ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ

ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ ΣΕ ΔΕΔΟΜΕΝΑ ΥΨΗΛΗΣ ΔΙΑΣΤΑΣΗΣ ΓΙΑ ΤΑ ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

ΠΡΙΣΚΑ ΕΥΑΓΓΕΛΙΑ-ΕΛΕΝΗ

Επιβλέπων: Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2014

ΠΕΡΙΛΗΨΗ

Υψηλής διάστασης στατιστικά προβλήματα προκύπτουν από διαφορετικούς τομείς της επιστημονικής έρευνας και τεχνολογικής ανάπτυξης. Η επιλογή μεταβλητών διαδραματίζει έναν κεντρικό ρόλο στη σύγχρονη στατιστική μάθηση και στις επιστημονικές ανακαλύψεις. Η περιγραφή των μεθόδων αντιμετώπισης ενός από τα πλέον σημαντικά προβλήματα της γραμμικής παλινδρόμησης: της επιλογής ενός υποσυνόλου από ανεξάρτητες μεταβλητές, έτσι ώστε αφενός μεν να υπάρχει εξοικονόμηση κόστους κατά την πρόβλεψη της εξαρτημένης μεταβλητής και αφετέρου να μην προκύπτει μεγάλη απώλεια στην αποτελεσματικότητα του μοντέλου πρόβλεψης. Σκοπός είναι η σωστή επιλογή των στατιστικά σημαντικών παραγόντων οι οποίοι επηρεάζουν την απόκριση. Θα παρουσιάσουμε ένα σύντομο απολογισμό της πρόσφατης εξέλιξης της θεωρίας, των μεθόδων και των εφαρμογών για την επιλογή μεταβλητών υψηλής διάστασης. Θα μελετήσουμε ποια είναι τα όρια της διάστασης που μπορούν να χειριστούν αυτές οι μέθοδοι, ποιος ο ρόλος των συναρτήσεων ποινής και ποιες είναι οι στατιστικές ιδιότητες που οδηγούν τις εξελίξεις του τομέα. Θα μελετήσουμε ακόμα ορισμένες πρόσφατες εξελίξεις στην υπερ-υψηλής διάστασης επιλογή μεταβλητών, με έμφαση στο ανεξάρτητο κρησάρισμα. Ένα πεδίο όπου εφαρμόζονται αρκετά αυτές οι μέθοδοι, και θα μελετήσουμε, είναι και τα γενικευμένα γραμμικά μοντέλα.

Στην παρούσα εργασία, αναλύεται ένα πλήθος μεθόδων, με εμβάθυνση στις νέες μεθόδους ποινικοποιημένης πιθανοφάνειας. Αυτές σε συνδυασμό με τα γενικευμένα γραμμικά μοντέλα εκτιμούν με ακρίβεια τους συντελεστές του μοντέλου και ταυτόχρονα επιλέγουν τις σημαντικές μεταβλητές.

Το πρώτο κεφάλαιο, αποτελεί μια εισαγωγή στο γενικό γραμμικό μοντέλο και στις βασικές τεχνικές εκτίμησης των παραμέτρων. Στη συνέχεια, αναφέρεται ένα πλήθος μεθόδων επιλογής μεταβλητών και παρουσιάζεται μια εκτενής ανάλυση και αξιολόγηση των νεών μεθόδων που βασίζονται στην εισαγωγή μιας συνάρτησης ποινής στην πιθανοφάνεια.

Το δεύτερο κεφάλαιο, πραγματεύεται τη μέθοδο σίγουρου κρησαρίσματος. Συγκεκριμένα, καταγράφονται οι βασικές έννοιες, οι επεκτάσεις του και οι έλεγχοι

του. Επίσης, παρουσιάζεται η διαδικασία που το μοντέλο αυτό συνδυάζεται με τις μεθόδους ποινικοποιημένης πιθανοφάνειας. Πολλές τεχνικές που συχνά χρησιμοποιούνται για την επιλογή μεταβλητών βασίζονται στο ανεξάρτητο κρησάρισμα. Η βασική ιδέα είναι να εφαρμόσουμε μια γρήγορη μέθοδο που θα κρησάρει τις μεταβλητές που έχουν ασθενή συσχέτιση με τη μεταβλητή απόκρισης. Η μέθοδος αυτή χρησιμοποιείται για να μειώσουμε τη διάσταση από υψηλή σε μια μέτριας κλίμακας που είναι μικρότερη από το μέγεθος του δείγματος.

Τέλος, στο τρίτο κεφάλαιο, παρουσιάζονται τα γενικευμένα γραμμικά μοντέλα, τα οποία είναι μια αξιοσημείωτη σύνθεση και επέκταση των γνωστών μοντέλων παλινδρόμησης, και πως αυτά συνδέονται με τις μεθόδους σίγουρου κρησαρίσματος. Γίνεται μια παρουσίαση της εκθετικής οικογένειας και αναφέρονται παραδείγματα προσομοίωσης αλλά και πραγματικών δεδομένων στα οποία φαίνεται η σημασία της μεθόδου του σίγουρου κρησαρίσματος.

ABSTRACT

High dimension statistical problems arising from different areas of scientific research and technological development. Variable selection plays a central role in modern statistical learning and scientific discoveries. The description of the methods, dealing with one of the most important problems, of linear regression: selecting a subset of independent variables so that on the one hand to save costs by predicting the dependent variable and on the other hand does not arise large loss in efficiency of the prediction model. The purpose is the right choice of statistically significant factors influencing the response. We will present a brief account of recent developments in the theory, methods and applications for the variable selection in high dimensional problems. We will study the limits of dimension, which can handle these methods, the role of penalty functions and the statistical properties leading to developments in the sector. We will consider even some recent developments in high dimension variables selection, with emphasis on sure independent screening. A field that several of these methods are applied is the generalized linear models. In this paper, we will analyze a number of methods, and provide insight on new methods penalized likelihood. These, combined with the generalized linear models accurately assess the factors of the model and simultaneously select the significant variables.

The first chapter is an introduction to the general linear model and to the basic parameter estimation techniques. Afterwards, many variable selection methods are mentioned and a thorough analysis and evaluation of the new methods, which are based in the introduction of a penalty in the likelihood function, is presented.

The second chapter deals with the sure independence screening. Specifically, the basic concepts are cited and its extensions. Furthermore, the procedure of the combination of this model with the penalized likelihood methods is presented. Many techniques are often used to select the independent variables based on sure independent screening. The basic idea is to implement a fast method to screen the variables that have a weak correlation with the response variable. This method is used to reduce the dimension of a moderate to high scale that is less than the sample size.

Finally, in the third chapter, the generalized linear models are presented, which are a remarkable synthesis and extension of known regression models, and how they are related sure independent screening. There is a presentation of the exponential family and we give examples of simulation and real data which show the importance of the method of Sure independent screening .

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών «Εφαρμοσμένες Μαθηματικές Επιστήμες», της Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου, υπό την επίβλεψη του Καθηγητή κ. Χρήστου Κουκουβίνου. Θα ήθελα λοιπόν να τον ευχαριστήσω θερμά για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο που ανταποκρίνεται στα επιστημονικά μου ενδιαφέροντα καθώς και για την αμέριστη συμπαράσταση και καθοδήγησή του καθ' όλη τη διάρκεια εκπόνησης.

Επίσης, ευχαριστώ τα υπόλοιπα μέλη της τριμελούς επιτροπής, τους για την αποτελεσματική συνεργασία και συμβολή τους στην ολοκλήρωση της παρούσας εργασίας.

Ευχαριστίες οφείλω και στον υποψήφιο διδάκτωρα Ανδρουλάκη Εμμανουήλ, για την πολύτιμη βοήθεια, τη συνεργασία που είχαμε, τις εύστοχες και σημαντικές συμβουλές καθόλη τη διάρκεια της εκπόνησης της εργασίας αυτής.

Ιδιαίτερες ευχαριστίες θέλω να εκφράσω προς την οικογένειά μου και κυρίως τους γονείς μου Χρήστο και Νίτσα για την ολόψυχη αγάπη και υποστήριξή τους όλα αυτά τα χρόνια. Αφιερώνω αυτή την εργασία στη μητέρα μου και στον πατέρα μου.

Αθήνα, Φεβρουάριος 2014

Πρίσκα Ευαγγελία- Ελένη

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1^{ο} 11
ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ 11
1.1. Εισαγωγή 11
1.1.1. Επιλογή μεταβλητών 13
1.2. Το γενικό γραμμικό μοντέλο παλινδρόμησης 15
1.2.1. Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο ελαχίστων τετραγώνων 16
1.2.2. Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο μέγιστης πιθανοφάνειας 18
1.3. Παλινδρόμηση κορυφογραμμής/διασέλου 19
1.3.1. Περιγραφή της μεθόδου 19
1.3.2 Παλινδρόμηση κορυφογραμμής/διασέλου και επιλογή μεταβλητών 26
1.4. Η διαδικασία PRESS 27
1.5. Μέθοδος επιλογής καλύτερου υποσυνόλου 27
1.5.1. Στατιστική συνάρτηση C_p – Mallows 30
1.5.2. Κριτήρια βασισμένα στην πιθανοφάνεια ή στην πληροφορία 33
1.6. Ποινικοποιημένα ελάχιστα τετράγωνα και ποινικοποιημένη πιθανοφάνεια 37
1.6.1 Εισαγωγή 37
1.6.2. Επιλογή μεταβλητών μέσω ποινικοποιημένων ελαχίστων τετραγώνων 39
1.6.3. Επιλογή μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας 46
1.6.4. Αριθμητικές συγκρίσεις 59
1.7. Μια επεξήγηση της διακλιμάκωσης 62
1.8. Μέθοδος LASSO 62
1.8.1. Περιγραφή της μεθόδου LASSO 63
1.8.2. Η περίπτωση του ορθοκανονικού πίνακα σχεδιασμού 64
1.8.3. Η γεωμετρία της LASSO 65
1.8.4. Τυπικά σφάλματα και εκτίμηση της παραμέτρου t. 67
1.8.5. Adaptive LASSO 68
1.9. Dantzig selector 69
1.10. Συμπεράσματα 72

ΚΕΦΑΛΑΙΟ 2ο 73
ΣΙΓΟΥΡΟ ΑΝΕΞΑΡΤΗΤΟ ΚΡΗΣΑΡΙΣΜΑ 73
2.1. Εισαγωγή 73
2.2. Μέθοδος Σίγουρου Κρησαρίσματος: μάθηση συσχέτισης 78
2.3. Σύνδεση με άλλες μεθόδους μείωσης διάστασης 82
2.4. Τεχνικές επιλογής μοντέλου βασισμένες στη SIS 83
2.5. Αριθμητικές μελέτες 84
2.5.1. Ανάλυση δεδομένων Λευχαιμίας 84
2.6. Κατάταξη Χαρακτηριστικού από τις Οριακές Χρησιμότητες. 86
2.7. Ποινικοποιημένη Ψευδοπιθανοφάνεια 87
2.8. Ομαδοποίηση και μετασχηματισμός των μεταβλητών εισόδου. 89
2.9. Μερικές επεκτάσεις της μάθησης συσχέτισης 91
2.9.1. Επαναληπτική SIS – Επαναληπτική μάθηση συσχέτισης 92
2.9.2. Επαναληπτική Επιλογή Χαρακτηριστικών 93
2.10. Μείωση των Ποσοστών Εσφαλμένης Επιλογής 96
2.10.1. Η πρώτη παραλλαγή της ISIS 97
2.10.2. Δεύτερη παραλλαγή της ISIS 101
2.11. Αριθμητικά στοιχεία 103
2.11.1. Προσομοιωμένο παράδειγμα 104
2.12. Συμπερασματικά σχόλια 105

ΚΕΦΑΛΑΙΟ 3ο 106
ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ 106
3.1. Εισαγωγή 106
3.2. Τυπική δομή για τα γενικευμένα γραμμικά μοντέλα 107
3.3. Στατιστική θεωρία για τα Γενικευμένα Γραμμικά Μοντέλα 112
3.3.1. Εκθετικές Οικογένειες 112
3.4. SIS στα γενικευμένα γραμμικά μοντέλα 118
3.4.1. Γενικευμένα Γραμμικά Μοντέλα 120
3.4.2. Ανεξάρτητο κρησάρισμα με MMLE 121
3.5. Ένα εκθετικό φράγμα της QMLE 122
3.6. Ιδιότητες Σίγουρου Κρησαρίσματος με MMLE 124
3.7. Ομοιόμορφη σύγκλιση και σίγουρο κρησάρισμα 128

3.8. Έλεγχος κατάταξης επιλογής σφαλμάτων. 131
3.9. Κρησάρισμα λόγου πιθανοφάνειας 133
3.10. Τελικές παρατηρήσεις 136
3.11. Αριθμητικά αποτελέσματα. 138

ΠΑΡΑΡΤΗΜΑ 155
Λογιστικές παλινδρομήσεις. 158
Γραμμικά μοντέλα 162

ΒΙΒΛΙΟΓΡΑΦΙΑ 166

ΚΕΦΑΛΑΙΟ 1ο

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

1.1. Εισαγωγή

Η ανάλυση παλινδρόμησης (*regression analysis*) εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας μέσω των τιμών της άλλης (ή των άλλων). Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε δύο είδη μεταβλητών: τις ανεξάρτητες ή επεξηγηματικές (independent variables / predictor) και τις εξαρτημένες ή απόκρισης (dependent / response variables). Σε πειραματικές έρευνες, ανεξάρτητη μεταβλητή X είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή, να καθορίσουμε τις τιμές της (η ονομασία ανεξάρτητη μεταβλητή παρόλο που χρησιμοποιείται συχνά δεν απεικονίζει ακριβώς την πραγματικότητα, επειδή στην πράξη οι μεταβλητές πρόβλεψης είναι σπάνια ανεξάρτητες μεταξύ τους). Εξαρτημένη μεταβλητή Y είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες μεταβλητές. Άρα, θεωρούμε ότι η ανάλυση παλινδρόμησης είναι μια τεχνική, η οποία εξετάζει την αλληλοεξάρτηση μεταξύ των μεταβλητών του συγκεκριμένου συνόλου δεδομένων. Προσδιορίζοντας τη σχέση των μεταβλητών λέμε ότι έχουμε προσδιορίσει ένα μοντέλο.

Ένα πολύ κοινό στατιστικό πρόβλημα είναι να μοντελοποιηθεί η σχέση μεταξύ ενός ή περισσοτέρων μεταβλητών εξόδου Y και των συναφών συμμεταβλητών (χαρακτηριστικών) $X_1,...,X_p$ που βασίζονται στο δείγμα μεγέθους n. Ένα χαρακτηριστικό γνώρισμα πολλών από τα σύγχρονα προβλήματα είναι ότι η διάσταση p είναι μεγάλη, δυνητικά πολύ μεγαλύτερη από το n και θεωρείται μαθηματικά ως μια συνάρτηση n, που αποκλίνει στο άπειρο. Η διάσταση μπορεί να είναι πολύ μεγαλύτερη από το μέγεθος του δείγματος. Όταν η διάσταση p είναι υπερβολικά υψηλή, συχνά θεωρείται ότι μόνο ένας μικρός αριθμός μεταβλητών μεταξύ των επεξηγηματικών μεταβλητών $\mathrm{X}_1,...,\mathrm{X}_p$ συμβάλλουν στην απόκριση, που οδηγεί στη σποραδικότητα του διανύσματος παραμέτρων β. Κατά συνέπεια, η

επιλογή μεταβλητών διαδραματίζει εξέχοντα ρόλο στη στατιστική μοντελοποίηση μεγάλων διαστάσεων.

Η ανάλυση δεδομένων υψηλής διάστασης γίνεται όλο και πιο συχνή και σημαντική σε διάφορους τομείς των επιστημών, της μηχανικής και των ανθρωπιστικών επιστημών, που κυμαίνονται από τη γονιδιωματική και τις επιστήμες υγείας στην οικονομία και τη μηχανική μάθηση. Χαρακτηρίζει πολλά σύγχρονα προβλήματα στη στατιστική (Hastie, Tibshirani and Friedman, 2009). Για παράδειγμα, στην ταξινόμηση ασθενειών με τη χρήση δεδομένων μικροσυστοιχιών γονιδιακής έκφρασης (Tibshirani et al. (2003), Fan and Ren (2006)), ο αριθμός των συστοιχιών είναι συνήθως της τάξης των δεκάδων ενώ ο αριθμός των προφίλ γονιδιακής έκφρασης είναι της τάξης των δεκάδων χιλιάδων. Στη μελέτη πρωτεϊνών-αλληλεπιδράσεις πρωτεϊνών, ο αριθμός των χαρακτηριστικών μπορεί να είναι της τάξης των εκατομυρρίων, ενώ το μέγεθος του δείγματος να είναι της τάξης των χιλιάδων. Άλλα παραδείγματα δεδομένων υψηλής διάστασης περιλαμβάνουν υψηλής ανάλυσης εικόνες, οικονομικά στοιχεία υψηλής συχνότητας, δεδομένα ηλεκτρονικού εμπορίου, δεδομένα αποθήκης, λειτουργικά δεδομένα και άλλα πολλά. Ο Donoho (2000) καταδεικνύει πιεστικά την ανάγκη για εξελίξεις στην ανάλυση δεδομένων υψηλής διάστασης και παρουσιάζει τα πλεονεκτήματα και τα μειονεκτήματα της υψηλής διάστασης. Οι Fan και Li (2006) έδωσαν μια περιεκτική εικόνα των στατιστικών προκλήσεων με την υψηλή διάσταση σε ένα ευρύ φάσμα θεμάτων, και ειδικότερα απέδειξαν ότι για την οργάνωση των στατιστικών προβλημάτων, οι παράμετροι του μοντέλου μπορούν να εκτιμηθούν εξίσου αν το καλύτερο μοντέλο είναι γνωστό εκ των προτέρων, όσο η διάσταση δεν είναι υπερβολικά υψηλή.

Η στατιστική ακρίβεια, η επεξηγηματικότητα του μοντέλου και η υπολογιστική πολυπλοκότητα είναι τρεις σημαντικοί παράγοντες οποιασδήποτε στατιστικής διαδικασίας. Σε συμβατικές μελέτες, ο αριθμός των παρατηρήσεων n είναι πολύ μεγαλύτερος από τον αριθμό των μεταβλητών ή των παραμέτρων p. Σε αυτές τις περιπτώσεις, κανένα από τα τρία στοιχεία δε χρειάζεται να θυσιαστεί για την αποτελεσματικότητα των άλλων. Οι παραδοσιακές μέθοδοι, παρόλα αυτά, αντιμετωπίζουν σημαντικές προκλήσεις όταν η διάσταση p είναι συγκρίσιμα μεγαλύτερη από το μέγεθος του δείγματος n. Αυτές οι προκλήσεις περιλαμβάνουν

πώς να σχεδιάσουμε στατιστικές διαδικασίες που είναι πιο αποτελεσματικές, υπολογιστικά αποδοτικές και ισχυρές.

Οι κύριοι στόχοι της υψηλής διάστασης παλινδρόμησης και ταξινόμησης, σύμφωνα με τον Bickel (2008), είναι:

- Η κατασκευή μιας (όσο το δυνατό) αποτελεσματικής μεθόδου για την πρόβλεψη μελλοντικών παρατηρήσεων
- Να αποκτήσει μια εικόνα για τη σχέση μεταξύ των χαρακτηριστικών και της απόκρισης για επιστημονικούς σκοπούς, και ενδεχομένως να κατασκευάσει μια πιο βελτιωμένη μέθοδο πρόβλεψης.

1.1.1. Επιλογή μεταβλητών

Το πρόβλημα της επιλογής μεταβλητών έχει πολλές πρακτικές εφαρμογές σε διάφορους τομείς, όπως οι κοινωνικές επιστήμες, η οικονομία, τα μετεωρολογικά φαινόμενα και πολλές άλλες. Ανά διαστήματα έχουν δημοσιευτεί αρκετά άρθρα σχετικά με την ανάλυση παλινδρόμησης αλλά και για το πρόβλημα της επιλογής μεταβλητών. Η επιλογή μεταβλητών διαδραματίζει έναν κεντρικό ρόλο στη σύγχρονη στατιστική μάθηση και στις επιστημονικές ανακαλύψεις, αφού έχει ως σκοπό να εντοπίσει όλες τις σημαντικές μεταβλητές των οποίων οι συντελεστές παλινδρόμησης δεν εξαφανίζονται και να παρέχει αποτελεσματικές εκτιμήσεις για αυτούς τους συντελεστές. Μια σημαντική προσπάθεια έγινε από την ομάδα Narula et al. το 2003, οι οποίοι παρουσίασαν μια σειρά από διαδικασίες οι οποίες παρουσιάζουν τη δυνατότητα να βρίσκουν μοντέλα χρησιμοποιώντας όσο λιγότερες μεταβλητές είναι δυνατόν και κάποια κριτήρια ώστε να είναι δυνατή η επιλογή του μοντέλου. Ένα άλλο σημαντικό άρθρο είναι του Bruce Ratner (2003) το οποίο μελέτησε τις μεθόδους επιλογής μεταβλητών στην παλινδρόμηση.

Σκοπός δηλαδή είναι να επιλεγούν οι παράγοντες που έχουν σημαντική επίδραση στην απόκριση Y. Με αυτόν τον τρόπο εξηγούμε τα δεδομένα με τον απλούστερο τρόπο, διότι οι περιττές προβλέψεις αφαιρούνται. Η αρχή της μετριοπάθειας (*Law of parsimony*) ορίζει ότι μεταξύ των διαφόρων πιθανών

εξηγήσεων για μια περίπτωση, η απλούστερη είναι και η καλύτερη. Με άλλα λόγια, χρησιμοποιούμε τον ελάχιστο αριθμό επεξηγηματικών μεταβλητών, οι οποίες περιγράφουν επαρκώς τα δεδομένα μας. Η εφαρμογή της στην ανάλυση παλινδρόμησης αποδεικνύει ότι το μικρότερο μοντέλο που ταιριάζει με τα δεδομένα είναι και το καλύτερο. Επιπροσθέτως, οι περιττοί προγνωστικοί παράγοντες θα προσθέσουν περισσότερα εμπόδια στον προσδιορισμό άλλων ποσοτήτων, οι οποίες μας ενδιαφέρουν. Αυτό θα έχει ως αποτέλεσμα να αυξηθούν οι βαθμοί ελευθερίας. Η συγγραμμικότητα (*collinearity*) δημιουργείται όταν έχουμε πολλές μεταβλητές και προσπαθούμε να κάνουμε την ίδια δουλειά. Εάν το μοντέλο πρόκειται να χρησιμοποιηθεί για πρόβλεψη μεταβλητών, μπορούμε να γλιτώσουμε χρόνο ή / και χρήμα όταν δε λαμβάνουμε υπόψην μας τις περιττές επεξηγηματικές μεταβλητές. Από την άλλη πλευρά, με τη μέθοδο αυτή καθίσταται δυνατός ο προσδιορισμός των ακραίων τιμών και των καθοριστικών σημείων (δηλαδή των σημείων που επηρεάζουν περισσότερο) ή τουλάχιστον να αποκλείσουμε κάποια προσωρινά. Επίσης αυτή η μέθοδος μας δίνει τη δυνατότητα να προσθέσουμε οποιοδήποτε μετασχηματισμό μεταβλητών ο οποίος φαίνεται να είναι κατάλληλος για τη συγκεκριμένη περίπτωση.

Το κυριότερο κριτήριο της επιλογής υποσυνόλου είναι το κριτήριο της μετριοπάθειας (*parsimony*): εάν τρεις μεταβλητές μπορούν να επεξηγήσουν ή να ικανοποιήσουν τη μεταβλητή *Y*, γιατί να χρησιμοποιήσουμε τέσσερις; (Σύμφωνα με τις σημειώσεις του Mandel (1989)).

Όταν λαμβάνουμε υπόψη μόνο τους στατιστικά σημαντικούς παράγοντες, ο αριθμός των οποίων είναι αρκετά μικρότερος σε σχέση με το αρχικό σύνολό τους (μια ιδιότητα γνωστή ως «αρχή της σποραδικότητας των επιδράσεων» (*sparsity-of-effects principle*) (Bickel, 1975), γλιτώνουμε χρόνο και χρήμα κατά τη διάρκεια μιας μελέτης. Έχουν προταθεί αρκετές μέθοδοι επιλογής μεταβλητών, εκ των οποίων οι περισσότερες αποτελούν αναπόσπαστο κομμάτι στατιστικών πακέτων. Η χρήση τους γίνεται ολοένα και πιο απαραίτητη, διότι το μέγεθος των δεδομένων που προκύπτουν έπειτα από διάφορες μελέτες, συνεχώς μεγαλώνει. Παρά το γεγονός ότι ο αριθμός των μεθόδων αυτών είναι μεγάλος, το πεδίο της επιλογής μεταβλητών βρίσκεται ακόμα υπό έρευνα και συνεχώς προτείνονται νεότερες και πιο βελτιωμένες μέθοδοι.

Το πρόβλημα της επιλογής μεταβλητών είναι αρκετά σύνηθες στο πλαίσιο των γραμμικών μοντέλων παλινδρόμησης (*linear regression models*) και των

γενικευμένων γραμμικών μοντέλων (*generalized linear models*) (McCullagh & Nelder, 1989) που θα μελετήσουμε διεξοδικά στο τρίτο κεφάλαιο.

1.2. Το γενικό γραμμικό μοντέλο παλινδρόμησης

Αρκετές φορές συναντάμε προβλήματα, για τα οποία υπάρχει η υποψία ότι οι τιμές κάποιας μεταβλητής εξαρτώνται από $k \geq 2$ επεξηγηματικές μεταβλητές. Το γενικό γραμμικό μοντέλο, το οποίο περιγράφει αυτή τη σχέση είναι

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i,\ i = 1, 2, ..., n \qquad (1.2.1)$$

όπου y_i είναι οι τιμές της απόκρισης, x_{ij} είναι οι τιμές των επεξηγηματικών μεταβλητών (υποθέτουμε, όπως και στο απλό γραμμικό μοντέλο, ότι οι μετρήσεις μας δεν υπόκεινται σε σφάλματα), β_j είναι οι άγνωστες παράμετροι του μοντέλου οι οποίες και πρέπει να εκτιμηθούν, ε_i είναι τα σφάλματα ή τα υπόλοιπα, τα οποία αποτελούν τυχαίες μεταβλητές, αντιπροσωπεύουν την απόκλιση από την προσεγγιστική τιμή και υποθέτουμε ότι ικανοποιούν τα παρακάτω:

- $E(\varepsilon_i) = 0$ **,** $\forall i$
- $Var(\varepsilon_i) = \sigma^2$, δηλαδή τα σφάλματα ικανοποιούν την υπόθεση της ομοιοσκεδαστικότητας.
- $Cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$, δηλαδή τα σφάλματα είναι ασυσχέτιστα.

