεκινήσαμε από ένα μικρότερο σχεδιασμό ($m = 18$) και κατασκευάσαμε ένα μεγαλύτερο σχεδιασμό ($m(= 18)$+interactions(=9)=27).

Πίνακας 13.6: Παράδειγμα 3: Ο αρχικός ισορροπημένος $\text{SSD}_{(10,18)}$

Εκτέλεση	x_a	x_b	x_c	x_d	x_e	x_f	x_g	x_h	x_i	x_j	x_k	x_l	x_m	x_n	x_o	x_p	x_q	x_r
1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1
2	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	-1
3	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1
4	1	-1	-1	-1	1	-1	1	1	-1	-1	-1	-1	1	-1	-1	1	1	1
5	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	-1	1	1
6	1	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	-1	1
7	1	1	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	-1
8	-1	1	1	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1
9	1	-1	1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Πίνακας 13.7: Παράδειγμα 3: Ο νέος μη ισορροπημένος $\text{SSD}_{(10,27)}$

Εκτέλεση	x_a	x_b	x_c	x_d	x_e	x_f	x_g	x_h	x_i	x_j	x_k	x_l	x_m	x_n	x_o	x_p	x_q	x_r
1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	-1	-1
2	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1	-1
3	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	1	-1
4	1	-1	-1	-1	1	-1	1	1	-1	-1	-1	-1	1	-1	-1	1	1	1
5	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	-1	1	1
6	1	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	-1	1
7	1	1	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1	-1
8	-1	1	1	-1	1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	-1
9	1	-1	1	1	-1	1	-1	-1	-1	-1	-1	1	1	1	-1	-1	-1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Εκτέλεση	x_{ar}	x_{bj}	x_{ck}	x_{dl}	x_{em}	x_{fn}	x_{go}	x_{hp}	x_{iq}
1	1	1	1	-1	1	-1	1	1	1
2	1	1	1	1	-1	1	-1	1	1
3	1	1	1	1	1	-1	1	-1	1
4	1	1	1	1	1	1	-1	1	-1
5	-1	1	1	1	1	1	1	-1	1
6	1	-1	1	1	1	1	1	1	-1
7	-1	1	-1	1	1	1	1	1	1
8	1	-1	1	-1	1	1	1	1	1
9	1	1	-1	1	-1	1	1	1	1
10	1	1	1	1	1	1	1	1	1

13.4.4 Σύγκριση Μη Ισορροπημένων και Ισορροπημένων Υπερκορεσμένων Σχεδιασμών

Σε αυτή την ενότητα, συγκρίνουμε τους κατασκευασμένους με τη βοήθεια υπολογιστών μη ισορροπημένους υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις, με εναλλακτικούς ισορροπημένους υπερκορεσμένους σχεδιασμούς που συναντάμε στη διεθνή βιβλιογραφία κατασκευασμένους με τη χρήση διαφορετικών κριτηρίων. Ο Πίνακας 13.8 παρουσιάζει τις τιμές από διάφορα κριτήρια βελτιστοποίησης (r_{max}, $E(s^2)$ (Booth & Cox), $E(s^2)$ (Marley & Woods)) καθώς και τις τιμές από το προτεινόμενο κριτήριο ($\alpha_{rhosquared}$), για ισορροπημένους υπερκορεσμένους σχεδιασμούς και τους μη ισορροπημένους ομόλογούς τους, για τρία υπερκορεσμένα πειράματα με διαφορετικό αριθμό πειραματικών εκτελέσεων και παραγόντων.

Αρχικά, συγκρίνουμε την απόδοση του κατασκευασμένου μη ισορροπημένου υπερκορεσμένου σχεδιασμού $\mathrm{SSD}_{(6,7)}$ με την απόδοση ισορροπημένων υπερκορεσμένων σχεδιασμών παραγόμενων μέσω διαφορετικών μεθόδων βελτιστοποίησης (συμπεριλαμβανομένου του $E(s^2)$ κριτηρίου των Booth και Cox, του D_2 κριτηρίου που προτάθηκε από τον Wu [229], του κριτηρίου $\max p_{cov}$ των Allen και Bernshteyn [4], καθώς επίσης συμπεριλαμβάνουμε το σχεδιασμό του Lin [148] παραγόμενο από μισά κλάσματα Plackett-Burman πινάκων). Έπειτα, η απόδοση του κατασκευασμένου μη ισορροπημένου υπερκορεσμένου σχεδιασμού $\mathrm{SSD}_{(6,7)}$ συγκρίνεται με έναν υπάρχον μη ισορροπημένο υπερκορεσμένο σχεδιασμό 6 εκτελέσεων και 7 παραγόντων $\mathrm{SSD}_{p_{cov}}$ κατασκευασμένο από τους Allen και Bernshteyn.

Πρώτον, είναι αξιοσημείωτο το πόσο όμοιες είναι οι αποδόσεις για τους ισορροπημένους υπερκορεσμένους σχεδιασμούς 6 εκτελέσεων και 7 παραγόντων, μολονότι παράγονται με διαφορετικούς τρόπους, όπως επίσης επισημάνθηκε από τους Allen και Bernshteyn. Ο κατασκευασμένος μη ισορροπημένος υπερκορεσμένος σχεδιασμός $\mathrm{SSD}_{(6,7)}$ έχει υψηλότερες τιμές για τα κριτήρια r_{max} και $E(s^2)$ (Booth & Cox), αλλά χαμηλότερες τιμές για τα κριτήρια ρ^2, $E(s^2)$ (Marley & Woods) και το προτεινόμενο $\alpha_{rhosquared}$ σε σύγκριση με το μη ισορροπημένο ομόλογό του σχεδιασμό $\mathrm{SSD}_{p_{cov}}$ των Allen και Bernshteyn. Οι τιμές των κριτηρίων για τον κατασκευασμένο μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\mathrm{SSD}_{(6,7)}$ και το μη ισορροπημένο ομόλογό του $\mathrm{SSD}_{p_{cov}}$ υπερβαίνουν τις αντίστοιχες τιμές για τους ισορροπημένους σχεδιασμούς για τα κριτήρια ρ^2, r_{max} και το $\alpha_{rhosquared}$.

Τώρα συγκρίνουμε τους δύο μη ισορροπημένους υπερκορεσμένους σχεδιασμούς 6 εκτελέσεων και 7 παραγόντων, δηλαδή, τους SSD_{BEDDO} και $\mathrm{SSD}_{p_{cov}}$.

Ο μη ισορροπημένος SSD_{BEDDO} έχει 4 $s_{ij} = 0$, 15 $s_{ij} = \pm 2$, 2 $s_{ij} = 4$ και $E(s^2) = 4.3810$ για X $(i, j = 1, \ldots m, i \neq j)$.

Ο μη ισορροπημένος $\mathrm{SSD}_{p_{cov}}$ έχει 21 $s_{ij} = \pm 2$ και $E(s^2) = 4.0000$ για X $(i, j = 1, \ldots m, i \neq j)$.

Ο μη ισορροπημένος SSD_{BEDDO} έχει 10 $s_{ij} = 0$, 16 $s_{ij} = \pm 2$, 2 $s_{ij} = 4$ και $E(s^2) = 3.4286$ για X $(i, j = 2, \ldots m + 1$, η πρώτη στήλη του X είναι $1_n)$.

Ο μη ισορροπημένος $\mathrm{SSD}_{p_{cov}}$ έχει 5 $s_{ij} = 0$, 21 $s_{ij} = \pm 2$, 2 $s_{ij} = 4$ και $E(s^2) = 4.1429$ για X $(i, j = 2, \ldots m + 1$, η πρώτη στήλη του X είναι $1_n)$.

Παρακάτω, συγκρίνουμε ως προς την απόδοσή τους, τον κατασκευασμένο μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\mathrm{SSD}_{(8,35)}$ με τον ισορροπημένο ομόλογό του, τον $E(s^2)$-βέλτιστο υπερκορεσμένο σχεδιασμό κατασκευασμένο από τον Cheng [47]. Οι τιμές των κριτηρίων για τον κατασκευασμένο μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\mathrm{SSD}_{(8,35)}$ υπερβαίνουν (ρ^2, r_{max} και $\alpha_{rhosquared}$) τις αντίστοιχες τιμές για τον ισορροπημένο ομόλογό του ή είναι χαμηλότερες ($E(s^2)$ (Booth & Cox), $E(s^2)$ (Marley & Woods)) από τις αντίστοιχες τιμές για τον ισορροπημένο ομόλογό του.

Τώρα, συγκρίνουμε ως προς την απόδοσή τους, τον κατασκευασμένο μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\mathrm{SSD}_{(10,27)}$ με τον ισορροπημένο ομόλογό του, τον κοντά στον $E(s^2)$-βέλτιστο υπερκορεσμένο σχεδιασμό κατασκευασμένο από τους Liu και Dean [153]. Οι τιμές των κριτηρίων για τον κατασκευασμένο μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\mathrm{SSD}_{(10,27)}$ είναι ίδιες (r_{max}) ή υπερβαίνουν (ρ^2, $\alpha_{rhosquared}$) τις αντίστοιχες τιμές για τον ισορροπημένο ομόλογό του ή είναι χαμηλότερες ($E(s^2)$ (Booth & Cox), $E(s^2)$ (Marley & Woods)) από τις αντίστοιχες τιμές για τον ισορροπημένο ομόλογό του.

Τα παραδείγματα αυτά στον Πίνακα 13.8, βάση του κριτηρίου $E(s^2)$, υποδεικνύουν ότι οι κατασκευασμένοι μέσω του BEDDO μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί είναι συγκρίσιμοι ($\mathrm{SSD}_{(6,7)}$) ή φαίνεται να υπερτερούν ($\mathrm{SSD}_{(8,35)}$, $\mathrm{SSD}_{(10,27)}$) σε σχέση με τους άλλους υπό εξέταση σχεδιασμούς. Αξίζει να σημειωθεί εδώ, ότι όλοι οι κατασκευασμένοι μέσω του BEDDO μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν χαμηλότερες τιμές για το κριτήριο $E(s^2)$ (Marley & Woods) σε σύγκριση με τους $E(s^2)$-βέλτιστους ή τους κοντά στους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς. Η χαλάρωση του περιορισμού της ισορροπίας επιτρέπει στους κατασκευασμένους μη ισορροπημένους υπερκορεσμένους σχεδια-

σμούς να βελτιώνονται σε σχέση με τις "$E(s^2)$" τιμές, καθώς επίσης να αποδίδουν αρκετά καλά σε σχέση και με τα υπόλοιπα κριτήρια βελτιστοποίησης. Οι κατασκευασμένοι μέσω του BEDDO μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί μπορεί ναι μεν να μην έχουν την ιδιότητα της ισορροπίας, αλλά ακόμα έχουν πλεονεκτήματα με την έννοια ότι είναι "σχεδόν" βέλτιστοι σε σχέση με αρκετά κριτήρια και οι επιδράσεις των παραγόντων είναι "σχεδόν" ορθογώνιες.

Πίνακας 13.8: Συγκριτική απόδοση ισορροπημένων ΥΣ και των αντίστοιχων μη ισορροπημένων ΥΣ

n	m	Κριτήριο	Ισορροπία	ρ^2	r_{max}	$E(s^2)$ (Booth & Cox)	$E(s^2)$ (Marley & Woods)	$\alpha_{rhosquared}$
6	7	Booth & Cox [a]	ΝΑΙ	0.1111	0.3333	4.0000	4.0000	0.4666
6	7	Wu [a]	ΝΑΙ	0.1111	0.3333	4.0000	4.0000	0.4666
6	7	Lin [a]	ΝΑΙ	0.1111	0.3333	4.0000	4.0000	0.4666
6	7	Allen & Bernshteyn [a]	ΝΑΙ	0.1111	0.3333	4.0000	4.0000	0.4666
6	7	Allen & Bernshteyn	ΟΧΙ	0.1501	0.4472	4.0000	4.1429	0.5527
6	7	BEDDO	ΟΧΙ	0.1270	0.7071	4.3810	3.4286	0.5045
8	35	Cheng [b]	ΝΑΙ	0.1176	0.5000	7.5294	7.5294	0.8235
8	35	BEDDO	ΟΧΙ	0.1216	0.5774	6.9647	7.1111	0.8289
10	27	Liu & Dean [a]	ΝΑΙ	0.0810	0.6000	8.1000	8.1000	0.7041
10	27	BEDDO	ΟΧΙ	0.0853	0.6000	7.0769	7.4286	0.7158

[a] near-optimal SSD (no $E(s^2)$-optimal SSD is available)
[b] $E(s^2)$-optimal SSD

13.5 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάσαμε μία μέθοδο βελτιστοποίησης, η οποία αποτελεί μία τεχνική προς τα πίσω απαλοιφής καθοδηγούμενη από ένα σχεδιασμό (backward elimination design-driven optimization-BEDDO), η οποία μελετά τη δομή ταυτόσημων επιδράσεων, είναι ικανή να εξαλείψει τους παράγοντες που ανιχνεύονται να είναι ταυτόσημοι μεταξύ τους μέσα στο σχεδιασμό, και τέλος επιτρέπει στον πειραματιστή να ανακτήσει ένα μη ισορροπημένο υπερκορεσμένο σχεδιασμό όσο το δυνατόν πιο κοντά σε έναν ορθογώνιο συμπεριλαμβανομένων των αλληλεπιδράσεων. Για την υλοποίηση της προτεινόμενης μεθόδου BEDDO, ενσωματώσαμε και συνδυάσαμε στατιστικά μέτρα βασισμένα στη συσχέτιση, τα οποία έχουν ληφθεί από δύο επιστημονικά πεδία, την κλασική θεωρία ελέγχου και τους σχεδιασμούς πειραμάτων, και παρουσιάσαμε επίσης ένα νέο κριτήριο βελτιστότητας το οποίο αποτελεί μία τροποποιημένη μορφή του Cronbach άλφα συντελεστή ($\alpha_{rhosquared}$), με πολύ καλές ιδιότητες δεδομένου ότι οι ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν παρόμοια απόδοση για τα κριτήρια $E(s^2)$ και $\alpha_{rhosquared}$. Ωστόσο, το κύριο πλεονέκτημα του τροποποιημένου κριτηρίου είναι η εύκολη ενσωμάτωσή του στη μέθοδο BEDDO και η άμεση αντιστοιχία του με τον κύριο στόχο της εξάλειψης των αλληλεπιδράσεων μικρής πληροφορήσεως το ταχύτερο δυνατόν καθώς και τον εντοπισμό των πιο πληροφοριακών παραγόντων του σχεδιασμού που δεν είναι αλληλένδετοι μεταξύ τους.

Με τον τρόπο αυτό, η BEDDO μέθοδος βελτιστοποίησης κάτω από το τροποποιημένο κριτήριο και με τη βοήθεια των υπολογιστών παράγει μία νέα κλάση μη ισορροπημένων υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις. Η νέα αυτή κλάση των μη ισορροπημένων υπερκορεσμένων σχεδιασμών που κατασκευάσαμε με τη βοήθεια των υπολογιστών είναι στην πραγματικότητα μερικώς ισορροπημένοι υπερκορεσμένοι σχεδιασμοί, με την έννοια ότι m_1 από τις m στήλες είναι ισορροπημένες (έχουν ίσο αριθμό "-1's" και "1's") και οι υπόλοιπες $(m-m_1)$ $(=m_2)$ είναι μη ισορροπημένες. Τέτοιου τύπου σχεδιασμοί με την ιδιότητα της μερικής ισορροπίας μπορεί να προκύψουν συχνά στην πράξη εφόσον η ύπαρξη των αλληλεπιδράσεων θεωρηθεί είτε στην κατασκευή είτε στην ανάλυση των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων. Οι μερικώς ισορροπημένοι σχεδιασμοί μπορεί να ενδιαφέρουν έναν πειραματιστή ο οποίος χρειάζεται σχεδιασμούς όσο το δυνατόν πιο ισορροπημένους και

μπορεί να είναι προτιμητέοι σε περιπτώσεις στις οποίες τα επίπεδα συγκεκριμένων παραγόντων σχετίζονται με υψηλά κόστη.

Συστήνουμε τους σχεδιασμούς αυτούς όχι μόνο για τις καλές στατιστικές τους ιδιότητες, αλλά επίσης για την εφαρμοσιμότητά τους σε ένα ευρύτερο σύνολο πραγματικών προβλημάτων της ζωής, όπου είναι πιθανόν να προκύπτουν τέτοιοι υπερκορεσμένοι σχεδιασμοί οι οποίοι δεν περιορίζονται στο να είναι ισορροπημένοι. Σε ένα πραγματικό σενάριο, μπορεί να προκύψει ένας μη ισορροπημένος υπερκορεσμένος σχεδιασμός που μπορεί να εφαρμοστεί σε οργανισμούς ή βιομηχανίες, όπου υπάρχει η ανάγκη για αποτελεσματικούς στατιστικούς υπολογισμούς και όταν ο κύριος στόχος είναι να εντοπιστούν σημαντικοί παράγοντες σε καταστάσεις στις οποίες μόνο ένας πολύ περιορισμένος αριθμός πειραματικών εκτελέσεων είναι διαθέσιμος. Αξίζει να σημειωθεί εδώ ότι ο αλγόριθμος BEDDO συμπεριλαμβανομένων των αλληλεπιδράσεων μπορεί να παράγει μεγαλύτερους σχεδιασμούς ξεκινώντας από μικρότερους σχεδιασμούς, μικρότερους σχεδιασμούς ξεκινώντας από μεγαλύτερους σχεδιασμούς ή σχεδιασμούς του ίδιου μεγέθους. Σχετικά με αυτή την παρατήρηση, από πρακτική σκοπιά, η μέθοδος BEDDO είναι ιδιαίτερα χρήσιμη για τις περιπτώσεις στις οποίες ο πειραματιστής υποθέτει ότι οι αλληλεπιδράσεις δεν είναι αμελητέες, και επιθυμεί να γνωρίζει εκ των προτέρων το σχεδιασμό που θα επιλέξει για περαιτέρω πειραματισμό καθώς επίσης και τον επιθυμητό αριθμό των παραγόντων που θα χρησιμοποιηθούν στο πείραμα.

ΚΕΦΑΛΑΙΟ **14**

Ανάλυση Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών με Αλληλεπιδράσεις

All models are wrong,
but some are useful.

—*George E.P. Box (1919–2013)*

Στο δέκατο τέταρτο αυτό κεφάλαιο μελετάμε τα πιθανά πλεονεκτήματα της χαλάρωσης της απαίτησης ισορροπίας στην ανάλυση των υπερκορεσμένων σχεδιασμών, δεδομένου ότι ναι μεν οι υπερκορεσμένοι σχεδιασμοί συνήθως κατασκευάζονται κάτω από την παραδοχή της ισορροπίας, αλλά κάτι τέτοιο φαίνεται να είναι άσχετο με τον τρόπο με τον οποίο αναλύονται τα δεδομένα τους, όπως άλλωστε επισημάνθηκε από τον Gilmour [92]. Ας σημειωθεί εδώ, ότι το πλαίσιο ανάλυσης μέχρι σήμερα για τους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς περιορίζεται στην κεντρική υπόθεση ότι τα πειραματικά δεδομένα προέρχονται από ένα γραμμικό μοντέλο. Στο παρόν κεφάλαιο, συγκρίνουμε μη ισορροπημένους υπερκορεσμένους σχεδιασμούς με ισορροπημένους ομόλογούς τους, κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων, αποκαλύπτοντας ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν πλεονεκτήματα όσον αφορά την απόδοσή τους, παρά την έλλειψη ισορροπίας. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου παρουσιάζονται στην επιστημονική εργασία [44].

14.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Στο Κεφάλαιο 13 αναφερθήκαμε διεξοδικά στο γεγονός ότι οι περισσότερες από τις μεθόδους κατασκευής και ανάλυσης των υπερκορεσμένων σχεδιασμών λειτουργούν κάτω από τη βασική παραδοχή-υπόθεση για τους σχεδιασμούς να είναι ισορροπημένοι. Επίσης, στο Κεφάλαιο 13 και στην ενότητα 13.4 αναφερθήκαμε στις πρόσφατες εργασίες οι οποίες αναφέρονται στα πιθανά πλεονεκτήματα του να χαλαρώσουμε την απαίτηση της ισορροπίας και μελετούν την κατασκευή ή την ανάλυση των μη ισορροπημένων υπερκορεσμένων σχεδιασμών, όπως για παράδειγμα οι εργασίες [4], [92], [118], [160] και [43].

Το πλαίσιο ανάλυσης μέχρι σήμερα για τους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς περιορίζεται στην κεντρική υπόθεση ότι τα πειραματικά δεδομένα προέρχονται από ένα γραμμικό μοντέλο. Οι Marley και Woods [160] σύγκριναν και αξιολόγησαν τις αποδόσεις διαφορετικών στρατηγικών ανάλυσης χρησιμοποιώντας ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς και μη ισορροπημένους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς, κάτω από την παραδοχή ενός γραμμικού μοντέλου. Όλοι οι προαναφερθέντες λόγοι, μας παρακίνησαν όπως είναι φυσικό, να μελετήσουμε τους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς για την ανάλυση δεδομένων, κάτω από την παραδοχή όμως των γενικευμένων γραμμικών μοντέλων, προκειμένου να αξιολογήσουμε την απόδοσή τους και να εξετάσουμε το εάν και το κατά πόσο η ιδιότητα της μη ισορροπίας επηρεάζει την αποτελεσματικότητά τους. Συγκεκριμένα, συγκρίνουμε την απόδοση (κάτω από τα γενικευμένα γραμμικά μοντέλα) και τις ιδιότητες βελτιστότητας των μη ισορροπημένων Bayesian D-βέλτιστων υπερκορεσμένων σχεδιασμών [118] σε σχέση με τους ισορροπημένους ομόλογούς τους. Επίσης, συγκρίνουμε την απόδοση (κάτω από τα γενικευμένα γραμμικά μοντέλα) και τις ιδιότητες βελτιστότητας των μη ισορροπημένων υπερκορεσμένων σχεδιασμών που περιλαμβάνουν αλληλεπιδράσεις, με ισορροπημένους ομόλογούς τους που υπάρχουν στη βιβλιογραφία. Σημειώνουμε σε αυτό το σημείο ότι δεν συμπεριλαμβάνουμε στη συγκριτική μελέτη μας τους μη ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς των Suen και Das [213] (που έχουν περιττό αριθμό εκτελέσεων εξ' ορισμού), δεδομένου ότι δεν υπάρχουν ισορροπημένοι ομόλογοι αυτών των σχεδιασμών (οι ισορροπημένοι σχεδιασμοί δεν μπορεί να έχουν περιττό αριθμό εκτελέσεων λόγω της ιδιότητας της ισορροπίας).

14.2 Ανάλυση Ισορροπημένων και Μη Ισορροπημένων Υπερκορεσμένων Σχεδιασμών

Στη μελέτη μας, η ανάλυση των ισορροπημένων και μη ισορροπημένων υπερκορεσμένων σχεδιασμών, κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων με διακριτή απόκριση, πραγματοποιήθηκε μέσω του αλγορίθμου της Θεωρίας Πληροφοριών (Information Theoretical Algorithm-ITA) όπως αυτός παρουσιάστηκε στην εργασία [7]. Ο ενδιαφερόμενος αναγνώστης μπορεί επίσης να ανατρέξει στο Κεφάλαιο 4 στην ενότητα 4.2.1 για περισσότερες λεπτομέρειες σχετικά με τα βήματα που εκτελεί ο αλγόριθμος ITA, ο οποίος μιμείται μία κατάσταση όπου οι επιδράσεις των παραγόντων μπορούν να ταξινομηθούν σε σχέση με την πιθανότητά τους να είναι ενεργές ή μη ενεργές.

Η απόφαση σχετικά με την τιμή κατωφλιού του w, η οποία στην πράξη καθορίζει το μέγιστο αριθμό των σημαντικών παραγόντων που μπορούν να αναγνωριστούν, παίζει σημαντικό ρόλο στην ανάπτυξη της μεθόδου. Στην πειραματική μελέτη μας θέσαμε το $w = m/2$, όπου m είναι ο αριθμός των στηλών του υπερκορεσμένου πίνακα σχεδιασμού, δηλαδή ο αριθμός των παραγόντων υπό εξέταση. Η απόφαση για το w έγινε βάση της αρχής της σποραδικότητας των επιδράσεων [23]. Για να ποσοτικοποιήσουμε τη σποραδικότητα στο σύνολο των δεδομένων μας κατά τη διάρκεια των πειραμάτων προσομοίωσης, χρησιμοποιήσαμε την ακόλουθη διαδικασία. Οι πραγματικά ενεργές μεταβλητές επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ ενδεχόμενων ενεργών παραγόντων. Για κάθε πείραμα, εκτιμήσαμε όλες τις

επιδράσεις των παραγόντων, οι οποίες στη συνέχεια επισημάνθηκαν ως ενεργές ή μη ενεργές. Έπειτα υπολογίζοντας το ποσοστό των εκτιμώμενων ενεργών επιδράσεων που εμφανίζονται σε κάθε πείραμα, εκτιμήσαμε ότι ο αριθμός των ενεργών επιδράσεων που μπορεί να αναγνωριστεί, δεν υπερβαίνει το $w = \frac{m}{2}$, όπου m είναι ο αριθμός των στηλών του πίνακα του υπερκορεσμένου σχεδιασμού.

Η απόφαση σχετικά με την τιμή κατωφλιού του SU μέτρου, η οποία καθορίζει το κατά πόσο ένας παράγοντας είναι σημαντικός ή όχι, παίζει επίσης σημαντικό ρόλο στην ανάπτυξη της μεθόδου. Για αυτόν το λόγο, πολλές διαφορετικές τιμές κατωφλιού (0.001, 0.01, 0.05, 0.1, 0.25, 0.4, median(**su**)) εξετάστηκαν στην εργασία [7] με στόχο να βρεθεί η βέλτιστη τιμή κατωφλιού του SU για την προτεινόμενη μέθοδο. Οι πρώτες έξι τιμές κατωφλιού ήταν σταθεροί αριθμοί, οι οποίοι επιλέχθηκαν σύμφωνα με τις τιμές του SU που παρατηρήθηκαν στα περισσότερα από τα πειράματα, ενώ η τιμή κατωφλιού median(**su**) βασίστηκε στις εκτιμώμενες τιμές του SU, συγκεκριμένα στο εύρος των τιμών του SU οι οποίες παρατηρήθηκαν στα πειράματα. Σε αυτή τη μελέτη, επίσης εξετάσαμε τις προαναφερθείσες τιμές κατωφλιού, προκειμένου να εξασφαλίσουμε τη βελτιστότητα της επιλογής της παραμέτρου SU για δεδομένα απόκρισης τα οποία αποτελούν μετρήσεις (counts). Τελικά, καθορίσαμε την τιμή κατωφλιού του SU να είναι ίση με το median(**su**), όπως επίσης καθορίζεται στις εργασίες [7] και [8], δεδομένου ότι η επιλογή αυτή παρατηρήθηκε να είναι η βέλτιστη μετά από δοκιμές προσομοίωσης, έχοντας ως αποτέλεσμα χαμηλότερες τιμές σφάλματος.

Εκτελέσαμε 1000 προσομοιώσεις για κάθε ένα από τα ακόλουθα σενάρια:

- Σενάριο I: παράχθηκαν 1000 σύνολα με δεδομένα απόκρισης Y από την Bernoulli κατανομή,
- Σενάριο II: παράχθηκαν 1000 σύνολα με δεδομένα απόκρισης Y από την Poisson κατανομή,
- Σενάριο III: παράχθηκαν 1000 σύνολα με δεδομένα απόκρισης Y από την αρνητική διωνυμική κατανομή,
- Σενάριο IV: παράχθηκαν 1000 σύνολα με δεδομένα απόκρισης Y από τη γεωμετρική κατανομή.

Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στο Κεφάλαιο 5 και στην ενότητα 5.2 για περισσότερες λεπτομέρειες σχετικά με τα γενικευμένα γραμμικά μοντέλα καθώς επίσης και τον τρόπο με τον οποίο μοντελοποιήσαμε τα δεδομένα απόκρισης για κάθε ένα από τα παραπάνω σενάρια.

Για κάθε μία από τις 1000 επαναλήψεις ακολουθήσαμε το παρακάτω πρωτόκολλο προσομοίωσης.

1. Δεδομένα απόκρισης Y παράχθηκαν για κάθε ένα από τα Σενάρια I, II, III και IV, αντίστοιχα. Από εδώ και στο εξής, ο $X = [1, x_1, x_2, \ldots, x_m]$ είναι ο $n \times (m+1)$ πίνακας του μοντέλου. Υποθέτουμε ότι κάθε ένας από τους m παράγοντες έχει δύο επίπεδα, ± 1. Η πρώτη στήλη του X είναι η $1_n = [1, \ldots, 1]^T$, με τη στήλη j να αντιστοιχεί στα επίπεδα του $(j-1)$-οστού παράγοντα για $j = 2, \ldots, m+1$.

2. Οι πραγματικά ενεργές μεταβλητές επιλέχθηκαν τυχαία από το σύνολο των $\{1, \ldots, m\}$ πιθανών ενεργών παραγόντων. Από τις $2, \ldots, m+1$ στήλες του X, p στήλες ανατέθηκαν σε ενεργούς παράγοντες τυχαία με το $p \leq \frac{n}{3}$ (όπως προτάθηκε από τους Marley και Woods [160]).

3. Για να αποκτήσουμε τους συντελεστές των ενεργών παραγόντων, ένα δείγμα μεγέθους p σχηματίστηκε από την ομοιόμορφη κατανομή $U(a = -8, b = 8)$, και τα πρόσημα ± 1 τυχαία κατανεμημένα σε κάθε αριθμό. Όταν ένας παραγόμενος συντελεστής έχει $|\beta| < 1$, τότε αντικαθίσταται από το 50% του μέγιστου παρατηρούμενου συντελεστή.

4. Για να αποκτήσουμε τους συντελεστές των μη ενεργών παραγόντων, ένα δείγμα μεγέθους *(m-p)* σχηματίστηκε από την ομοιόμορφη κατανομή $U(a = -0.2, b = 0.2)$, και τα πρόσημα ± 1 τυχαία κατανεμημένα σε κάθε αριθμό.

Οι Box και Meyer [23], και πρόσφατα οι Marley και Woods [160], επισήμαναν ότι η αποτελεσματικότητα των υπερκορεσμένων σχεδιασμών στην ανίχνευση ενεργών παραγόντων απαιτεί τον αριθμό των ενεργών παραγόντων να είναι μικρός, μια αρχή γνωστή ως αρχή της σποραδικότητας των επιδράσεων. Ειδικότερα, οι Marley και Woods πρότειναν τις ακόλουθες κατευθυντήριες γραμμές για την αποδοτική χρήση των υπερκορεσμένων σχεδιασμών. Πρώτον, η αναλογία των παραγόντων προς τον αριθμό των πειραματικών εκτελέσεων θα πρέπει να είναι μικρότερη από 2, και δεύτερον ο αριθμός των πειραματικών εκετλέσεων θα πρέπει να είναι τουλάχιστον τρεις φορές ο αριθμός των ενεργών παραγόντων. Οι Marley και Woods πρότειναν ότι κάτω από αυτές τις δύο προϋποθέσεις, οι υπερκορεσμένοι σχεδιασμοί είναι πιο πιθανό να είναι επιτυχείς. Ωστόσο, οι συγγραφείς σημειώνουν ότι σε μια υπό εξέταση περίπτωση στην οποία η πρώτη προϋπόθεση δεν ισχύει ($m/n \geq 2$), ο υπερκορεσμένος σχεδιασμός εξακολουθούσε να έχει καλή απόδοση. Τέτοιες περιπτώσεις εξετάζονται επίσης στη δική μας μελέτη προσομοιώσεων ($\mathrm{SSD}_{(12,24)}$, $\mathrm{SSD}_{(10,27)}$, $\mathrm{SSD}_{(8,35)}$). Επιπλέον, αξίζει να σημειωθεί εδώ ότι κάποιοι πειραματιστές δεν επιθυμούν ένα μεγάλο αριθμό από μη ισορροπημένες στήλες σε μία μελέτη, ωστόσο, κάποιοι άλλοι πειραματιστές μπορεί να είναι πιο πρόθυμοι να δεχτούν περισσότερες μη ισορροπημένες στήλες σε αντάλλαγμα με ένα ελαφρώς καλύτερο aliasing μεταξύ των επιδράσεων ενδιαφέροντος, όπως άλλωστε επισημάνθηκε από τον Marley [159]. Για τους παραπάνω λόγους, προκειμένου να αξιολογήσουμε την απόδοση της προτεινόμενης μεθόδου, μελετήσαμε αρκετά υπερκορεσμένα πειράματα με διαφορετικό μέγεθος, διαφορετικό επίπεδο υπερκορεσμού και ισορροπίας.

Δεδομένου ότι οι συνθήκες στην πράξη είναι συνήθως διαφορετικές από εκείνες των προσομοιώσεων και ότι ο πειραματιστής δεν ξέρει πόσοι και ποιοί παράγοντες μπορεί να είναι ενεργοί, προκειμένου να εξετάσουμε το πόσο ευαίσθητα είναι τα αποτελέσματα στην επιλογή και τον αριθμό των ενεργών στηλών-παραγόντων, θεωρήσαμε την τυχαία ανάθεση των ενεργών παραγόντων στις στήλες, από τη στιγμή που είναι σημαντικό να αποφύγουμε την οποιαδήποτε μεροληψία επιλογής. Η επιλογή της ομοιόμορφης κατανομής είναι ευεργετική για δύο λόγους. Πρώτον, εξασφαλίζει στο βήμα 3 ότι οι επιδράσεις διαφόρων μεγεθών θεωρούνται ταυτόχρονα (μικρότερου, μεσαίου αλλά και μεγαλύτερου μεγέθους επιδράσεις). Δεύτερον, εξασφαλίζει στο βήμα 4 ότι οι συντελεστές που λαμβάνονται για τους μη ενεργούς παράγοντες είναι σχετικά μικροί και "σχεδόν μηδενικοί", δεδομένου ότι οι συντελεστές των μη ενεργών παραγόντων δεν θα πρέπει να ορίζονται ακριβώς στο μηδέν, αντ'αυτού θα πρέπει να είναι μικροί τυχαίοι αριθμοί. Γενικά, η επιλογή των κατανομών στα προηγούμενα βήματα 3 και 4 εξασφαλίζει το διαχωρισμό μεταξύ των συντελεστών των ενεργών και ανενεργών παραγόντων. Για το λόγο αυτό, θεωρήσαμε ένα ευρύ φάσμα προσομειωμένων μοντέλων που ήταν διαφορετικά από την άποψη αυτή.

Προκειμένου να αξιολογήσουμε την προτεινόμενη μέθοδο και να συγκρίνουμε τα αποτελέσματα, εκτιμήσαμε τα σφάλματα Τύπου Ι και Τύπου ΙΙ. Αυτό που μας ενδιαφέρει να ελέγξουμε στα πειράματα προσομείωσης είναι το κόστος του να δηλώσουμε μία ανενεργή επίδραση ως ενεργή (σφάλμα Τύπου Ι) καθώς και το κόστος του να δηλώσουμε μία ενεργή επίδραση ως ανενεργή (σφάλμα Τύπου ΙΙ).