Από τη σχέση (1.2.1) φαίνεται ότι η απόκριση y_i (response) είναι μια γραμμική συνάρτηση των συντελεστών παλινδρόμησης (regression coefficients) β_j, με $j = 1, 2, ..., k$.

Το παραπάνω μοντέλο ονομάζεται γραμμικό μοντέλο και είναι το μοντέλο παλινδρόμησης που χρησιμοποιείται περισσότερο. Ένα γραμμικό μοντέλο ονομάζεται *απλό γραμμικό μοντέλο* όταν έχει μια μόνο επεξηγηματική μεταβλητή, ενώ όταν έχει περισσότερες ονομάζεται *γενικό ή πολλαπλό γραμμικό μοντέλο*.

Ένα από τα κύρια ζητήματα στο γραμμικό μοντέλο είναι να καθορίσουμε το σύνολο των επεξηγηματικών μεταβλητών που θα εισαχθούν στο μοντέλο έτσι ώστε να εξηγήσουμε καλύτερα τις αλλαγές στην κατανομή της μεταβλητής απόκρισης. Το πρόβλημα αυτό αναφέρεται ως **πρόβλημα επιλογής μεταβλητών**.

Η εκτίμηση των παραμέτρων του μοντέλου, δηλαδή των β_j, με $j = 1, 2, ..., k$ μπορεί να γίνει με δύο μεθόδους:

- Τη μέθοδο των ελαχίστων τετραγώνων (*least squares method*).
- Τη μέθοδο μέγιστης πιθανοφάνειας (*maximum likelihood method*).

1.2.1. Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο ελαχίστων τετραγώνων

Η παλιότερη μορφή της γραμμικής παλινδρόμησης είναι η μέθοδος των ελαχίστων τετραγώνων, η οποία δημοσιεύθηκε από τον Legendre (1805) και από τον Gauss (1809). Ο όρος «ελάχιστα τετράγωνα» προέκυψε από τον Legendre. Ο Legendre και ο Gauss εφάρμοσαν αυτή τη μέθοδο για να προσδιορίσουν από τις αστρονομικές προβλέψεις τις τροχιές των σωμάτων σε σχέση με τον ήλιο. Ο Euler εργάστηκε στο ίδιο πρόβλημα το 1748 χωρίς όμως να έχει επιτυχία. Ο Gauss δημοσίευσε μια πιο βελτιωμένη θεωρία για τη μέθοδο των ελαχίστων τετραγώνων το 1821, συμπεριλαμβάνοντας μια έκδοση η οποία είναι πολύ γνωστή σήμερα ως θεώρημα του Gauu-Markov. Το θεώρημα αυτό, αποτελεί θεμελιώδες θεώρημα στον τομέα των γραμμικών μοντέλων.

Η μέθοδος των ελαχίστων τετραγώνων μπορεί να χρησιμοποιηθεί για να εκτιμήσεις τις παραμέτρους σε ένα μοντέλο γραμμικής παλινδρόμησης ανεξάρτητα από το είδος της κατανομής της μεταβλητής απόκρισης.

Έστω ότι $n>k$. Για να εκτιμηθούν οι παράμετροι β_j του μοντέλου, χρησιμοποιείται η μέθοδος ελαχίστων τετραγώνων (*least squares method*). Αυτή η μέθοδος συνίσταται στην ελαχιστοποίηση του αθροίσματος τετραγώνων των σφαλμάτων.

$$S(\beta_0,\beta_1,...,\beta_k)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{k}\beta_j x_{ij}\right)^2,$$

οπότε τελικά προκύπτουν οι εκτιμητές $\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k$. Είναι προτιμότερο να γράψουμε των εξίσωση (1.2.1) υπό την μορφή πινάκων, δηλαδή

$$\underset{\sim}{Y}=\underset{\sim}{X}\underset{\sim}{\beta}+\underset{\sim}{\varepsilon},$$

όπου

$$\underset{\sim}{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{\sim}{X}=\begin{pmatrix} 1 & x_{11}\cdots & x_{1k} \\ 1 & x_{21}\cdots & x_{2k} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1}\cdots & x_{nk} \end{pmatrix} \quad \underset{\sim}{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \underset{\sim}{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Το $\underset{\sim}{Y}$ είναι ένα $n\times 1$ διάνυσμα των παρατηρήσεων, ο $\underset{\sim}{X}$ είναι ένας $n\times k$ πίνακας επεξηγηματικών μεταβλητών, το $\underset{\sim}{\beta}$ είναι ένα $k\times 1$ διάνυσμα συντελεστών παλινδρόμησης και το $\underset{\sim}{\varepsilon}$ είναι ένα $n\times 1$ διάνυσμα τυχαίων σφαλμάτων. Συνεπώς, για να βρεθεί η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\underset{\sim}{\beta}}$, πρέπει να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των σφαλμάτων

$$S(\underset{\sim}{\beta})=\sum_{i=1}^{n}\varepsilon_i^2=\varepsilon'\varepsilon=(\underset{\sim}{Y}-\underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{Y}-\underset{\sim}{X}\underset{\sim}{\beta}).$$

Έτσι, προκύπτει ότι

$$\hat{\underset{\sim}{\beta}}=(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}.$$

Αναφέρουμε και κάποιες ιδιότητες της εκτιμήτριας ελαχίστων τετραγώνων.

Καταρχήν,

$$\begin{aligned} E(\hat{\underset{\sim}{\beta}}) &= E\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}\right] \\ &= E\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'(\underset{\sim}{X}\underset{\sim}{\beta}+\underset{\sim}{\varepsilon})\right] \\ &= E\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta}+(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{\varepsilon}\right]=\underset{\sim}{\beta}, \end{aligned}$$

καθότι

$$E(\underset{\sim}{\varepsilon})=\underset{\sim}{0}$$

και

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{I}.$$

Άρα το $\hat{\underset{\sim}{\beta}}$ αποτελεί αμερόληπτη εκτιμήτρια (*unbiased estimator*) του $\underset{\sim}{\beta}$. Επίσης έχουμε ότι

$$\begin{aligned} Var(\hat{\underset{\sim}{\beta}}) &= Var\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}\right] \\ &= \left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\right]Var(\underset{\sim}{Y})\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\right]' \\ &= \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}. \end{aligned}$$

1.2.2. Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο μέγιστης πιθανοφάνειας

Αν στις βασικές υποθέσεις του απλού γραμμικού μοντέλου προσθέσουμε και το γεγονός ότι τα σφάλματα είναι κανονικά κατανεμημένα, τότε δεν είναι μόνο ασυσχέτιστα αλλά κατ΄ανάγκη είναι και ανεξάρτητα. Χρησιμοποιώντας διανύσματα, γράφουμε $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2\underset{\sim}{\mathrm{I}})$, δηλαδή το $\underset{\sim}{\varepsilon}$ ακολουθεί τη *n*-διάστατη πολυμεταβλητή κανονική κατανομή με $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ και $Var(\underset{\sim}{\varepsilon}) = \sigma^2\underset{\sim}{\mathrm{I}}$. Τότε η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\underset{\sim}{\beta}}$, ταυτίζεται με την εκτιμήτρια μέγιστης πιθανοφάνειας. Όσον αφορά την τελευταία μέθοδο, ισχύουν τα παρακάτω:

Η **μέθοδος μέγιστης πιθανοφάνειας**, προτάθηκε από τον R.A. Fisher (Aldrich, 1997). Συγκεκριμένα, υποθέτουμε ότι ένας πληθυσμός έχει άγνωστη παράμετρο $\underset{\sim}{\theta} = (\theta_1, \theta_2, ..., \theta_k) \in \Theta$ και η συνάρτηση πυκνότητας πιθανότητας είναι $f(\underset{\sim}{x} \mid \underset{\sim}{\theta})$ κι εμείς θέλουμε να εκτιμήσουμε την παράμετρο $\underset{\sim}{\theta}$.

Θεωρούμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ από τον πληθυσμό. Αν $f(x_1 \mid \underset{\sim}{\theta}), f(x_2 \mid \underset{\sim}{\theta}), ..., f(x_n \mid \underset{\sim}{\theta})$ είναι η συνάρτηση πυκνότητας πιθανότητας κάθε τιμής του τυχαίου δείγματος, τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας των μεταβλητών $X_1, X_2, ..., X_n$ είναι

$$f(x_1, x_2, ..., x_n \mid \underset{\sim}{\theta}) = f(x_1 \mid \underset{\sim}{\theta}) f(x_2 \mid \underset{\sim}{\theta}) \cdots f(x_n \mid \underset{\sim}{\theta}) \qquad (1.2.2.1).$$

Στην περίπτωση συγκεκριμένων παρατηρήσεων $x_1, x_2, ..., x_n$ τυχαίου δείγματος, η (1.2.2.1) είναι συνάρτηση μόνο της παραμέτρου $\underset{\sim}{\theta}$ και συμβολίζεται ως εξής:

$$L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n) = f(x_1 \mid \underset{\sim}{\theta}) f(x_2 \mid \underset{\sim}{\theta}) \cdots f(x_n \mid \underset{\sim}{\theta}) = \prod_{i=1}^{n} f(x_i \mid \underset{\sim}{\theta}) \qquad (1.2.2.2).$$

Η (1.2.2.2) καλείται **συνάρτηση πιθανοφάνειας** (*likelihood function*) του τυχαίου δείγματος $X_1, X_2, ..., X_n$ και εκφράζει το πόσο «πιθανοφανείς», δηλαδή πόσο σύμφωνες με το συγκεκριμένο δείγμα είναι οι διάφορες τιμές της παραμέτρου $\underset{\sim}{\theta}$.

Η μέθοδος μέγιστης πιθανοφάνειας συνίσταται στην επιλογή της τιμής $\underset{\sim}{\theta}$ η οποία μεγιστοποιεί τη συνάρτηση πιθανοφάνειας,

$$L(\hat{\underset{\sim}{\theta}} \mid x_1, x_2, ..., x_n) = \sup_{\underset{\sim}{\theta} \in \Theta} L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n).$$

Η τιμή $\hat{\underset{\sim}{\theta}}$ καλείται **εκτιμήτρια μέγιστης πιθανοφάνειας** της $\underset{\sim}{\theta}$. Μεγιστοποίηση της $L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n)$ σημαίνει μεγιστοποίηση της πιθανότητας εμφάνισης των τιμών $x_1, x_2, ..., x_n$ στο δείγμα $X_1, X_2, ..., X_n$. Η τιμή αυτή $\hat{\underset{\sim}{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$, βρίσκεται με λύση των εξισώσεων

$$\frac{\partial \log L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n)}{\partial \theta_r} = 0, \quad r = 1, 2, ..., k.$$

Φυσικά, για να είναι η λύση αυτή πράγματι σημείο μεγίστου, θα πρέπει ο Εσσιανός πίνακας $\left[\frac{\partial^2 \log L(\underset{\sim}{\theta})}{\partial \theta_i \partial \theta_j} \right]_{k \times k}$ να είναι γνήσια αρνητικός για $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}$.

Η εκτίμηση μέγιστης πιθανοφάνειας έχει όλες τις ιδιότητες καλής εκτιμήτριας, δηλαδή είναι αμερόληπτη, συνεπής, αποτελεσματική και επαρκής. Για αυτό και αυτή η μέθοδος είναι η καλύτερη μέθοδος εκτίμησης παραμέτρων.

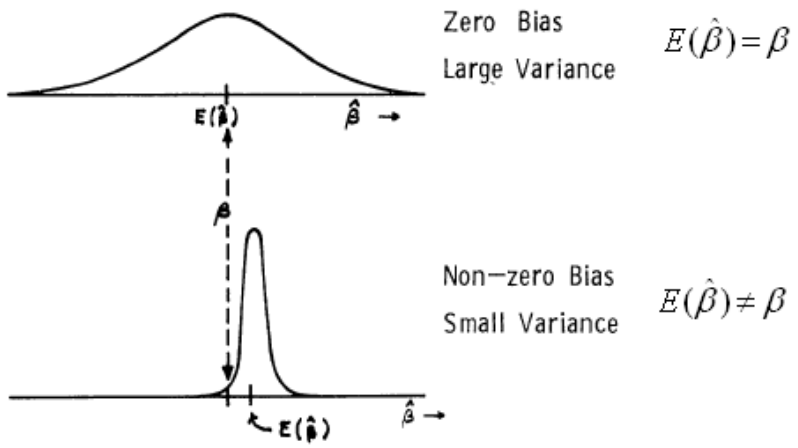
1.3. Παλινδρόμηση κορυφογραμμής/διασέλου

1.3.1. Περιγραφή της μεθόδου

Ενα αρκετά συχνό πρόβλημα που συναντάται σε ένα μοντέλο πολλαπλής παλινδρόμησης, είναι η ύπαρξη συσχέτισης μεταξύ των προβλεπουσών μεταβλητών του μοντέλου. Αυτό σημαίνει ότι κάποιες στήλες στον πίνακα X των παρατηρήσεων είναι γραμμικά εξαρτημένες. Αποτέλεσμα αυτού είναι ο πίνακας $X'X$ να είναι

ιδιάζων $(\det(X'X)=0)$ και τότε ο πίνακας $(X'X)^{-1}$ δεν υπάρχει. Οπότε έχουμε το γνωστό πρόβλημα πολυσυγγραμικότητας (*multicollinearity*) και έτσι προκύπτουν μη σταθερές εκτιμήσεις των παραμέτρων του μοντέλου. Για παράδειγμα, οι εκτιμήσεις μπορεί να έχουν λάθος πρόσημο ή να προκύψουν μεγαλύτερες από ότι οι φυσικές ή πρακτικές υποθέσεις θεωρούν λογικό. Επίσης, μικρές αλλαγές στα δεδομένα του πίνακα X θα επιφέρουν μεγάλες αλλαγές στις τιμές των εκτιμηθέντων συντελεστών παλινδρόμησης. Οπότε, στην περίπτωση που εφαρμοστεί η μέθοδος ελαχίστων τετραγώνων, παρουσία πολυσυγγραμικότητας, οι συντελεστές που θα προκύψουν θα έχουν μεγάλες απόλυτες τιμές και θα είναι αρκετά ασταθείς.

Το μειονέκτημα της μεθόδου ελαχίστων τετραγώνων είναι η απαίτηση ότι το $\hat{\beta}$ πρέπει να είναι αμερόληπτος εκτιμητής του β. Σύμφωνα με το θεώρημα Gauss-Markov ο εκτιμητής ελαχίστων τετραγώνων έχει την ελάχιστη διασπορά στην κλάση των αμερόληπτων γραμμικών εκτιμητών, χωρίς όμως να εξασφαλίζεται ότι θα είναι είναι και μικρή η τιμή της. Στο παρακάτω γράφημα 1.3.1.1 (περίπτωση όπου $E(\hat{\beta})=\beta$) απεικονίζεται η κατανομή του αμερόληπτου εκτιμητή $\hat{\beta}$.



<u>Γράφημα 1.3.1.1</u>: Κατανομή του $\hat{\beta}$

Η διασπορά του $\hat{\beta}$ είναι μεγάλη, με αποτέλεσμα τα διαστήματα εμπιστοσύνης για το β να έχουν μεγάλο εύρος, το οποίο με τη σειρά του οδηγεί σε μια αρκετά ασταθή σημειακή εκτίμηση $\hat{\beta}$.

Ένας τρόπος αντιμετώπισης του προαναφερθέντος προβλήματος, είναι να μη ληφθεί υπόψην η απαίτηση της αμεροληψίας. Οπότε, υποθέτουμε ότι μπορούμε να βρούμε ένα μεροληπτικό εκτιμητή, έστω $\hat{\beta}^*$, που να έχει μικρότερη διασπορά από το $\hat{\beta}$. Το μέσο τετραγωνικό σφάλμα (*mean squared error*) του $\hat{\beta}^*$ είναι

$$MSE(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2 = Var(\hat{\beta}^*) + \left[E(\hat{\beta}^*) - \beta\right]^2$$

ή

$$MSE(\hat{\beta}^*) = Var(\hat{\beta}^*) + (bias)^2 .$$

Επιτρέποντας ένα ποσό μεροληψίας στο $\hat{\beta}^*$, η $Var(\hat{\beta}^*)$ μπορεί να γίνει μικρή ώστε τελικά το $MSE(\hat{\beta}^*)$ να είναι μικρότερο από τη διασπορά του αμερόληπτου εκτιμητή $\hat{\beta}$. Στο παραπάνω γράφημα 1.3.1.1 (περίπτωση όπου $E(\hat{\beta}) \neq \beta$), διαφαίνεται η περίπτωση όπου η διασπορά του μεροληπτικού εκτιμητή $\hat{\beta}^*$ είναι σαφώς μικρότερη της διασποράς του αμερόληπτου $\hat{\beta}$.

Έχουν αναπτυχθεί αρκετές μέθοδοι για την εύρεση μεροληπτικών εκτιμητών για τους συντελεστές παλινδρόμησης. Μια εξ΄ αυτών είναι η παλινδρόμηση κορυφογραμμής ή διασέλου (*ridge regression*). Ιστορικά, η «διαδικασία ίχνους κορυφογραμμής» (*ridge trace procedure*), προτάθηκε απο τον A.E. Hoerl το 1962 και συζητήθηκε εκτενώς από τους A.E. Hoerl και R.W. Kennard (Hoerl & Kennard, (1970), Hoerl et al. (1975)). Ο εκτιμητής διασέλου $\hat{\underset{\sim}{\beta}}_R$ προσδιορίζεται από την εξίσωση

$$(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})\hat{\underset{\sim}{\beta}}_R = \underset{\sim}{X}'\underset{\sim}{Y}$$

ή

$$\hat{\underset{\sim}{\beta}}_R = (\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{Y},$$

όπου $k \geq 0$ είναι μια σταθερά η οποία επιλέγεται από τον αναλυτή και ονομάζεται παράμετρος μεροληψίας (*biasing parameter*). Να παρατηρήσουμε ότι στην περίπτωση όπου $k = 0$, ο παραπάνω εκτιμητής συμπίπτει με αυτόν της μεθόδου ελαχίστων τετραγώνων. Ουσιαστικά, η παλινδρόμηση διασέλου ελαχιστοποιεί την ποσότητα

$$\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_j \beta_j^2,$$

ή ισοδύναμα την ποσότητα $\sum_{i=1}^{N}\left(y_i - \sum_j \beta_j x_{ij}\right)^2$, υπό τον περιορισμό (*constraint*) $\sum_j \beta_j^2 \leq t$.

Συνεχίζουμε, αναφέροντας κάποιες ιδιότητες του εκτιμητή διασέλου. Καταρχήν, είναι ένας γραμμικός μετασχηματισμός του εκτιμητή ελαχίστων τετραγώνων, καθότι

$$\hat{\underset{\sim}{\beta}}_R = (\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} = (\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}(\underset{\sim}{X}'\underset{\sim}{X})\hat{\underset{\sim}{\beta}} = \underset{\sim}{Z}_k\hat{\underset{\sim}{\beta}}.$$

Οπότε από τη στιγμή που $E(\hat{\underset{\sim}{\beta}}_R) = E(\underset{\sim}{Z}_k\hat{\underset{\sim}{\beta}}) = \underset{\sim}{Z}_k\underset{\sim}{\beta}$, ο $\hat{\underset{\sim}{\beta}}_R$ είναι ένας μεροληπτικός εκτιμητής του $\underset{\sim}{\beta}$. Επίσης, έχουμε ότι $Var(\hat{\underset{\sim}{\beta}}_R) = \sigma^2(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}$

και το μέσο τετραγωνικό σφάλμα είναι

$$MSE(\hat{\underset{\sim}{\beta}}_R) = Var(\hat{\underset{\sim}{\beta}}_R) + (bias)^2$$

$$= \sigma^2 Tr\left[(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}\right] + k^2\underset{\sim}{\beta}'(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-2}\underset{\sim}{\beta}$$

$$= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + k)^2} + k^2\underset{\sim}{\beta}'(\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-2}\underset{\sim}{\beta},$$

όπου $\lambda_1, \lambda_2, ..., \lambda_p$ είναι οι ιδιοτιμές του πίνακα $\underset{\sim}{X}'\underset{\sim}{X}$. Για $k > 0$, η μεροληψία του $\hat{\underset{\sim}{\beta}}_R$ αυξάνεται, καθώς αυξάνεται το k, ενώ αντιθέτως η διασπορά μειώνεται.

Το πρόβλημα έγκειται στην επιλογή του κατάλληλου k, ώστε η μείωση στη διασπορά να είναι μεγαλύτερη από την αύξηση (του τετραγώνου) της μεροληψίας. Εαν αυτό επιτευχθεί, τότε το $MSE(\hat{\underset{\sim}{\beta}}_R)$ θα είναι μικρότερο από τη διασπορά του εκτιμητή ελαχίστων τετραγώνων $\hat{\beta}$. Οι Hoerl και Kennard απέδειξαν ότι όντως υπάρχει μη μηδενική τιμή του k, με την ως άνω ιδιότητα, προϋποθέτοντας ότι το $\underset{\sim}{\beta}'\underset{\sim}{\beta}$ είναι φραγμένο. Το άθροισμα τετραγώνων των σφαλμάτων είναι

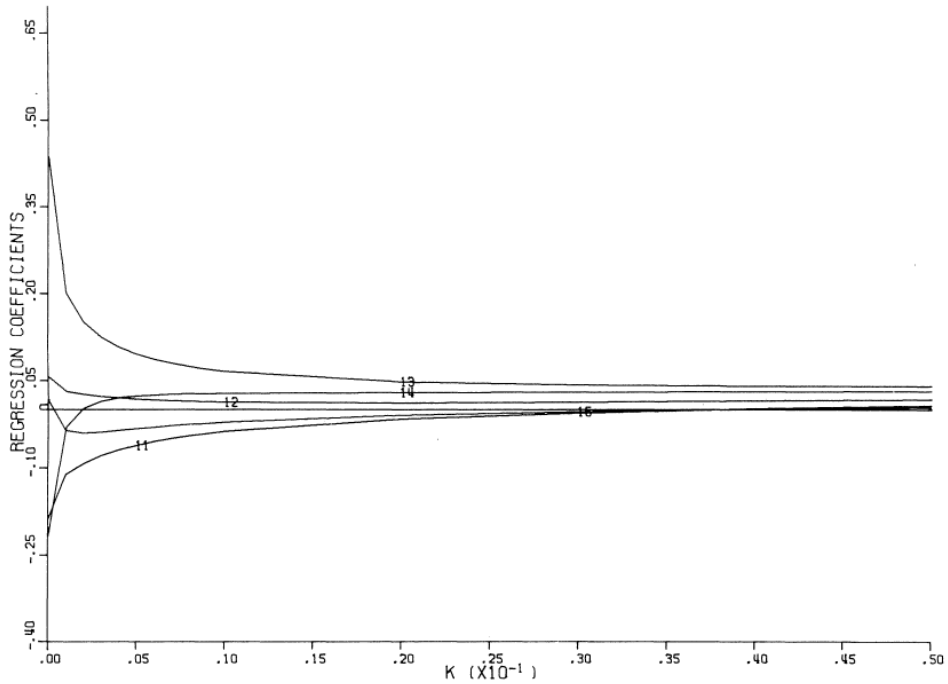
$$SSE = (\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}_R)'(\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}_R)$$

$$= (\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}})'(\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}) + (\hat{\underset{\sim}{\beta}}_R - \hat{\underset{\sim}{\beta}})'(\underset{\sim}{X}'\underset{\sim}{X})(\hat{\underset{\sim}{\beta}}_R - \hat{\underset{\sim}{\beta}}).$$

Όσο αυξάνεται το k αυξάνεται και το SSE. Δεδομένου τώρα ότι το συνολικό άθροισμα τετραγώνων (SST) είναι σταθερό, προκύπτει ότι ο συντελεστής προσδιορισμού R^2 μειώνεται καθώς αυξάνεται το k. Να υπενθυμίσουμε ότι

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Συνεπώς, η παλινδρόμηση διασέλου μπορεί να μην παρέχει το καλύτερο μοντέλο, αλλά από τη στιγμή που μας ενδιαφέρει να αποκτήσουμε ένα σταθερό σύνολο από εκτιμήσεις των παραμέτρων, αυτό δεν μας απασχόλει ιδιαίτερα. Όσον αφορά τώρα την τιμή του k που πρέπει να επιλεγεί αυτό, κατά τους Hoerl και Kennard, γίνεται παρατηρώντας το ίχνος κορυφογραμμής. Πρόκειται για την γραφική παράσταση των συντελεστών παλινδρόμησης για διάφορες τιμές του k, όπου συνήθως $k \in [0,1]$. Παρουσία της πολυσυγγραμικότητας, η αστάθεια των συντελεστών παλινδρόμησης είναι εμφανής. Για κάποια τιμή του k, οι εκτιμήσεις $\hat{\underset{\sim}{\beta}}_R$ σταθεροποιούνται. Επί παραδείγματι, ας παρατηρήσουμε το παρακάτω γράφημα 1.3.1.2 (Marquardt & Snee, 1975).