14.3 Σύγκριση Ισορροπημένων ΥΣ και Μη Ισορροπημένων Bayesian D-Βέλτιστων ΥΣ

Στην ενότητα αυτή, συγκρίνουμε τους μη ισορροπημένους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς με ισορροπημένους ομόλογούς τους, οι οποίοι έχουν τον ίδιο αριθμό πειραματικών εκτελέσεων και παραγόντων, προκειμένου να αξιολογήσουμε την από-

δοσή τους για την ανάλυση των δεδομένων κάτω από τα γενικευμένα γραμμικά μοντέλα με διακριτή απόκριση και να εξετάσουμε τις ιδιότητες βελτιστότητάς τους.

14.3.1 Ανάλυση Δεδομένων Μέσω Γενικευμένων Γραμμικών Μοντέλων

Για την αξιολόγηση της απόδοσης, θεωρούμε τρία υπερκορεσμένα πειράματα με διαφορετικό μέγεθος, διαφορετικό επίπεδο υπερκορεσμού και ισορροπίας. Θεωρούμε τρεις μη ισορροπημένους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς ($\mathrm{SSD}_{(12,16)}$, $\mathrm{SSD}_{(12,18)}$ $\mathrm{SSD}_{(12,24)}$), όπως αυτοί πινακοποιήθηκαν από τους Jones, Lin και Nachtsheim [118]. Οι Bayesian D-βέλτιστοι υπερκορεσμένοι σχεδιασμοί είναι μη ισορροπημένοι επειδή ο σταθερός όρος είναι πρωταρχικός όρος, και μπορεί να εκτιμηθεί ξεχωριστά από τις κύριες επιδράσεις. Στη συνέχεια συγκρίνουμε αυτούς τους μη ισορροπημένους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς με ισορροπημένους ομόλογούς τους, με τον ίδιο αριθμό πειραματικών εκτελέσεων και παραγόντων, όπως αυτοί πινακοποιήθηκαν από τον Nguyen [172]. Τα ληφθέντα αποτελέσματα σύμφωνα με το θεωρούμενο υπερκορεσμένο σχεδιασμό και το εξεταζόμενο Σενάριο παρουσιάζονται στον Πίνακα 14.1. Στην πρώτη στήλη του Πίνακα 14.1, δίνεται ο αριθμός που αντιστοιχεί σε κάθε χρησιμοποιούμενο μοντέλο. Στις επόμενες τέσσερις στήλες, δίνεται ο ισορροπημένος ή μη ισορροπημένος υπερκορεσμένος σχεδιασμός, καθώς επίσης ο αριθμός m των παραγόντων και ο αριθμός n των πειραματικών εκτελέσεων του σχεδιασμού. Στην έκτη στήλη δίνεται το θεωρούμενο Σενάριο. Οι τελευταίες δύο στήλες που ονομάζονται ως "Type I" και "Type II" παρουσιάζουν τις μέσες τιμές των ποσοστών των σφαλμάτων Τύπου I και Τύπου I, αντίστοιχα, πάνω σε 1000 προσομοιώσεις σύμφωνα με το υπό εξέταση Σενάριο.

Παρατηρούμε από τον Πίνακα 14.1 ότι για τον ίδιο αριθμό εκτελέσεων n και παραγόντων m, ισορροπημένοι και μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν παρόμοια απόδοση με μικρές διαφορές όσον αφορά τα σφάλματα Τύπου I και Τύπου II. Συγκεκριμένα, η μελέτη έδειξε ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί αποδίδουν πολύ καλά και παρουσιάζουν ελαφρώς χαμηλότερα σφάλματα Τύπου II σε συνδυασμό με ελαφρώς υψηλότερα σφάλματα Τύπου I, σε σύγκριση με τους ισορροπημένους ομόλογους υπερκορεσμένους σχεδιασμούς που εξετάστηκαν. Αυτό πρακτικά σημαίνει ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί τείνουν να δηλώνουν σε μεγαλύτερο ποσοστό ανενεργές επιδράσεις να είναι ενεργές και αντίστοιχα σε μικρότερο ποσοστό ενεργές επιδράσεις να είναι ανενεργές συγκρινόμενοι με τους ισορροπημένους υπερκορεσμένους σχεδιασμούς ίδιου μεγέθους.

Επιπρόσθετα, παρατηρούμε από τον Πίνακα 14.1 ότι για κάθε θεωρούμενο υπερκορεσμένο σχεδιασμό (ισορροπημένο ή μη ισορροπημένο), οι τιμές των σφαλμάτων Τύπου I και Τύπου II, αντίστοιχα, διαφέρουν ελάχιστα κάτω από τα Σενάρια I, II, III και IV. Τα αποτελέσματα αυτά υποδεικνύουν ότι η προτεινόμενη μέθοδος είναι αρκετά εύρωστη όσον αφορά τη μοντελοποίηση της απόκρισης κάτω από τα γενικευμένα γραμμικά μοντέλα. Όπως προαναφέραμε, οι Marley και Woods πρότειναν ότι για την αποτελεσματική χρήση των υπερκορεσμένων σχεδιασμών, η αναλογία των παραγόντων προς τον αριθμό των πειραματικών εκτελέσεων θα πρέπει να είναι μικρότερη από 2. Όταν στη μελέτη προσομοίωσης θεωρούμε τους υπερκορεσμένους σχεδιασμούς $\mathrm{SSD}_{(12,16)}$ και $\mathrm{SSD}_{(12,18)}$, η παραπάνω συνθήκη ισχύει, και παρατηρούμε ότι αυτοί οι υπερκορεσμένοι σχεδιασμοί υπερέχουν ξεκάθαρα (όσον αφορά τις μέσες τιμές των σφαλμάτων Τύπου I και Τύπου II) από τους σχεδιασμούς όπως ο $\mathrm{SSD}_{(12,24)}$ για τον οποίο η παραπάνω συνθήκη δεν ισχύει. Ωστόσο, ακόμα και σε αυτή την περίπτωση για την οποία ισχύει ότι $m/n = 2$, οι υπερκορεσμένοι σχεδιασμοί εξακολουθούν να έχουν αρκετά καλή απόδοση, δεδομένου ότι οι μέσες τιμές σφάλματος διατηρούνται σε χαμηλά επίπεδα. Οι συνολικές μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II όλων των 24 μοντέλων που θεωρούνται στον Πίνακα 14.1 είναι 0.22 και 0.08, αντίστοιχα, πράγμα που σημαίνει ότι η προτεινόμενη μέθοδος έχει σημαντικά χαμηλότερες τιμές σφαλμάτων Τύπου II, και υψηλότερες τιμές σφάλματος Τύπου I. Σε γενικές γραμμές, ο Πίνακας 14.1 δείχνει ότι η προτεινόμενη

μέθοδος έχει την τάση να δηλώνει σε ένα υψηλότερο ποσοστό ανενεργές επιδράσεις να είναι ενεργές και σε πολύ χαμηλότερο ποσοστό ενεργές επιδράσεις να είναι ανενεργές. Έτσι, η προτεινόμενη μέθοδος θα μπορούσε να χαρακτηριστεί ως συντηρητική υπό αυτή την έννοια.

Πίνακας 14.1: Συγκριτική απόδοση των μη ισορροπημένων Bayesian D-βέλτιστων ΥΣ με τους ισορροπημένους ομόλογούς τους

Μοντέλο	ΥΣ ΑΠΟ	m	n	Ισορροπία	Σενάριο	Type I	Type II
1	Jones et al. (2008)	16	12	ΟΧΙ	I	0.17	0.03
2	Jones et al. (2008)	16	12	ΟΧΙ	II	0.19	0.06
3	Jones et al. (2008)	16	12	ΟΧΙ	III	0.23	0.07
4	Jones et al. (2008)	16	12	ΟΧΙ	IV	0.22	0.06
Average						0.2025	0.0550
5	Nguyen (1996)	16	12	ΝΑΙ	I	0.17	0.04
6	Nguyen (1996)	16	12	ΝΑΙ	II	0.19	0.07
7	Nguyen (1996)	16	12	ΝΑΙ	III	0.22	0.09
8	Nguyen (1996)	16	12	ΝΑΙ	IV	0.21	0.07
Average						0.1975	0.0675
9	Jones et al. (2008)	18	12	ΟΧΙ	I	0.21	0.04
10	Jones et al. (2008)	18	12	ΟΧΙ	II	0.23	0.07
11	Jones et al. (2008)	18	12	ΟΧΙ	III	0.25	0.09
12	Jones et al. (2008)	18	12	ΟΧΙ	IV	0.23	0.07
Average						0.2300	0.0675
13	Nguyen (1996)	18	12	ΝΑΙ	I	0.21	0.05
14	Nguyen (1996)	18	12	ΝΑΙ	II	0.22	0.08
15	Nguyen (1996)	18	12	ΝΑΙ	III	0.24	0.11
16	Nguyen (1996)	18	12	ΝΑΙ	IV	0.22	0.09
Average						0.2225	0.0825
17	Jones et al. (2008)	24	12	ΟΧΙ	I	0.23	0.06
18	Jones et al. (2008)	24	12	ΟΧΙ	II	0.24	0.07
19	Jones et al. (2008)	24	12	ΟΧΙ	III	0.26	0.12
20	Jones et al. (2008)	24	12	ΟΧΙ	IV	0.26	0.11
Average						0.2475	0.0900
21	Nguyen (1996)	24	12	ΝΑΙ	I	0.22	0.07
22	Nguyen (1996)	24	12	ΝΑΙ	II	0.23	0.09
23	Nguyen (1996)	24	12	ΝΑΙ	III	0.26	0.14
24	Nguyen (1996)	24	12	ΝΑΙ	IV	0.25	0.12
Average						0.2400	0.1050

14.3.2 Ιδιότητες Βελτιστότητας

Σε αυτή την ενότητα συγκρίνουμε τους προαναφερθέντες μη ισορροπημένους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς με τους ισορροπημένους ομόλογούς τους, χρησιμοποιώντας διαφορετικά κριτήρια βελτιστοποίησης. Εκτός από το κριτήριο $E(s^2)$, ένα εύλογο κριτήριο για να συγκρίνουμε υπερκορεσμένους σχεδιασμούς είναι η ελαχιστοποίηση του $max_{i<j}|r_{ij}|$, όπου r_{ij} είναι η συσχέτιση δύο στηλών c_i και c_j. Η μεγαλύτερη απόλυτη τιμή της συσχέτισης r_{ij} μεταξύ όλων των ζευγών των στηλών συμβολίζεται με το r_{max}. Η συσχέτιση μεταξή δύο στηλών c_i και c_j ενός υπερκορεσμένου σχεδιασμού υπολογίζεται από την κλασική φόρμουλα του Pearson. Αξίζει να σημειωθεί εδώ ότι στην περίπτωση ενός ισορροπημένου υπερκορεσμένου σχεδιασμού η συσχέτιση ορίζεται ως $r_{ij} = \frac{c_i' c_j}{n}$. Δεδομένου ότι στην παρούσα μελέτη θεωρούμε Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς, θεωρούμε επιπλέον ως κριτήριο την αντικειμενική συνάρτηση Bayesian D-βελτιστότητας, $D_{Bayes} = |X^{\mathrm{T}}X + \mathbf{R}|$, όπου $X^{\mathrm{T}}X$ είναι ο πίνακας πληροφορίας του υπερκορεσμένου σχεδιασμού και $\mathbf{R}$ είναι ο πίνακας πληροφορίας των παραμέτρων της θεωρηθείσας εκ των προτέρων κατανομής. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει για περισσότερες πληροφορίες σχεικά με την αντικειμενική συνάρτηση Bayesian D-βελτιστότητας στην εργασία [118]. Ο Πίνακας 14.2 παρουσιάζει τις διάφορες τιμές των κριτηρίων βελτιστοποίησης (r_{max}, $E(s^2)$ (Booth & Cox), D_{Bayes}), οι οποίες προέκυψαν για τους τρεις μη ισορροπημένους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς και τους ισορροπημένους ομόλογούς τους.

Παρατηρούμε από τον Πίνακα 14.2 ότι οι τιμές των κριτηρίων για τους μη ισορροπημένους

Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς είναι ίδιες, μεγαλύτερες ή μικρότερες από τις αντίστοιχες τιμές για τους ισορροπημένους υπερκορεσμένους σχεδιασμούς. Σε μία από τις τρεις περιπτώσεις ($\text{SSD}_{(12,18)}$) οι τιμές είναι ακριβώς οι ίδιες, και στις άλλες δύο περιπτώσεις ($\text{SSD}_{(12,16)}$, $\text{SSD}_{(12,24)}$) υπάρχουν μέτριες διαφορές. Σημειώνουμε εδώ ότι σε μία από τις τρεις περιπτώσεις ($\text{SSD}_{(12,16)}$) η τιμή του κριτηρίου $E(s^2)$ του Bayesian D-βέλτιστου υπερκορεσμένου σχεδιασμού είναι μικρότερη σε σύγκριση με την αντίστοιχη τιμή του ισορροπημένου ομόλογού του. Αυτό οφείλεται στο γεγονός ότι η χαλάρωση του περιορισμού της ισορροπίας επιτρέπει στους μη ισορροπημένους Bayesian D-βέλτιστους υπερκορεσμένους σχεδιασμούς να βελτιώνονται σε σχέση με τις "$E(s^2)$" τιμές, όπως άλλωστε επισημαίνεται στην εργασία [118].

Πίνακας 14.2: Κριτήρια βελτιστότητας μη ισορροπημένων Bayesian D-βέλτιστων ΥΣ και των αντίστοιχων ισορροπημένων ΥΣ

n	m	ΥΣ ΑΠΟ	Ισορροπία	[a] r_{max}	[a] $E(s^2)$ (Booth & Cox)	[b] D_{Bayes} (Jones et al. 2008)
12	16	Nguyen (1996)	ΝΑΙ	0.3333	5.200	6.32
12	16	Jones et al. (2008)	ΟΧΙ	0.4800	4.800	6.37
12	18	Nguyen (1996)	ΝΑΙ	0.3333	5.96	4.85
12	18	Jones et al. (2008)	ΟΧΙ	0.3333	5.96	4.85
12	24	Nguyen (1996)	ΝΑΙ	0.3333	7.83	2.004
12	24	Jones et al. (2008)	ΟΧΙ	0.6700	7.83	2.008

[a] the smaller-the-better
[b] the larger-the-better

14.4 Σύγκριση Ισορροπημένων ΥΣ και Μη Ισορροπημένων ΥΣ με Αλληλεπιδράσεις

Στην ενότητα αυτή, συγκρίνουμε τους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις με ισορροπημένους ομόλογούς τους, οι οποίοι έχουν τον ίδιο αριθμό πειραματικών εκτελέσεων και παραγόντων, προκειμένου να αξιολογήσουμε την απόδοσή τους για την ανάλυση των δεδομένων κάτω από τα γενικευμένα γραμμικά μοντέλα με διακριτή απόκριση και να εξετάσουμε τις ιδιότητες βελτιστότητάς τους.

14.4.1 Ανάλυση Δεδομένων Μέσω Γενικευμένων Γραμμικών Μοντέλων

Για την αξιολόγηση της απόδοσης, θεωρούμε τρία υπερκορεσμένα πειράματα με διαφορετικό μέγεθος, διαφορετικό επίπεδο υπερκορεσμού και ισορροπίας. Αρχικά, θεωρούμε ένα μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\text{SSD}_{(6,7)}$ από την εργασία [43], ένα μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\text{SSD}_{(6,7)}$ από την εργασία [4], και συγκρίνουμε την απόδοσή τους με ισορροπημένους υπερκορεσμένους σχεδιασμούς 6 εκτελέσεων και 7 παραγόντων, παραγόμενων μέσω διαφορετικών μεθόδων βελτιστοποίησης και συνδυαστικών ιδιοτήτων (συμπεριλαμβανομένου του $E(s^2)$ κριτηρίου των Booth και Cox, του D_2 κριτηρίου που προτάθηκε από τον Wu [229], του κριτηρίου $\max p_{cov}$ των Allen και Bernshteyn [4], καθώς επίσης συμπεριλαμβάνουμε το σχεδιασμό του Lin [148] παραγόμενο από μισά κλάσματα Plackett-Burman πινάκων). Τα ληφθέντα αποτελέσματα σε συμφωνία με τους υπερκορεσμένους σχεδιασμούς των 6 εκτελέσεων και 7 παραγόντων παρουσιάζονται στον Πίνακα 14.3. Έπειτα, συγκρίνουμε ως προς την απόδοσή τους, το μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\text{SSD}_{(10,27)}$ από την εργασία [43], με τον ισορροπημένο ομόλογό του, τον κοντά στον $E(s^2)$-βέλτιστο υπερκορεσμένο σχεδιασμό κατασκευασμένο από τους Liu και Dean [153]. Τα ληφθέντα αποτελέσματα σε συμφωνία με τους υπερκορεσμένους σχεδιασμούς των 10 εκτελέσεων και 27 παραγόντων παρουσιάζονται στον Πίνακα 14.4. Στο τέλος, συγκρίνουμε ως προς την απόδοσή τους, το μη ισορροπημένο υπερκορεσμένο σχεδιασμό $\text{SSD}_{(8,35)}$ από την

εργασία [43], με τον ισορροπημένο ομόλογό του, τον $E(s^2)$-βέλτιστο υπερκορεσμένο σχεδιασμό κατασκευασμένο από τον Cheng [47]. Τα ληφθέντα αποτελέσματα σε συμφωνία με τους υπερκορεσμένους σχεδιασμούς των 8 εκτελέσεων και 35 παραγόντων παρουσιάζονται στον Πίνακα 14.5.

Στην πρώτη στήλη των Πινάκων 14.3, 14.4 και 14.5, δίνεται ο αριθμός που αντιστοιχεί σε κάθε χρησιμοποιούμενο μοντέλο. Στις επόμενες τέσσερις στήλες, δίνεται ο ισορροπημένος ή μη ισορροπημένος υπερκορεσμένος σχεδιασμός, καθώς επίσης ο αριθμός m των παραγόντων και ο αριθμός n των πειραματικών εκτελέσεων του σχεδιασμού. Στην έκτη στήλη δίνεται το θεωρούμενο Σενάριο. Οι τελευταίες δύο στήλες των Πινάκων 14.3, 14.4 και 14.5 που ονομάζονται ως "Type I" και "Type II" παρουσιάζουν τις μέσες τιμές των ποσοστών των σφαλμάτων Τύπου I και Τύπου I, αντίστοιχα, πάνω σε 1000 προσομοιώσεις σύμφωνα με το υπό εξέταση Σενάριο.