<u>Γράφημα 1.3.1.2</u>: Ίχνος κορυφογραμμής

Βλέπουμε ότι έχουμε μεγάλες αλλαγές στους συντελεστές παλινδρόμησης για μικρές τιμές του k, αλλά παρόλα αυτά οι συντελεστές σταθεροποιούνται αρκετά γρήγορα, όσο αυτό αυξάνει. Σκοπός είναι να επιλεγεί μια (σχετικά μικρή) τιμή του k, όπου και να επιτυγχάνεται αυτή η σταθερότητα. Τελικά, αυτό θα έχει ως αποτέλεσμα να προκύψει ένα σύνολο σταθερών εκτιμητών με μικρότερο μέσο τετραγωνικό σφάλμα σε σχέση με του εκτιμητές ελαχίστων τετραγώνων, καθότι η ελάττωση που επιτυγχάνουν στη διασπορά σφάλματος θα είναι μεγαλύτερη από την αντιστάθμιση για την μεροληψία που εισάγεται. Στο παραπάνω γράφημα 1.3.1.2, μια καλή επιλογή του k θα ήταν κοντά στο 0.012.

Εναλλακτικά, οι Hoerl, Kennard και Baldwin (1975) πρότειναν ως μια κατάλληλη επιλογή του k την εξής:

$$k = \frac{p\hat{\sigma}^2}{\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}}} \qquad (1.3.1.1),$$

όπου τα $\hat{\underset{\sim}{\beta}}$ και $\hat{\sigma}^2$ βρίσκονται βάσει της μεθόδου ελαχίστων τετραγώνων και p είναι ο αριθμός των παραμέτρων στο μοντέλο, χωρίς να υπολογίζουμε το β_0. Μέσω προσομοιώσεων έδειξαν ότι ο εκτιμητής διασέλου που προκύπτει βελτιώνει σημαντικά το MSE. Οι Mcdonald και Galarneau (1975) πρότειναν το k να υπολογίζεται κατά τέτοιον τρόπο ώστε

$$\hat{\underset{\sim}{\beta}}_R{}'\hat{\underset{\sim}{\beta}}_R = \hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}} - \hat{\sigma}^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}.$$

Να σημειωθεί όμως ότι η μέθοδός τους έχει το μειονέκτημα ότι το k μπορεί να προκύψει αρνητικό. Επίσης, ο Mallows (1973) πρότεινε μια διαδικασία επιλογής του k με γραφικό τρόπο, βασιζόμενος σε μια τροποποίηση του γνωστού στατιστικού C_p.

Επιπλέον, πέρα από τη χρησιμοποίηση του ίχνους κοριφυγραμμής, πρέπει ο αναλυτής να λαμβάνει υπόψην του και το λεγόμενο VIF (*Variance Inflation Factor*) για την επιλογή του k. Πρόκειται για ένα μέτρο ανίχνευσης του μεγέθους της πολυσυγγραμμικότητας. Το VIF για τον συντελεστή $\hat{\beta}_i$ δίδεται από τον τύπο

$$VIF_i = \frac{1}{1 - R_i^2},$$

όπου R_i^2 είναι ο συντελεστής συσχέτισης της μεταβλητής x_i (γενικά τα VIF είναι τα διαγώνια στοιχεία του αντιστρόφου του πίνακα συσχέτισης (*correlation matrix*)). Συγκεκριμένα, το VIF είναι ένας δείκτης μέτρησης του κατά πόσο αυξήθηκε η διασπορά ενός συντελεστή λόγω της πολυσυγγραμμικότητας. Η τετραγωνική του ρίζα, δείχνει πόσο μεγαλύτερο είναι το τυπικό σφάλμα σε σχέση με αυτό που θα έπρεπε να είναι αν η μεταβλητή αυτή ήταν ασυσχέτιστη με τις υπόλοιπες. Για παράδειγμα, αν το VIF μιας μεταβλητής είναι 5.27, άρα $\sqrt{VIF} = 2.3$, αυτό σημαίνει ότι το τυπικό σφάλμα του συντελεστή της μεταβλητής είναι 2.3 φορές μεγαλύτερο από τη τιμή που θα είχαμε κανονικά. Γενικά το VIF τείνει να μειώνεται αρκετά όσο το k απομακρύνεται από το 0 και έπειτα παρουσιάζει μια μέτρια αλλαγή. Τελικά, ο στόχος είναι να γίνει μια επιλογή του k στο σημείο όπου το VIF έχει αρκετά μικρή

τιμή, ενώ συγχρόνως οι εκτιμώμενοι συντελεστές έχουν κάπως σταθεροποιηθεί (παρατηρώντας και το ίχνος κορυφογραμμής).

1.3.2. Παλινδρόμηση κορυφογραμμής/διασέλου και επιλογή μεταβλητών

Οι συνήθεις αλγόριθμοι επιλογής μεταβλητών δεν αποδίδουν όταν τα δεδομένα χαρακτηρίζονται από σοβαρής μορφής πολυσυγγραμικότητα. Παρόλα αυτά, αν με τη χρήση μεροληπτικών εκτιμητών έχει επιτευχθεί μια σχεδόν ορθογωνιότητα, μπορεί να γίνει πλέον επιλογή μεταβλητών. Οι Hoerl και Kennard (1970) πρότειναν ότι το ίχνος κορυφογραμμής μπορεί να χρησιμοποιηθεί ως οδηγός για το σκοπό αυτό. Συγκεκριμένα αναφέρουν τους κάτωθι κανόνες για την αφαίρεση μεταβλητών από το μοντέλο:

- Αφαιρούνται οι μεταβλητές που εμφανίζουν σταθερότητα, αλλά έχουν μικρή προβλεπτική ικανότητα, δηλαδή μεταβλητές με μικρούς κανονικοποιημένους συντελεστές.
- Αφαιρούνται οι μεταβλητές με ασταθείς συντελεστές που δε διατηρούν την προβλεπτική τους ικανότητα, δηλαδή με συντελεστές που συγκλίνουν στο μηδέν.
- Αφαιρούνται μια η περισσότερες από τις εναπομείνουσες μεταβλητές με ασταθείς συντελεστές.

Οπότε το υποσύνολο των υπόλοιπων μεταβλητών, χρησιμοποιείται στο τελικό μοντέλο. Στη συνέχεια, είναι εφικτό να εξεταστεί αν οι μεταβλητές αυτές αποτελούν ένα ορθογώνιο ή σχεδόν ορθογώνιο υποσύνολο. Αυτό επιτυγχάνεται κάνοντας τη γραφική παράσταση των συναρτήσεων $\hat{\underset{\sim}{\beta}}_R'(k)\hat{\underset{\sim}{\beta}}_R(k)$ και $\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}}/(1+k)^2$, συναρτήσει του k. Στην περίπτωση που ισχύει η ορθογωνιότητα (ή σχεδόν ορθογωνιότητα), θα πρέπει να είναι αρκετά όμοια τα γραφήματα.

1.4. Η διαδικασία PRESS

Η διαδικασία επιλογής μεταβλητών *PRESS* (*Prediction Error Sum of Squares*), προτάθηκε από τον D.M. Allen (1971,1974). Έστω ότι p είναι ο αριθμός των παραμέτρων του μοντέλου, συμπεριλαμβανομένου του β_0 και ότι υπάρχουν συνολικά n παρατηρήσεις. Ακολουθούνται τα παρακάτω βήματα:

- Για δοσμένο i, διαγράφουμε την i παρατήρηση.
- Προσαρμόζουμε το μοντέλο στις υπόλοιπες $n-1$ παρατηρήσεις.
- Προβλέπουμε το $\hat{y}_{i(i)}$, την παρατήρηση δηλαδή που παραλείψαμε.
- Υπολογίζουμε το σφάλμα πρόβλεψης $y_i - \hat{y}_{i(i)}$.
- Επαναλαμβάνουμε τα προηγούμενα βήματα, παραλείποντας μια προς μια όλες τις παρατηρήσεις.
- Για κάθε ένα από τα προκύπτοντα μοντέλα παλινδρόμησης, υπολογίζουμε το παρακάτω άθροισμα τετραγώνων των σφαλμάτων πρόβλεψης

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2 .$$

- Επιλέγουμε το «καλύτερο» μοντέλο παλινδρόμησης, προτιμώντας κάποιο με όσο γίνεται μικρή τιμή του $PRESS$, σε συνδυασμό όμως με μικρό αριθμό μεταβλητών. Οπότε θα έχουμε ένα συγκριτικά μικρό προβλεπόμενο άθροισμα τετραγώνων αλλά το μοντέλο δεν θα περιλαμβάνει πάρα πολλές ανεξάρτητες μεταβλητές.

1.5. Μέθοδος επιλογής καλύτερου υποσυνόλου

Υπάρχουν περιπτώσεις που καθιστούν αναγκαία την επιλογή ενός μικρού υποσυνόλου από ένα μεγαλύτερο σύνολο μεταβλητών οι οποίες χρησιμοποιούνται για την πρόβλεψη μια εξαρτημένης μεταβλητής. Για παράδειγμα, η διαδικασία πρόβλεψης της εξαρτημένης μεταβλητής Y μπορεί να θεωρείται οικονομικά ασύμφορη και αρκετά χρονοβόρα εφόσον χρησιμοποιηθούν όλες οι τιμές που έχουμε

για τις ανεξάρτητες μεταβλητές. Για το λόγο αυτό αναζητούμε ένα μικρό υποσύνολο μεταβλητών που θα μπορεί με αρκετή ακρίβεια να προβλέψει την τιμή της μεταβλητής *Y*. Η επιλογή αυτού του συνόλου θα μπορούσε προφανώς στη συνέχεια να χρησιμοποιηθεί για κάποια μελλοντική πρόβλεψη εφόσον τα δεδομένα είναι αντιπροσωπευτικά των συνθηκών υπό τις οποίες θα γίνει η πρόβλεψη. Από την άλλη πλευρά, στην προσπάθειά μας να κατανοήσουμε την επίδραση που έχει η μια μεταβλητή πάνω στην άλλη, ιδιαίτερα όταν τα δεδομένα που έχουμε συλλέξει είναι μέσω παρατήρησης ή μέσω έρευνας και όχι δεδομένα που προήλθαν μέσω πειραμάτων, μπορεί να είναι επιθυμητό να μην χρησιμοποιήσουμε πολλές μεταβλητές οι οποίες έχουν κάποια ισχυρή επίδραση μεταξύ τους.

Κάποιες φορές τα δεδομένα για την πρόβλεψη που θέλουμε να κάνουμε έχουν ήδη συλλεχθεί για προηγούμενες προβλέψεις ή για άλλους σκοπούς με αποτέλεσμα να μην υπάρχει επιπλέον κόστος εάν συμπεριληφθούν στο μοντέλο πρόβλεψης. Σε άλλες περιπτώσεις μπορεί να υπάρχει σημαντικό επιπλέον κόστος εάν χρησιμοποιηθούν όλα τα δεδομένα που έχει ως αποτέλεσμα την εξέταση του κόστους σε σχέση με την ακρίβεια των προβλέψεων.

Η παραδοσιακή ιδέα της μεθόδου επιλογής καλύτερου υποσυνόλου που μπορεί να θεωρηθεί ως μια ειδική μορφή ποινικοποιημένης πιθανοφάνειας, είναι υπολογιστικά πολύ ακριβή για πολλές σύγχρονες εφαρμογές. Άλλες μορφές μεθόδων ποινικοποιημένης πιθανοφάνειας έχουν αναπτυχθεί με επιτυχία κατά την τελευταία δεκαετία για να αντιμετωπίσει την υψηλή διάσταση. Έχουν εφαρμοστεί ευρέως για την ταυτόχρονη επιλογή σημαντικών μεταβλητών και την εκτίμηση των επιπτώσεών τους στην υψηλής διάστασης στατιστική συμπερασματολογία. Σε αυτό το κεφάλαιο, θα παρουσιάσουμε την πρόσφατη εξέλιξη της θεωρίας, των μεθόδων και των εφαρμογών σε υψηλής διάστασης επιλογή μεταβλητών, τα όρια της διάστασης που μπορούν να χειριστούν τέτοιες μέθοδοι και το ρόλο της συνάρτησης ποινής.

Μερικούς από τους λόγους για τη χρήση μόνο ενός υποσυνόλου διαθέσιμων μεταβλητών πρόβλεψης (σύμφωνα με το Miler, 1984) είναι:

- Να εκτιμήσουμε ή να προβλέψουμε με μικρό κόστος τη μείωση του αριθμού των μεταβλητών από τα δεδομένα που πρόκειται να συλλεχθούν.
- Να προβλέψουμε με περισσότερη ακρίβεια καταργώντας ανούσιες μεταβλητές.

- Να περιγράψουμε μια σειρά πολυμεταβλητών δεδομένων με μετριότητα (parsinomy).
- Να εκτιμήσουμε τους συντελεστές παλινδρόμησης με μικρότερα σφάλματα (ειδικά όταν μερικοί από τους προγνωστικούς παράγοντες σχετίζονται πολύ μεταξύ τους) , (BurakEksioglu, RizaDemirer, IsmailCapar, 2005).

Η **μέθοδος επιλογής καλύτερου υποσυνόλου** (*best subset selection method*) χρησιμοποιεί διάφορα κριτήρια βάσει των οποίων επιλέγεται το καλύτερο υποσύνολο μεταβλητών, το οποίο και θα οδηγήσει στο σωστότερο αντίστοιχο μοντέλο. Τα κριτήρια αυτά χωρίζονται σε τέσσερις μεγάλες κατηγορίες (Koukouvinos, C., Mylona, K. and Vonta, F. (2008))

- Κριτήρια πρόβλεψης (*prediction criteria*).
- Κριτήρια βασισμένα στην πιθανοφάνεια ή στην πληροφορία (*likelihood or information based criteria*).
- Κριτήρια που προκύπτουν και κριτήρια που καθοδηγούνται, από τα δεδομένα (*Data-reuse and data driven criteria*).
- Κριτήρια μεγιστοποίησης των Μπεϋζιανών εκ των υστέρων πιθανοτήτων (*maximizing Bayesian posterior probabilities*).

Τα κριτήρια πρόβλεψης βασίζονται στα σφάλματα πρόβλεψης. Τα πιο γνωστά είναι το κριτήριο *FPE* (*Final Prediction Error*, (Akaike, 1969)) και το κριτήριο C_p-*Mallows* (Mallows, 1973). Τα κριτήρια που προκύπτουν από τα δεδομένα ή αλλιώς *bootstrap* μέθοδοι, εισήχθησαν από τον Efron και παρέχουν εύκολους και αποτελεσματικούς τρόπους για να εκτιμήσουμε την αναμενόμενη συνολική διαφορά μεταξύ του σωστού και του υποψήφιου μοντέλου. Εφαρμογές αυτών των τεχνικών μπορούν να βρεθούν στο (Linhart & Zucchini, 1986). Ο Breiman (1992) πρότεινε ένα τέτοιο κριτήριο και από τις προσομοιώσεις που πραγματοποίησε έδειξε ότι είναι καλύτερο από το C_p. Όσον αφορά τα κριτήρια καθοδηγούμενα από τα δεδομένα ή μέθοδοι διασταυρωμένης επικύρωσης (*Cross-validation methods*) αυτά βασίζονται στο διαχωρισμό των δεδομένων σε δύο υποσύνολα, όπου το ένα χρησιμοποιείται για την επιλογή του μοντέλου και το άλλο

για τον καθορισμό της προβλεπτικής ικανότητας του μοντέλου. Μια τέτοια μέθοδος, είναι η γενικευμένη διασταυρωμένη επικύρωση (*Generalized Cross-validation* $-GCV$), η οποία επιλέγει το μοντέλο με την καλύτερη μέση προβλεπτική ικανότητα, χρησιμοποιώντας διάφορους τρόπους διαχωρισμού του αρχικού συνόλου δεδομένων. Τέλος, οι Μπεϋζιανές μέθοδοι, χρησιμοποιούν μια ιεραρχική εκ των προτέρων κατανομή (*hierarchical prior*) ώστε να αναθέσουν μεγαλύτερες εκ των υστέρων πιθανότητες σε περισσότερο υποσχόμενα μοντέλα.

1.5.1. Στατιστική συνάρτηση C_p – Mallows

Ένα άλλο μέτρο για την αξιολόγηση της καταλληλότητας του μοντέλου, είναι η στατιστική συνάρτηση **C_p – Mallows**. Βασίζεται στην ακρίβεια πρόβλεψης του μοντέλου και πιο συγκεκριμένα στο **Μέσο Τετραγωνικό Σφάλμα**, το οποίο ορίζεται γενικώς για την εκτιμήτρια $\hat{\theta}$ μιας παραμέτρου θ από τη σχέση

$$\begin{aligned}
\text{ΜΤΣ}(\hat{\theta}) &= \text{E}(\hat{\theta}-\theta)^2 \\
&= \text{E}[\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta]^2 \\
&= \text{E}[\hat{\theta}-E(\hat{\theta})]^2 + 2E[\hat{\theta}-E(\hat{\theta})]\,\text{E}[E(\hat{\theta})-\theta] + E[E(\hat{\theta})-\theta]^2 \\
&= V(\hat{\theta}) + [E(\hat{\theta})-\theta]^2 \\
&= V(\hat{\theta}) + bias^2
\end{aligned}$$

Η παραπάνω σχέση μας δίνει και το λόγο που στρέφουμε την προσοχή μας στο *ΜΤΣ*. Το *ΜΤΣ* συνδέεται με τη μεροληψία (*bias*) εκτιμητριών παραμέτρων ενός μοντέλου που δεν έχει οριστεί σωστά και δίνει μεροληπτικές εκτιμήτριες. Επιπλέον, η διαδικασία επιλογής μεταβλητών δεν καταλήγει στο σωστό υποσύνολο μεταβλητών κάθε φορά λόγω δειγματοληπτικού σφάλματος ή μεταβλητότητας. Άρα πρέπει να λάβουμε υπόψη τη μεροληψία.

Για τον προσδιορισμό της C_p βασιζόμαστε στο συνολικό *ΜΤΣ* της πρόβλεψης δηλαδή στο

$$\sum_{i=1}^{n} \text{MT}\Sigma(\hat{\text{y}}_i) = \sum_{i=1}^{n} E(\hat{\text{y}}_i - E(\text{y}_i))^2 \ .$$

Έστω οι προβλέψεις $\hat{\text{y}}_1, ..., \hat{\text{y}}_n$ του υπό εξέταση μοντέλου με p όρους (p = πλήθος επεξηγηματικών μεταβλητών + 1), όχι απαραιτήτως το μοντέλο που περιέχει όλες τις διαθέσιμες υποψήφιες επεξηγηματικές μεταβλητές. Τότε το συνολικό *ΜΤΣ* διαιρεμένο με σ^2 ακολουθώντας την προηγούμενη διαδικασία, δίνεται από τη σχέση

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}\left[E(\text{y}_i) - E(\hat{\text{y}}_i)\right]^2 + \sum_{i=1}^{n} V(\hat{\text{y}}_i)\right\} = \frac{SB(\text{p})}{\sigma^2} + \frac{\sum_{i=1}^{n} V(\hat{\text{y}}_i)}{\sigma^2}$$

όπου $SB(\text{p}) = \sum bias^2$.

Ο δεύτερος όρος απλοποιείται ως εξής:

$$\sum_{i=1}^{n} V(\hat{\text{y}}_i) = tr(V(\hat{\text{y}})) = tr(V(\text{Hy})) = tr(\text{H}V(\text{y})\text{H}')$$

$$= tr(\text{H}(\sigma^2 \text{I})\text{H}') = \sigma^2 tr(\text{H}^2) = \sigma^2 tr(\text{H}) = \sigma^2 p \ .$$

Προκειμένου να βρούμε μια έκφραση για το $SB(\text{p})$, εξετάζουμε το άθροισμα τετραγώνων των υπολοίπων για το μοντέλο με τους p όρους, δηλαδή το

$$SSE(\text{p}) = \sum_{i=1}^{n} (\text{y}_i - \hat{\text{y}}_i)^2 \ .$$

Η αναμενόμενη τιμή του $SSE(\text{p})$ μπορεί να υπολογιστεί ως εξής

$$E(SSE(\text{p})) = \sum_{i=1}^{n} \text{E}\left[(\text{y}_i - \hat{\text{y}}_i)^2\right]$$

$$= \sum_{i=1}^{n}\left[\sigma^2 + V(\hat{\text{y}}_i) + (\mu_i - \nu_i)^2 - 2\sigma^2 h_{ii}\right], \text{ με } \nu_i = E(\hat{\text{y}}_i)$$

$$= n\sigma^2 + \sum_{i=1}^{n} V(\hat{\text{y}}_i) + \sum_{i=1}^{n} (\mu_i - \nu_i)^2 - 2\sigma^2 \sum_{i=1}^{n} h_{ii}$$

$$= n\sigma^2 + p\sigma^2 + SB(\text{p}) - 2\sigma^2 p$$

$$= (\mathrm{n}-\mathrm{p})\sigma^2 + SB(\mathrm{p})$$

αφού

$$\mathrm{E}\left[(\mathrm{y}_i - \hat{\mathrm{y}}_i)^2\right] = E\left[[(\mathrm{y}_i - \mu_i) - (\hat{\mathrm{y}}_i - \mu_i)]^2\right] \qquad \text{με } \mu_i = E(\mathrm{y}_i)$$

$$= E\left[(\mathrm{y}_i - \mu_i)^2\right] + E\left[(\hat{\mathrm{y}}_i - \mu_i)^2\right] - 2E\left[(\mathrm{y}_i - \mu_i)(\hat{\mathrm{y}}_i - \mu_i)\right]$$

$$= \sigma^2 + E\left[\{(\hat{\mathrm{y}}_i - \nu_i) - (\mu_i - \nu_i)\}^2\right] - 2E\left[(\mathrm{y}_i - \mu_i)(\hat{\mathrm{y}}_i - \nu_i + \nu_i - \mu_i)\right]$$

$$= \sigma^2 + E\left[(\hat{\mathrm{y}}_i - \nu_i)^2\right] + E\left[(\mu_i - \nu_i)^2\right] - 2E\left[(\hat{\mathrm{y}}_i - \nu_i) + (\mu_i - \nu_i)\right]$$

$$= \sigma^2 + E\left[(\hat{\mathrm{y}}_i - \nu_i)^2\right] + E\left[(\mu_i - \nu_i)^2\right] - 2E\left[(\hat{\mathrm{y}}_i - \nu_i) + (\mu_i - \nu_i)\right] - 2E\left[(\mathrm{y}_i - \mu_i)(\hat{\mathrm{y}}_i - \nu_i)\right] - 2E\left[(\mathrm{y}_i - \mu_i)(\nu_i - \mu_i)\right]$$

$$= \sigma^2 + V(\hat{\mathrm{y}}_i) + (\mu_i - \nu_i)^2 - 2(\mu_i - \nu_i)E(\hat{\mathrm{y}}_i - \nu_i) - 2E\left[(\mathrm{y}_i - \mu_i)(\hat{\mathrm{y}}_i - \nu_i)\right] - 2(\nu_i - \mu_i)E(\mathrm{y}_i - \mu_i)$$

$$= \sigma^2 + V(\hat{\mathrm{y}}_i) + (\mu_i - \nu_i)^2 - 2\,\mathrm{cov}(\mathrm{y}_i, \hat{y}_i)$$

$$= \sigma^2 + V(\hat{\mathrm{y}}_i) + (\mu_i - \nu_i)^2 - 2\sigma^2 h_{ii}, \qquad \text{όπου } \nu_i = E(\hat{\mathrm{y}}_i).$$

Αντικαθιστώντας το $SB(\mathrm{p})$ στη σχέση Γ_p λαμβάνουμε

$$\Gamma_p = \frac{1}{\sigma^2}(\mathrm{E}(SSE(\mathrm{p})) - (\mathrm{n}-\mathrm{p})\sigma^2) + \frac{p\sigma^2}{\sigma^2} = \frac{\mathrm{E}(SSE(\mathrm{p}))}{\sigma^2} + 2p - n$$

Η ποσότητα Γ_p εκτιμάται από την

$$C_p = \frac{SSE(\mathrm{p})}{\hat{\sigma}^2} + 2p - n$$

Δηλαδή το $\mathrm{E}(SSE(\mathrm{p}))$ έχει αντικατασταθεί από το παρατηρούμενο άθροισμα τετραγώνων των υπολοίπων του υπό εξέταση μοντέλου και το σ^2 από μια καλή εκτίμηση του $\hat{\sigma}^2$. Συνήθως χρησιμοποιείται το μοντέλο που περιέχει όλες τις υποψήφιες μεταβλητές x_j (και όχι το τρέχον μοντέλο με τους p όρους), δηλαδή

$\hat{\sigma}^2 = \frac{SSE(\mathrm{p}')}{n - p'}$, όπου p' = το πλήθος όλων των διαθέσιμων επεξηγηματικών μεταβλητών +1.

Αν η μεροληψία στο μοντέλο με τις p μεταβλητές είναι αμελητέα, τότε $SB(\mathrm{p}) \simeq 0$ και

$$E\left[C_p \middle| Bias = 0\right] = \frac{(\mathrm{n}-\mathrm{p})\sigma^2}{\sigma^2} + 2p - n = p$$

Επομένως, ως βέλτιστο μοντέλο θεωρείται εκείνο για το οποίο ισχύει $C_p \simeq p$.

Ειδικότερα, αν υπάρχουν περισσότερα από ένα μοντέλο με $C_p \simeq p$, προτιμότερο είναι εκείνο με το μικρότερο p, δηλαδή αυτό που περιέχει λιγότερες μεταβλητές.

Η τιμή του κριτηρίου C_p αυξάνεται καθώς εισάγεται μια νέα ανεξάρτητη μεταβλητή στο γραμμικό μοντέλο. Χρησιμοποιώντας το κριτήριο C_p επιδιώκουμε να προσδιορίσουμε εκείνα τα γραμμικά μοντέλα για τα οποία η τιμή του κριτηρίου είναι μικρή και είναι κοντά στην τιμή του p. Τα σύνολα με μικρή τιμή του C_p έχουν μικρή τιμή και για το μέσο τετραγωνικό σφάλμα και όταν η τιμή του C_p είναι κοντά στο p τότε η μεροληψία για το μοντέλο είναι μικρή.

1.5.2. Κριτήρια βασισμένα στην πιθανοφάνεια ή στην πληροφορία

Τα κριτήρια αυτά μπορούν να θεωρηθούν ως σχεδόν αμερόληπτοι εκτιμητές της αναμενόμενης συνολικής διαφοράς μεταξύ του σωστού και του υποψήφιου μοντέλου. Συνήθως έχουν δύο όρους, όπου ο πρώτος αποτελεί έναν μεροληπτικό εκτιμητή και ο δεύτερος είναι μια διόρθωση ή ποινή (*penalty*) όπως λέγεται και έτσι ο εκτιμητής γίνεται σχεδόν αμερόληπτος. Στη συνέχεια, θα αναλύσουμε τα πιο γνωστά κριτήρια, συγκεκριμένα το AIC (*Akaike Information Criterion*) και το BIC (*Bayesian Information Criterion*) καθώς και μια σειρά άλλων επίσης σημαντικών κριτηρίων. Στα επόμενα, l_j θα είναι η μέγιστη τιμή του λογάριθμου της συνάρτησης

πιθανοφάνειας του προσαρμοσμένου μοντέλου, p_j το πλήθος των συμμεταβλητών στο μοντέλο και τέλος n θα είναι το μέγεθος του δείγματος.

Το κριτήριο **AIC** προτάθηκε από τον Hirotsugu Akaike (1974) με το όνομα «το κριτήριο πληροφοριών» (information criterion). Το κριτήριο αυτό στηρίζεται στην ιδέα της εντροπίας των πληροφοριών. Στην πραγματικότητα προσφέρεται ως ένα σχετικό μέτρο για τα στοιχεία που χάνονται σε ένα πραγματικό μοντέλο. Θα μπορούσαμε να πούμε ότι το κριτήριο αυτό περιγράφει τη σχέση μεταξύ της ακρίβειας και της πολυπλοκότητας του μοντέλου. Όμως το κριτήριο αυτό δε μπορεί να περιγράψει εάν το μοντέλο που μελετάμε ταιριάζει ικανοποιητικά με τα δεδομένα μας. (Γενικά, με τον όρο κριτήριο πληροφορίας εννοούμε έναν μηχανισμό που χρησιμοποιεί τα δεδομένα ενός προβλήματος δίνοντας σε κάθε υποψήφιο μοντέλο μια τιμή που το χαρακτηρίζει.)

Το AIC αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με το όσο το δυνατό μικρότερο αριθμό παραμέτρων. Στη γενική περίπτωση ορίζεται από τη σχέση

$$AIC(j) = -2(l_j - p_j),$$

Σκοπός είναι να επιλεχθεί το μοντέλο που δίνει τη μικρότερη τιμή του κριτηρίου. Μια ισοδύναμη μορφή του AIC, θεωρώντας ότι τα σφάλματα είναι ανεξάρτητα και κανονικά κατανεμημένα, είναι

$$AIC(j) = n[\ln(2\pi RSS / n) + 1] + 2p_j,$$

όπου n το μέγεθος του δείγματος και RSS το άθροισμα τετραγώνων των σφαλμάτων. Επίσης, στην περίπτωση των γενικευμένων γραμμικών μοντέλων (Konishi & Kitagama, 2008), που θα μελετήσουμε παρακάτω, παίρνει τη μορφή

$$AIC(j) = -2\sum_{i=1}^{n}\left\{\left[y_i \underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}} - b(\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}})\Big/\alpha(\hat{\varphi}) + c(y_i,\hat{\varphi})\right]\right\} - 2(p_j + 2),$$

όπου $\alpha(\cdot)$ και $c(\cdot)$ είναι συναρτήσεις της παραμέτρου κλίμακας φ.

Τελικά, η εισαγωγή επιπλέον παραμέτρων στο μοντέλο μειώνει την τιμή του AIC μόνο αν αυτές βελτιώνουν την προσαρμογή του μοντέλου.

Οι Hurvich και Tsai (1989) ασχολήθηκαν με την περίπτωση όπου έχουμε μικρά δείγματα, και πρότειναν μια βελτίωση του AIC για την περίπτωση αυτή, οπότε έχουμε το κριτήριο

$$AIC_c = AIC + \frac{2(p_j+1)(p_j+2)}{n-p_j-2}.$$

Με το κριτήριο αυτό, επιλέγονται μικρότερα υποσύνολα μεταβλητών όταν ο αρχικός τους συνολικός αριθμός είναι μεγαλύτερος σε σχέση με το μέγεθος του δείγματος n. Επίσης, για μεγάλο n, έχει παρόμοια συμπεριφορά με το AIC. Να σημειώσουμε ότι στις προσομοιώσεις που έκαναν απεδείχθη η πολύ καλύτερη απόδοση και αποτελεσματικότητα του κριτηρίου, συγκρινόμενο με πλήθος άλλων κριτηρίων, όσον αφορά την επιλογή του σωστού μεγέθους μοντέλου και στις περιπτώσεις όπου έχουμε μικρό δείγμα και αριθμό μεταβλητών.

Τα προτερήματα αλλά και τα μειονεκρήματα του AIC, καθώς και όλων των κριτηρίων πληροφορίας που χρησιμοποιούν τη μέγιστη πιθανοφάνεια, απορρέουν από τη συσχέτιση του AIC με τις εκτιμήτριες μέγιστης πιθανοφάνειας.

Με βάση την εξίσωση του Bayes, ο Schwartz (1978) ανέπτυξε ένα άλλο κριτήριο, το οποίο το ονόμασε **κριτήριο πληροφορίας του Bayes** επειδή είναι κριτήριο πληροφοριών με βάση τη Bayesian Μέθοδο. Ουσιαστικά, είναι ένα κριτήριο για την επιλογή μοντέλων μεταξύ μιας ομάδας παραμετρικών μοντέλων με διαφορετικούς αριθμούς παραμέτρων. Διαλέγοντας ένα μοντέλο για τη βελτιστοποίηση του *BIC* είναι μια μορφή κανονικοποίησης.

Το κριτήριο BIC δίνεται από τον τύπο

$$BIC(j) = l_j - (1/2)p_j \ln(n).$$

Το μοντέλο που επιλέγεται είναι αυτό που δίνει τη μεγαλύτερη τιμή του κριτηρίου. Να σημειώσουμε επίσης, ότι υπάρχει μια κλίμακα (*Raftery's scale*), που ανάλογα με την απόλυτη διαφορά στις τιμές του κριτηρίου BIC μεταξύ δύο μοντέλων κρίνεται κατά πόσο το ένα μοντέλο είναι καλύτερο από το άλλο (Hardin & Hilbe, 2007). Η κλίμακα αυτή αναγράφεται στον παρακάτω Πίνακα 1.5.1.1.

|Διαφορά των τιμών BIC|	Ένδειξη
0-2	Ασθενής (*Weak*)
2-8	Θετική (*Positive*)
6-10	Ισχυρή (*Strong*)
>10	Πολύ Ισχυρή (*Very Strong*)

Πίνακας 1.5.1.1 : Κλίμακα Raftery

Συνεχίζουμε με το κριτήριο του *Rissanen* (1978), σύμφωνα με το οποίο αναζητείται το μοντέλο που μεγιστοποιεί την ποσότητα

$$RC(j) = nl_j + \sum_{i=1}^{p_j} \ln\left(\theta_i^2 \frac{\partial^2 l_j}{\partial \theta_i^2}\right) + (p_j + 1)\ln(n+2) + \ldots,$$

όπου θ_i είναι οι παράμετροι του μοντέλου. Το κριτήριο αυτό είχε χρησιμοποιηθεί για την προσαρμογή αυτοπαλινδρομούμενων μοντέλων κινητού μέσου (*Autoregressive Moving Average Models - ARMA*) και ο τελευταίος όρος ήταν $2\ln(r+1)(s+1)$, όπου r και s οι τάξεις του αυτοπαλινδρομούμενου και του κινητού μέσου μέρους του μοντέλου αντίστοιχα, με $p_j = r + s$.

Θα αναφέρουμε τώρα ένα ακόμα κριτήριο, αρκετά πρόσφατο, αυτό των Bryant και Cordero-Brana, οι οποίοι το 2000 (Bryant & Cordero-Brana, 2000) βασίστηκαν στο *MDL* (*Minimum Description Length*) για την κατασκευή του κριτηρίου. Συγκεκριμένα, όρισαν ως *MDL* μια κλάσης M μοντέλων ως την κλάση όλων των γραμμικών υποσυνόλων των μεταβλητών, δηλαδή

$$MDL(Y \mid M, g) = -\ln f(Y, \hat{\underset{\sim}{\beta}}) - \ln g(\hat{\underset{\sim}{\beta}}),$$

όπου Y είναι τα δεδομένα, f είναι η πυκνότητα πιθανότητας (της κανονικής), $\hat{\underset{\sim}{\beta}}$ είναι το διάνυσμα των συντελεστών παλινδρόμησης και g είναι το μήκος του μηνύματος που απαιτείται για την κωδικοποίηση των βέλτιστων τιμών των

συντελεστών. Συνεπώς, ο τελευταίος όρος θεωρείται μια ποινή στην αύξηση των μεταβλητών. Το κριτήριο Bryant και Cordero-Brana συνίσταται στην ελαχιστοποίηση της ποσότητας

$$MDL = (n/2)\ln\left(\frac{RSS}{n}\right) - \ln\Gamma((n-p)/2) + p\ln(R_0/\sqrt{\pi}) + \ln\ln(\sigma_{\max}/\sigma_{\min})^2 + G(n)\ ,$$

όπου $G(n) = \frac{n}{2}\left(\frac{n-1}{n}\ln n + \ln\pi\right)$ και R_0 είναι ο αριθμός των τυπικών αποκλίσεων, εύρους όσοι και οι κανονικοποιημένοι συντελεστές παλινδρόμησης (π.χ. αν το εύρος είναι ± 5 τυπικές αποκλίσεις, το $R_0 = 10$), στους οποίους έχει υπολογισθεί το *MDL* και $\sigma_{\min} < \hat{\sigma} < \sigma_{\max}$ όλων των υπό θεώρηση μοντέλων.

1.6. Ποινικοποιημένα ελάχιστα τετράγωνα και ποινικοποιημένη πιθανοφάνεια

1.6.1. Εισαγωγή

Οι πιο γνωστές και συχνότερα χρησιμοποιούμενες μέθοδοι επιλογής μεταβλητών, είναι ως γνωστόν η κατά βήματα απαλοιφή (*stepwise deletion*) και η μέθοδος επιλογής καλύτερου υποσυνόλου (*best subset selection*). Έχουν όμως το μειονέκτημα ότι αγνοούν τα στοχαστικά σφάλματα που εμφανίζονται κατά τη διαδικασία της επιλογής μεταβλητών καθώς και ότι είναι υπολογιστικά χρονοβόρες. Οι Fan και Li (2001), πρότειναν μια καινούργια μεθοδολογία, βασισμένη στα **ποινικοποιημένα ελάχιστα τετράγωνα** (*penalized least squares*), η οποία διατηρεί τις καλές ιδιότητες της παλινδρόμησης κορυφογραμμής αλλά και της μεθόδου επιλογής καλύτερου υποσυνόλου. Η μεθοδολογία τους αυτή, επεκτείνεται και σε μοντέλα βασισμένα στη πιθανοφάνεια, όπως π.χ. στην περίπτωση όπου έχουμε δίτιμη απόκριση (*binary response*). Μια γνωστή οικογένεια τέτοιων μοντέλων είναι τα γενικευμένα γραμμικά μοντέλα. Ουσιαστικά τώρα, αυτό που τελικά επιτυγχάνεται, είναι ότι ταυτόχρονα γίνεται και εκτίμηση των παραμέτρων του μοντέλου και μηδενισμός κάποιων, άρα ικανοποιείται ο σκοπός της επιλογής μεταβλητών.

Η διαδικασία της ποινικοποίησης χρησιμοποιείται συχνά στην επιλογή μεταβλητών και συνίσταται στην εισαγωγή κάποιων συναρτήσεων ποινής (*penalty functions*), οι οποίες πρέπει να έχουν τις ακόλουθες τρεις ιδιότητες (Fan and Li, 2001):

1. Σποραδικότητα: Ο εκτιμητής αποτελέσματος θέτει αυτόματα τους μικρούς εκτιμηθέντες συντελεστές στο μηδέν για να επιτύχει την επιλογή μεταβλητών και να μειώσει την πολυπρλοκότητα του μοντέλου.
2. Αμεροληψία: Ο εκτιμητής αποτελέσματος είναι σχεδόν αμερόληπτος, ειδικά όταν οι πραγματικοί συντελεστές β_j είναι μεγάλοι, για να να μειώσει τη μεροληψία.
3. Συνέχεια: Ο εκτιμητής αποτελέσματος είναι συνεχής στα δεδομένα για να μειώσει την αστάθεια στην πρόβλεψη του μοντέλου (Breiman, 1996).

Η παλινδρόμηση bridge που προτάθηκε από τους Frank και Friedman (1993), και η μέθοδος LASSO που προτάθηκε από τον Tibshirani (1996) είναι μέλη της μεθόδου των ποινικοποιημένων ελαχίστων τετραγώνων, με τη διαφορά ότι οι σχετικές με τις μεθόδους αυτές συναρτήσεις ποινής L_q δεν ικανοποιούν όλες τις προαναφερθείσες απαιτήσεις.

Όπως αναφέραμε και προηγουμένως, η καινούργια μέθοδος επεκτάθηκε και σε μοντέλα βασισμένα στη πιθανοφάνεια (*likelihood-based models*). Η διαφορά σε σχέση με τις παραδοσιακές μεθόδους (όπου συνήθως χρησιμοποιείται τετραγωνική συνάρτηση ποινής), είναι ότι οι νέες συναρτήσεις ποινής είναι συμμετρικές, κυρτές στο $(0,\infty)$ και διακατέχονται από ιδιομορφίες (*singularities*) στην αρχή. Να σημειωθεί, ότι σε αντίθεση με τις παραδοσιακές μεθόδους επιλογής μεταβλητών, η νέα μέθοδος έχει ισχυρό θεωρητικό υπόβαθρο. Επίσης, στην εργασία τους, οι Fan και Li (2001) πρότειναν ένα αρκετά αποδοτικό αλγόριθμο βελτιστοποίησης της ποινικοποιημένης πιθανοφάνειας ο οποίος οδηγεί στην εκτίμηση των παραμέτρων και στον υπολογισμό του τυπικού σφάλματος. Δόθηκε μια συγκεκριμένη φόρμουλα υπολογισμού του σφάλματος για τους εκτιμηθέντες συντελεστές χρησιμοποιώντας τη μέθοδο *sandwich*. Η μέθοδος αυτή έχει δοκιμαστεί και είναι αρκετά ακριβής για

πρακτικούς σκοπούς ακόμα και στην περίπτωση μέτριου μεγέθους δείγματος. Οι προτεινόμενες αυτές διαδικασίες επιλογής συγκρινόμενες με άλλες μεθόδους επιλογής μεταβλητών δίνουν πάντα καλύτερα και ορθότερα αποτελέσματα.

Συνεχίζοντας την περιγραφή των χαρακτηριστικών των μεθόδων αυτών, αναφέρουμε το μεγαλύτερο πλεονέκτημά τους. Συγκεκριμένα, επιλέγουν τις σημαντικές μεταβλητές και εκτιμούν τους συντελεστές τους ταυτόχρονα. Οπότε μπορούν να αναπτυχθούν οι δειγματικές ιδιότητες (*sampling properties*) των μεθόδων. Στην συνέχεια παρουσιάζουμε πως οι δείκτες σύγκλισης (*rates of convergence*) των προτεινόμενων εκτιμητών της ποινικοποιημένης πιθανοφάνειας (*penalized likelihood estimators*) εξαρτώνται από την παράμετρο κανονικοποίησης. Πρέπει να σημειώσουμε ότι οι εκτιμητές ποινικοποιημένης πιθανοφάνειας, έχουν τόσο καλή απόδοση όσον αφορά την επιλογή του σωστού μοντέλου, όσο και η διαδικασία προβλεψιμότητας (*oracle procedure*), αρκεί να έχει επιλεγεί σωστά η παράμετρος κανονικοποίησης (*regularization parameter*). Σαν να ήταν δηλαδή γνωστό από την αρχή το σωστό υπομοντέλο (*submodel*). Αυτό πρακτικά, σημαίνει ότι όταν οι σωστές παράμετροι του μοντέλου έχουν κάποιες μηδενικές συνιστώσες, αυτές εκτιμώνται από τη μέθοδο ως μηδενικές με πιθανότητα να τείνει στη μονάδα. Ενώ όσον αφορά τις μη μηδενικές συνιστώσες, αυτές εκτιμώνται τόσο καλά όπως όταν είναι γνωστό το σωστό υπομοντέλο. Αυτό προφανώς αυξάνει την ακρίβεια εκτίμησης τόσο των μηδενικών όσο και των μη μηδενικών συνιστωσών. Οπότε και υπερτερούν της μεθόδου εκτίμησης μέγιστης πιθανοφάνειας.

1.6.2. Επιλογή μεταβλητών μέσω ποινικοποιημένων ελαχίστων τετραγώνων

Θεωρούμε το γνωστό γραμμικό μοντέλο

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

όπου $\underset{\sim}{Y}$ είναι ένα $n \times 1$ διάνυσμα των παρατηρήσεων, $\underset{\sim}{X}$ ένας $n \times d$ πίνακας επεξηγηματικών μεταβλητών, $\underset{\sim}{\beta}$ ένα $d \times 1$ διάνυσμα συντελεστών παλινδρόμησης και το $\underset{\sim}{\varepsilon}$ είναι ένα $n \times 1$ διάνυσμα τυχαίων σφαλμάτων. Όπως και στην περίπτωση

του μοντέλου γραμμικής παλινδρόμησης, υποθέτουμε ότι τα y_i είναι υπό συνθήκη ανεξάρτητα, δοθέντων των x_{ij}. Επίσης, υποθέτουμε και ότι οι στήλες του πίνακα $\underset{\sim}{X}$ είναι ορθοκανονικές (*orthonormal*). Ο υπολογισμός της εκτιμήτριας γίνεται μέσω της ελαχιστοποίησης της ποσότητας

$$\| \underset{\sim}{Y} - \underset{\sim}{X} \underset{\sim}{\beta} \|^2 ,$$

η οποία ισοδυναμεί με την ποσότητα

$$\| \underset{\sim}{\hat{\beta}} - \underset{\sim}{\beta} \|^2 ,$$

όπου

$$\underset{\sim}{\hat{\beta}} = \underset{\sim}{X}' \underset{\sim}{Y}$$

είναι η OLS (*ordinary least squares*) εκτιμήτρια. Θέτοντας τώρα ως

$$\underset{\sim}{z} = \underset{\sim}{X}' \underset{\sim}{Y}$$

και έστω ότι

$$\underset{\sim}{\hat{Y}} = \underset{\sim}{X} \underset{\sim}{X}' \underset{\sim}{Y} ,$$

μια μορφή των ποινικοποιημένων ελαχίστων τετραγώνων είναι η εξής:

$$\frac{1}{2} \| \underset{\sim}{Y} - \underset{\sim}{X} \underset{\sim}{\beta} \|^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|) = \frac{1}{2} \| \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} \|^2 + \frac{1}{2} \sum_{j=1}^{d} (z_j - \beta_j)^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|) \quad (1.6.2.1).$$

Να σημειωθεί ότι οι συναρτήσεις ποινής p_j στην (1.6.2.1) δεν είναι απαραίτητα οι ίδιες για όλα τα j. Για παράδειγμα μπορεί να θέλουμε να κρατήσουμε ορισμένες σημαντικές μεταβλητές σε ένα παραμετρικό μοντέλο και για αυτό το λόγο να μη θέλουμε να ποινικοποιήσουμε τις αντίστοιχες παραμέτρους τους. Για ευκολία όμως, θεωρούμε ότι οι συναρτήσεις ποινής είναι οι ίδιες για όλους τους συντελεστές, και θα συμβολίζονται ως $p(|\cdot|)$. Επίσης, αντί $\lambda p(|\cdot|)$ θα χρησιμοποιούμε το συμβολισμό $p_\lambda(|\cdot|)$, δείχνοντας έτσι ότι το $p(|\cdot|)$ εξαρτάται από το λ.

Το πρόβλημα ελαχιστοποίησης της (1.6.2.1) είναι ισοδύναμο με την ελαχιστοποίηση των συνιστωσών. Οπότε θεωρούμε το παρακάτω πρόβλημα ελαχίστων τετραγώνων

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \qquad (1.6.2.2).$$

Στη συνεχεία, χρησιμοποιώντας τη *Hard* συνάρτηση ποινής (βλ. σχήμα 1.6.2.1(α))

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda),$$

προκύπτει η *Hard* εκτιμήτρια (Antoniadis & Fan (2001), Fan (1997)) (βλ. σχήμα 1.6.2.2α).

$$\hat{\theta} = zI(|z| > \lambda) \qquad (1.6.2.3)$$

Με άλλα λόγια, η λύση της (1.6.2.1) είναι

$$z_j I(|z_j| > \lambda)$$

η οποία συμπίπτει με την επιλογή καλύτερου υποσυνόλου και την κατά βήματα πρόσθεση και απαλοιφή στους ορθοκανονικούς σχεδιασμούς. Σημειώνουμε επιπλέον πως η συνάρτηση ποινής *Hard* είναι ομαλότερη από την συνάρτηση ποινής εντροπίας (*entropy penalty*)

$$p_\lambda(|\theta|) = \left(\frac{\lambda^2}{2}\right) I(|\theta| \neq 0),$$

η οποία και αυτή οδηγεί στη λύση (1.6.2.3).

Μια συνάρτηση ποινής για να είναι καλή, πρέπει να δίνει εκτιμητές με τις ακόλουθες ιδιότητες:

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

- Σποραδικότητα: Ο εκτιμητής που προκύπτει πρέπει να αποτελεί κανόνα περιορισμού (*thresholding rule*), ώστε οι εκτιμηθέντες συντελεστές με μικρή τιμή, να μηδενίζονται. Έτσι, μειώνεται η πολυπλοκότητα του μοντέλου.
- Συνέχεια: Ο εκτιμητής που προκύπτει πρέπει είναι συνεχής. Αποφεύγεται κατά αυτόν τον τρόπο η αστάθεια στην πρόβλεψη του μοντέλου.

Θα εξηγήσουμε τώρα τις παραπάνω ιδιότητες. Καταρχήν, η πρώτη παράγωγος της $\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)$ ως προς θ είναι

$$\operatorname{sgn}(\theta)\{|\theta| + p'_\lambda(|\theta|)\} - z\,.$$

Παρατηρούμε ότι όταν $p'_\lambda(|\theta|) = 0$ για μεγάλο $|\theta|$, τότε ο εκτιμητής που προκύπτει είναι ίσος με z όταν το $|z|$ είναι επαρκώς μεγάλο. Για αυτό το λόγο, όταν η πραγματική παράμετρος $|\theta|$ είναι μεγάλη, η τιμή $|z|$ είναι και αυτή μεγάλη και με μεγάλη πιθανότητα. Οπότε, ο PLS (*penalized least squares*) εκτιμητής είναι

$$\hat{\theta} = z\,,$$

ο οποίος και είναι σχεδόν αμερόληπτος. Δηλαδή, η προϋπόθεση $p'_\lambda(|\theta|) = 0$ για μεγάλο $|\theta|$, είναι μια επαρκής προϋπόθεση για την αμεροληψία μιας μεγάλης πραγματικής παραμέτρου. Όσον αφορά τη δεύτερη ιδιότητα, για να αποτελεί ο προκύπτων εκτιμητής κανόνα περιορισμού, πρέπει να ισχύει ότι

$$\min_\theta \{|\theta| + p'_\lambda(|\theta|)\} > 0\,.$$

Το παρακάτω γράφημα 1.6.2.3 παρέχει περισσότερες εξηγήσεις σχετικά με αυτό. Όταν τώρα

$$|z| < \min_{\theta \neq 0} \{|\theta| + p'_\lambda(|\theta|)\}$$

η παράγωγος της $\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)$ είναι θετική για όλα τα θετικά θ και αρνητική για όλα τα αρνητικά θ. Οπότε σε αυτήν την περίπτωση, ο PLS εκτιμητής $\hat{\theta}$ είναι

μηδέν. Όταν όμως $|z| > \min_{\theta \neq 0}\{|\theta| + p'_\lambda(|\theta|)\}$, δύο διασταυρώσεις (*crossings*) μπορούν να υπάρξουν, όπως φαίνεται και στο σχήμα 1.6.2.1. Η μεγαλύτερη είναι ο PLS εκτιμητής. Αυτό συνεπάγεται ότι ικανή και αναγκαία συνθήκη για την ύπαρξη συνέχειας είναι το $\min_{\theta}\{|\theta| + p'_\lambda(|\theta|)\}$ να πετυχαίνεται στο μηδέν. Από αυτό αντιλαμβανόμαστε πως η συνάρτηση ποινής που ικανοποιεί τις ιδιότητες της σποραδικότητας και της συνέχειας, πρέπει να είναι ιδιάζουσα (*singular*) στην αρχή.