Παρατηρούμε από τους Πίνακες 14.3, 14.4 και 14.5, ότι για τον ίδιο αριθμό εκτελέσεων n και παραγόντων m, ισορροπημένοι και μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν παρόμοια απόδοση με μικρές διαφορές όσον αφορά τα σφάλματα Τύπου I και Τύπου II. Συγκεκριμένα, η μελέτη έδειξε ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί αποδίδουν πολύ καλά και παρουσιάζουν ελαφρώς χαμηλότερα σφάλματα Τύπου II σε συνδυασμό με ελαφρώς υψηλότερα σφάλματα Τύπου I, σε σύγκριση με τους ισορροπημένους ομόλογους υπερκορεσμένους σχεδιασμούς που εξετάστηκαν. Αυτό πρακτικά σημαίνει ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί τείνουν να δηλώνουν σε μεγαλύτερο ποσοστό ανενεργές επιδράσεις να είναι ενεργές και αντίστοιχα σε μικρότερο ποσοστό ενεργές επιδράσεις να είναι ανενεργές συγκρινόμενοι με τους ισορροπημένους υπερκορεσμένους σχεδιασμούς ίδιου μεγέθους.

Επιπρόσθετα, παρατηρούμε από τους Πίνακες 14.3, 14.4 και 14.5, ότι για κάθε θεωρούμενο υπερκορεσμένο σχεδιασμό (ισορροπημένο ή μη ισορροπημένο), οι τιμές των σφαλμάτων Τύπου I και Τύπου II, αντίστοιχα, διαφέρουν ελάχιστα κάτω από τα Σενάρια I, II, III και IV. Τα αποτελέσματα αυτά υποδεικνύουν ότι η προτεινόμενη μεθοδος είναι αρκετά εύρωστη όσον αφορά τη μοντελοποίηση της απόκρισης κάτω από τα γενικευμένα γραμμικά μοντέλα. Όπως προαναφέραμε, οι Marley και Woods πρότειναν ότι για την αποτελεσματική χρήση των υπερκορεσμένων σχεδιασμών, η αναλογία των παραγόντων προς τον αριθμό των πειραματικών εκτελέσεων θα πρέπει να είναι μικρότερη από 2. Όταν στη μελέτη προσομοίωσής μας θεωρούμε τον υπερκορεσμένο σχεδιασμό $\mathrm{SSD}_{(6,7)}$ η παραπάνω συνθήκη ισχύει, και παρατηρούμε ότι αυτοί οι υπερκορεσμένοι σχεδιασμοί υπερέχουν ξεκάθαρα (όσον αφορά τις μέσες τιμές των σφαλμάτων Τύπου I και Τύπου II) από τους σχεδιασμούς όπως ο $\mathrm{SSD}_{(10,27)}$ και ο $\mathrm{SSD}_{(8,35)}$ για τους οποίους η παραπάνω συνθήκη δεν ισχύει. Ωστόσο, ακόμα και σε αυτή την περίπτωση για την οποία ισχύει ότι $m/n > 2$, οι υπερκορεσμένοι σχεδιασμοί εξακολουθούν να έχουν αρκετά καλή απόδοση, δεδομένου ότι οι μέσες τιμές σφάλματος Τύπου II διατηρούνται σε χαμηλά επίπεδα. Οι συνολικές μέσες τιμές για τα σφάλματα Τύπου I και Τύπου II όλων των μοντέλων που θεωρούνται στους Πίνακες 14.3, 14.4 και 14.5, είναι 0.30 και 0.09, αντίστοιχα, πράγμα που σημαίνει ότι η προτεινόμενη μέθοδος έχει σημαντικά χαμηλότερες τιμές σφαλμάτων Τύπου II, και υψηλότερες τιμές σφάλματος Τύπου I. Σε γενικές γραμμές, οι Πίνακες 14.3, 14.4 και 14.5, δείχνουν ότι η προτεινόμενη μέθοδος έχει την τάση να δηλώνει σε ένα υψηλότερο ποσοστό ανενεργές επιδράσεις να είναι ενεργές και σε πολύ χαμηλότερο ποσοστό ενεργές επιδράσεις να είναι ανενεργές. Έτσι, η προτεινόμενη μέθοδος θα μπορούσε να χαρακτηριστεί ως συντηρητική υπό αυτή την έννοια.

Πίνακας 14.3: Συγκριτική απόδοση των ισορροπημένων και μη ισορροπημένων $\text{SSDs}_{(6,7)}$

Μοντέλο	ΥΣ ΑΠΟ	m	n	Ισορροπία	Σενάριο	Type I	Type II
1	Chatterjee et al. (2014)	7	6	OXI	I	0.20	0.01
2	Chatterjee et al. (2014)	7	6	OXI	II	0.21	0.02
3	Chatterjee et al. (2014)	7	6	OXI	III	0.26	0.04
4	Chatterjee et al (2014)	7	6	OXI	IV	0.26	0.05
Average						0.2325	0.0300
5	Allen & Bernshteyn (2003)	7	6	OXI	I	0.19	0.01
6	Allen & Bernshteyn (2003)	7	6	OXI	II	0.21	0.02
7	Allen & Bernshteyn (2003)	7	6	OXI	III	0.27	0.04
8	Allen & Bernshteyn (2003)	7	6	OXI	IV	0.26	0.05
Average						0.2325	0.0300
9	Booth & Cox (1962)	7	6	NAI	I	0.19	0.02
10	Booth & Cox (1962)	7	6	NAI	II	0.20	0.04
11	Booth & Cox (1962)	7	6	NAI	III	0.26	0.06
12	Booth & Cox (1962)	7	6	NAI	IV	0.25	0.07
Average						0.2250	0.0475
13	Wu (1993)	7	6	NAI	I	0.19	0.02
14	Wu (1993)	7	6	NAI	II	0.20	0.04
15	Wu (1993)	7	6	NAI	III	0.26	0.06
16	Wu (1993)	7	6	NAI	IV	0.25	0.07
Average						0.2250	0.0475
17	Lin (1993)	7	6	NAI	I	0.19	0.02
18	Lin (1993)	7	6	NAI	II	0.20	0.04
19	Lin (1993)	7	6	NAI	III	0.26	0.06
20	Lin (1993)	7	6	NAI	IV	0.25	0.07
Average						0.2250	0.0475
21	Allen & Bernshteyn (2003)	7	6	NAI	I	0.19	0.02
22	Allen & Bernshteyn (2003)	7	6	NAI	II	0.20	0.04
23	Allen & Bernshteyn (2003)	7	6	NAI	III	0.26	0.06
24	Allen & Bernshteyn (2003)	7	6	NAI	IV	0.25	0.07
Average						0.2250	0.0475

Πίνακας 14.4: Συγκριτική απόδοση των ισορροπημένων και μη ισορροπημένων $\text{SSDs}_{(10,27)}$

Μοντέλο	ΥΣ ΑΠΟ	m	n	Ισορροπία	Σενάριο	Type I	Type II
1	Chatterjee et al. (2014)	27	10	OXI	I	0.29	0.07
2	Chatterjee et al. (2014)	27	10	OXI	II	0.32	0.07
3	Chatterjee et al. (2014)	27	10	OXI	III	0.32	0.10
4	Chatterjee et al. (2014)	27	10	OXI	IV	0.32	0.08
Average						0.3125	0.0800
5	Liu & Dean (2004)	27	10	NAI	I	0.29	0.08
6	Liu & Dean (2004)	27	10	NAI	II	0.31	0.09
7	Liu & Dean (2004)	27	10	NAI	III	0.32	0.12
8	Liu & Dean (2004)	27	10	NAI	IV	0.31	0.10
Average						0.3075	0.0975

Πίνακας 14.5: Συγκριτική απόδοση των ισορροπημένων και μη ισορροπημένων $\text{SSDs}_{(8,35)}$

Μοντέλο	ΥΣ ΑΠΟ	m	n	Ισορροπία	Σενάριο	Type I	Type II
1	Chatterjee et al. (2014)	35	8	OXI	I	0.34	0.12
2	Chatterjee et al. (2014)	35	8	OXI	II	0.36	0.13
3	Chatterjee et al. (2014)	35	8	OXI	III	0.37	0.14
4	Chatterjee et al. (2014)	35	8	OXI	IV	0.36	0.15
Average						0.3575	0.1350
5	Cheng (1997)	35	8	NAI	I	0.34	0.14
6	Cheng (1997)	35	8	NAI	II	0.35	0.15
7	Cheng (1997)	35	8	NAI	III	0.37	0.17
8	Cheng (1997)	35	8	NAI	IV	0.35	0.16
Average						0.3525	0.1550

14.4.2 Ιδιότητες Βελτιστότητας

Σε αυτή την ενότητα συγκρίνουμε τους προαναφερθέντες μη ισορροπημένους υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις με τους ισορροπημένους ομόλογούς τους, χρησιμοποιώντας διαφορετικά κριτήρια βελτιστοποίησης. Ο Πίνακας 14.6 παρουσιάζει τις διάφορες τιμές κριτηρίων βελτιστοποίησης (r_{max}, $E(s^2)$ (Booth & Cox), $E(s^2)$ (Mar-

ley & Woods)), οι οποίες προέκυψαν για τους τρεις μη ισορροπημένους υπερκορεσμένους σχεδιασμούς και τους ισορροπημένους ομόλογούς τους.

Πρώτον, είναι αξιοσημείωτο το πόσο όμοιες είναι οι αποδόσεις για τους ισορροπημένους υπερκορεσμένους σχεδιασμούς 6 εκτελέσεων και 7 παραγόντων, μολονότι παράγονται με διαφορετικούς τρόπους. Παρατηρούμε επίσης από τον Πίνακα 14.6 ότι οι τιμές των κριτηρίων για τους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις διαφέρουν ελάχιστα από τις αντίστοιχες τιμές για τους ισορροπημένους υπερκορεσμένους σχεδιασμούς, εκτός από το κριτήριο r_{max} για τους $\mathrm{SSD}_{(6,7)}$ και $\mathrm{SSD}_{(10,27)}$. Τα παραδείγματα αυτά στον Πίνακα 14.6, υποδεικνύουν ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί που περιλαμβάνουν αλληλεπιδράσεις είναι συγκρίσιμοι με τους ισορροπημένους ομόλογούς τους σε σχέση με το $E(s^2)$ ή r_{max} κριτήριο βελτιστοποίησης. Ας σημειωθεί εδώ ότι η χαλάρωση του περιορισμού της ισορροπίας επιτρέπει στους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς να βελτιώνονται σε σχέση με τις "$E(s^2)$" τιμές, καθώς επίσης να αποδίδουν αρκετά καλά σε σχέση και με τα υπόλοιπα κριτήρια βελτιστοποίησης. Οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί που περιλαμβάνουν αλληλεπιδράσεις μπορεί ναι μεν να μην έχουν την ιδιότητα της ισορροπίας, αλλά ακόμα έχουν πλεονεκτήματα με την έννοια ότι είναι "σχεδόν" βέλτιστοι σε σχέση με αρκετά κριτήρια και οι επιδράσεις των παραγόντων είναι "σχεδόν" ορθογώνιες.

Πίνακας 14.6: Κριτήρια βελτιστότητας μη ισορροπημένων ΥΣ με αλληλεπιδράσεις και των αντίστοιχων ισορροπημένων ΥΣ

n	m	Κριτήριο	Ισορροπία	r_{max}	$E(s^2)$ (Booth & Cox)	$E(s^2)$ (Marley & Woods)
6	7	Booth & Cox (1962) [a]	ΝΑΙ	0.3333	4.0000	4.0000
6	7	Wu (1993) [a]	ΝΑΙ	0.3333	4.0000	4.0000
6	7	Lin (1993) [a]	ΝΑΙ	0.3333	4.0000	4.0000
6	7	Allen & Bernshteyn (2003) [a]	ΝΑΙ	0.3333	4.0000	4.0000
6	7	Allen & Bernshteyn (2003)	ΟΧΙ	0.4472	4.0000	4.1429
6	7	Chatterjee et al. (2014)	ΟΧΙ	0.7071	4.3810	3.4286
10	27	Liu & Dean (2004) [a]	ΝΑΙ	0.6000	8.1000	8.1000
10	27	Chatterjee et al. (2014)	ΟΧΙ	0.6000	7.0769	7.4286
8	35	Cheng (1997) [b]	ΝΑΙ	0.5000	7.5294	7.5294
8	35	Chatterjee et al. (2014)	ΟΧΙ	0.5774	6.9647	7.1111

[a] the smaller-the-better
[b] the larger-the-better

14.5 Συμπεράσματα

Οι υπερκορεσμένοι σχεδιασμοί συνήθως κατασκευάζονται κάτω από την παραδοχή της ισορροπίας, αλλά κάτι τέτοιο φαίνεται να είναι άσχετο με τον τρόπο με τον οποίο αναλύονται τα δεδομένα τους [92]. Ας σημειωθεί εδώ, ότι το πλαίσιο ανάλυσης μέχρι σήμερα για τους μη ισορροπημένους υπερκορεσμένους σχεδιασμούς περιορίζεται στην κεντρική υπόθεση ότι τα πειραματικά δεδομένα προέρχονται από ένα γραμμικό μοντέλο. Στο παρόν κεφάλαιο, συγκρίναμε μη ισορροπημένους υπερκορεσμένους σχεδιασμούς με ισορροπημένους ομόλογούς τους κάτω από την παραδοχή των γενικευμένων γραμμικών μοντέλων με διακριτή απόκριση, αποκαλύπτοντας ότι οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί έχουν πλεονεκτήματα όσον αφορά την απόδοσή τους, παρά την έλλειψη ισορροπίας. Οι πιο φουσκωμένες τιμές των σφαλμάτων Τύπου Ι δεν αποτελούν πρόβλημα δεδομένου ότι κάτω από την αρχή της σποραδικότητας των επιδράσεων, είναι φυσικό και πολύ πιθανό να προκύψουν σφάλματα Τύπου Ι, όπως αναφέρθηκε και από τον Lin [149]. Άλλωστε, οι υπερκορεσμένοι σχεδιασμοί κυρίως χρησιμοποιούνται για να εντοπιστούν οι παράγοντες που πρέπει να εξεταστούν σε περαιτέρω αναλύσεις, και ως εκ τούτου, τα χαμηλά ποσοστά σφάλματος Τύπου ΙΙ είναι ιδιαίτερα επιθυμητά. Η προτεινόμενη μέθοδος επιτυγχάνει κάτι τέτοιο πολύ ικανοποιητικά.

Τα αποτελέσματα του παρόντος κεφαλαίου αποτελούν αποδείξεις ότι οι μη ισορροπημένοι

υπερκορεσμένοι σχεδιασμοί είναι συγκρίσιμοι με τους ισορροπημένους ομόλογούς τους, και μπορεί να αποτελούν ένα χρήσιμο εργαλείο για τα πειράματα κρησαρίσματος. Τα αποτελέσματα του παρόντος κεφαλαίου επίσης παρέχουν τις ακόλουθες κατευθυντήριες γραμμές για τη χρήση των ισορροπημένων και μη ισορροπημένων υπερκορεσμένων σχεδιασμών κάτω από τα γενικευμένα γραμμικά μοντέλα με διακριτή απόκριση.

1. Ισορροπημένοι και μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί του ίδιου μεγέθους φαίνεται να έχουν παρόμοια ικανότητα να ανιχνεύουν ενεργούς ή μη ενεργούς παράγοντες.

2. Οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί αποδίδουν πολύ καλά, και παρουσιάζουν ελαφρώς χαμηλότερα σφάλματα Τύπου ΙΙ σε συνδυασμό με ελαφρώς υψηλότερα σφάλματα Τύπου Ι σε σύγκριση με τους ισορροπημένους ομόλογούς τους ίδιου μεγέθους.

3. Για αμφότερες τις δύο κλάσεις των υπερκορεσμένων σχεδιασμών (ισορροπημένοι και μη ισορροπημένοι), η απόδοση χειροτερεύει καθώς ο λόγος των παραγόντων προς τις πειραματικές εκτελέσεις αυξάνεται, αν και καλά αποτελέσματα (χαμηλά σφάλματα Τύπου ΙΙ) προκύπτουν ακόμα και στις περιπτώσεις για τις οποίες $m/n \geq 2$. Ωστόσο, σε κάποες περιπτώσεις για τις οποίες $m/n > 2$, οι μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί φαίνεται να υπερτερούν των ισορροπημένων ομολόγων τους όσον αφορά κάποια κριτήρια βελτιστοποίησης (για τους $\mathrm{SSD}_{(10,27)}$ και $\mathrm{SSD}_{(8,35)}$ παρατηρήσαμε χαμηλότερες τιμές του $E(s^2)$). Για την περίπτωση στην οποία $m/n = 2$ ($\mathrm{SSD}_{(12,24)}$) παρατηρήσαμε σχεδόν ταυτόσημες τιμές για τα κριτήρια $E(s^2)$ και D_{Bayes}.

4. Ισορροπημένοι και μη ισορροπημένοι υπερκορεσμένοι σχεδιασμοί μπορεί να έχουν διαφορετικές ιδιότητες βελτιστότητας, αλλά αυτό φαίνεται να είναι άσχετο και να μην επηρεάζει τον τρόπο με τον οποίο τα δεδομένα των υπερκορεσμένων σχεδιασμών αναλύονται κάτω από τα γενικευμένα γραμμικά μοντέλα με διακριτή απόκριση.