Είναι γνωστό πως η συνάρτηση ποινής L_2

$$p_\lambda(|\theta|) = \lambda |\theta|^2$$

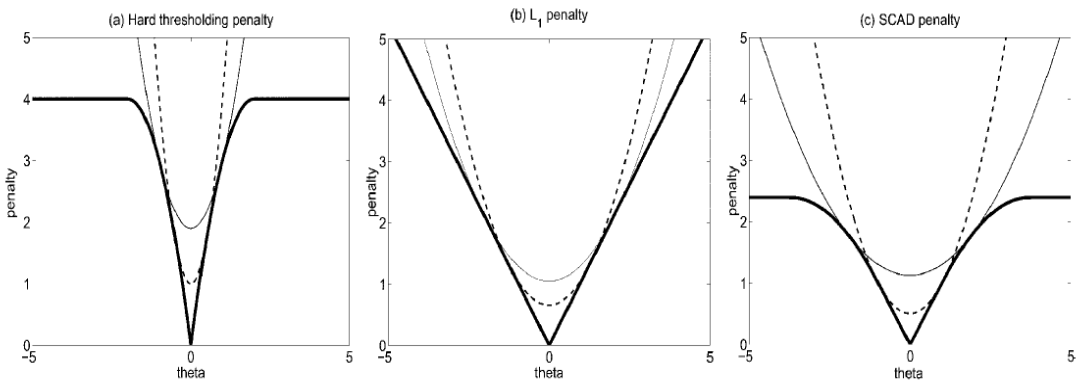
οδηγεί στην παλινδρόμηση κορυφογραμμής. Η συνάρτηση ποινής L_1, οδηγεί στον *soft* οριακό κανόνα

$$\hat{\theta}_j = \mathrm{sgn}(z_j)(|z_j| - \lambda)_+ ,$$

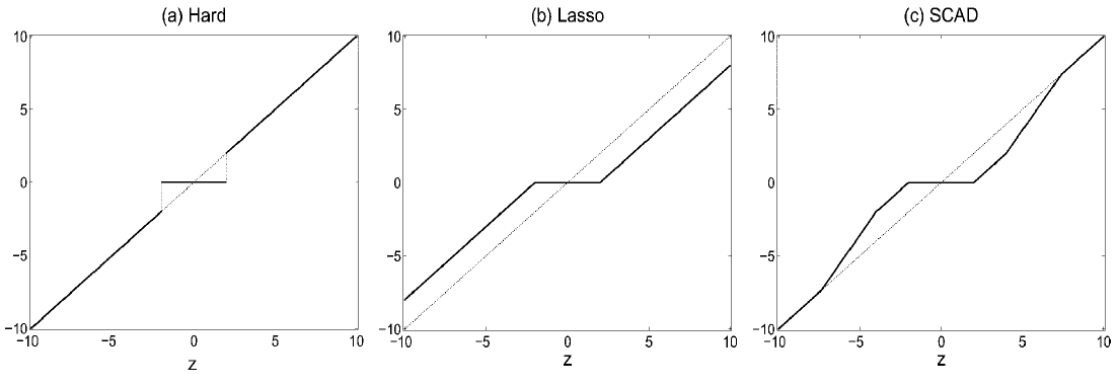
που προτάθηκε από τους Donoho και Johnstone (1994). Η $LASSO$ που προτείνεται από τον Tibshirani (1996, 1997) είναι ο PLS εκτιμητής με συνάρτηση ποινής τη L_1. Επίσης, η L_q συνάρτηση ποινής

$$p_\lambda(|\theta|) = \lambda |\theta|^q$$

οδηγεί στην παλινδρόμηση *bridge* (Frank & Friedman (1993), Fu (1998)). Η λύση είναι συνεχής μόνο για $q \geq 1$. Παρόλα αυτά, όταν $q > 1$, δεν παράγεται μια σποραδική λύση (βλ. σχήμα 1.6.2.4(a)). Η μόνη συνεχής λύση με κανόνα περιορισμού σε αυτή την οικογένεια συναρτήσεων είναι με τη συνάρτηση ποινής L_1, αυτό όμως προκύπτει μεταβάλλοντας τον εκτιμητή κατά μια σταθερά λ, άρα χάνεται και η αμεροληψία (βλ. σχήμα 1.6.2.2(b)). Επίσης για $0 \leq q < 1$, δεν ικανοποιείται η συνθήκη της συνέχειας.

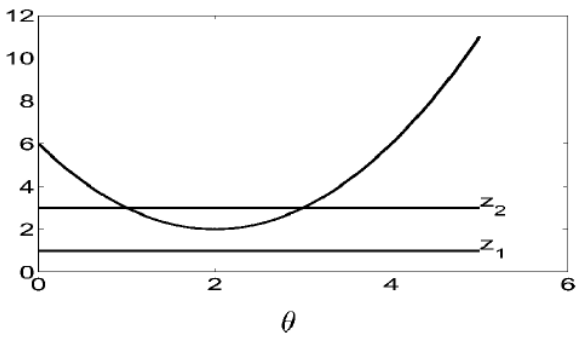


<u>Σχήμα 1.6.2.1</u>: (a) Οι τρεις συναρτήσεις ποινής και οι τετραγωνικές τους προσεγγίσεις.

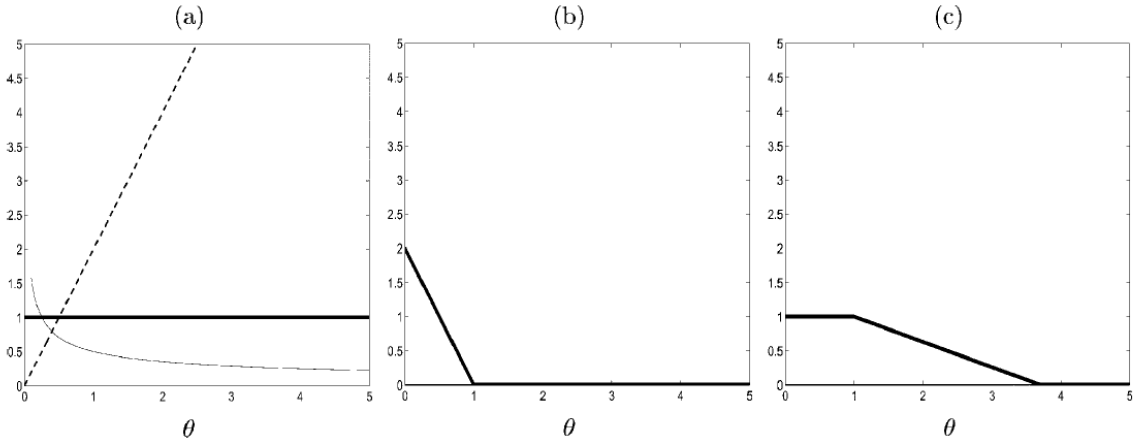


<u>Σχήμα 1.6.2.2</u>: Οι εκτιμήτριες (thresholding functions) (a) Hard, (b) Soft ήLASSO

και (c) Scad, όπου για την τελευταία λ=2 και a=3.7.



<u>Σχήμα 1.6.2.3</u>: Η συνάρτηση $\theta + p_\lambda(|\theta|)$ ως προς θ.



<u>Σχήμα 1.6.2.4</u>: Οι συναρτήσεις $p'_\lambda(|\theta|)$ ως προς θ, για (a) τις συναρτήσεις ποινής L_q, (b) τη Hard συνάρτηση ποινής και (c) τη SCAD. Στο (a), η παχιά γραμμή αντιστοιχεί στην L_1, η διακεκομμένη στην $L_{0.5}$ και η λεπτή γραμμή στην L_2 συνάρτηση ποινής.

1.6.2.1. Η συνάρτηση ποινής

Οι συναρτήσεις ποινής L_q με $q>1$ δεν ικανοποιεί τη συνθήκη σποραδικότητας, ενώ οι συναρτήσεις ποινής L_1 δεν ικανοποιούν τη συνθήκη αμεροληψίας, και η κοίλη συνάρτηση ποινής L_q με $0 \le q < 1$ δεν ικανοποιεί τη συνθήκη συνέχειας. Με άλλα λόγια, καμία από τις συναρτήσεις ποινής L_q δεν ικανοποιεί και τις τρεις συνθήκες συγχρόνως. Για αυτό το λόγο, οι Fan (1997) και Fan και Li (2001) εισήγαγαν μια συνεχής και διαφορίσιμη συνάρτηση ποινής, τη $SCAD$ (*Smoothly Clipped Absolute Deviation penalty*) (Σχήμα 1.6.2.1(c)), της οποίας η παράγωγος ορίζεται ως

$$p'_\lambda(\theta) = \lambda\left\{ I(\theta \le \lambda) + \frac{(\alpha\lambda - \theta)_+}{(\alpha-1)\lambda} I(\theta > \lambda) \right\}, \text{ για κάποιο } a > 2 \text{ και } \theta > 0$$

όπου $p_\lambda(0) = 0$ και συχνά χρησιμοποιείται το $a = 3.7$ (μέσω της Μπεϋζιανής ανάλυσης ρίσκου (*Bayesian risk analysis*). Ικανοποιεί και τις τρεις προαναφερθείσες ιδιότητες.

Η συγκεκριμένη συνάρτηση δεν ποινικοποιεί υπερβολικά τις μεγάλες τιμές του θ και δίνει μια συνεχής λύση, την

$$\hat{\theta} = \begin{cases} \operatorname{sgn}(z)(|z| - \lambda)_+, & |z| \le 2\lambda \\ \{(\alpha - 1)z - \operatorname{sgn}(z)\alpha\lambda\} / (\alpha - 2), & 2\lambda < |z| \le \alpha\lambda \\ z, & |z| > \alpha\lambda \end{cases} \quad (1.6.2.1.1)$$

Η λύση αυτή δόθηκε από τον Fan (1997) ο οποίος έκανε μια εκτενής συζήτηση για την περίπτωση των κυματοσυναρτήσεων (*wavelets*).

Η λύση (1.6.2.1.1) έχει δύο άγνωστες παραμέτρους, α και λ. Στην πράξη θα μπορούσαμε να υπολογίσουμε το βέλτιστο ζεύγος (α, λ) βάσει κάποιων κριτηρίων, όπως της διασταυρωμένης επικύρωσης και της γενικευμένης διασταυρωμένης επικύρωσης. Κάτι που μπορεί να είναι υπολογιστικά χρονοβόρο.

1.6.3. Επιλογή μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας

Η μέχρι στιγμής αναπτυχθείσα μεθοδολογία, μπορεί να εφαρμοσθεί σε πλήθος στατιστικών μοντέλων, όπως γραμμικά μοντέλα παλινδρόμησης (*linear regression models*), εύρωστα γραμμικά μοντέλα (*robust linear models*) και γενικευμένα γραμμικά μοντέλα βασισμένα στην πιθανοφάνεια (*likelihood-based generalized linear models*). Θα θεωρούμε πλέον ότι ο πίνακας σχεδιασμού $\underset{\sim}{X} = (x_{ij})$ είναι κανονικοποιημένος, ώστε κάθε στήλη να έχει μέση τιμή 0 και διασπορά 1.

Στο κλασικό μοντέλο παλινδρόμησης οι εκτιμητές ελαχίστων τετραγώνων παράγονται με την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων. Οπότε η (1.6.2.1) μπορεί να επεκταθεί για την περίπτωση όπου ο πίνακας σχεδιασμού δεν είναι ορθοκανονικός (*orthonormal*). Μια ισοδύναμη μορφή της (1.6.2.1) είναι

$$\frac{1}{2}(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (1.6.3.1).$$

Ελαχιστοποιώντας την (1.6.3.1) ως προς $\underset{\sim}{\beta}$, οδηγούμαστε σε έναν εκτιμητή ποινικοποιημένων ελαχίστων τετραγώνων του $\underset{\sim}{\beta}$.

Είναι γνωστό τώρα ότι ο OLS εκτιμητής δεν είναι εύρωστος. Μπορούμε όμως να θεωρήσουμε τη συνάρτηση ψ του Huber (1981), οπότε αντί της ελαχιστοποίησης της (1.6.3.1), μπορούμε να ελαχιστοποιήσουμε την

$$\sum_{i=1}^{n}\psi(|y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta}|) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \qquad (1.6.3.2)$$

ως προς $\underset{\sim}{\beta}$, ώστε να πάρουμε έναν εύρωστο ποινικοποιημένο εκτιμητή του $\underset{\sim}{\beta}$.

Στην περίπτωση των γενικευμένων γραμμικών μοντέλων, γίνεται συμπερασματολογία βάσει των εκάστοτε υποβοσκουσών συναρτήσεων πιθανοφάνειας. Με τη βοήθεια τώρα του ποινικοποιημένου εκτιμητή μέγιστης πιθανοφάνειας, μπορούμε να επιλέξουμε σημαντικές μεταβλητές. Έχουμε τα εξής: Καταρχήν, έστω ότι τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ έχουν συλλεχθεί ανεξάρτητα. Με δεδομένα τα $\underset{\sim}{x}_i$, η Y_i έχει συνάρτηση πιθανοφάνειας

$$f_i(g(\underset{\sim}{x}_i'\underset{\sim}{\beta}), y_i),$$

όπου g είναι μια γνωστή συνάρτηση σύνδεσης. Έστω ότι

$$l_i = \log f_i$$

είναι ο λογάριθμος της πιθανοφάνειας του Y_i. Οπότε μπορούμε να ορίσουμε την ποινικοποιημένη πιθανοφάνεια ως

$$\sum_{i=1}^{n} l_i(g(\underset{\sim}{x}_i'\underset{\sim}{\beta}), y_i) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|).$$

Η μεγιστοποίηση της ως άνω συνάρτησης, είναι ισοδύναμη με την ελαχιστοποίηση της

$$-\sum_{i=1}^{n} l_i(g(\underset{\sim}{x}_i'\underset{\sim}{\beta}), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (1.6.3.3)$$

ως προς $\underset{\sim}{\beta}$. Αν αυτό γίνει για κάποια οριακή παράμετρο λ, θα πάρουμε τον ποινικοποιημένο εκτιμητή μέγιστης πιθανοφάνειας (*penalized maximum likelihood estimator*).

1.6.3.1. Δειγματοληπτικές και προβλεπτικές ιδιότητες

Σε αυτήν την ενότητα θα αναπτύξουμε την ασυμπτωτική θεωρία του μη κοίλου εκτιμητή ποινικοποιημένης πιθανοφάνειας. Έστω

$$\underset{\sim}{\beta}_0 = (\beta_{10},...,\beta_{d0})' = (\underset{\sim}{\beta}'_{10}, \underset{\sim}{\beta}'_{20})'.$$

Χωρίς βλάβη της γενικότητας, θεωρούμε ότι

$$\underset{\sim}{\beta}_{20} = \underset{\sim}{0}.$$

Έστω ότι $I(\underset{\sim}{\beta}_0)$ είναι ο πίνακας πληροφορίας του Fisher (*Fisher information matrix*) και έστω $I_1(\underset{\sim}{\beta}_{10}, \underset{\sim}{0})$ η πληροφορία κατά Fisher, γνωρίζοντας ότι $\underset{\sim}{\beta}_{20} = \underset{\sim}{0}$. Αρχικά θα δείξουμε ότι υπάρχει ένας εκτιμητής ποινικοποιημένης πιθανοφάνειας που συγκλίνει στο

$$O_p(n^{-1/2} + \alpha_n)$$

όπου

$$\alpha_n = \max\left\{ p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0 \right\} \qquad (1.6.3.1.1).$$

Αυτό σημαίνει ότι για τις *Hard* και *SCAD* συναρτήσεις ποινής, ο εκτιμητής ποινικοποιημένης πιθανοφάνειας είναι $\sqrt{n}$-συνεπής (*root-nconsistent*) αν $\lambda_n \to 0$. Επιπλέον θα δείξουμε ότι για τον εκτιμητή αυτόν πρέπει να ισχύει ότι

$$\hat{\underset{\sim}{\beta}}_2 = \underset{\sim}{0}$$

και ότι το $\hat{\underset{\sim}{\beta}}_1$ είναι ασυμπτωτικά της κανονικής κατανομής με πίνακα συνδιασποράς I_1^{-1}, αν $n^{1/2}\lambda_n \to \infty$. Αυτό συνεπάγεται ότι ο εκτιμητής ποινικοποιημένης πιθανοφάνειας συμπεριφέρεται τόσο καλά όσο αν ήταν γνωστό ότι $\underset{\sim}{\beta}_{20} = \underset{\sim}{0}$.

Αυτή η προβλεπτική συμπεριφορά του εκτιμητή σχετίζεται άμεσα με το φαινόμενο υπερ-αποδοτικότητας, (*superefficiency phenomenon*). Έστω το απλούστερο γραμμικό μοντέλο παλινδρόμησης

$$\underset{\sim}{Y} = \underset{\sim}{1}_n \mu + \underset{\sim}{\varepsilon} ,$$

όπου

$$\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, I_n) .$$

Ένας υπερ-αποδοτικός εκτιμητής για το μ είναι

$$\delta_n = \begin{cases} \bar{Y}, & |\bar{Y}| \geq n^{-1/4} \\ c\bar{Y}, & |\bar{Y}| < n^{-1/4} \end{cases} .$$

Αν θέσουμε $c = 0$, τότε το δ_n συμπίπτει με τον *Hard* εκτιμητή με παράμετρο $\lambda_n = n^{-1/4}$. Αυτός ο εκτιμητής υπολογίζει ακριβώς την παράμετρο στο 0 χωρίς να την υπολογίζει σε οποιοδήποτε άλλο σημείο.

Ας γενικεύσουμε τώρα το αποτέλεσμα, θεωρώντας ότι η ποινικοποίηση πραγματοποιείται σε κάθε συνιστώσα του $\underset{\sim}{\beta}$. Η περίπτωση όπου κάποιες συνιστώσες δεν ποινικοποιούνται, όπως για παράδειγμα η διασπορά στο γραμμικό μοντέλο, δεν παρουσιάζει κάποιο πρόβλημα. Έστω λοιπόν

$$\underset{\sim}{V}_i = (\underset{\sim}{X}_i, Y_i) \text{, με } i = 1, \ldots, n$$

και ότι $L(\underset{\sim}{\beta})$ είναι ο λογάριθμος της πιθανοφάνειας των παρατηρήσεων $\underset{\sim}{V}_1, \ldots, \underset{\sim}{V}_n$. Έστω επίσης ότι

$$Q(\underset{\sim}{\beta}) = L(\underset{\sim}{\beta}) - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) ,$$

είναι η ποινικοποιημένη συνάρτηση πιθανοφάνειας. Θα αναφέρουμε στη συνέχεια τα σχετικά θεωρήματα και λήμματα των Fan και Li (2001), αλλά πρωτίστως θα αναφέρουμε κάποιες απαραίτητες υποθέσεις κανονικότητας (*regularity conditions*):

(A) Οι παρατηρήσεις $\underset{\sim}{V}_i$ είναι i.i.d. με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V}, \underset{\sim}{\beta})$. Η $f(\underset{\sim}{V}, \underset{\sim}{\beta})$ έχει μια κοινή βάση και το μοντέλο είναι αναγνωρίσιμο (*identifiable*). Επίσης, η πρώτη και η δεύτερη λογαριθμημένη παράγωγος της f ικανοποιεί τις εξισώσεις

$$E_{\underset{\sim}{\beta}}\left[\frac{\partial \log f(\underset{\sim}{V}, \underset{\sim}{\beta})}{\partial \beta_j}\right] = 0 \text{, για } j = 1, \ldots, d$$

και

$$I_{jk}(\underset{\sim}{\beta}) = E_{\underset{\sim}{\beta}}\left[\frac{\partial}{\partial \beta_j}\log f(\underset{\sim}{V}, \underset{\sim}{\beta})\frac{\partial}{\partial \beta_k}\log f(\underset{\sim}{V}, \underset{\sim}{\beta})\right] = E_{\underset{\sim}{\beta}}\left[-\frac{\partial^2}{\partial \beta_j \partial \beta_k}\log f(\underset{\sim}{V}, \underset{\sim}{\beta})\right].$$

(B) Ο πίνακας πληροφορίας του Fisher

$$I(\underset{\sim}{\beta}) = E\left\{\left[\frac{\partial}{\partial \underset{\sim}{\beta}}\log f(\underset{\sim}{V}, \underset{\sim}{\beta})\right]\left[\frac{\partial}{\partial \underset{\sim}{\beta}}\log f(\underset{\sim}{V}, \underset{\sim}{\beta})\right]'\right\}$$

είναι πεπερασμένος και θετικά ορισμένος στο $\underset{\sim}{\beta} = \underset{\sim}{\beta}_0$.

(C) Υπάρχει ένα ανοικτό υποσύνολο ω του Ω το οποίο περιέχει την πραγματική παράμετρο $\underset{\sim}{\beta}_0$ τέτοιο ώστε για σχεδόν όλα τα $\underset{\sim}{V}$, η συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V}, \underset{\sim}{\beta})$ επιδέχεται τις παραγώγους τρίτης τάξης

$$\frac{\partial f(\underset{\sim}{V}, \underset{\sim}{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_l} \text{, για όλα τα } \underset{\sim}{\beta} \in \omega.$$

Επίσης, υπάρχουν συναρτήσεις M_{jkl} τέτοιες ώστε

$$\left|\frac{\partial^3}{\partial\beta_j\partial\beta_k\partial\beta_l}\log f(\underset{\sim}{V},\underset{\sim}{\beta})\right| \le M_{jkl}(\underset{\sim}{V}) \text{, για όλα τα } \underset{\sim}{\beta} \in \omega,$$

όπου $m_{jkl} = E_{\underset{\sim}{\beta}_0}[M_{jkl}] < \infty$, $\forall\, j,k,l$.

Θεώρημα 1.6.3.1.1. Έστω ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ είναι i.i.d. (*independent and identically distributed*), κάθε ένα με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V},\underset{\sim}{\beta})$ και ότι ικανοποιούν τις παραπάνω υποθέσεις (A)-(C). Αν

$$\max\left\{|\, p''_{\lambda_n}(|\,\beta_{j0}\,|)\,|: \beta_{j0} \ne 0\right\} \to 0,$$

τότε υπάρχει ένα τοπικό μέγιστο $\hat{\underset{\sim}{\beta}}$ του $Q(\underset{\sim}{\beta})$ τέτοιο ώστε

$$\|\,\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}_0\,\| = O_p(n^{-1/2} + \alpha_n),$$

με το α_n να δίνεται από την (1.6.3.1.1). Από το θεώρημα αυτό είναι προφανές ότι με μια σωστή επιλογή του λ_n θα υπάρξει ένας $\sqrt{n}$-συνεπής ποινικοποιημένος εκτιμητής. Θα δείξουμε τώρα ότι ο εκτιμητής αυτός έχει την ιδιότητα της σποραδικότητας $\hat{\underset{\sim}{\beta}}_2 = \underset{\sim}{0}$.

Λήμμα 1.6.3.1.1. Έστω πάλι ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ είναι i.i.d., κάθε ένα με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V},\underset{\sim}{\beta})$ και ότι ικανοποιούν τις υποθέσεις (A)-(C). Έστω ότι

$$\liminf_{n\to\infty} \liminf_{\theta\to 0_+} p'_{\lambda_n}(\theta)/\lambda_n > 0 \qquad (1.6.3.1.2).$$

Αν $\lambda_n \to 0$ και $\sqrt{n}\lambda_n \to \infty$ όσο το $n \to \infty$, τότε με πιθανότητα που τείνει στο 1, για κάθε δοσμένο $\underset{\sim}{\beta}_1$ που ικανοποιεί

$$\| \underset{\sim}{\beta}_1 - \underset{\sim}{\beta}_{10} \| = O_p(n^{-1/2})$$

και για κάθε σταθερά C, ισχύει ότι

$$Q\left\{\begin{pmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{0} \end{pmatrix}\right\} = \max_{\|\underset{\sim}{\beta}_2\| \le Cn^{-1/2}} Q\left\{\begin{pmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_2 \end{pmatrix}\right\}.$$

Ορίζουμε τώρα ως

$$\Sigma = diag\left\{ p''_{\lambda_n}(|\beta_{10}|,...,(|\beta_{s0}|)\right\}$$

και

$$\underset{\sim}{b} = \left(p'_{\lambda_n}(|\beta_{10}|)\operatorname{sgn}(\beta_{10}),..., p'_{\lambda_n}(|\beta_{s0}|)\operatorname{sgn}(\beta_{s0})\right)'.$$

Θεώρημα 1.6.3.1.2 (Προβλεπτική ιδιότητα)

Θεωρούμε ξανά ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ είναι i.i.d, κάθε ένα με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V},\underset{\sim}{\beta})$ και ότι ικανοποιούν τις υποθέσεις (A)-(C). Έστω επίσης ότι η συνάρτηση ποινής $p_{\lambda_n}(|\theta|)$ ικανοποιεί τη συνθήκη (1.6.3.1.2). Αν $\lambda_n \to 0$ και $\sqrt{n}\lambda_n \to \infty$ όσο το $n \to \infty$, τότε με πιθανότητα που τείνει στο 1, οι $\sqrt{n}$-συνεπείς εκτιμητές $\hat{\underset{\sim}{\beta}} = \begin{pmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{pmatrix}$, του Θεωρήματος 1.6.3.1.1, πρέπει να ικανοποιούν τα παρακάτω:

- Σποραδικότητα (*sparsity*):

$$\hat{\underset{\sim}{\beta}}_2 = \underset{\sim}{0}.$$

- Ασυμπτωτική κανονικότητα (*assymptotic normality*):

$$\sqrt{n}\left(I_1(\underset{\sim}{\beta}_{10}) + \Sigma\right)\left\{\hat{\underset{\sim}{\beta}}_1 - \underset{\sim}{\beta}_{10} + \left(I_1(\underset{\sim}{\beta}_{10}) + \Sigma\right)^{-1}\underset{\sim}{b}\right\} \to \mathrm{N}\left\{\underset{\sim}{0}, I_1(\underset{\sim}{\beta}_{10})\right\},$$

όπου $I_1(\underset{\sim}{\beta}_{10}) = I_1(\underset{\sim}{\beta}_{10}, \underset{\sim}{0})$ η πληροφορία κατά Fisher, γνωρίζοντας ότι $\underset{\sim}{\beta}_2 = \underset{\sim}{0}$.

Συνεπώς, ο ασυμπτωτικός πίνακας συνδιασποράς του $\hat{\underset{\sim}{\beta}}_1$ είναι

$$\frac{1}{n}\left\{I_1(\underset{\sim}{\beta}_{10})+\Sigma\right\}^{-1} I_1(\underset{\sim}{\beta}_{10})\left\{I_1(\underset{\sim}{\beta}_{10})+\Sigma\right\}^{-1},$$

και για τις συναρτήσεις ποινής που αναπτύχθηκαν στην ενότητα 1.6.2, είναι προσεγγιστικά ίσος με

$$\frac{1}{n} I_1^{-1}(\underset{\sim}{\beta}_{10}) \text{ αν το } \lambda_n \to 0 .$$

Πρέπει να σημειωθεί ότι για τις *SCAD* και *Hard* συναρτήσεις ποινής, αν $\lambda_n \to 0$ τότε $\alpha_n = 0$. Οπότε βάσει του Θεωρήματος 1.6.3.1.2, όταν $\sqrt{n}\lambda_n \to \infty$, οι αντίστοιχοι εκτιμητές ποινικοποιημένης πιθανοφάνειας έχουν την προβλεπτική ιδιότητα (*oracle property*) και συμπεριφέρονται τόσο καλά όσο και οι εκτιμητές μέγιστης πιθανοφάνειας, όσον αφορά την εκτίμηση του $\underset{\sim}{\beta}_1$, δεδομένου ότι $\underset{\sim}{\beta}_2 = \underset{\sim}{0}$. Παρόλα αυτά, για την L_1 συνάρτηση ποινής, ισχύει ότι $\alpha_n = \lambda_n$. Οπότε, η $\sqrt{n}$-συνέπεια απαιτεί $\lambda_n = O_p(n^{-1/2})$. Όμως, η προβλεπτική ιδιότητα του Θεωρήματος 1.6.3.1.2 απαιτεί $\sqrt{n}\lambda_n \to \infty$. Οι δύο αυτές συνθήκες για τη *LASSO* δεν ικανοποιούνται ταυτόχρονα. Συνεπώς, δεν ισχύει η προβλεπτική ιδιότητα για την L_1 συνάρτηση ποινής. Αντιθέτως, για την L_q συνάρτηση ποινής, με $q<1$, η προβλεπτική ιδιότητα ισχύει αν έχουμε επιλέξει το σωστό λ_n.

Θα συνεχίσουμε, κάνοντας μια αναφορά περί των συνθηκών κανονικότητας (A)-(C), όσον αφορά τα γενικευμένα γραμμικά μοντέλα. Με μια *canonical link*, η κατανομή του Y δεδομένου ότι $\underset{\sim}{X} = \underset{\sim}{x}$, ανήκει στην *canonical* εκθετική οικογένεια, που θα μελετήσουμε παρακάτω, με συνάρτηση πυκνότητας πιθανότητας

$$f(y,\underset{\sim}{x},\underset{\sim}{\beta}) = c(y)\exp\left\{\frac{y\underset{\sim}{x}'\underset{\sim}{\beta} - b(\underset{\sim}{x}'\underset{\sim}{\beta})}{\alpha(\varphi)}\right\}.$$

Προφανώς, η συνθήκη (A) ικανοποιείται. Ο πίνακας πληροφορίας του Fisher είναι

$$I(\underset{\sim}{\beta}) = E\left\{b''(\underset{\sim}{x}'\underset{\sim}{\beta})\underset{\sim}{x}\underset{\sim}{x}'\right\} / \alpha(\varphi) .$$

Οπότε αν το $E\left\{b''(\underset{\sim}{x}'\underset{\sim}{\beta})\underset{\sim}{x}\underset{\sim}{x}'\right\}$ είναι πεπερασμένο και θετικά ορισμένο, τότε ισχύει και η συνθήκη (B). Επίσης, αν για όλα τα $\underset{\sim}{\beta}$ σε κάποια γειτονιά του $\underset{\sim}{\beta}_0$, ισχύει ότι

$$|b^{(3)}(\underset{\sim}{x}'\underset{\sim}{\beta})| \leq M_0(\underset{\sim}{x})$$

για κάποια συνάρτηση $M_0(\underset{\sim}{x})$ που ικανοποιεί

$$E_{\underset{\sim}{\beta}_0}\left\{M_0(\underset{\sim}{x})X_j X_k X_l\right\} < \infty \ \forall\, j,k,l\,,$$

τότε ισχύει και η συνθήκη (C). Για γενικότερες συναρτήσεις σύνδεσης, παρόμοιες υποθέσεις πρέπει να ικανοποιούνται ώστε να ισχύουν οι συνθήκες (A)-(C). Τα αποτελέσματα των Θεωρημάτων 1.6.3.1.1 και 1.6.3.1.2 μπορούν να προκύψουν και για τις περιπτώσεις των ποινικοποιημένων ελαχίστων τετραγώνων (1.6.3.1) και της ποινικοποιημένης εύρωστης γραμμικής παλινδρόμησης (1.6.3.2).

1.6.3.2. Ο προτεινόμενος αλγόριθμος

Είναι δύσκολο να λυθεί το πρόβλημα ποινικοποιημένης πιθανοφάνειας όταν η συνάρτηση ποινής p_λ είναι μη κυρτή. Παρόλα αυτά, οι Fan και Lv (2009) είναι σε θέση να δώσουν τις συνθήκες κάτω από τις οποίες ο εκτιμητής μέγιστης πιθανοφάνειας υπάρχει και είναι μοναδικός.

Ο Tibshirani (1996) πρότεινε έναν αλγόριθμο για την επίλυση του προβλήματος ελαχίστων τετραγώνων της $LASSO$, ενώ ο Fu (1998) πρότεινε έναν “shooting” αλγόριθμο για την μέθοδο $LASSO$. Οι Fan και Li (2001) πρότειναν έναν αλγόριθμο με ενιαία και αποτελσματική τοπική τετραγωνική προσεγγίση (*local quadratic approximations-LQA*) για τη βελτιστοποίηση ποινικοποιημένης πιθανοφάνειας. Ο πρώτος όρος των (1.6.3.1), (1.6.3.2) και (1.6.3.3) μπορεί να θεωρηθεί ως μια συνάρτηση απώλειας (*loss function*) του $\underset{\sim}{\beta}$. Ας την ονομάσουμε $l(\underset{\sim}{\beta})$. Οπότε οι (1.6.3.1), (1.6.3.2) και (1.6.3.3) μπορούν να γραφούν σε μια ενιαία μορφή ως

$$l(\underset{\sim}{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

Οι συναρτήσεις ποινής L_1 και $SCAD$ είναι ιδιάζουσες στην αρχή και δεν έχουν συνεχείς παραγώγους δεύτερης τάξης. Παρόλα αυτά, μπορούν να προσεγγισθούν τοπικά από μια τετραγωνική συνάρτηση ως ακολούθως: Για δοσμένες αρχικές συνθήκες $\beta^* = \left(\beta_1^*, ..., \beta_p^*\right)^T$, η συνάρτηση ποινής p_λ μπορεί να προσεγγιστεί τοπικά από μια τετραγωνική συνάρτηση όπως

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^*|) + \frac{1}{2}\left\{\frac{p'_\lambda(|\beta_j^*|)}{|\beta_j^*|}\right\}\left(\beta_j^2 - (\beta_j^*)^2\right), \qquad (1.6.3.2.1)$$

για $\beta_j \approx \beta_j^*$.

Με αυτό και με μία LQA στη log-πιθανοφάνεια, η ποινικοποιημένη πιθανοφάνεια μετατρέπεται σε ένα πρόβλημα ελαχίστων τετραγώνων που δέχεται μια λύση κλειστού τύπου. Για να αποφευχθεί η αριθμητική αστάθεια, θέτει τον εκτιμημένο συντελεστή $\hat{\beta}_j = 0$ αν το β_j^* είναι πολύ κοντά στο 0, το οποίο συνεπάγεται με διαγραφή της j-οστής μεταβλητής από το τελικό μοντέλο. Είναι σαφές ότι η τιμή 0 είναι μια κατάσταση απορρόφησης της LQA με την έννοια ότι μόλις ένας συντελεστής συρρικνωθεί στο 0, θα παραμένει στο 0 και στις επόμενες επαναλήψεις. Το οποίο αποτελεί και το μόνο μειονέκτημα αυτής της προσέγγισης.

Η LQA ενημερώνει την τετραγωνική προσέγγιση σε κάθε βήμα κατά τη διάρκεια της επανάληψης, η οποία επιταχύνει τη σύγκλιση του αλγορίθμου.

Το σχήμα 1.6.2.1 δείχνει τις συναρτήσεις ποινής L_1 και $SCAD$ καθώς και τις προσεγγίσεις τους βάσει της (1.6.3.2.1), για δύο διαφορετικές τιμές του β_j^*. Αν τώρα η $l(\underset{\sim}{\beta})$ είναι η L_1 συνάρτηση απώλειας, όπως στην (1.6.3.2), τότε δεν έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $\underset{\sim}{\beta}$. Παρόλα αυτά, η ποσότητα $\psi(|y - \underset{\sim}{x}'\underset{\sim}{\beta}|)$ στην (1.6.3.2) μπορεί κατά ανάλογο τρόπο να προσεγγισθεί από την

$$\left\{\psi(y - \underset{\sim}{x}'\underset{\sim}{\beta}_0)/(y - \underset{\sim}{x}'\underset{\sim}{\beta}_0)^2\right\}(y - \underset{\sim}{x}'\underset{\sim}{\beta})^2,$$

αρκεί η αρχική τιμή $\underset{\sim}{\beta}_0$ του $\underset{\sim}{\beta}$ να είναι αρκετά κοντά στην τιμή ελαχιστοποίησης. Όταν κάποια από τα υπόλοιπα $|y - \underset{\sim}{x}'\underset{\sim}{\beta}_0|$ είναι μικρά, η προσέγγιση αυτή δεν είναι καλή.

Υποθέτουμε στη συνέχεια ότι ο λογάριθμος της πιθανοφάνειας έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $\underset{\sim}{\beta}$. Συνεπώς, είναι εφικτό ο πρώτος όρος της (1.6.3.2.1) να προσεγγισθεί από μια τετραγωνική συνάρτηση. Οπότε, το πρόβλημα ελαχιστοποίησης (1.6.3.2.1) μπορεί να υποβιβασθεί σε ένα τετραγωνικό πρόβλημα ελαχιστοποίησης (*quadratic minimization problem*) και ο αλγόριθμος Newton-Raphson μπορεί να χρησιμοποιηθεί. Πράγματι, η (1.6.3.2.1) προσεγγίζεται (εκτός από έναν σταθερό όρο) από την ποσότητα

$$l(\underset{\sim}{\beta}_0) + \nabla l(\underset{\sim}{\beta}_0)'(\underset{\sim}{\beta} - \underset{\sim}{\beta}_0) + \frac{1}{2}(\underset{\sim}{\beta} - \underset{\sim}{\beta}_0)'\nabla^2 l(\underset{\sim}{\beta}_0)(\underset{\sim}{\beta} - \underset{\sim}{\beta}_0) + \frac{1}{2}n\underset{\sim}{\beta}'\sum_\lambda(\underset{\sim}{\beta}_0)\underset{\sim}{\beta} \qquad (1.6.3.2.3),$$

όπου

$$\nabla l(\underset{\sim}{\beta}_0) = \frac{\partial l(\underset{\sim}{\beta}_0)}{\partial \underset{\sim}{\beta}},$$

$$\nabla^2 l(\underset{\sim}{\beta}_0) = \frac{\partial^2 l(\underset{\sim}{\beta}_0)}{\partial \underset{\sim}{\beta}\partial \underset{\sim}{\beta}'},$$

$$\sum_\lambda(\underset{\sim}{\beta}_0) = diag\left\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\right\}.$$

Το τετραγωνικό πρόβλημα ελαχιστοποίησης (1.6.3.2.3), έχει ως λύση την

$$\hat{\underset{\sim}{\beta}}_1 = \hat{\underset{\sim}{\beta}}_0 - \left\{\nabla^2 l(\underset{\sim}{\beta}_0) + n\sum_\lambda(\underset{\sim}{\beta}_0)\right\}^{-1}\left\{\nabla l(\underset{\sim}{\beta}_0) + n\sum_\lambda(\underset{\sim}{\beta}_0)\underset{\sim}{\beta}_0\right\}.$$

Όταν επέλθει σύγκλιση του αλγορίθμου, ο εκτιμητής ικανοποιεί τη συνθήκη

$$\frac{\partial l(\hat{\underset{\sim}{\beta}}_0)}{\partial \beta_j} + np'_\lambda(|\hat{\beta}_{j0}|)\,\mathrm{sgn}(\hat{\beta}_{j0}) = 0,$$

η οποία αποτελεί την εξίσωση ποινικοποιημένης πιθανοφάνειας, για τα μη μηδενικά στοιχεία του $\hat{\underset{\sim}{\beta}}_0$. Συγκεκριμένα, για το πρόβλημα ποινικοποιημένων ελαχίστων

τετραγώνων (1.6.3.1), η λύση βρίσκεται με επαναληπτικό (*iterative*) υπολογισμό της παλινδρόμησης κορυφογραμμής

$$\underset{\sim}{\beta}_1 = \left\{ \underset{\sim}{X}'\underset{\sim}{X} + n\sum\nolimits_{\lambda}(\underset{\sim}{\beta}_0) \right\}^{-1} \underset{\sim}{X}'\underset{\sim}{Y} .$$

Ομοίως, η λύση της (1.4.3.2) προκύπτει με επαναληπτικό υπολογισμό της

$$\underset{\sim}{\beta}_1 = \left\{ \underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{X} + \frac{1}{2} n\sum\nolimits_{\lambda}(\underset{\sim}{\beta}_0) \right\}^{-1} \underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{Y} ,$$

όπου

$$\underset{\sim}{W} = diag\left\{ \psi(|y_1 - \underset{\sim}{x}_1'\underset{\sim}{\beta}_0|)/(y_1 - \underset{\sim}{x}_1'\underset{\sim}{\beta}_0)^2, ..., \psi(|y_n - \underset{\sim}{x}_n'\underset{\sim}{\beta}_0|)/(y_n - \underset{\sim}{x}_n'\underset{\sim}{\beta}_0)^2 \right\}.$$

Όπως και στην περίπτωση του εκτιμητή μέγιστης πιθανοφάνειας, έχοντας μια καλή αρχική τιμή $\underset{\sim}{\beta}_0$, η μονοβηματική διαδικασία μπορεί να είναι εξίσου αποδοτική όσο και η πλήρως επαναληπτική διαδικασία όπου παίρνουμε τον εκτιμητή ποινικοποιημένης πιθανοφάνειας, κάνοντας χρήση του αλγόριθμου Newton-Raphson. Αν τώρα θεωρήσουμε ως $\underset{\sim}{\beta}^{(k-1)}$ μια καλή αρχική τιμή στο k βήμα, ο επόμενος επαναληπτικός υπολογισμός μπορεί να θεωρηθεί ως μονοβηματική διαδικασία, άρα ο προκύπτων εκτιμητής εξακολουθεί να μπορεί να είναι το ίδιο αποδοτικός όσο αυτός που θα προέκυπτε με την πλήρως επαναληπτική μέθοδο (για λεπτομέρειες σχετικά με τη διαφορά του εκτιμητή μέγιστης πιθανοφάνειας και των εκτιμητών *k*-βημάτων, (Robinson, 1988)). Συμπερασματικά, ο εκτιμητής που θα προκύψει με τον αλγόριθμο που αναφέραμε κάνοντας λίγες επαναλήψεις, μπορεί να θεωρηθεί ως εκτιμητής ενός βήματος και θα έχει την ίδια απόδοση. Οπότε βάσει αυτού του σκεπτικού, δεν χρειάζεται να επαναλάβουμε τον αλγόριθμο μέχρι να επέλθει σύγκλιση, αρκεί οι αρχικές εκτιμήσεις να είναι καλές. Ως αρχικές εκτιμήσεις τώρα, μπορούν να δοθούν αυτές του πλήρους μοντέλου, αρκεί να μην είναι υπερβολικά παραμετροποιημένες.

1.6.3.3. Υπολογισμός του τυπικού σφάλματος

Τα τυπικά σφάλματα των εκτιμηθέντων παραμέτρων μπορούν άμεσα να υπολογισθούν, λόγω του ότι γίνεται ταυτόχρονη εκτίμηση παραμέτρων και επιλογή μεταβλητών. Ο *sandwich* τύπος μπορεί να χρησιμοποιηθεί για την εκτίμηση της συνδιασποράς του $\hat{\underset{\sim}{\beta}}_1$, η μη εξαφανισμένη συνιστώσα του $\hat{\underset{\sim}{\beta}}$. Οπότε έχουμε,

$$\operatorname{cov}(\hat{\underset{\sim}{\beta}}_1) = \left\{\nabla^2 l(\hat{\underset{\sim}{\beta}}_1) + n\sum\nolimits_{\lambda}(\hat{\underset{\sim}{\beta}}_1)\right\}^{-1} \operatorname{cov}\left\{\nabla l(\hat{\underset{\sim}{\beta}}_1)\right\}\left\{\nabla^2 l(\hat{\underset{\sim}{\beta}}_1) + n\sum\nolimits_{\lambda}(\hat{\underset{\sim}{\beta}}_1)\right\}^{-1} \qquad (1.6.3.3.1).$$

Ο τύπος αυτός είναι αρκετά ακριβής και για μέτρια μεγέθη δειγμάτων.

Όταν χρησιμοποιείται η L_1 συνάρτηση απώλειας στην εύρωστη παλινδρόμηση, πρέπει να πραγματοποιηθούν κάποιες τροποποιήσεις στον αλγόριθμο καθώς επίσης και στον αντίστοιχο *sandwich* τύπο. Στην περίπτωση όπου $\psi(x) = |x|$, τα διαγώνια στοιχεία του $\underset{\sim}{W}$ είναι

$$\{|r_i|^{-1}\}, \text{ με } r_i = y_i - \underset{\sim}{x}_i{}'\underset{\sim}{\beta}_0 \text{ και } i = 1,\ldots,n.$$

Οπότε για μια δοθείσα τιμή του $\underset{\sim}{\beta}_0$, όταν κάποια από τα υπόλοιπα $\{r_i\}$ είναι κοντά στο 0, αυτά τα σημεία αποκτούν πολύ βάρος. Για αυτό το λόγο αντικαθίσταται το βάρος με

$$(\alpha_n + |r_i|^{-1}).$$

Στις εφαρμογές που έκαναν οι Fan και Li, χρησιμοποίησαν ως α_n το $2n^{-1/2}$ *quantile* των απολύτων τιμών των υπολοίπων, $\{|r_i|\}$. Οπότε το α_n άλλαζε σε κάθε επανάληψη.

1.6.3.4. Έλεγχος της σύγκλισης του αλγορίθμου

Οι Fan και Li απέδειξαν, με χρήση του προγράμματος MATLAB, ότι όντως ο αλγόριθμος που πρότειναν συγκλίνει στη σωστή λύση. Συγκεκριμένα, χρησιμοποίησαν ένα διάνυσμα $\underset{\sim}{\beta}$ διάστασης 100, αποτελούμενο από 50 μηδενικά

και 50 μη μηδενικά στοιχεία που δημιουργήθηκαν από την κατανομή $N(0,5^2)$. Επίσης χρησιμοποίησαν έναν 100×100 ορθοκανονικό πίνακα σχεδιασμού, λόγω του ότι τα ποινικοποιημένα ελάχιστα τετράγωνα (PLS) έχουν τότε μαθηματική λύση κλειστής μορφής, οπότε και ήταν εφικτή η σύγκρισή της με αυτήν της αλγοριθμικής μεθόδου τους. Το διάνυσμα των αποκρίσεων $\underset{\sim}{Y}$ δημιουργήθηκε βάσει του γραμμικού μοντέλου $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$. Τα αποτελέσματα ήταν τα εξής: Το MATLAB χρειάστηκε 0.27, 0.39 και 0.16 sec για να επέλθει σύγκλιση όσον αφορά τα PLS με τη $SCAD$, L_1 και $Hard$ συνάρτηση ποινής αντίστοιχα. Επίσης, ο αριθμός των επαναλήψεων ήταν 30, 30 και 5 αντίστοιχα. Να σημειωθεί, ότι στη δέκατη επανάληψη, ο PLS εκτιμητής ήταν ήδη αρκετά κοντά στη σωστή τιμή.

1.6.4. Αριθμητικές συγκρίσεις

Στην ενότητα αυτή, θα συγκρίνουμε την απόδοση των προτεινόμενων μεθόδων με τις ήδη υπάρχουσες και θα ελέγξουμε την ακρίβεια της μεθόδου εύρεσης του τυπικού σφάλματος. Επίσης θα αναφέρουμε και κάποιες μελέτες προσομοίωσης (*simulation studies*) που έκαναν οι Fan και Li χρησιμοποιώντας τις ποινικοποιημένες μεθόδους.

1.6.4.1. Σφάλμα πρόβλεψης και σφάλμα μοντέλου

Το σφάλμα πρόβλεψης (*prediction error*) ορίζεται ως το μέσο σφάλμα στην πρόβλεψη του Y, δεδομένου νέου $\underset{\sim}{x}$ (που προφανώς δεν χρησιμοποιήθηκε στην κατασκευή της εξίσωσης πρόβλεψης). Υπάρχουν δύο περιπτώσεις, το X να είναι τυχαίο (*random*) και το X να είναι ελεγχόμενο (*controlled*). Στην πρώτη περίπτωση, τόσο το Y όσο και το $\underset{\sim}{x}$ είναι τυχαία επιλεγμένα. Στην δεύτερη περίπτωση, ο πίνακας σχεδιασμού επιλέγεται από τους πειραματιστές και μόνο το Y είναι τυχαίο. Στο εξής θα θεωρούμε ότι το X είναι τυχαίο.

Σε αυτήν την περίπτωση, τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ θεωρούνται τυχαίο δείγμα από κάποια κατανομή. Τότε, αν $\hat{\mu}(\underset{\sim}{x})$ είναι η πρόβλεψη βάσει των δεδομένων που έχουμε στην κατοχή μας, το σφάλμα πρόβλεψης ορίζεται ως

$$PE(\hat{\mu}) = E\{Y - \hat{\mu}(\underset{\sim}{x})\}^2 .$$

Ο παραπάνω τύπος μπορεί να αναλυθεί ως

$$PE(\hat{\mu}) = E\{Y - E(Y \mid \underset{\sim}{x})\}^2 + E\{E(Y \mid \underset{\sim}{x}) - \hat{\mu}(\underset{\sim}{x})\}^2 .$$

Ο πρώτος όρος είναι το σφάλμα πρόβλεψης λόγω του θορύβου στα δεδομένα και ο δεύτερος λόγω της έλλειψης προσαρμογής (*lack of fit*) του μοντέλου. Αυτός ο δεύτερος όρος ονομάζεται σφάλμα μοντέλου (*model error*) και συμβολίζεται ως $ME(\hat{\mu})$. Να σημειώσουμε ότι αν $Y = \underset{\sim}{x}'\underset{\sim}{\beta} + e$, με $E(e \mid \underset{\sim}{x}) = 0$, τότε

$$ME(\hat{\mu}) = (\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})' E(\underset{\sim}{x}\underset{\sim}{x}')(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}) .$$

1.6.4.2. Επιλογή των οριακών παραμέτρων

Οι Fan και Li, προκειμένου να εκτιμήσουν τη ρυθμιστική (*tuning*) παράμετρο $\underset{\sim}{\theta}$, όπου $\underset{\sim}{\theta} = (\lambda, \alpha)$ για τη $SCAD$ συνάρτηση ποινής και τη $\underset{\sim}{\theta} = \lambda$ για τη $LASSO$ και $Hard$, χρησιμοποίησαν δύο μεθόδους. Την πενταπλή (*fivefold*) διασταυρωμένη επικύρωση και τη γενικευμένη διασταυρωμένη επικύρωση. Θα αναπτύξουμε τις δύο αυτές διαδικασίες για την περίπτωση των γραμμικών μοντέλων παλινδρόμησης. Η επέκταση των διαδικασιών αυτών σε εύρωστα γραμμικά μοντέλα παλινδρόμησης καθώς και γραμμικά μοντέλα βασισμένα στην πιθανοφάνεια, δεν εμπεριέχει ιδιαίτερες δυσκολίες.

Στη μέθοδο της πενταπλής διασταυρωμένης επικύρωσης, συμβολίζουμε ως T το σύνολο των δεδομένων και ως $T - T^{\nu}$ και T^{ν} το σύνολο εκπαίδευσης (*training set*) και το σύνολο ελέγχου (*test set*) αντίστοιχα, με $\nu = 1, \ldots, 5$. Για κάθε $\underset{\sim}{\theta}$ και ν, βρίσκουμε τον εκτιμητή $\hat{\underset{\sim}{\beta}}^{(\nu)}(\underset{\sim}{\theta})$ του $\underset{\sim}{\beta}$, χρησιμοποιώντας το σύνολο εκπαίδευσης $T - T^{\nu}$. Εν συνεχεία, εφαρμόζουμε το κριτήριο της διασταυρωμένης επικύρωσης

$$CV(\underset{\sim}{\theta}) = \sum_{\nu=1}^{5} \sum_{(y_k, \underset{\sim}{x}_k) \in T^{\nu}} \left\{ y_k - \underset{\sim}{x}_k' \hat{\underset{\sim}{\beta}}^{(\nu)}(\underset{\sim}{\theta}) \right\}^2$$

και βρίσκουμε το $\hat{\underset{\sim}{\theta}}$ που ελαχιστοποιεί το $CV(\underset{\sim}{\theta})$.

Στη μέθοδο της γενικευμένης διασταυρωμένης επικύρωσης, μετατρέπουμε τη λύση ως

$$\underset{\sim}{\beta}_1(\underset{\sim}{\theta}) = \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n \sum\nolimits_{\lambda} (\underset{\sim}{\beta}_0) \right\}^{-1} \underset{\sim}{X}' \underset{\sim}{Y}.$$

Οπότε η προσαρμοσμένη τιμή $\hat{\underset{\sim}{Y}}$ του $\underset{\sim}{Y}$ είναι

$$\underset{\sim}{X} \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n \sum\nolimits_{\lambda} (\underset{\sim}{\beta}_0) \right\}^{-1} \underset{\sim}{X}' \hat{\underset{\sim}{Y}}$$

και μπορούμε να θεωρήσουμε ως πίνακα προβολής τον

$$\underset{\sim}{P}_{\underset{\sim}{X}} \{ \hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta}) \} = \underset{\sim}{X} \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n \sum\nolimits_{\lambda} (\hat{\underset{\sim}{\beta}}) \right\}^{-1} \underset{\sim}{X}'.$$

Ορίζοντας τώρα το πλήθος των σημαντικών παραμέτρων στην προσαρμογή του ποινικοποιημένου μοντέλου ελαχίστων τετραγώνων ως

$$e(\underset{\sim}{\theta}) = tr[\underset{\sim}{P}_{\underset{\sim}{X}} \{ \hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta}) \}],$$

το κριτήριο της γενικευμένης διασταυρωμένης επικύρωσης είναι

$$GCV(\underset{\sim}{\theta}) = \frac{1}{n} \frac{\| \underset{\sim}{Y} - \underset{\sim}{X} \underset{\sim}{\beta}(\underset{\sim}{\theta}) \|^2}{\{1 - e(\underset{\sim}{\theta}) / n\}^2}$$

και

$$\hat{\underset{\sim}{\theta}} = \arg\min_{\theta} \{ GCV(\underset{\sim}{\theta}) \}.$$

1.7. Μια επεξήγηση της διακλιμάκωσης

Ας υποθέσουμε ότι για ένα συγκεκριμένο σύνολο δεδομένων έχουμε υπολογίσει τους συντελεστές παλινδρόμησης, βάσει της μεθόδου ελαχίστων τετραγώνων. Έστω ότι κάποιοι εξ αυτών έχουν τιμή γύρω στο 100 ενώ κάποιοι άλλοι γύρω στο 0.01. Τότε μια προσαύξηση κατά μια μονάδα στο πρώτο σύνολο συντελεστών θα έχει ελάχιστη επιρροή στην προσαρμογή ενώ η ίδια ενέργεια θα επιφέρει μεγάλη επίδραση στους συντελεστές με τις μικρές τιμές. Αν όμως προσαρμόσουμε τα ελάχιστα τετράγωνα βάζοντας ένα φράγμα στο άθροισμα των απολύτων τιμών των συντελεστών, τότε οι μεγάλοι συντελεστές θα συρρικνωθούν σημαντικά, ενώ οι μικροί θα παραμείνουν σχεδόν ανεπηρέαστοι. Οπότε και γίνεται μια διακλιμάκωση (*scaling*) στις τιμές των συντελεστών, κάτι που θα οδηγήσει σε καλύτερες εκτιμήσεις αυτών.