ΚΕΦΑΛΑΙΟ **15**

Κάτω Φράγματα του Μέτρου $E(S^2)$ για Μερικώς Ισορροπημένους Σχεδιασμούς Δύο Επιπέδων

The shortest path between two truths in the real domain
passes through the complex domain.

—*Jacques S. Hadamard (1865–1963)*

Στο δέκατο πέμπτο αυτό κεφάλαιο μελετάμε τα πιθανά πλεονεκτήματα της χαλάρωσης της απαίτησης ισορροπίας στην κατασκευή υπερκορεσμένων σχεδιασμών, δεδομένου ότι οι υπερκορεσμένοι σχεδιασμοί συνήθως κατασκευάζονται κάτω από την παραδοχή της ισορροπίας. Παρουσιάζουμε ένα νέο κάτω φράγμα του κριτηρίου βελτιστοποίησης $E(s^2)$ για μερικώς ισορροπημένους σχεδιασμούς, και το εφαρμόζουμε σε μερικώς ισορροπημένους, μη ισορροπημένους και ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς που υπάρχουν στη βιβλιογραφία. Επίσης, παρέχουμε νέους $E(s^2)$-βέλτιστους μερικώς ισορροπημένους σχεδιασμούς οι οποίοι προκύπτουν από την εφαρμογή του νέου αυτού κάτω φράγματος. Τα ερευνητικά αποτελέσματα αυτού του κεφαλαίου παρουσιάζονται στην επιστημονική εργασία [45].

15.1 Ερευνητικό Πρόβλημα - Κινητήριος Ιδέα

Πρόσφατες επιστημονικές εργασίες αναφέρονται στα πιθανά πλεονεκτήματα της χαλάρωσης της απαίτησης της ισορροπίας και μελετούν την κατασκευή ή ακόμα και την ανάλυση των μη ισορροπημένων υπερκορεσμένων σχεδιασμών, όπως για παράδειγμα οι εργασίες [4], [92], [118], [160] και [43]. Στο Κεφάλαιο 13, εφαρμόζοντας τη μέθοδο BEDDO και χαλαρώνοντας τον περιορισμό της ισορροπίας, κατασκευάσαμε μη ισορροπημένους υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις, σχεδιασμούς με καλές στατιστικές ιδιότητες που συγκρίνονται ευνοϊκά με ισορροπημένους ομόλογούς τους για πολλά κριτήρια βελτιστοποίησης, συμπεριλαμβανομένου του $E(s^2)$. Η νέα αυτή κλάση των μη ισορροπημένων υπερκορεσμένων σχεδιασμών που κατασκευάσαμε με τη βοήθεια των υπολογιστών είναι στην πραγματικότητα μερικώς ισορροπημένοι υπερκορεσμένοι σχεδιασμοί, με την έννοια ότι m_1 από τις m στήλες είναι ισορροπημένες (έχουν ίσο αριθμό "-1's" και "1's") και οι υπόλοιπες $(m - m_1)$ $(= m_2)$ στήλες είναι μη ισορροπημένες. Τέτοιου τύπου σχεδιασμοί με την ιδιότητα της μερικής ισορροπίας μπορεί να προκύψουν συχνά στην πράξη εφόσον η ύπαρξη των αλληλεπιδράσεων θεωρηθεί είτε στην κατασκευή είτε στην ανάλυση των υπερκορεσμένων σχεδιασμών κύριων επιδράσεων. Οι μερικώς ισορροπημένοι σχεδιασμοί μπορεί να ενδιαφέρουν έναν πειραματιστή ο οποίος χρειάζεται σχεδιασμούς όσο το δυνατόν πιο ισορροπημένους και μπορεί να είναι προτιμητέοι σε περιπτώσεις στις οποίες τα επίπεδα συγκεκριμένων παραγόντων σχετίζονται με υψηλά κόστη. Οι παραπάνω λόγοι, όπως είναι φυσικό, μας παρακίνησαν να ψάξουμε νέα κάτω φράγματα για το $E(s^2)$ κριτήριο για τη νέα αυτή κλάση των μερικώς ισορροπημένων σχεδιασμών.

15.2 Το Κριτήριο Βελτιστοποίησης $E(s^2)$

Η δομή ενός υπερκορεσμένου σχεδιασμού είναι πολύ σημαντική, και η κατασκευή τους έχει προσελκύσει το ενδιαφέρον πολλών ερευνητών τα τελευταία χρόνια. Για την αξιολόγηση των κατασκευασμένων υπερκορεσμένων σχεδιασμών και τη μέτρηση της αποτελεσματικότητάς τους ορισμένα κριτήρια βελτιστοποίησης είναι απαραίτητα. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στην εργασία [91] για μια λεπτομερή αναθεώρηση των κριτηρίων βελτιστοποίησης τα οποία αναπτύχθηκαν στη βιβλιογραφία σχετικά με τους υπερκορεσμένους σχεδιασμούς δύο επιπέδων.

Σε αυτό το κεφάλαιο, μελετούμε το κριτήριο βελτιστοποίησης $E(s^2)$ των Booth και Cox [21], το οποίο ορίζεται ως $E(s^2) = \sum s_{ij}^2/m(m-1)$, όπου $\sum s_{ij}^2$ είναι δύο φορές το άθροισμα των τετραγώνων των στοιχείων πάνω από τη διαγώνιο του πίνακα πληροφορίας $X^{\mathrm{T}}X$ του υπερκορεσμένου σχεδιασμού, όπου ο πίνακας σχεδιασμού λαμβάνεται από το δοθέν σχεδιασμό αντικαθιστώντας τα "0's" με "-1's". Για έναν ισορροπημένο σχεδιασμό με m παράγοντες σε δύο επίπεδα ο κάθε ένας και N πειραματικές εκτελέσεις $(m \geq N)$, ο Nguyen [172], και οι Tang και Wu [215], ανεξάρτητα, παρείχαν ένα κάτω φράγμα

$$E(s^2) \geq \frac{N^2(m-N+1)}{(N-1)(m-1)} \tag{15.1}$$

για το $E(s^2)$. Οι Tang και Wu έδειξαν ότι το κάτω φράγμα για το $E(s^2)$ μπορεί να επιτευχθεί εάν και μόνο εάν $m = q(N-1)$ και $N \equiv 0 \pmod 4$ ή για $m = 2q(N-1)$ και $N \equiv 2 \pmod 4$ για κάποιο θετικό ακέραιο q. Άρα, το N είναι άρτιος αριθμός. Ένας σχεδιασμός καλείται $E(s^2)$-βέλτιστος όταν αυτό το φράγμα επιτυγχάνεται. Ο Cheng [47] έδειξε ότι με τη διαγραφή ή την προσθήκη ενός ή δύο ορθογώνιων (ή κοντά ορθογώνιων για $N \equiv 2 \pmod 4$) στηλών, ο σχεδιασμός εξακολουθεί να είναι $E(s^2)$-βέλτιστος. Εάν το $N = 8$ το παραπάνω αποτέλεσμα ισχύει επίσης με την προσθήκη ή τη διαγραφή τριών στηλών. Πρόσφατα βελτιωμένα κάτω φράγματα του $E(s^2)$ για ισορροπημένους σχεδιασμούς δόθηκαν από τους Butler et al. [34],

Liu και Hickernell [152], Bulutoglu και Cheng [30], καθώς επίσης και από τους Ryan και Bulutoglu [202]. Το καλύτερο πρόσφατα γνωστό κάτω φράγμα στην περίπτωση αυτή δόθηκε από τους Das et al. [60], καθώς επίσης οι Nguyen και Cheng [173] γενίκευσαν το κάτω φράγμα της σχέσης (15.1) έτσι ώστε να είναι εφαρμόσιμο σε άρτιο και περιττό αριθμό πειραματικών εκτελέσεων.

Όλα τα προαναφερθέντα φράγματα ισχύουν για ισορροπημένους σχεδιασμούς. Για μη ισορροπημένους σχεδιασμούς τα παραπάνω φράγματα δεν είναι εφαρμόσιμα. Ακόμα κι αν υπήρξε πολλή ερευνητική δουλειά για το κάτω φράγμα του κριτηρίου $E(s^2)$ για ισορροπημένους σχεδιασμούς, οι εργασίες για το κάτω φράγμα του κριτηρίου $E(s^2)$ για μη ισορροπημένους σχεδιασμούς είναι ακόμα σε πρώιμο στάδιο, αν και ορισμένες νέες προσεγγίσεις έχουν αναπτυχθεί τα τελευταία χρόνια. Οι Booth και Cox πρότειναν το κριτήριο $E(s^2)$ το οποίο επιλέγει ένα σχεδιασμό ελαχιστοποιώντας το άθροισμα των τετραγώνων των εσωτερικών γινομένων μεταξύ όλων των πιθανών ζευγών των στηλών του X. Εάν η στήλη των 1's συμπεριληφθεί στον X, τότε οι Marley και Woods [160] επέκτειναν αυτό τον ορισμό για να συμπεριλάβουν το εσωτερικό γινόμενο της πρώτης στήλης με κάθε άλλη στήλη του X, δηλαδή,

$$E(s^2) = \sum_{i \neq j} s_{ij}^2/((m+1)m), \tag{15.2}$$

όπου το s_{ij} είναι το i,j-οστό στοιχείο του πίνακα $X^{\mathrm{T}}X$ $(i, j = 1, \ldots m+1)$. Οι δύο ορισμοί είναι ισοδύναμοι για τους ισορροπημένους σχεδιασμούς. Οι Bulutoglu και Ryan [32] πρότειναν ένα βελτιωμένο κάτω φράγμα για το $E(s^2)$, εφαρμόσιμο σε μη ισορροπημένους υπερκορεσμένους σχεδιασμούς δύο επιπέδων με περιττό αριθμό πειραματικών εκτελέσεων. Το καλύτερο πρόσφατα γνωστό κάτω φράγμα, στην περίπτωση των μη ισορροπημένων υπερκορεσμένων σχεδιασμών δύο επιπέδων με περιττό αριθμό πειραματικών εκτελέσεων, δόθηκε πρόσφατα από τους Suen και Das [213]. Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει για περισσότερες λεπτομέρειες σχετικά με την ιστορική εξέλιξη αυτών των φραγμάτων στην εργασία [123].

Στο παρόν κεφάλαιο, παρουσιάζουμε κάποια γενικά αποτελέσματα βελτιστοποίησης για μερικώς ισορροπημένους σχεδιασμούς και παρέχουμε ένα νέο κάτω φράγμα (lower bound-LB) για το κριτήριο βελτιστοποίησης $E(s^2)$. Αυτό το νέο $E(s^2)$ κάτω φράγμα (LB) που προτείνουμε είναι αναπόφευκτα εφαρμόσιμο μόνο σε άρτιο αριθμό πειραματικών εκτελέσεων, δεδομένου ότι οι μερικώς ισορροπημένοι σχεδιασμοί δεν μπορούν να έχουν περιττό αριθμό πειραματικών εκτελέσεων εξαιτίας της ιδιότητας της μερικής ισορροπίας. Το νέο αυτό $E(s^2)$ κάτω φράγμα (LB) αναπτύχθηκε για την αξιολόγηση μερικώς ισορροπημένων υπερκορεσμένων σχεδιασμών, αλλά επίσης μπορεί να εφαρμοστεί στην περίπτωση ισορροπημένων υπερκορεσμένων σχεδιασμών δύο επιπέδων, καθώς επίσης και στην περίπτωση μη ισορροπημένων υπερκορεσμένων σχεδιασμών δύο επιπέδων με άρτιο αριθμό πειραματικών εκτελέσεων.

15.2.1 Συμβολισμοί και Προκαταρκτικά

Ας θεωρήσουμε έναν υπερκορεσμένο σχεδιασμό N εκτελέσεων και m παραγόντων ο κάθε ένας σε δύο επίπεδα "0" και "1". Ας υποθέσουμε ότι ο σχεδιασμός είναι μερικώς ισορροπημένος, με την έννοια ότι m_1 από τις m στήλες είναι ισορροπημένες (έχουν ίσο αριθμό "0's" και "1's") και οι υπόλοιπες $(m - m_1)$ $(= m_2)$ στήλες είναι μη ισορροπημένες. Ας σημειωθεί εδώ ότι από τη στιγμή που ο σχεδιασμός είναι μερικώς ισορροπημένος, το N πρέπει να είναι πολλαπλάσιο του 2, έστω $N = 2n$. Θεωρούμε έναν ακέραιο u μεταξύ του 1 και $(n-1)$. Για $1 \leq l \leq u$, ας υποθέσουμε ότι μεταξύ των m_2 μη ισορροπημένων στηλών, k_{l1} στήλες έχουν $(n+l)$ "1's" και $(n-l)$ "0's" και k_{l2} στήλες μεταξύ των m_2 έχουν $(n-l)$ "1's" και $(n+l)$ "0's". Έστω $k_l = k_{l1} + k_{l2}$. Τότε είναι εύκολο να παρατηρηθεί ότι $\sum_{l=1}^{u} k_l = m_2$. Χωρίς απώλεια της γενικότητας, ας υποθέσουμε ότι οι πρώτες m_1 στήλες του δοθέντος σχεδιασμού είναι ισορροπημένες. Επιπλέον, μεταξύ των τελευταίων m_2 μη

ισορροπημένων στηλών, $m_1 + \sum_{w=1}^{l-1} k_w + 1$ έως $m_1 + \sum_{w=1}^{l-1} k_w + k_{l1}$ στήλες έχουν $(n+l)$ "1's" και $(n-l)$ "0's", καθώς επίσης $m_1 + \sum_{w=1}^{l-1} k_w + k_{l1} + 1$ έως $m_1 + \sum_{w=1}^{l} k_w$ στήλες έχουν $(n-l)$ "1's" και $(n+l)$ "0's". Πρέπει να σημειωθεί ότι όταν $m_2 = 0$, ο σχεδιασμός γίνεται ένας ισορροπημένος υπερκορεσμένος σχεδιασμός N πειραματικών εκτελέσεων.

Συμβολίζουμε με $\mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$ την κλάση των μερικώς ισορροπημένων υπερκορεσμένων σχεδιασμών N εκτελέσεων και με $\mathcal{D}_0(N, 2^m)$ την κλάση των ισορροπημένων υπερκορεσμένων σχεδιασμών N εκτελέσεων.

Παράδειγμα 15.1. *Ας θεωρήσουμε τον ακόλουθο ισορροπημένο υπερκορεσμένο σχεδιασμό 4 εκτελέσεων $d \in \mathcal{D}_0(4, 2^4)$ όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Για αυτόν το σχεδιασμό, $m_1 = m = 4$ στήλες είναι ισορροπημένες (έχουν ίσο αριθμό "0's" και "1's") και $m_2 = 0$.

Παράδειγμα 15.2. *Ας θεωρήσουμε τον ακόλουθο μερικώς ισορροπημένο υπερκορεσμένο σχεδιασμό 4 εκτελέσεων $d \in \mathcal{D}(4, 3, 1(k_1 = 1), 2^4)$ όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Αυτός ο σχεδιασμός είναι μερικώς ισορροπημένος με την έννοια ότι οι πρώτες $(m-1)$ από τις $m = 4$ στήλες είναι ισορροπημένες (έχουν ίσο αριθμό "0's" και "1's") και η τελευταία στήλη είναι μη ισορροπημένη έχοντας ένα "0" και τρεις "1's". Άρα, $m_1 = 3$ και $\sum_{l=1}^{u} k_l = m_2 = 1$.

Για $1 \le j \le m$, $\alpha = 0, 1$, συμβολίζουμε με n_α^j τον αριθμό των φορών που ο παράγοντας F_j εμφανίζεται στο επίπεδο α. Για $1 \le j_1 < j_2 \le m$, $\alpha, \beta = 0, 1$, συμβολίζουμε με $n_{\alpha\beta}^{j_1 j_2}$ τον αριθμό των φορών που οι παράγοντες F_{j_1} και F_{j_2} εμφανίζονται στα επίπεδα α και β αντίστοιχα. Τώρα ορίζουμε με

$$\begin{aligned} \phi(j) &= 2\sum_{\alpha=0}^{1} (n_\alpha^j)^2, 1 \le j \le m, \text{ και} \\ \phi(j_1 j_2) &= 4\sum_{\alpha=0}^{1}\sum_{\beta=0}^{1} (n_{\alpha\beta}^{j_1 j_2})^2, \ 1 \le j_1 < j_2 \le m. \end{aligned}$$

Ορισμός 15.1. *Για έναν υπερκορεσμένο σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$, ορίζουμε την τιμή του $E(s^2)$ ως*

$$E(s^2) = \frac{1}{m(m-1)} \sum_{j_1=1}^{m} \sum_{j_2(\neq j_1)=1}^{m} \phi(j_1 j_2) - \frac{2}{m} \sum_{j=1}^{m} \phi(j) + N^2.$$

Παρατήρηση 15.1. *Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}_0(N, 2^m)$, ο Ορισμός 15.1 γίνεται*

$$E(s^2) \;=\; \frac{1}{m(m-1)} \sum_{j_1=1}^{m} \sum_{j_2(\neq j_1)=1}^{m} \phi(j_1 j_2) - N^2,$$

που είναι ακριβώς ίδιος με τον ορισμό των Booth και Cox (1962).

Ορισμός 15.2. *Ένας υπερκορεσμένος σχεδιασμός $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$ θα λέγεται $E(s^2)$-βέλτιστος εάν επιτυγχάνει την ελάχιστη τιμή του $E(s^2)$.*

15.2.2 Νέο Κάτω Φράγμα για το $E(s^2)$ Κριτήριο Βελτιστοποίησης

Σε αυτή την ενότητα παρέχουμε ένα νέο κάτω φράγμα για το $E(s^2)$ κριτήριο βελτιστοποίησης.