1.8. Μέθοδος LASSO

Το κίνητρο για τη δημιουργία της LASSO (*Least Absolute Shrinkage and Selection Operator*) προήλθε από μια ενδιαφέρουσα πρόταση του Breiman (1993). Έχει το πλεονέκτημα ότι προκαλεί τον μηδενισμό κάποιων συντελεστών κι έτσι δίνει ερμηνεύσιμα μοντέλα. Αυτό συμβαίνει διότι ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολοίπων, υπό τον περιορισμό το άθροισμα των απολύτων τιμών των συντελεστών να είναι μικρότερο από μια σταθερά. Με τις προσομοιώσεις που έκανε ο Tibshirani (1996), αποδείχθηκε ότι η *LASSO* παρουσιάζει παρόμοιες ιδιότητες με τη μέθοδο επιλογής καλύτερου υποσυνόλου, όσον αφορά την ερμηνεία των μοντέλων που προκύπτουν καιπαρουσιάζει την ίδια σταθερότητα όπως η μέθοδος της παλινδρόμησης διασέλου. Η LASSO είναι εφαρμόσιμη σε πολλά μοντέλα. Βρίσκει εφαρμογή και στα γενικευμένα γραμμικά μοντέλα.

1.8.1. Περιγραφή της μεθόδου LASSO

Θεωρούμε ότι έχουμε τα δεδομένα $(\mathrm{x}^i, \mathrm{y}_i)$, i=1,2,…,N, όπου $\underset{\sim}{\mathrm{x}}^i = (\mathrm{x}_{i1},...,\mathrm{x}_{ip})^T$ είναι οι ανεξάρτητες μεταβλητές και y_i οι αποκρίσεις. Ως συνήθως, υποθέτουμε ότι είτε οι παρατηρήσεις είναι ανεξάρτητες ή ότι οι αποκρίσεις y_i είναι υπό συνθήκη ανεξάρτητες από τα x_{ij} που μας δόθηκαν. Επίσης, υποθέτουμε ότι τα x_{ij} είναι κανονικοποιημένα, έτσι ώστε $\frac{\sum_i x_{ij}}{N} = 0$ και $\frac{\sum_i x_{ij}^{\ 2}}{N} = 1$.

Έστω $\underset{\sim}{\hat{\beta}} = (\hat{\beta}_1,...,\hat{\beta}_p)^T$, τότε οι εκτιμητές *LASSO* ορίζονται ως εξής:

$$(\hat{\alpha}, \underset{\sim}{\hat{\beta}}) = \mathrm{argmin}\left\{\sum_{i=1}^{N}(\mathrm{y}_i - \alpha - \sum_j \beta_j x_{ij})^2\right\}, \qquad (1.8.1.1)$$

με τον περιορισμό ότι $\sum_j |\beta_j| \le t$, όπου $t \ge 0$.

Η t ονομάζεται ρυθμιστική παράμετρος (*tuning parameter*).

Έτσι για όλα τα t ισχύει ότι $\hat{\alpha} = \bar{y}$. Μπορούμε να κάνουμε απαλοιφή του α, υποθέτοντας χωρίς βλάβη της γενικότητας ότι $\bar{y} = 0$.

Ο υπολογισμός της λύσης της εξίσωσης $(\hat{\alpha}, \underset{\sim}{\hat{\beta}}) = \mathrm{argmin}\left\{\sum_{i=1}^{N}(\mathrm{y}_i - \alpha - \sum_j \beta_j x_{ij})^2\right\}$ είναι ένα πρόβλημα τετραγωνικού προγραμματισμού με γραμμικούς ανισωτικούς περιορισμούς.

Η παράμετρος t ελέγχει το μέγεθος της συρρίκνωσης, στο οποίο υπόκεινται οι συντελεστές. Έστω $\hat{\beta}_j^0$ είναι οι εκτιμητές ελαχίστων τετραγώνων και $t_0 = \sum_j |\hat{\beta}_j^0|$, τότε οι τιμές των *t*, οι οποίες είναι μικρότερες από t_0 θα προκαλέσει συρρίκνωση των λύσεων προς το μηδέν και ορισμένοι από τους συντελεστές μπορεί να γίνουν ακριβώς μηδέν. Για παράδειγμα, αν $t = \frac{t_0}{2}$, τότε το αποτέλεσμα θα είναι σχεδόν ίδιο

με το να βρούμε το καλύτερο υποσύνολο μεγέθους $\frac{p}{2}$. Επίσης, αξίζει να σημειωθεί ότι ο πίνακας σχεδιασμού δεν είναι απαραίτητο να είναι πλήρους βαθμού.

1.8.2. Η περίπτωση του ορθοκανονικού πίνακα σχεδιασμού

Καλύτερη εικόνα σχετικά με τη φύση της συρρίκνωσης μπορεί να μας δώσει ο ορθοκανονικός πίνακας σχεδιασμού. Έστω $\underset{\sim}{X}$ ένας $n \times p$ πίνακας σχεδιασμού, ο οποίος είναι ορθοκανονικός, δηλαδή ισχύει ότι $\underset{\sim}{X}^T \underset{\sim}{X} = I$. Οι λύσεις της εξίσωσης $(\hat{\alpha}, \hat{\underset{\sim}{\beta}}) = \text{argmin}\left\{\sum_{i=1}^{N}(\text{y}_i - \alpha - \sum_j \beta_j x_{ij})^2\right\}$ είναι

$$\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+,$$

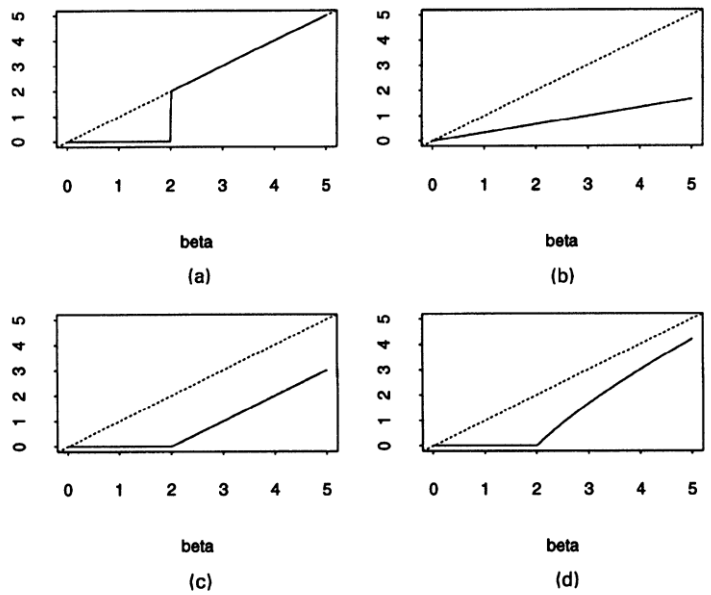
όπου το γ καθορίζεται από τη σχέση $\sum |\beta_j| = t$.

Στην περίπτωση ορθοκανονικού πίνακα σχεδιασμού, η μέθοδος επιλογής καλύτερου υποσυνόλου μεγέθους k, επιλέγει τους k μεγαλύτερους συντελεστές κατά απόλυτη τιμή και ταυτόχρονα θέτει τους υπόλοιπους ίσους με μηδέν. Για κάποια επιλογή του λ, αυτό είναι ισοδύναμο με το να θέσουμε $\hat{\beta}_j = \hat{\beta}_j^0$ όταν $|\hat{\beta}_j^0| > \lambda$ και $\hat{\beta}_j = 0$ όταν $|\hat{\beta}_j^0| \le \lambda$.

Η παλινδρόμηση διασέλου (*ridge regression*) ελαχιστοποιεί την ποσότητα $\sum_{i=1}^{N}(\text{y}_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum \beta_j^2$, ή ισοδύναμα ελαχιστοποιεί την $\sum_{i=1}^{N}(\text{y}_i - \sum_j \beta_j x_{ij})^2$, δεδομένου ότι $\sum \beta_j^2 \le t$.

Οι λύσεις (*ridge solutions*) που προκύπτουν είναι $\frac{1}{1+\gamma}\hat{\beta}_j^0$, όπου το γ εξαρτάται από το λ ή το t.

Στο παρακάτω σχήμα , βλέπουμε τη μορφή των λύσεων αυτών.

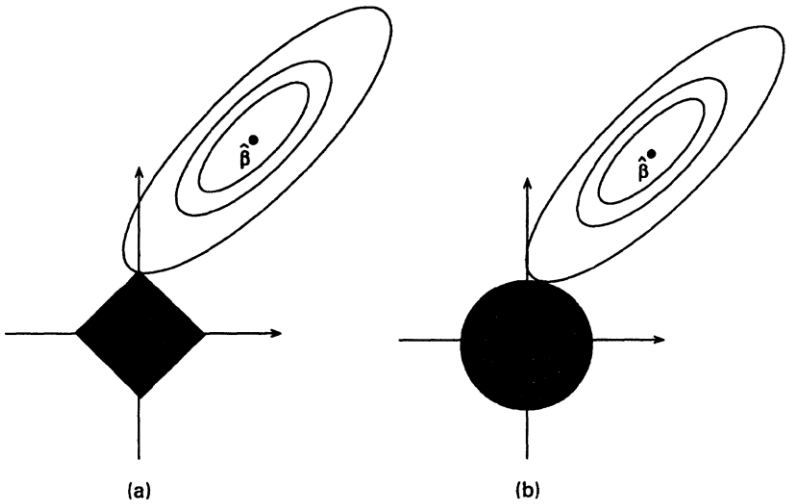


<u>Σχήμα.1.8.2.1</u>: (a) Μέθοδος επιλογής καλύτερου υποσυνόλου, (b) παλινδρόμηση διασέλου, (c) μέθοδος LASSO και (d) μέθοδος garrote : —— μορφή της συρρίκνωσης των συντελεστών στην περίπτωση του ορθοκανονικού πίνακα σχεδιασμού.

1.8.3. Η γεωμετρία της LASSO

Είναι φανερό από το σχήμα 1.8.2.1 για ποιο λόγο η μέθοδος LASSO δίνει συχνά συντελεστές με τιμή ακριβώς μηδέν. Τα ερωτήματα, όμως, που τίθενται είναι γιατί αυτό συμβαίνει και στη γενική περίπτωση κατά την οποία ο πίνακας σχεδιασμού δεν είναι ορθογώνιος, καθώς επίσης για ποιο λόγο η *ridge regression* (παλινδρόμηση διασέλου), η οποία χρησιμοποιεί τον περιορισμό $\sum \beta_j^2 \leq t$, έχει την ίδια ιδιότητα.

Η εξήγηση προκύπτει κατά κάποιο τρόπο από το Σχήμα 1.8.3.1, στην περίπτωση κατά την οποία p=2.



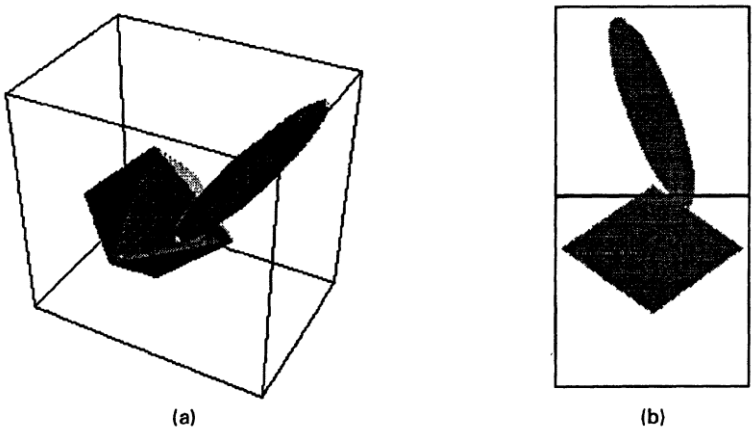
<u>Σχήμα 1.8.3.1</u>: (a) Γεωμετρία της LASSO για p=2, (b) γεωμετρία της παλινδρόμησης διασέλου για p=2.

Το κριτήριο $\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2$ ισοδυναμεί με την τετραγωνική συνάρτηση $(\underset{\sim}{\beta} - \hat{\underset{\sim}{\beta}}^0)^T(\underset{\sim}{X}^T \underset{\sim}{X})(\underset{\sim}{\beta} - \hat{\underset{\sim}{\beta}}^0)$ (συν μια σταθερά).

Είναι κεντραρισμένα στους *OLS* εκτιμητές. Επίσης, η περιοριστική περιοχή είναι ο χρωματισμένος ρόμβος. Η λύση που δίνει η *LASSO*, είναι το πρώτο σημείο όπου τα περιγράμματα ακουμπούν το ρόμβο. Αυτό κάποιες φορές συμβαίνει σε κάποια γωνία, οπότε έχουμε και την περίπτωση μηδενικού συντελεστή. Στο σχήμα Η(b) φαίνεται η εναλλακτική περίπτωση της παλινδρόμησης διασέλου, όπου δεν υπάρχουν γωνίες για να συναντήσουν οι ελλείψεις, οπότε και είναι αρκετά δύσκολο να προκύψουν μηδενικοί συντελεστές.

Συνεχίζουμε, θέτοντας το εξής ερώτημα: Υπάρχει περίπτωση τα πρόσημα των εκτιμηθέντων συντελεστών της μεθόδου *LASSO* να διαφέρουν από αυτά των *OLS* εκτιμητών $\hat{\beta}_j^0$; Από τη στιγμή που είναι κανονικοποιημένες οι μεταβλητές, για την περίπτωση όπου $p = 2$, οι κύριοι άξονες των ελλείψεων είναι σε κλίση $\pm 45^o$ ως προς τους άξονες συντεταγμένων και μπορεί να δειχθεί ότι οι ελλείψεις πρέπει να συναντήσουν το ρόμβο στο τεταρτημόριο που περιέχει το $\hat{\underset{\sim}{\beta}}^0$. Αν παρόλα αυτά το $p > 2$ και υπάρχει έστω και μια μέτρια συσχέτιση στα δεδομένα, αυτό δεν είναι

σίγουρο ότι ισχύει. Στο επόμενο σχήμα I φαίνεται ένα παράδειγμα στις τρεις διαστάσεις. Με τη βοήθεια της κάτοψης (b) βλέπουμε ότι η έλλειψη ακουμπά την περιοριστική περιοχή σε οκτημόριο διαφορετικό από αυτό όπου βρίσκεται το κέντρο της. Οπότε, παρόλο που η μέθοδος garotte διατηρεί το πρόσημο κάθε $\hat{\beta}_j^0$, αυτό είναι δυνατόν να μην ισχύει για τη *LASSO*.



<u>Σχήμα I</u>: (a) Περίπτωση όπου η LASSO εκτιμήτρια βρίσκεται σε οκτημόριο διαφορετικό από εκείνο της εκτιμήτριας ελαχίστων τετραγώνων, (b) κάτοψη.

1.8.4. Τυπικά σφάλματα και εκτίμηση της παραμέτρου t.

Οι εκτιμητές *LASSO* είναι μια μη γραμμική και μη διαφορίσιμη συνάρτηση των τιμών της απόκρισης. Αυτό έχει ως αποτέλεσμα να είναι δύσκολη η ακριβής εκτίμηση των τυπικών σφαλμάτων τους. Ένας τρόπος αντιμετώπισης του προβλήματος αυτού, είναι με χρήση της μεθόδου *bootstrap*: είτε μπορούμε να θεωρήσουμε το t σταθερό, είτε μπορούμε να βελτιστοποιήσουμε ως προς t για κάθε δείγμα *bootstrap*. Σταθεροποίηση του t, ουσιαστικά σημαίνει να επιλέξουμε το καλύτερο υποσύνολο και έπειτα να χρησιμοποιήσουμε το τυπικό σφάλμα των ελαχίστων τετραγώνων για το υποσύνολο αυτό.

Μια προσεγγιστική κλειστής μορφής εκτίμηση μπορεί να αποκτηθεί γράφοντας τη συνάρτηση ποινής $\sum_j |\beta_j|$ ως $\sum_j \beta_j^2 / |\beta_j|$. Συνεπώς, βάσει της

LASSO εκτιμήτριας $\hat{\underset{\sim}{\beta}}$, μπορούμε να προσεγγίσουμε τη λύση με χρήση παλινδρόμησης διασέλου (*ridge regression*) της μορφής

$$\underset{\sim}{\beta}^* = (\underset{\sim}{X}'\underset{\sim}{X} + \lambda \underset{\sim}{W}^-)^{-1}\underset{\sim}{X}'\underset{\sim}{Y},$$

όπου $\underset{\sim}{W}$ είναι ο διαγώνιος πίνακας με στοιχεία $|\hat{\beta}_j|$, ο $\underset{\sim}{W}^-$ είναι ο γενικευμένος αντίστροφος του $\underset{\sim}{W}$ και το λ έχει επιλεγεί κατά τρόπον ώστε $\sum_j |\beta_j|^* = t$. Ο πίνακας συνδιασποράς των εκτιμητών μπορεί τότε να προσεγγισθεί από την ποσότητα

$$(\underset{\sim}{X}'\underset{\sim}{X} + \lambda \underset{\sim}{W}^-)^{-1}\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X} + \lambda \underset{\sim}{X}^-)^{-1}\hat{\sigma}^2,$$

όπου $\hat{\sigma}^2$ είναι η εκτίμηση της διασποράς του σφάλματος. Η δυσκολία της μεθοδολογίας αυτής είναι ότι δίνει εκτιμώμενη διασπορά ίση με 0 για τις μεταβλητές με $\hat{\beta}_j = 0$.

Όσον αφορά τώρα την εκτίμηση της παραμέτρου t, ο Tibshirani χρησιμοποίησε τρεις μεθόδους. Αυτές είναι η διασταυρωμένη επικύρωση, η γενικευμένη διασταυρωμένη επικύρωση και έναν αναλυτικό αμερόληπτο εκτιμητή ρίσκου του Stein. Αξίζει να σημειωθεί ότι οι δύο πρώτες μέθοδοι είναι εφαρμόσιμες στην περίπτωση που ο πίνακας σχεδιασμού είναι τυχαίος, οπότε και υποθέτουμε ότι οι παρατηρήσεις $(\underset{\sim}{X}, \underset{\sim}{Y})$ έχουν δημιουργηθεί από μια άγνωστη κατανομή. Η τελευταία μέθοδος προϋποθέτει συγκεκριμένο $\underset{\sim}{X}$. Παρόλα αυτά, σε πραγματικά προβλήματα, συχνά δεν υπάρχει καθαρή διάκριση μεταξύ των δύο περιπτώσεων και απλά προτιμάμε τη βολικότερη μέθοδο.

1.8.5. Adaptive LASSO

Η LASSO στο Tibshirani (1996) έχει χρησιμοποιηθεί ευρέως εξαιτίας της κυρτότητάς της. Όμως, δημιουργεί μεροληψία εκτίμησης. Το πρόβλημα αυτό επισημάνθηκε από τους Fan & Li (2001) και επισήμως φαίνεται από το Zou (2006) ακόμη και σε μια πεπερασμένη ρύθμιση παραμέτρων. Για να ξεπεραστεί αυτό το

πρόβλημα μεροληψίας, ο Zou (2006) πρότεινε μια προσαρμοσμένη LASSO και ο Meinshausen (2007) πρότεινε μια «χαλαρή» (relaxed) LASSO.

Η ιδέα στο Zou (2006) είναι να χρησιμοποιηθεί μια προσαρμοσμένη σταθμισμένη l_1- ποινή στο ποινικοποιημένο πρόβλημα ελαχίστων τετραγώνων. Ειδικότερα, εισήγαγε τον όρο της ποινικοποίησης

$$\lambda \sum_{j=1}^{d} \omega_j \left| \beta_j \right|,$$

όπου $\lambda \geq 0$ είναι μια παράμετρος κανονικοποίησης και $\omega = \left(\omega_1, ..., \omega_d\right)^T$ είναι ένα γνωστό σταθμισμένο διάνυσμα. Πρότεινε επίσης τη χρήση του σταθμισμένου διανύσματος $\hat{\omega} = 1 \big/ \left|\hat{\beta}\right|^{\gamma}$, όπου $\gamma \geq 0$, η δύναμη είναι κατανοητή κατά συνιστώσες και $\hat{\beta}$ είναι μια ρίζα n συνεχούς εκτιμήτριας. Ωστόσο, το μηδέν είναι μια απορροφητική κατάσταση της προσαρμοσμένης LASSO.

Η περίπτωση $\gamma = 1$ είναι στενά συνδεδεμένη με τη μη-αρνητική garrote στο Breiman (1995). Ο Zou (2006) έδειξε επίσης ότι η προσαρμοσμένη LASSO μπορεί να λυθεί από τον αλγόριθμο LARS, ο οποίος είχε προταθεί στο Efron et al. (2004). Χρησιμοποιώντας την ίδια πεπερασμένη παράμετρο set-up όπως και στους Knight & Fu (2000), Zou (2006) διαπιστώθηκε ότι η προσαρμοσμένη LASSO έχει τις oracle ιδιότητες καθώς η παράμετρος συντονισμού επιλέγεται με τέτοιο τρόπο ώστε $\lambda / n^{1/2} \to 0$ και $\lambda n^{(\gamma-1)/2} \to \infty$ καθώς $n \to \infty$.

1.9. Dantzig selector

Σε πολλές σημαντικές στατιστικές εφαρμογές, ο αριθμός των μεταβλητών ή παραμέτρων p είναι πολύ μεγαλύτερος από τον αριθμό των παρατηρήσεων n. Για παράδειγμα, στη ραδιολογία και στη βιοϊατρική απεικόνιση υπάρχουν πολλές περιπτώσεις στις οποίες $p \gg n$. Υποθέτουμε ότι έχουμε παρατηρήσεις για τις οποίες ισχύει $Y = X\beta + z$, όπου $\beta \in \mathbb{R}$ είναι ένα διάνυσμα παραμέτρων, X είναι ένας

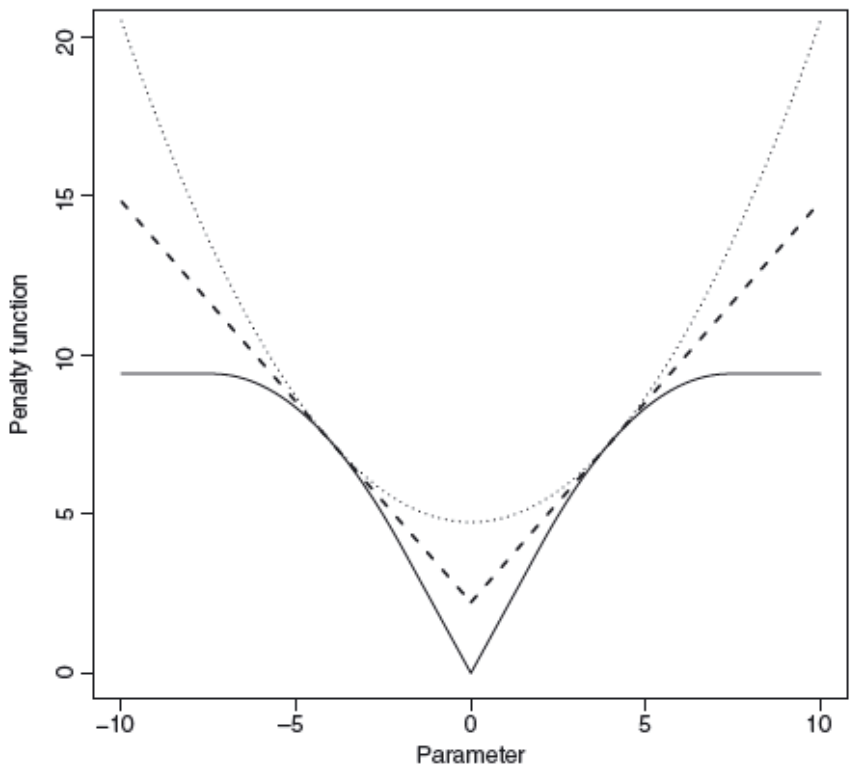
πίνακας με πολύ λιγότερες ενδεχομένως γραμμές από στήλες ($n \ll p$) και $z \sim N(0, \sigma^2 \mathrm{I}_n)$ είναι ένα διάνυσμα ανεξάρτητων κανονικών τυχαίων μεταβλητών. Για να υπολογίσουμε το β, εισάγουμε ένα νέο εκτιμητή, που προτάθηκε από τους Candes & Tao (2007) για να ανακτήσει ένα αραιό υψηλής διάστασης διάνυσμα παραμέτρων στο γραμμικό μοντέλο, ο οποίος λέγεται επιλογέας Dantzig (*Dantzig selector*) και είναι μια λύση του προβλήματος της l_1 κανονικοποίησης $\min_{\tilde{\beta} \in \mathbb{R}^p} \left\| \tilde{\beta} \right\|_{l_1}$ δεδομένου ότι $\left\| X_r^* \right\|_\infty \leq (1 + \mathrm{t}^{-1}) \sqrt{2 \log p} \cdot \sigma$ όπου r είναι το διάνυσμα υπολοίπων $Y - X\beta$ και t είναι μια θετική scalar.