Λήμμα 15.1. *Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$, έχουμε ότι*

$$\phi(j) \;=\; \begin{cases} N^2, & \text{εάν } 1 \le j \le m_1, \\ N^2 + (2l)^2, & m_1 + \sum_{w=1}^{l-1} k_w + 1 \le j \le m_1 + \sum_{w=1}^{l} k_w. \end{cases}$$

Απόδειξη. Η απόδειξη ότι το $\phi(j) = N^2$, $1 \le j \le m_1$, προκύπτει από το γεγονός ότι, για μια ισορροπημένη στήλη, έχουμε

$$n_0^j = n_1^j = \frac{N}{2}.$$

Τώρα για τα υπόλοιπα, από τη στιγμή που $m_1 + \sum_{w=1}^{l-1} k_w + 1 \le j \le m_1 + \sum_{w=1}^{l} k_w$, η στήλη που αντιστοιχεί στο j-οστό παράγοντα θα έχει είτε

1. $(n-l)$ "1's" και $(n+l)$ "0's", ή
2. $(n+l)$ "1's" και $(n-l)$ "0's",

όπου $n = N/2$.

Στην περίπτωση (1), έχουμε ότι

$$\phi(j) \;=\; 2(n-l)^2 + 2(n+l)^2 \;=\; (2n)^2 + (2l)^2 \;=\; N^2 + (2l)^2.$$

Για την περίπτωση (2) η απόδειξη είναι παρόμοια.

□

Για $1 \le i_1 < i_2 \le N$, $1 \le j \le m$, ορίζουμε ότι

$$c_{i_1 i_2}^j \;=\; \begin{cases} 1, & \text{εάν οι καταχωρήσεις των γραμμών } i_1 \text{ και } i_2 \\ & \text{της } j\text{-οστής στήλης συμπίπτουν,} \\ 0, & \text{σε διαφορετική περίπτωση,} \end{cases}$$

και $c_{i_1 i_2} = \sum_{j=1}^{m} c_{i_1 i_2}^j$. Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$, έστω X ο πίνακας σχεδιασμού που προκύπτει από το δοθέν σχεδιασμό αντικαθιστώντας τα "0"s με "-1"s.

Λήμμα 15.2. *Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$, έχουμε ότι*

$$\sum_{i_1=1}^{N} \sum_{i_2(\neq i_1)=1}^{N} c_{i_1 i_2} = 2\left[\sum_{l=1}^{u} k_l l^2 + mn(n-1)\right].$$

Απόδειξη. Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$,

$$XX^T = \begin{pmatrix} m & 2c_{12}-m & \cdots & 2c_{1N}-m \\ 2c_{12}-m & m & \cdots & 2c_{2N}-m \\ \vdots & \vdots & \vdots & \vdots \\ 2c_{1N}-m & 2c_{2N}-m & \cdots & m \end{pmatrix}$$

Ως εκ τούτου,

$$1_N^T XX^T 1_N = 2\sum_{i_1=1}^{N} \sum_{i_2(\neq i_1)=1}^{N} c_{i_1 i_2} - mN(N-2).$$

Εύκολα αποδεικνύεται ότι

$$\begin{aligned} 1_N^T XX^T 1_N &= \sum_{j=1}^{m} (n_1^j - n_0^j)^2 \\ &= \sum_{j=1}^{m} \phi(j) - mN^2 = 4\sum_{l=1}^{u} k_l l^2. \end{aligned}$$

□

Λήμμα 15.3. *Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$ έχουμε ότι*

$$E(s^2) = \frac{1}{m(m-1)}\left[4\sum_{i_1=1}^{N} \sum_{i_2(\neq i_1)=1}^{N} c_{i_1 i_2}^2 - 8m\sum_{l=1}^{u} k_l l^2 + 4m^2 N - m(m+1)N^2\right].$$

Απόδειξη.

$$\begin{aligned} \sum_{j_1=1}^{m} \sum_{j_2(\neq j_1)=1}^{m} \phi(j_1 j_2) &= \operatorname{tr}\left[(XX^T)(XX^T)\right] + 2(m-1)\sum_{j=1}^{m} \phi(j) - m^2 N^2 \\ &= 4\sum_{i_1=1}^{N} \sum_{i_2(\neq i_1)=1}^{N} c_{i_1 i_2}^2 - 4m\sum_{i_1=1}^{N} \sum_{i_2(\neq i_1)=1}^{N} c_{i_1 i_2} + 2(m-1)\sum_{j=1}^{m} \phi(j) \\ &= 4\sum_{i_1=1}^{N} \sum_{i_2(\neq i_1)=1}^{N} c_{i_1 i_2}^2 - 8\sum_{l=1}^{u} k_l l^2 + 4m^2 N - 2mN^2. \end{aligned}$$

Η απόδειξη έπειτα ακολουθεί άμεσα.

□

Έστω w ο μεγαλύτερος ακέραιος που περιέχεται μέσα στο $(S/N(N-1))$, όπου $S = 2\left[\sum_{l=1}^{u} k_l l^2 + mn(n-1)\right]$. Ορίζουμε το $\theta = S - N(N-1)w$ και το $\theta^* = Sw + \theta(w+1)$. Έπειτα έχουμε το ακόλουθο Λήμμα.

Λήμμα 15.4. *Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$, έχουμε ότι*

$$\sum_{i_1=1}^{N} \sum_{i_2(\neq i_1)=1}^{N} c_{i_1 i_2}^2 \quad \geq \quad \theta^*.$$

Στο ακόλουθο θεώρημα παρουσιάζεται το κάτω φράγμα για το κριτήριο $E(s^2)$ για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$.

Θεώρημα 15.1. *Για οποιοδήποτε σχεδιασμό $d \in \mathcal{D}(N, m_1, m_2(k_1, k_2, \cdots, k_u), 2^m)$, έχουμε ότι*

$$E(s^2) \quad \geq \quad LB,$$

όπου

$$LB \quad = \quad \frac{1}{m(m-1)} \left[4\theta^* - 8m \sum_{l=1}^{u} k_l l^2 + 4m^2 N - m(m+1)N^2 \right].$$

15.2.3 Νέοι $E(s^2)$-Βέλτιστοι Μερικώς Ισορροπημένοι Σχεδιασμοί

Σε αυτή την ενότητα, παρέχουμε νέους $E(s^2)$-βέλτιστους μερικώς ισορροπημένους σχεδιασμούς οι οποίοι λαμβάνονται με την εφαρμογή του νέου κάτω φράγματος. Από τώρα και στο εξής, το "+" συμβολίζει το "+1", και το "−" συμβολίζει το "−1".

Παράδειγμα 15.3. *Θεωρούμε τον ακόλουθο σχεδιασμό $d \in \mathcal{D}(4, 3, 1(k_1 = 1), 2^4)$ όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} + & + & + & + \\ - & + & - & + \\ + & - & - & + \\ - & - & + & - \end{pmatrix}.$$

Για αυτόν το σχεδιασμό το $LB = E(s^2) = 2$.

Παράδειγμα 15.4. *Θεωρούμε τον ακόλουθο σχεδιασμό $d \in \mathcal{D}(8, 7, 1(k_1 = 1), 2^8)$ όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} + & + & + & + & + & + & + & + \\ - & + & - & + & - & + & - & + \\ + & - & - & + & + & - & - & + \\ - & - & + & + & - & - & + & + \\ + & + & + & - & - & - & - & + \\ - & + & - & - & + & - & + & - \\ + & - & - & - & - & + & + & - \\ - & - & + & - & + & + & - & - \end{pmatrix}.$$

Για αυτόν το σχεδιασμό το $LB = E(s^2) = 2.1429$.

Παράδειγμα 15.5. *Θεωρούμε τον ακόλουθο σχεδιασμό $d \in \mathcal{D}(8, 7, 1(k_1 = 1), 2^8)$ όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} + & + & + & + & + & + & + & + \\ - & + & - & + & - & + & - & + \\ + & - & - & + & + & - & - & + \\ - & - & + & + & - & - & + & + \\ + & + & + & - & - & - & - & + \\ - & + & - & - & + & - & + & + \\ + & - & - & - & - & + & + & - \\ - & - & + & - & + & + & - & - \end{pmatrix}.$$

Για αυτόν το σχεδιασμό το $LB = E(s^2) = 1.7143$.

Παράδειγμα 15.6. *Θεωρούμε τον ακόλουθο σχεδιασμό* $d \in \mathcal{D}(6, 10, 1(k_1 = 1), 2^{11})$ *όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} + & + & + & + & - & - & - & + & + & - & + \\ - & + & + & - & - & + & + & - & + & + & + \\ + & - & + & - & + & + & - & + & - & + & + \\ + & + & - & + & + & + & + & - & - & - & + \\ - & - & - & - & + & - & + & + & + & - & - \\ - & - & - & + & - & - & - & - & - & + & - \end{pmatrix}.$$

Για αυτόν το σχεδιασμό το $LB = E(s^2) = 4.4364$.

Παράδειγμα 15.7. *Θεωρούμε τον ακόλουθο σχεδιασμό* $d \in \mathcal{D}(8, 14, 1(k_1 = 1), 2^{15})$ *όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} - & + & - & - & + & - & - & - & - & - & - & - & - & + & + \\ - & - & + & - & - & + & - & + & - & - & + & - & + & - & + \\ + & + & + & + & + & + & - & - & + & - & - & + & + & - & + \\ + & + & + & + & - & - & + & + & + & - & + & - & - & + & + \\ + & - & - & + & - & + & + & - & - & + & - & - & - & - & + \\ - & + & - & - & - & + & + & + & + & + & - & + & + & + & - \\ - & - & + & + & + & - & + & - & - & + & + & + & + & + & - \\ + & - & - & - & + & - & - & + & + & + & + & + & - & - & - \end{pmatrix}.$$

Για αυτόν το σχεδιασμό το $LB = E(s^2) = 5.4095$

Παράδειγμα 15.8. *Θεωρούμε τον ακόλουθο σχεδιασμό* $d \in \mathcal{D}(8, 21, 1(k_1 = 1), 2^{22})$ *όπου οι στήλες αντιστοιχούν στους παράγοντες και οι γραμμές είναι οι συνδυασμοί των επιπέδων ή οι πειραματικές εκτελέσεις:*

$$d = \begin{pmatrix} + & + \\ + & + & + & + & + & + & + & + & + & - & - & - & - & - & - & - & - & - & - & - & - & + \\ + & + & + & - & - & - & - & - & - & + & + & + & + & + & + & - & - & - & - & - & - & + \\ - & - & - & + & + & + & - & - & - & + & + & + & - & - & - & + & + & + & - & - & - & + \\ + & - & - & - & - & - & + & + & - & + & - & - & + & - & - & + & + & - & + & + & - & + \\ - & + & - & + & - & - & + & - & - & - & - & - & - & + & + & + & - & + & + & - & + & - \\ - & - & + & - & + & - & - & - & + & - & + & - & + & - & - & + & - & + & + & + & + & - \\ - & - & - & - & - & + & - & + & + & - & - & + & + & - & + & - & - & + & - & + & + & - \end{pmatrix}.$$

Για αυτόν το σχεδιασμό το $LB = E(s^2) = 6.5974$.

15.2.4 Εφαρμογή του Νέου Κάτω Φράγματος σε Υπάρχοντες Σχεδιασμούς

Σε αυτή την ενότητα, εφαρμόζουμε το νέο κάτω φράγμα του κριτηρίου βελτιστοποίησης $E(s^2)$ σε διάφορους τύπους σχεδιασμών που υπάρχουν στη βιβλιογραφία, δηλαδή, μερικώς ισορροπημένους, μη ισορροπημένους και ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς.

Ο Πίνακας 15.1 παρουσιάζει τις τιμές των κριτηρίων και τα αποτελέσματα για τρεις μερικώς ισορροπημένους υπερκορεσμένους σχεδιασμούς που περιλαμβάνουν αλληλεπιδράσεις, όπως αυτοί δίνονται από τους Chatterjee, Koukouvinos και Parpoula [43]. Επίσης παρουσιάζουμε τις τιμές των κριτηρίων και τα αποτελέσματα για έναν $(6, 7)$ μερικώς ισορροπημένο υπερκορεσμένο σχεδιασμό και έναν $(10, 11)$ μη ισορροπημένο υπερκορεσμένο σχεδιασμό, όπως αυτοί δίνονται από τους Allen και Bernshteyn [4]. Στον Πίνακα 15.2, εξετάζουμε ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς, για τους οποίους ισχύει ότι $m = 2(n-1)$, όπως αυτοί δίνονται από τους Koukouvinos et al. [130] με τιμές (n, m): $(6, 10)$, $(8, 14)$, $(10, 18)$, $(12, 22)$, $(14, 26)$, $(16, 30)$, $(18, 34)$, $(20, 38)$ και $(22, 42)$. Επίσης εξετάζουμε τον $(6, 10)$ ισορροπημένο $E(s^2)$-βέλτιστο υπερκορεσμένο σχεδιασμό του Lin [148] και τον $(12, 22)$ ισορροπημένο $E(s^2)$-βέλτιστο υπερκορεσμένο σχεδιασμό των Tang και Wu [215]. Σημειώνουμε εδώ ότι το $E(s^2)$ υπολογίζεται από τον Ορισμό 15.1 στον Πίνακα 15.1. Στον Πίνακα 15.2, το $E(s^2)$ υπολογίζεται από τον ορισμό που δίνεται στην Παρατήρηση 15.1, που είναι ακριβώς ίδιος με τον ορισμό των Booth και Cox (1962).

Πίνακας 15.1: Τιμές κριτηρίων για μερικώς ισορροπημένους (PB) και μη ισορροπημένους (U) υπερκορεσμένους σχεδιασμούς

n	m	ΥΣ ΑΠΟ	Ισορροπία	$E(s^2)$	LB
6	7	Chatterjee et al. (2014)	PB	4.3810	2.4762
8	35	Chatterjee et al. (2014)	PB	6.9647	6.4000
10	27	Chatterjee et al. (2014)	PB	7.0769	6.6667
6	7	Allen & Bernshteyn (2003)	PB	4.0000	1.7143
10	11	Allen & Bernshteyn (2003)	U	20.5818	17.9636

Πίνακας 15.2: Τιμές κριτηρίων για ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς

n	m	ΥΣ ΑΠΟ	$E(s^2)$	LB
6	10	Koukouvinos et al. (2008)	4.0000	4.0000
8	14	Koukouvinos et al. (2008)	4.9231	4.9231
10	18	Koukouvinos et al. (2008)	5.8824	5.8824
12	22	Koukouvinos et al. (2008)	6.8571	6.8571
14	26	Koukouvinos et al. (2008)	7.8400	7.8400
16	30	Koukouvinos et al. (2008)	8.8276	8.8276
18	34	Koukouvinos et al. (2008)	9.8182	9.8182
20	38	Koukouvinos et al. (2008)	10.8108	10.8108
22	42	Koukouvinos et al. (2008)	11.8049	11.8049
6	10	Lin (1993)	4.0000	4.0000
12	22	Tang & Wu (1997)	6.8571	6.8571

15.3 Συμπεράσματα

Σε αυτό το κεφάλαιο, παρουσιάσαμε ένα νέο κάτω φράγμα του κριτηρίου βελτιστοποίησης $E(s^2)$ για μερικώς ισορροπημένους σχεδιασμούς, και μέσα από πολλά παραδείγματα, δείξαμε την εφαρμογή αυτού του φράγματος σε μερικώς ισορροπημένους, μη ισορροπημένους και ισορροπημένους $E(s^2)$-βέλτιστους υπερκορεσμένους σχεδιασμούς που υπάρχουν στη βιβλιογραφία. Επίσης, παρέχουμε νέους $E(s^2)$-βέλτιστους μερικώς ισορροπημένους σχεδιασμούς, οι οποίοι προκύπτουν από την εφαρμογή του νέου κάτω φράγματος. Ο νέος ορισμός του

$E(s^2)$ ισοδυναμεί με τον ορισμό των Booth και Cox (1962) για την ισορροπημένη περίπτωση, και το νέο αυτό κάτω φράγμα είναι επίσης εφαρμόσιμο σε ισορροπημένους σχεδιασμούς δύο επιπέδων ($m_2 = 0$). Επιπλέον, το νέο αυτό κάτω φράγμα είναι εφαρμόσιμο σε μη ισορροπημένους σχεδιασμούς δύο επιπέδων ($m_1 = 0$) με άρτιο αριθμό πειραματικών εκτελέσεων. Συστήνουμε τους νεόυς αυτούς $E(s^2)$-βέλτιστους μερικώς ισορροπημένους σχεδιασμούς, όχι μόνο για τις πολύ καλές στατιστικές τους ιδιότητες, αλλά και για την εφαρμοσιμότητά τους σε ένα ευρύτερο σύνολο πραγματικών προβλημάτων της ζωής, όπου είναι πιθανόν να προκύπτουν σχεδιασμοί οι οποίοι δεν περιορίζονται στο να είναι ισορροπημένοι [159]. Οι μερικώς ισορροπημένοι σχεδιασμοί μπορεί να ενδιαφέρουν έναν πειραματιστή ο οποίος χρειάζεται σχεδιασμούς όσο το δυνατόν πιο ισορροπημένους και μπορεί να είναι προτιμητέοι σε περιπτώσεις στις οποίες τα επίπεδα συγκεκριμένων παραγόντων σχετίζονται με υψηλά κόστη.

Βιβλιογραφία

[1] B. Abraham, H. Chipman and K. Vijayan, Some risks in the construction and analysis of supersaturated designs, *Technometrics*, 41 (1999), 135-141.

[2] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd edn., Wiley, New York, 2007.

[3] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19 (1974), 716-723.

[4] T.T. Allen and M. Bernshteyn, Supersaturated designs that maximize the probability of identifying active factors, *Technometrics*, 45 (2003), 90-97.

[5] A. Antoniadis, Wavelets in statistics: a review (with discussion), *Journal of the Italian Statistical Association*, 6 (1997), 97-144.

[6] N. Balakrishnan, C. Koukouvinos and C. Parpoula, Analysis of a supersaturated design using entropy prior complexity for binary responses via generalized linear models, *Statistical Methodology*, 79 (2012) 478-485.

[7] N. Balakrishnan, C. Koukouvinos and C. Parpoula, An information theoretical algorithm for analyzing supersaturated designs for a binary response, *Metrika*, 76 (2013), 1-18.

[8] N. Balakrishnan, C. Koukouvinos and C. Parpoula, Analyzing supersaturated designs for discrete responses via generalized linear models, *Statistical Papers*, (2014), DOI: 10.1007/s00362-013-0569-z.

[9] N. Balakrishnan, C. Koukouvinos and C. Parpoula, On the computation of entropy prior complexity and marginal prior distribution for the Bernoulli model, *Journal of Statistical Theory and Practice*, 9 (2015), 59-72.

[10] V. Balasubramanian, Statistical inference, Occam's Razor, and statistical mechanics on the space of probability distributions, *Neural Computation*, 9 (1997), 349-368.

[11] S.D. Beattie, D.K.F. Fong and D.K.J. Lin, A two-stage Bayesian model selection strategy for supersaturated designs, *Technometrics*, 44 (2002), 55-63.

[12] Y. Benjamini and Y. Hochberg, The adaptive control of the false discovery rate in multiple hypotheses testing, *Journal of Educational and Behavioral Statistics*, 25 (2000), 60-83.

[13] C.H. Bennett, On the nature and origin of complexity in discrete, homogeneous locally-interacting systems, *Foundations of Physics*, 16 (1986), 585-592.

[14] J.F. Bercher, On some entropy functionals derived from Rényi information divergence, *Information Sciences*, 178 (2008), 2489-2506.

[15] A.L Berger, S. Della Pietra and V.J. Della Pietra, A maximum-entropy approach to natural language processing, *Computational Linguistics*, 22 (1996), 39-71.

[16] W. Bialek, I. Nemenman and N. Tishby, Predictability, complexity, and learning, *Neural Computation*, 13 (2001), 2409-2463.

[17] J. Biesiada and W. Duch, Feature selection for high-dimensional data: a Pearson redundancy based filter, In: M. Kurzynski et al. (Eds.) *Computer Recognitions Systems 2*, Vol. 45, pp.242-249, Springer, Berlin, Heidelberg, 2007.

[18] N. Blunch, *Introduction to Structural Equation Modelling Using SPSS and Amos*, SAGE Publications Ltd.N., 2008.

[19] M. Bogdan, J.K. Ghosh and R.W. Doerge, Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci, *Genetics*, 167 (2004), 989-999.

[20] M. Bogdan, J. K. Ghosh and M. Z. Szatkowska, Selecting explanatory variables with the modified version of the Bayesian information criterion, *Quality and Reliability Engineering International*, 24 (2008), 627-641.

[21] K.H.V. Booth and D.R. Cox, Some systematic supersaturated designs, *Technometrics*, 4 (1962), 489-495.

[22] G.E.P. Box and J.S. Hunter, The 2^{k-p} fractional factorial designs, *Technometrics*, 3 (1961), 311-351, 449-458.

[23] G.E.P. Box and R.D. Meyer, An analysis for unreplicated fractional factorials, *Technometrics*, 28 (1986), 11-18.

[24] G.E.P. Box and R.D. Meyer, Finding the active factors in fractionated screening experiments, *Journal of Quality Technology*, 25 (1993), 94-105.

[25] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, 30 (1997), 1145-1159.

[26] P.S. Bradley and O.L Mangasarian, Feature selection via concave minimization and support vector machines, In: Shavlik, J. (Eds.) *Machine Learning Proceedings of the Fifteenth International Conference (ICML '98)*, pp.82-90, Morgan Kaufmann, San Fransisco, CA, 1998.

[27] L. Breiman, Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, 24 (1996), 2350-2383.

[28] J.G. Brookshear, *Theory of Computation: Formal Languages, Automata, and Complexity*, The Benjamin-Cummings Publishing Company, Redwood City, California, 1989.

[29] M.C. Bueso, G. Qian and J.M. Angulo, Stochastic complexity and model selection from incomplete data, *Journal of Statistical Planning and Inference*, 76 (1999), 273-284.

[30] D.A. Bulutoglu and C.S. Cheng, Construction of $\mathrm{E}(s^2)$-optimal supersaturated designs, *The Annals of Statistics*, 32 (2004), 1662-1678.

[31] D.A. Bulutoglu, Cyclicly constructed $\mathrm{E}(s^2)$-optimal supersaturated designs, *Journal of Statistical Planning and Inference*, 137 (2007), 2413-2428.

[32] D.A. Bulutoglu and K.J. Ryan, $\mathrm{E}(s^2)$-optimal supersaturated designs with good minmax properties when N is odd, *Journal of Statistical Planning and Inference*, 138 (2008), 1754-1762.

[33] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2 (1998), 121-167.

[34] N. Butler, R. Mead, K.M. Eskridge and S.G. Gilmour, A general method of constructing E(s^2)-optimal supersaturated designs, *Journal of the Royal Statistical Society: Series B*, 63 (2001), 621-632.

[35] N. Butler, Minimax 16-run supersaturated designs, *Statistics and Probability Letters*, 73 (2005), 139-145.

[36] N. Butler, Two-level supersaturated designs for 2^k runs and other cases, *Journal of Statistical Planning and Inference*, 139 (2009), 23-29.

[37] C. Cachin, *Entropy Measures and Unconditional Security in Cryptography*, Swiss Federal Institute of Technology, Zurich, PhD Thesis, 1997.

[38] A.C. Cameron and P.K. Trivedi, *Regression Analysis of Count Data*, Cambridge University Press, New York, 1998.

[39] 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.

[40] R.G. Catalán, J Garay and R. López-Ruiz, Features of the extension of a statistical measure of complexity for continuous systems, *Physical Reviews E*, 66 (2002), 011102(6).

[41] A. Caticha, Information and entropy. In: K.H. Knuth (Eds.) *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Vol. 30, AIP, New York, 2007.

[42] S.B. Charles, A comparison of marginal likelihood computation methods, In: W. Härdle and B. Rönz (Eds.) *COMPSTAT 2002: Proceedings in Computational Statistics*, pp. 111-117, Springer-Verlag, Berlin, Heidelberg, 2002.

[43] K. Chatterjee, C. Koukouvinos and C. Parpoula, *Computer-aided unbalanced supersaturated designs involving interactions*, (submitted for publication).

[44] K. Chatterjee, C. Koukouvinos and C. Parpoula, On the analysis of unbalanced two-level supersaturated designs via generalized linear models, (submitted for publication).

[45] K. Chatterjee, C. Koukouvinos and C. Parpoula, A lower bound to $E(s^2)$-optimality measure for partially balanced two-level designs, (submitted for publication).

[46] J. Chen and D.K.J. Lin, On the identifiability of a supersaturated design, *Journal of Statistical Planning and Inference*, 72 (1998), 99-107.

[47] C.S. Cheng, E(s^2)-optimal supersaturated designs, *Statistica Sinica*, 7 (1997), 929-939.

[48] C.S. Cheng and B. Tang, Upper bounds on the number of columns in supersaturated designs, *Biometrika*, 88 (2001), 1169-1174.

[49] H.A. Chipman, Bayesian variable selection with related predictors, *The Canadian Journal of Statistics*, 24 (1996), 17-36.

[50] A. Christmann and S. Van Aelst, Robust estimation of Cronbach's alpha, *Journal of Multivariate Analysis*, 97 (2006), 1660-1674.

[51] J.M. Cortina, What is coefficient alpha? An examination of theory and applications, *Journal of Applied Psychology*, 78 (1993), 98-104.

[52] A. Cossari, Applying Box-Meyer method for analyzing supersaturated designs, *Quality Technology and Quantitative Management*, 5 (2008), 393-401.

[53] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, 2nd edn., Hoboken, New Jersey, John Wiley and Sons, 2006.

[54] D.R. Cox and D.V. Hinkley, *Theoretical Statistics*, Chapman and Hall, 1974.

[55] 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.

[56] L.J. Cronbach, Coefficient alpha and the internal structure of tests, *Psychometrika*, 16 (1951), 297-334.

[57] J.P. Crutchfield and K. Young, Inferring statistical complexity, *Physical Review Letters*, 63 (1989), 105-108.

[58] C. Czado and A.E. Raftery, Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors, *Statistical Papers*, 47 (2006), 419-442.

[59] Z. Daróczy, Generalized information functions, *Information and Control*, 16 (1970), 36-51.

[60] A. Das, A. Dey, L.Y. Chan and K. Chatterjee, $E(s^2)$-optimal supersaturated designs, *Journal of Statistical Planning and Inference*, 138 (2008), 3749-3757.

[61] M. Dash, H. Liu and H. Motoda, Consistency based feature selection, In: *Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining*, pp.98-109, Springer-Verlag, 2000.

[62] L. Davis, *Handbook of Genetic Algorithms*, Van Nostrand, Reinhold, 1991.

[63] A.M. Dean and S.M. Lewis, Comparison of group screening strategies for factorial experiments, *Computational Statistics and Data Analysis*, 39 (2002), 287-297.

[64] S. Della Pietra, V.J. Della Pietra and J.D. Lafferty, Inducing features of random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (1997), 380-393.

[65] E.R. DeLong, D. M. DeLong and D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics*, 44 (1988), 837-845.

[66] L.Y. Deng, D.J.K. Lin and J. Wang, Marginally oversaturated designs, *Communications in Statistics-Theory and Methods*, 25 (1996), 2557-2573.

[67] R. Dorfman, The detection of defective members of large populations, *Annals of Mathematical Science*, 14 (1943), 436-440.

[68] D. Draguljić, D.C. Woods, A.M. Dean, S.M. Lewis and A.E. Vine, Screening strategies in the presence of interactions, *Technometrics*, 56 (2014), 1-16.

[69] N.R. Draper and F. Pukelsheim, An overview of design of experiments, *Statistical Papers*, 37 (1996), 1-32.

[70] J. Dunn, Model complexity: the fit to random data reconsidered, *Psychological Research*, 63 (2000), 174-182.

[71] P. Economou and C. Caroni, *Statistical Regression Models*, Symeon Publications, Athens, 2010 (in Greek).

[72] D.J. Edwards and R.W. Mee, Supersaturated designs: are our results significant? *Computational Statistics and Data Analysis*, 55 (2011), 2652-2664.

[73] D. Erdman, L. Jackson and A. Sinko, Zero-inflated Poisson and zero-inflated negative binomial models using the COUNTREG Procedure, Paper 322-2008, *SAS Institute Inc.*, Cary, NC, 2008.

[74] K.M. Eskridge, S.G. Gilmour, R. Mead, N.A. Butler and D.A. Travnicek, Large supersaturated designs, *Journal of Statistical Computation and Simulation*, 74 (2004), 525-542.

[75] J. Fan, Comments on "Wavelets in statistics: a review" by A. Antoniadis, *Journal of the Italian Statistical Association*, 6 (1997), 131-138.

[76] 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.

[77] J. Fan and R. Li, Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics*, 30 (2002), 74-99.

[78] J. Fan and R. Li, Statistical challenges with high dimensionality: feature selection in knowledge discovery, In: M.Sanz-Sole et al. (Eds.) *Proceedings of the International Congress of Mathematicians*, Vol.III, pp.595-622, European Mathematical Society, *Zürich*, 2006.

[79] J. Fan, Y. Chen, H. M. Chan, P. Tam and Y.Ren, Removing intensity effects and identifying significant genes for Affymetrix arrays in MIF-suppressed neuroblastoma cells, In: *Proceedings of the National Academy of Sciences of the United States of America*, 103 (2005), 17751-17756.

[80] D.P. Feldman and J.P. Crutchfield, Measures of statistical complexity, *Physics Letters A*, 238 (1998), 244-252.

[81] F. Fleuret, Fast binary feature selection with conditional mutual information, *Journal of Machine Learning Research*, 5 (2004), 1531-1555.

[82] M.F. Franklin, Constructing tables of minimum aberration p^{n-m} designs, *Technometrics* 26 (1984), 225-232.

[83] A. Fries and W.G. Hunter, Minimum aberration $2^{\text{k-p}}$ designs, *Technometrics*, 22 (1980), 601-608.

[84] G. Fung and O.L. Mangasarian, A feature selection newton method for support vector machine classification, *Computational Optimization and Applications*, 28 (2004), 185-202.

[85] S. Furuichi, Information theoretical properties of Tsallis entropies, *Journal of Mathematical Physics*, 47 (2006), p.023302.

[86] A. Genkin, D.D. Lewis and D. Madigan, Large-scale Bayesian logistic regression for text categorization, *Technometrics*, 49 (2007), 291-304.

[87] S. Georgiou, C. Koukouvinos and P. Mantas, A new method for the construction of two-level E(s^2)-optimal supersaturated designs, *Journal of Statistical Theory and Applications*, 5 (2006), 403-415.

[88] S.D. Georgiou, On the construction of E(s^2)-optimal supersaturated designs, *Metrika*, 69 (2008), 189-198.

[89] S.D. Georgiou, Modelling by supersaturated designs, *Computational Statistics and Data Analysis*, 53 (2008), 428-435.

[90] S.D. Georgiou, D. Draguljic and A. Dean, An overview of two-level supersaturated designs with cyclic structure, *Journal of Statistical Theory and Practice*, 3 (2009), 489-504.

[91] S.D. Georgiou, Supersaturated designs: a review of their construction and analysis, *Journal of Statistical Planning and Inference*, 144 (2014), 92-109.

[92] S.G. Gilmour, *Supersaturated Designs in Factor Screening*, In: S.M. Lewis and A.M. Dean (Eds.) *Screening*, pp.169-190, Springer, NewYork, 2006.

[93] C. Gini, Variabilita e Mutabilita: contributo allo studio delle distribuzioni e relazioni stati-stiche, *Studi Economico-Giuridici dell'Univ. di Cagliari*, 3 (1912), 1-158.

[94] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA., 1989.

[95] L. Golshani, E. Pasha and G. Yari, Some properties of Rényi entropy and Rényi entropy rate, *Information Sciences*, 179 (2009), 2426-2433.

[96] P.D. Grünwald, MDL tutorial. In: P.D. Grünwald, I.J. Myung and M. A. Pitt (Eds.) *Advances in Minimum Description Length: Theory and Applications*, MIT Press, Boston, MA, 2005.

[97] P.D. Grünwald, *The Minimum Description Length Principle*, The MIT Press, Cambridge, MA, 2007.

[98] S. Gupta and P. Kohli, Analysis of supersaturated designs: a review, *Journal of Indian Society of Agricultural Statistics*, 62 (2008), 156-168.

[99] V.K. Gupta, R. Parsad, B. Kole and L. Bhar, Computer-generated efficient two-level supersaturated designs, *Journal of Indian Society of Agricultural Statistics*, 62 (2008), 183-194.

[100] V.K. Gupta, P. Singh, B. Kole and R. Parsad, Addition of runs to a two-level supersaturated design, *Journal of Statistical Planning and Inference*, 140 (2010), 2531-2535.

[101] I. Guyon and A. Elisseeff, An Introduction to variable and feature Selection, *Journal of Machine Learning Research*, 3 (2003), 1157-1182.

[102] P. Hall and J. Hannan, On stochastic complexity and nonparametric density estimation, *Biometrika*, 75 (1988), 705-714.

[103] M.A. Hall, *Correlation Based Feature Selection for Machine Learning*, Department of Computer Science, Waikato, Hamilton, New Zeland, PhD Thesis, 1999.

[104] M. Hamada and C.F.J. Wu, Analysis of designed experiments with complex aliasing, *Journal of Quality Technology*, 24 (1992), 130-137.

[105] J.A. Hanley and B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, 143 (1982), 29-36.

[106] A.J. Hansen and B. Yu, Model selection and the principle of minimum description length, *Journal of the American Statistical Association*, 96 (2001), 746-774.

[107] R.V.L. Hartley, Transmission of information, *Bell System Technical Journal*, 7 (1928), 535-563.

[108] J. Havrda and F. Charvát, Quantification method of classification processes: concept of structural α-entropy, *Kybernetika*, 3 (1967), 30-35.

[109] J.M. Hilbe, *Negative Binomial Regression*, Cambridge University Press, New York, 2008.

[110] R. Hocking, The analysis and selection of variables in linear regression, *Biometrics*, 32 (1976), 1-49.

[111] D. R. Holcomb, D. C. Montgomery and W. M. Carlyle, Analysis of supersaturated designs, *Journal of Quality Technology*, 35 (2003), 13-27.

[112] D. R. Holcomb, D. C. Montgomery and W. M. Carlyle, The use of supersaturated experiments in turbine engine development, *Quality Engineering*, 19 (2007), 17-27.

[113] C.S. Hong and B.J. Kim, Mutual information and redundancy for categorical data, *Statistical Papers*, 52 (2011), 17-31.

[114] J.E. Hopcroft, R. Motwani and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 3rd edn., Addison-Wesley, Reading, MA, 2000.

[115] D.-H. Jang, C.M. Anderson-Cook and Y. Kim, Correlation-based r-plot for evaluating supersaturated designs, *Quality Reliability Engineering International*, 30 (2014), 503-512.

[116] E.T. Jaynes, *Probability Theory-The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[117] R. Jenssen and T. Eltoft, A new information theoretic analysis of sum-of-squared-error kernel clustering, *Neurocomputing*, 72 (2008), 23-31.

[118] B.A. Jones, D.K.J. Lin and C.J.Nachtsheimc, Bayesian D-optimal supersaturated designs, *Journal of Statistical Planning and Inference*, 138 (2008), 86-92.

[119] B.A. Jones, W. Li, C.J. Nachtheim and K.Q. Ye, Model-robust supersaturated and partially supersaturated designs, *Journal of Statistical Planning and Inference*, 139 (2009), 45-53.

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

[121] A.I. Khuri, B. Mukherjee, B.K. Sinha and M. Ghosh, Design issues for generalized linear models: a review, *Statistical Science*, 21 (2006), 376-399.

[122] K. Kira and L. Rendell, The feature selection problem: traditional methods and a new algorithm, In: *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp.129-134, AAAI Press/The MIT Press, Menlo Park, 1992.

[123] B. Kole, J. Gangwani, V.K. Gupta and R. Parsad, Two level supersaturated designs:a review. *Journal of Statistical Theory and Practice*, 4 (2010), 598-608.

[124] C. Koukouvinos, *Linear Models and Designs*, N.T.U.A. Press, Athens, 2010 (in Greek).

[125] C. Koukouvinos, *Statistical Designs*, N.T.U.A. Press, Athens, 2010 (in Greek).

[126] C. Koukouvinos, P. Mantas and K. Mylona, A general construction of E(s^2)-optimal large supersaturated designs, *Metrika*, 68 (2008), 99-110.

[127] C. Koukouvinos, E. Massou, K. Mylona and C. Parpoula, Analyzing supersaturated designs with entropic measures, *Journal of Statistical Planning and Inference*, 141 (2011), 1307-1312.

[128] C. Koukouvinos and K. Mylona, A general construction of E(s^2)-optimal supersaturated designs via supplementary difference sets, *Metrika*, 70 (2009), 257-265.

[129] 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.

[130] C. Koukouvinos, K. Mylona and D.E. Simos, E(s^2)-optimal and minimax-optimal supersaturated designs via multi-objective simulated annealing, *Journal of Statistical Planning and Inference*, 138 (2008), 1639-1646.

[131] C. Koukouvinos, K. Mylona and D.E. Simos, A hybrid SAGA algorithm for the construction of E(s^2)-optimal cyclic supersaturated designs, *Journal of Statistical Planning and Inference*, 139 (2009), 478-485.

[132] C. Koukouvinos and A. Papaioannou, *Information and Coding Theory*, N.T.U.A. Press, Athens, 2008 (in Greek).

[133] C. Koukouvinos and C. Parpoula, Analyzing supersaturated designs by means of an information based criterion, *Communications in Statistics-Simulation and Computation*, 41 (2012), 44-57.

[134] C. Koukouvinos and C. Parpoula, Variable selection and computation of the prior probability of a model via ROC curves methodology, *Journal of Data Science*, 10 (2012), 653-672.

[135] C. Koukouvinos and C. Parpoula, A penalized wrapper method for screening main effects and interactions in supersaturated designs, *Quality and Reliability Engineering International*, (2014), DOI: 10.1002/qre.1679.

[136] C. Koukouvinos, C. Parpoula and D.E. Simos, Genetic algorithm and data mining techniques for design selection in databases, In: *Proceedings of the 2013 IEEE International Conference on Availability, Reliability and Security*, pp.743-746, Regensburg, 2013, DOI: 10.1109/ARES.2013.98.

[137] C. Koukouvinos and S. Stylianou, A method for analyzing supersaturated designs, *Communications in Statistics-Simulation and Computation*, 34 (2005), 929-937.

[138] G. Kuder and M. Richardson, The theory of estimation of test reliability, *Psychometrika*, 2 (1937), 151-160.

[139] N. Lachiche and P. Flach, Improving accuracy and cost of the two-class and multi-class probabilistic classifiers using ROC curves, In: T. Fawcett and N. Mishra (Eds.) *Proceedings of the 20th International Conference on Machine Learning*, pp. 416-423, AAAI Press, 2003.

[140] M.D. Lee, Generating additive clustering models with minimal stochastic complexity, *Journal of Classification*, 19 (2002), 69-85.

[141] S.M. Lewis and A.M. Dean, Detection of interactions in experiments on large numbers of factors (with discussion), *Journal of the Royal Statistical Society: Series B*, 63 (2001), 633-672.

[142] M. Li and P.M.B. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 1993.

[143] W.W. Li and C.F.J. Wu, Columnwise-pairwise algorithms with applications to the construction of supersaturated designs, *Technometrics*, 39 (1997), 171-179.

[144] R. Li and D.K.J. Lin, Data analysis in supersaturated designs, *Statistics and Probability Letters*, 59 (2002), 135-144.

[145] R. Li and D.K.J. Lin, Analysis methods for supersaturated designs: some comparisons, *Journal of Data Science*, 1 (2003), 249-260.

[146] R. Li, D.K.J. Lin and B. Li, Statistical inference in massive data sets, *Applied Stochastic Models in Business and Industry*, 29 (2013), 399-409

[147] P. Li, S. Zao and Z. Zhang, A cluster analysis selection strategy for supersaturated designs, *Computational Statistics and Data Analysis*, 54 (2010), 1605-1612.

[148] D.K.J. Lin, A new class of supersaturated designs, *Technometrics*, 35 (1993), 28-31.

[149] D.K.J. Lin, Generating systematic supersaturated designs, *Technometrics*, 37 (1995), 213-225.

[150] D.K.J. Lin, Spotlight interaction effects in main-effect plans: a supersaturated design approach, *Quality Engineering*, 11 (1998), 133-139.

[151] M. Liu and R. Zhang, Construction of E(s^2) optimal supersaturated designs using cyclic BIBDs, *Journal of Statistical Planning and Inference*, 91 (2000), 139-150.

[152] M.-Q. Liu and F.J. Hickernell, $E(s^2)$-optimality and minimum discrepancy in 2-level supersaturated designs, *Statistica Sinica*, 12 (2002), 931-939.

[153] Y.F. Liu and A. Dean, k-circulant supersaturated designs, *Technometrics*, 46 (2004), 32-43.

[154] Y.F. Liu, S. Ruan and A.M. Dean, Construction and analysis of Es^2 efficient supersaturated designs, *Journal of Statistical Planning and Inference*, 137 (2007), 1516-1529.

[155] Y.F. Liu, M.Q. Liu and R. Zhang Construction of multi-level supersaturated designs via Kronecker product, *Journal of Statistical Planning and Inference*, 137 (2007), 2984-2992.

[156] R. López-Ruiz, H.L. Mancini and X. Calbet, A statistical measure of complexity, *Physics Letters A*, 209 (1995), 321-326.

[157] X. Lu and Y. Meng, A new method in the construction of two-level supersaturated designs, *Journal of Statistical Planning and Inference*, 86 (2000), 229-238.

[158] X. Lu and X. Wu, A strategy of searching active factors in supersaturated screening experiments, *Journal of Quality Technology*, 36 (2004), 392-399.

[159] C. Marley, *Screening experiments using supersaturated designs with application to industry*, University of Southampton, Faculty of Social and Human Sciences School of Mathematics, PhD Thesis, 2010.

[160] 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.

[161] P. McCullagh and J.A. Nelder, *Generalized linear models*, 2nd edn., Chapman and Hall, London, 1989.

[162] A.J. Miller, *Subset Selection in Regression*, 2nd edn., Chapman and Hall/CRC, Boca Raton, 2002.

[163] D.C. Montgomery, *Design and Analysis of Experiments*, 5th edn., John Wiley and Sons, New York, 2004.

[164] D.C. Montgomery, E.A. Peck and G.G Vining, *Introduction to Linear Regression Analysis*, 4th edn., Wiley, Hoboken, 2006.

[165] M.A. Moore and H.H. Epps, Accelerated weathering of marine fabrics, *Journal of Testing and Evaluation*, 20 (1992), 139-143.

[166] M.D. Morris, An overview of group factor screening, In: A.M. Dean and S.M. Lewis (Eds.) *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, pp.191-206, New York, Springer, 2006.

[167] 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.

[168] I.J. Myung and M.A. Pitt, Applying Occam's razor in modeling cognition: A Bayesian approach, *Psychonomic Bulletin & Review*, 4 (1997), 79-95.

[169] I.J. Myung, The importance of complexity in model selection, *Journal of Mathematical Psychology.* 44 (2000), 190-204.

[170] I.J., Myung, V. Balasubramanian and M.A. Pitt, Counting probability distributions: differential geometry and model selection, In: *Proceedings of the National Academy of Sciences*, 97 (2000), 11170-11175.

[171] J. Nelder and R. Wedderburn, Generalized linear models, *Journal of the Royal Statistical Society: Series A*, 135 (1972), 370-384.

[172] N.K. Nguyen, An algorithmic approach to constructing supersaturated designs, *Technometrics*, 38 (1996), 69-73.

[173] N.K. Nguyen and C.S. Cheng, New E(s^2)-optimal supersaturated designs constructed from incomplete block designs, *Technometrics*, 50 (2008), 26-31.

[174] N. Niki, M. Iwata, H. Hashiguchi and S. Yamata, Optimal selection and ordering of columns in supersaturated designs, *Journal of Statistical Planning and Inference*, 141 (2011), 2449-2462.

[175] H. Nyquist, Certain factors affecting telegraph speed, *Bell System Technical Journal*, 3 (1924), 324-352.

[176] C. Parpoula, K. Drosou, C. Koukouvinos and K. Mylona, A new variable selection approach inspired by supersaturated designs given a large-dimensional dataset, *Journal of Data Science*, 12 (2014), 35-52.

[177] R.L. Pearson, Karl Pearson and the chi-squared test, *International Statistical Review*, 51 (1983), 59-72.

[178] 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.

[179] M.S. Pepe, Receiver operating characteristic methodology, *Journal of the American Statistical Association*, 95 (2000), 308-311.

[180] M.S. Pepe, An interpretation for ROC curve and inference using GLM procedures, *Biometrics*, 56 (2000), 352-359.

[181] M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Science Series, Oxford University Press, 2003.

[182] F.K.H. Phoa, Y.H. Pan and H. Xu, Analysis of supersaturated designs via Dantzing selector, *Journal of Statistical Planning and Inference*, 139 (2009), 2362-2372.

[183] F.K.H. Phoa, W.K. Wong and H. Xu, The need of considering the interactions in the analysis of screening designs, *Journal of Chemometrics*, 23 (2009), 545-553.

[184] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes in C*, Cambridge University Press, Cambridge, England, 1988.

[185] C. Pumplün, S. Rüping, K. Morik and C. Weihs, *D-optimal plans in observational studies*, Technical Report 44/2005, SFB 475, Complexity reduction in multivariate data structures, Technische Universität Dortmund, 44221 Dortmund, Germany, 2005, URL http://www.statistik.tu-dortmund.de/sfb-tr2005.html.

[186] C. Pumplün, C. Weihs and A. Preusser, Experimental design for variable selection in data bases, In: C. Weihs and W. Gaul (Eds.) *Classification-The Ubiquitous Challenge*, pp. 192-199, Springer, Berlin, 2005.

[187] J.R. Quinlan, Induction of decision trees, *Machine Learning*, 1 (1986), 81-106.

[188] C.R. Rao and Y. Wu, On model selection (with discussion), In: P. Lahiri (Eds.) *Institute of Mathematical Statistical Lecture Notes-Monograph Series*, 38 (2001), 1-64.

[189] A. Rényi, On measures of information and entropy, In: *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pp.547-561, 1961.

[190] J. Rissanen, A universal prior for integers and estimation by minimum description length, *The Annals of Statistics*, 11 (1983), 416-431.

[191] J. Rissanen, Stochastic complexity and modeling, *The Annals of Statistics*, 14 (1986), 1080-1100.

[192] J. Rissanen, Stochastic complexity (with discussion), *Journal of the Royal Statistical Society: Series B*, 49 (1987), 223-265.

[193] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific Publishing Company, Singapore, 1989.

[194] J. Rissanen, Fisher information and stochastic complexity, *IEEE Transactions on Information Theory*, 42 (1996), 40-47.

[195] J. Rissanen, Complexity and Information in Modeling, In: K. Velupillai (Eds.) *Computability, Complexity and Constructivity in Economic Analysis*, Chapter IV, Blackwell Publishings, Oxford, 2005.

[196] J. Rissanen, *Information and Complexity in Statistical Modeling*, Springer-Verlag, New York, 2007.

[197] J. Rissanen, *Optimal Estimation of Parameters*, Cambridge University Press, Cambridge, UK, 2012.

[198] S. Roman, *Introduction to Coding and Information Theory*, Springer-Verlag, Heidelberg, 1997.

[199] S. Rosset, Model selection via the AUC, In: *Proceedings of the 21st International Conference on Machine Learning*, p.89, Banff, Alberta, Canada, 2004.

[200] S. Rosset and J. Zhu, Piecewise linear regularized solution paths, *The Annals of Statistics*, 35 (2007), 1012-1030.

[201] S. Rüping and C. Weihs, *Kernelized design of experiments*, Technical Report, Sonderforschungsbereich 475, Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, (2009), No. 02, http://hdl.handle.net/10419/36602.

[202] 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.

[203] F.E. Satterthwaite, Random balance experimentation (with discussions), *Technometrics*, 1 (1959), 111-137.

[204] J. Schiffner, and C. Weihs, *D-optimal plans for variable selection in data bases*, Technical Report // Sonderforschungsbereich 475, Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, No.14, (2009), http://hdl.handle.net/10419/41052.

[205] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics*, 6 (1978), 461-464.

[206] P.R. Scinto, R.G. Wilkinson and D.K.J. Lin, Screening for fuel economy: a case study of supersaturated designs in practice, *Quality Engineering*, 23 (2011), 15-25.

[207] G. Seymour, *Predictive Inference: An Introduction*, 1st edn., Chapman and Hall, New York, 1993.

[208] C.E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, 27 (1948), 379-423, 623-656.

[209] B. Sierra, E. Lazkano, I. Inza, M. Merino, P. Y. Larraraga and J. Quiroga, Prototype selection and feature subset selection by estimation of distribution algorithms: a case study in the survival of cirrhotic patients treated with TIPS, In: A.L. Rector et al. (Eds.) *Lecture Notes in Artificial Intelligence*, Vol. 2101, pp.20-29, Springer, Berlin, 2001.

[210] P. Smyth and R.M. Goodman, An information theoretic approach to rule induction from databases, *IEEE Transactions on Knowledge and Data Engineering*, 4 (1992), 301-316.

[211] D.J. Spiegelhalter, N.G. Best, B.P. Carlin and A. Van der Linde, Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society: Series B*, 64 (2002), 583-639.

[212] N.J. Srivastava, Designs for searching for non-negligible effects, In: *A Survey of Statistical Designs and Linear Models*, pp.507-519, North-Holland, Amsterdam, 1975.

[213] C.S. Suen and A. Das, $E(s^2)$-optimal supersaturated designs with odd number of runs, *Journal of Statistical Planning and Inference*, 140 (2010), 1398-1409.

[214] N.M. Svrakic, O. Nesic, M.R.K. Dasu, D. Herndon and J.R. Perez-Polo, Statistical approach to DNA chip analysis, *Recent Progress in Hormone Research*, 58 (2003), 75-93.

[215] B. Tang and C.F.J. Wu, A method for constructing supersaturated designs and its Es^2-optimality, *Canadian Journal of Statistics*, 25 (1997), 191-201.

[216] R.J. Tibshirany, Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society: Series B*, 58 (1996), 267-288.

[217] C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics, *Journal of Statistical Physics*, 52 (1988), 479-487.

[218] A. Van der Linde, A Bayesian view of model complexity, *Statistica Neerlandica*, 66 (2012), 253-271.

[219] W. Vanpaemel, Measuring model complexity with the prior predictive, In: Y. Bengio (Eds.) *Advances in Neural Information Processing Systems (NIPS)*, Vol. 22, pp.1919-1927, Curran Associates, Inc., Red Hook, New York, 2009.

[220] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[221] A.E. Vine, S.M. Lewis and A.M. Dean, Two-stage group screening in the presence of noise factors and unequal probabilities of active effects, *Statistica Sinica*, 15 (2005), 871-888.

[222] A.E. Vine, S.M. Lewis, A.M. Dean and D. Brunson, A critical assessment of two-stage group screening through industrial experimentation, *Technometrics*, 50 (2008), 15-25.

[223] K.F., Wallis, *A Note on the Calculation of Entropy from Histograms*, Unpublished paper, University of Warwick, UK, 2006.

[224] L. Wang, J. Zhu and H. Zou, The doubly regularized support vector machine, *Statistica Sinica*, 16 (2006), 589-615.

[225] L. Wang, J. Zhu and H. Zou, Hybrid huberized support vector machines for microarray classification and gene selection, *Bioinformatics*, 24 (2008), 412-419.

[226] G.S. Watson, A study of the group screening method, *Technometrics*, 3 (1961), 371-388.

[227] 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.

[228] I.H. Witten and E. Frank, *Data Mining: Practical Machine learning Tools and Techniques with Java Implementations*, 2nd edn., Morgan Kaufmann Publishers, San Francisco, 2005.

[229] C.F.J. Wu, Construction of supersaturated designs through partially aliased interactions, *Biometrika*, 80 (1993), 661-669.

[230] C.F.J. Wu and M. Hamada, *Experiments: Planning, Analysis, and Optimization*, 2nd edn., John Wiley and Sons, New York, 2009.

[231] S. Yamada and D.K.J. Lin, Supersaturated designs including an orthogonal base, *Canadian Journal of Statistics*, 25 (1997), 203-213.

[232] S. Yamata, Selection of active factors by stepwise regression in the data analysis of supersaturated design, *Quality Engineering*, 16 (2004), 501-513.

[233] G.-B. Ye, Y. Chen and X. Xie, Efficient variable selection in support vector machines via the alternating direction method of multipliers, In: *Proceedings of the* 14th *International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*, Vol.15, pp.832-840, Fort Lauderdale, FL, USA, JMLR W&CP, 2011.

[234] L. Yu and H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, In: *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pp. 856-863, Washington DC, 2003.

[235] 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 (2006), 235-248.

[236] J. Zhu, S. Rosset, T.Hastie and R. Tibshirani, 1-norm support vector machines, In: *Advances in Neural Information Processing Systems 16, Proceedings of the 2003 Conference*, pp.1-8, Vancouver and Whistler, British Columbia, Canada, 2003.

[237] K. Zografos, Entropy and divergence measures for mixed variables, In: F. Vonta et al. (Eds.) *Statistical Models and Methods for Biomedical and Technical Systems*, Statistics for Industry and Technology, pp.519-534, Birkhauser, 2008.

[238] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B*, 67 (2005), 301-320.