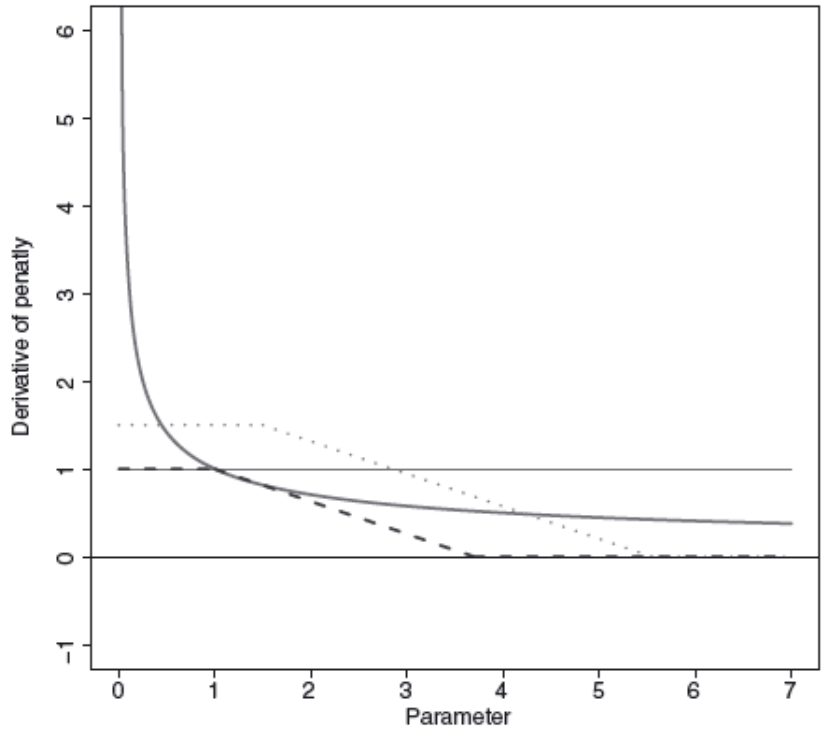
Έστω ότι το X υπακούει σε μια ενιαία αρχή αβεβαιότητας και η πραγματική παράμετρος β είναι αρκετά αραιή, τότε με πολύ μεγάλη πιθανότητα ισχύει ότι:

$$\left\| \hat{\beta} - \beta \right\|_{l_2}^2 \leq C^2 \cdot 2 \log p \cdot \left(\sigma^2 + \sum_i \min(\beta_i^2, \sigma^2) \right)$$

Τα αποτελέσματά μας είναι μη ασυμπτωτικά και δίνουμε τιμές για σταθερό C. Ακόμη και αν το n είναι πολύ μικρότερο του p, ο εκτιμητής μας επιτυγχάνει μια απώλεια σε ένα λογαριθμικό παράγοντα του ιδανικού μέσου τετραγωνικού σφάλματος θα επιτευχθεί με ένα *oracle* το οποίο παρέχει πλήρη πληροφόρηση σχετικά με το ποιες συμμεταβλητές είναι μη μηδενικές και η οποία ήταν πάνω από το επίπεδο θορύβου.

Στην πολυμεταβλητή παλινδρόμηση και από την άποψη της επιλογής μοντέλου, το αποτέλεσμα μας λέει ότι είναι δυνατόν να επιλεγεί το καλύτερο υποσύνολο μεταβλητών με την επίλυση ενός πολύ απλού κυρτού προγράμματος (*convex program*) το οποίο στην πραγματικότητα μπορεί εύκολα να αναδιατυπωθεί σαν ένα βολικό γραμμικό πρόγραμμα.





<u>Σχήμα 1.9.1</u>. (a) SCAD ποινή (——) και οι τοπικές γραμμικές (- - - -) και τετραγωνικές $(\cdots)$ προσεγγίσεις και (b) $p'_\lambda(\cdot)$ για την ποινικοποιημένη L_1 (——), SCAD με λ=1 (- - - -) και λ=1.5 $(\cdots)$ και προσαρμοσμένη LASSO (——) με γ=0.5.

1.10. Συμπεράσματα

Οι μέθοδοι που πρότειναν οι Fan και Li, αποδεδειγμένα έχουν πολύ καλή απόδοση όσον αφορά την επιλογή σημαντικών μεταβλητών. Ο *sandwich* τύπος που κατασκεύασαν για την εκτίμηση των τυπικών σφαλμάτων είναι επίσης αρκετά αποτελεσματικός και ο αλγόριθμος υλοποίησης της όλης μεθόδου υποστηρίζεται από στατιστική θεωρία, με αποτέλεσμα οι εκτιμητές που κατασκευάζονται να έχουν καλές στατιστικές ιδιότητες. Σε σύγκριση με τη μέθοδο επιλογής καλύτερου υποσυνόλου, η οποία είναι αρκετά χρονοβόρα, οι νέες μέθοδοι δίνουν αποτελέσματα αρκετά πιο γρήγορα. Το μεγάλο πλεονέκτημά τους είναι η ταυτόχρονη επιλογή σημαντικών μεταβλητών και η εκτίμηση των συντελεστών, κάτι που γίνεται βελτιστοποιώντας μια ποινικοποιημένη πιθανοφάνεια. Αυτό έχει ως αποτέλεσμα και την ακριβή εκτίμηση των τυπικών σφαλμάτων. Επίσης, απέδειξαν ότι η συνάρτηση ποινής *SCAD*, έχει την καλύτερη απόδοση στην επιλογή σημαντικών μεταβλητών, χωρίς να δημιουργείται μεροληψία, εν αντιθέσει με τη *LASSO* μέθοδο του Tibshirani (1996) όπου χρησιμοποιείται η L_1 συνάρτηση ποινής. Η όλη διαδικασία της ποινικοποιημένης πιθανοφάνειας μπορεί εύκολα να επεκταθεί και σε άλλα πεδία της Στατιστικής, όπως η Ανάλυση Επιβίωσης.

ΚΕΦΑΛΑΙΟ 2ο

ΣΙΓΟΥΡΟ ΑΝΕΞΑΡΤΗΤΟ ΚΡΗΣΑΡΙΣΜΑ

2.1. Εισαγωγή

Πολλές τεχνικές που συχνά χρησιμοποιούνται για την επιλογή μεταβλητών βασίζονται στο ανεξάρτητο κρησάρισμα. Η βασική ιδέα είναι να εφαρμόσουμε μια γρήγορη αλλά ακατέργαστη μέθοδο μείωσης της διάστασης σε ένα μέτριο μέγεθος (συνήθως κάτω από το μέγεθος του δείγματος) που θα κρησάρει τις μεταβλητές που έχουν ασθενή συσχέτιση με τη μεταβλητή απόκρισης, και μετά θα δείξουμε ότι οι κρησαρισμένες μεταβλητές είναι όντως ασυσχέτιστες, με πιθανότητα που τείνει στο 1. Μετά από αυτό, εφαρμόζουμε μια πιο εκλεπτυσμένη και σχετικά χαμηλής διάστασης τεχνική, όπως μια ποινικοποιημένη μέθοδος πιθανοφάνειας που βασίζεται στο ομαλό «ψαλίδισμα» της απόλυτης απόκλισης (SCAD) ποινής μπορεί να εφαρμοστεί για να εκτελέσει την τελική επιλογή χαρακτηριστικών και εκτίμηση παραμέτρων ταυτόχρονα. Το ανεξάρτητο κρησάρισμα επιστρατεύει εκείνα τα χαρακτηριστικά που έχουν την καλύτερη οριακή χρησιμότητα, το οποίο αντιστοιχεί στη μεγαλύτερη οριακή συσχέτιση με την απόκριση στο πλαίσιο της παλινδρόμησης ελαχίστων τετραγώνων. Αυτό μειώνει συγχρόνως τη διάσταση και επιβεβαιώνει την ακρίβεια εκτίμησης. Για τα προβλήματα μεγάλης διάστασης, παρουσιάστηκε το «σίγουρο» κρησάρισμα (sure screening) και προτάθηκε μια μέθοδος πάνω σε αυτό, που βασίζεται στη μάθηση συσχέτισης και καλείται **Sure Independence Screening (SIS)**. Η μέθοδος αυτή χρησιμοποιείται για να μειώσουμε τη διάσταση από υψηλή σε μια μέτριας κλίμακας που είναι μικρότερη από το μέγεθος του δείγματος. Υπολογίζουμε το βαθμό στον οποίο η διάσταση μπορεί να μειωθεί από τη μέθοδο του ανεξάρτητου κρησαρίσματος, που εξαρτάται από τις αλληλεπιδράσεις του πίνακα συνδιακύμανσης των συμμεταβλητών και των πραγματικών παραμέτρων.

Ωστόσο, η SIS διαδικασία στους Fan & Lv (2008) περιορίζει μόνο τα απλά γραμμικά μοντέλα και τα τεχνικά τους επιχειρήματα εξαρτώνται σε μεγάλο βαθμό από τις κοινές υποθέσεις κανονικότητας και δεν μπορούν εύκολα να επεκταθούν ακόμη και μέσα στο πλαίσιο του γραμμικού μοντέλου. Αυτό περιορίζει σημαντικά τη χρήση της στην πράξη που αποκλείει τις κατηγορικές μεταβλητές. Οι Huang, Horowitz & Ma (2008) ερεύνησαν ακόμα την οριακή *bridge* παλινδρόμηση στο κανονικό γραμμικό μοντέλο και τα επιχειρήματά τους εξαρτώνται επίσης σε μεγάλο βαθμό από τις ρητές εκφράσεις των ελαχίστων τετραγωνικών εκτιμητών και *bridge* παλινδρόμηση.

Το ανεξάρτητο κρησάρισμα χρησιμοποιείται ευρέως για την επιλογή γονιδίων ή την κατάταξη των ασθενειών στη βιοπληροφορική. Σε αυτές τις εφαρμογές, τα γονίδια ή οι πρωτεΐνες καλούνται *στατιστικά σημαντικές* αν οι εκφράσεις που σχετίζονται στην ομάδα θεραπείας διαφέρουν σημαντικά από την ομάδα ελέγχου, με αποτέλεσμα μια μεγάλη και ενεργή βιβλιογραφία σχετική με πολλά προβλήματα δοκιμών. Τα επιλεγμένα χαρακτηριστικά χρησιμοποιούνται συχνά για την κατάταξη των ασθενειών (Tibshirani et al. (2003) και Fan & Ren (2006)). Αυτή η μέθοδος κρησαρίσματος είναι πράγματι μια μορφή ανεξάρτητου κρησαρίσματος και έχει «δικαιολογηθεί» από τους Fan & Fan (2008) κάτω από κάποιες ιδανικές καταστάσεις. Είναι εύκολο να κατασκευαστούν χαρακτηριστικά τα οποία είναι οριακά ασύνδετα, αλλά σχετίζονται από κοινού με την απόκριση. Αυτά τα χαρακτηριστικά θα προβληθούν από μεθόδους ανεξάρτητης μάθησης, όπως η t-δοκιμή δύο δειγμάτων. Με άλλα λόγια, τα γονίδια τα οποία έχουν προβληθεί έξω από τις στατιστικές δοκιμές μπορούν πράγματι να είναι σημαντικά στην κατάταξη των ασθενειών και την κατανόηση μοριακών μηχανισμών της νόσου. Άλλα παραδείγματα περιλαμβάνουν την κατάταξη οριακών συσχετίσεων (Fan & Lv, 2008) ή την επιλογή χαρακτηριστικών χρησιμοποιώντας ένα δείγμα δύο t-test σε υψηλής διάστασης ταξινόμηση (Tibshirani et al, 2003). Μέσα στο πλαίσιο του γραμμικού μοντέλου, οι Fan & Lv (2008) έδειξαν ότι η διαδικασία της απλής κατάταξης συσχετίσης κατέχει την ιδιότητα του σίγουρου ανεξάρτητου κρησαρίσματος υπό ορισμένες προϋποθέσεις (στο πλαίσιο του γραμμικού μοντέλου με κανονικές συμμεταβλητές και αποκρίσεις) και μια σημαντική μεθοδολογική επέκτασή της, που καλείται Επαναληπτικό Σίγουρο

Ανεξάρτητο Κρησάρισμα (ISIS), καλύπτει περιπτώσεις όπου οι συνθήκες κανονικότητας μπορεί να αποτύχουν, για παράδειγμα αν ένα χαρακτηριστικό είναι οριακά ασυσχέτιστο, αλλά σχετίζεται από κοινού με την απόκριση, ή την αντίστροφη κατάσταση όπου ένα χαρακτηριστικό είναι από κοινού ασυσχέτιστο αλλά έχει υψηλότερη οριακή συσχέτιση από κάποια σημαντικά χαρακτηριστικά. Χονδρικά, η ISIS λειτουργεί με διαδοχικά επαναλαμβανόμενη εκτέλεση επιλογής χαρακτηριστικών για να επιστρατεύσει ένα μικρό αριθμό χαρακτηριστικών, υπολογίζοντας τα υπόλοιπα με βάση το μοντέλο εφοδιασμένο με αυτά τα προσλαμβανόμενα χαρακτηριστικά, και στη συνέχεια, χρησιμοποιώντας τα υπόλοιπα που έχουν δουλέψει ως μεταβλητή απόκρισης για να συνεχίσει την πρόσληψη νέων χαρακτηριστικών. Το κρίσιμο βήμα είναι να υπολογίσουμε τα υπόλοιπα που δουλεύουν, το οποίο είναι εύκολο για το πρόβλημα παλινδρόμησης ελαχίστων τετραγώνων αλλά δεν είναι προφανές για άλλα προβλήματα. Η βελτιωμένη απόδοση της ISIS έχει τεκμηριωθεί στους Fan & Lv (2008).

Θα επεκτείνουμε, λοιπόν, την ISIS και θα προτείνουμε μια γενικότερη εκδοχή της ανεξάρτητης μάθησης με την κατάταξη των εκτιμητών της μέγιστης οριακής πιθανοφάνειας ή την ίδια τη μέγιστη οριακή πιθανοφάνεια στα γενικευμένα γραμμικά μοντέλα ως ειδική περίπτωση. Ακόμη και στο σύνολο των ελαχίστων τετραγώνων, η νέα μέθοδος βελτιώνει την ISIS επιτρέποντας τη διαγραφή χαρακτηριστικών στην επαναληπτική διαδικασία. Η τεχνική μάς επιτρέπει να επιλέξουμε τα σημαντικά χαρακτηριστικά σε υψηλής διάστασης ταξινόμηση όπου η δύο δειγμάτων *t*-μέθοδος που χρησιμοποιείται ευρέως αποτυγχάνει. Μια νέα τεχνική εισάγεται για να μειώσει το ποσοστό εσφαλμένης επιλογής κατά τη φάση του κρησαρίσματος χαρακτηριστικών. Σε ένα αρκετά γενικό πλαίσιο, η μάθηση συσχέτισης φαίνεται ότι έχει την ιδιότητα του σίγουρου κρησαρίσματος ακόμα και για διάσταση που αυξάνεται εκθετικά. Έτσι, οι συνθήκες κάτω από τις οποίες η ανεξάρτητη μάθηση διαθέτει την ιδιότητα του σίγουρου κρησαρίσματος είναι εκπληκτικά απλές. Αυτό δικαιολογεί την εφαρμογή μιας τέτοιας απλής μεθόδου σε ένα ευρύ φάσμα.

Μια ιδιαίτερα δημοφιλής οικογένεια των μεθόδων βασίζεται στη μέθοδο των ποινικοποιημένων ελαχίστων τετραγώνων ή γενικότερα στην ποινικοποιημένη ψευδο-πιθανοφάνεια, που αναλύσαμε στο πρώτο κεφάλαιο. Παραδείγματα

περιλαμβάνουν τη LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), Dantzig selector (Candes & Tao, 2007), και των σχετικών μεθόδων τους (Zou (2006), Zou & Li (2008)). Θεωρητικές μελέτες αυτών των μεθόδων επικεντρώνονται στην επιμονή (Greenshtein & Ritov (2004), van de Geer (2008)), στη συνέπεια και στις ιδιότητες Oracle (Fan & Li (2001), Zou (2006)). Αυτές οι μέθοδοι έχουν προσελκύσει ένα μεγάλο αριθμό θεωρητικής μελέτης και αλγοριθμικής ανάπτυξης πρόσφατα. Ωστόσο, ο υπολογισμός που συνδέεται με αυτές τις μεθόδους καθιστά δύσκολο να εφαρμοστεί άμεσα στα υπερυψηλής διάστασης στατιστικά προβλήματα μάθησης, τα οποία συνεπάγονται τις ταυτόχρονες προκλήσεις της υπολογιστικής σκοπιμότητας, στατιστικής ακρίβειας και αλγοριθμικής σταθερότητας (Fan, Samworth & Wu, (2009)).

Ο πρώτος στόχος είναι να επεκτείνουμε τις SIS και ISIS σε πολύ πιο γενικά μοντέλα. Μια πρόκληση εδώ είναι να δημιουργήσουμε τον κατάλληλο ορισμό των υπολοίπων (*residuals*) σε αυτό το πλαίσιο. Περιγράφουμε μια διαδικασία που παρακάμπτει αποτελεσματικά αυτό το θέμα και ως εκ τούτου επιτρέπει την επιθυμητή επέκταση της ISIS. Στην πραγματικότητα, η μέθοδός μας βελτιώνει ακόμα και την αρχική ISIS των Fan & Lv (2008), υπό την έννοια ότι επιτρέπει τη διαγραφή μεταβλητής στη διαδικασία πρόσληψης. Η μεθοδολογία μας εφαρμόζεται σε ένα πολύ γενικό πλαίσιο ψευδο-πιθανοφάνειας, στο οποίο ο στόχος είναι να βρούμε το διάνυσμα παραμέτρων $\beta = (\beta_1, ..., \beta_p)^T$ το οποίο είναι αραιό και ελαχιστοποιεί την αντικειμενική συνάρτηση της μορφής

$$Q(\beta_0, \beta) = \sum_{i=1}^{n} L(\mathrm{Y}_i, \beta_0 + x_i^T \beta),$$

όπου τα (x_i^T, Y_i) είναι το διάνυσμα συμμεταβλητών και απόκρισης για το *i*-οστό στοιχείο. Σημαντικές εφαρμογές αυτής της μεθοδολογίας περιλαμβάνουν τα ακόλουθα:

1. Γενικευμένα Γραμμικά Μοντέλα: Όλα τα γενικευμένα γραμμικά μοντέλα, συμπεριλαμβανομένης της λογιστικής παλινδρόμησης και των Poisson λογαριθμικών-γραμμικών μοντέλων, ταιριάζουν πολύ φυσικά στο μεθοδολογικό μας πλαίσιο και για αυτό θα τα μελετήσουμε στο επόμενο κεφάλαιο. Σημειώνουμε ότι τα μοντέλα λογιστικής παλινδρόμησης αποδίδουν μια δημοφιλή προσέγγιση για τη

μελέτη των προβλημάτων ταξινόμησης. Παρακάτω θα παρουσιάσουμε προσομοιώσεις στις οποίες η προσέγγισή μας συγκρίνεται ευνοϊκά με την ανταγωνιστική τεχνική LASSO (Tibshirani, 1996).

2. Ταξινόμηση. Άλλες κοινές προσεγγίσεις για την ταξινόμηση υποθέτουν ότι η απόκριση παίρνει τιμές στο διάστημα $\{-1,1\}$ και ταιριάζει επίσης στο πλαίσιό μας. Για παράδειγμα, υποστηρίζουμε ότι το διάνυσμα ταξινομητών μηχανής (Vapnik, 1995) χρησιμοποιεί τη συνάρτηση απώλειας $\mathrm{L}(\mathrm{Y}_i, \beta_0 + x_i^T\beta) = \{1 - \mathrm{Y}_i(\beta_0 + x_i^T\beta)\}_+$, ενώ ο ενισχυμένος αλγόριθμος Adaboost (Freud & Schapire, 1997) χρησιμοποιεί τη $\mathrm{L}(\mathrm{Y}_i, \beta_0 + x_i^T\beta) = \exp\{-\mathrm{Y}_i(\beta_0 + x_i^T\beta)\}$

3. Στιβαρή τοποθέτηση: Στον καθορισμό ενός υψηλής διάστασης γραμμικού μοντέλου, καλό είναι να είμαστε προσεκτικοί σχετικά με τη σχέση μεταξύ των χαρακτηριστικών και της απόκρισης. Έτσι, αντί της συμβατικής συνάρτησης απώλειας ελαχίστων τετραγώνων, μπορεί να προτιμήσουμε μια ισχυρή συνάρτηση απώλειας, όπως η L_1 απώλεια $\mathrm{L}(\mathrm{Y}_i, \beta_0 + x_i^T\beta) = \left|\mathrm{Y}_i - \beta_0 - x_i^T\beta\right|$ ή την απώλεια Huber (Huber, 1964), η οποία επίσης ταιριάζει στο πλαίσιό μας.

Οποιαδήποτε μέθοδος κρησαρίσματος, από προεπιλογή, έχει ένα μεγάλο ποσοστό εσφαλμένης επιλογής (*FSR*), δηλαδή πολλά ασήμαντα χαρακτηριστικά επιλέγονται μετά το κρησάρισμα. Ένας δεύτερος στόχος είναι να παρουσιάσουμε δύο παραλλαγές της μεθοδολογίας SIS, οι οποίες μειώνουν σημαντικά το *FSR*. Και οι δύο βασίζονται στο διαχωρισμό των δεδομένων σε (συνήθως) δύο ομάδες. Η πρώτη έχει την επιθυμητή ιδιότητα, ότι δηλαδή σε προβλήματα υψηλής διάστασης η πιθανότητα επιλογής εσφαλμένων ασήμαντων χαρακτηριστικών είναι μικρή. Έτσι, αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη ως μέσο γρήγορου εντοπισμού χαρακτηριστικών που πρέπει να συμπεριληφθούν στο τελικό μοντέλο. Η δεύτερη μέθοδος είναι λιγότερο επιθετική, και για το γραμμικό μοντέλο έχει την ίδια ιδιότητα σίγουρου κρησαρίσματος, όπως η πρωτότυπη τεχνική SIS. Οι εφαρμογές των προτεινόμενων μεθόδων μας απεικονίζονται παρακάτω.

2.2. Μέθοδος Σίγουρου Κρησαρίσματος: μάθηση συσχέτισης

Με τον όρο σίγουρο κρησάρισμα εννοούμε την ιδιότητα όπου όλες οι σημαντικές μεταβλητές επιβιώνουν μετά την εφαρμογή της διαδικασίας κρησαρίσματος μεταβλητών με πιθανότητα που τείνει στο 1. Η μέθοδος για τη μείωση της διάστασης είναι επιθυμητή αν έχει την ιδιότητα του σίγουρου κρησαρίσματος. Θα παρουσιάσουμε μια απλή μέθοδο σίγουρου κρησαρίσματος που χρησιμοποιεί τη μάθηση συσχέτισης. Εστιάζουμε σε κάθε μεταβλητή εισόδου έτσι ώστε η παρατηρούμενη μέση τιμή να είναι 0 και περιορίζουμε κάθε επεξηγηματική μεταβλητή έτσι ώστε η τυπική απόκλιση του δείγματος να είναι 1. Έστω $M_* = \{1 \le i \le p : \beta_i \ne 0\}$ είναι το πραγματικό αραιό μοντέλο με μη-αραιό μέγεθος $s = |M_*|$. Οι υπόλοιπες $\mathrm{p}-\mathrm{s}$ μεταβλητές μπορούν να συσχετιστούν με τη μεταβλητή απόκρισης μέσω της σύνδεσης με τις επεξηγηματικές μεταβλητές που περιέχονται στο μοντέλο.

Έστω $\omega = (\omega_1, ..., \omega_p)^T$ ένα p-διάστασης διάνυσμα που προκύπτει από την κατά συνιστώσες παλινδρόμηση

$$\omega = X^T y \tag{2.2.1}$$

όπου, με μικρή κατάχρηση της σημειογραφίας, τα $n \times p$ στοιχεία του πίνακα X, τυποποιούνται κατά στήλη. Ως εκ τούτου, το ω είναι πράγματι ένα διάνυσμα των οριακών συσχετίσεων των επεξηγηματικών μεταβλητών με τη μεταβλητή απόκρισης να αλλάζει μέγεθος λόγω της τυπικής απόκλισης της απόκρισης.

Για οποιοδήποτε δοσμένο $\gamma \in (0,1)$, ταξινομούμε τα p μεγέθη του φορέα του διανύσματος ω σε φθίνουσα σειρά και ορίζουμε το υπομοντέλο

$\mathrm{M}_\gamma = \{1 \le i \le p : |\omega_i|$ να είναι ανάμεσα στο πρώτο $[\gamma_n]$ που είναι το μεγαλύτερο από όλα$\}$ όπου $[\gamma_n]$ είναι το ακέραιο μέρος του γ_n.

Αυτός είναι ένας τρόπος για να συρρικνώσουμε ολόκληρο το μοντέλο $\{1, ..., \mathrm{p}\}$ στο υπομοντέλο M_γ με μέγεθος $d = [\gamma_n] < n$.

Μια τέτοια μάθηση συσχέτισης κατατάσσει τη σημαντικότητα των χαρακτηριστικών σύμφωνα με την οριακή συσχέτιση με τη μεταβλητή απόκρισης και φιλτράρει αυτές που έχουν αδύναμες οριακές συσχετίσεις με τη μεταβλητή απόκρισης. Αυτή τη μέθοδο κρησαρίσματος συσχέτισης την ονομάζουμε SIS, αφού κάθε χαρακτηριστικό χρησιμοποιείται ανεξάρτητα ως μια επεξηγηματική μεταβλητή ώστε να αποφασίσει πόσο χρήσιμο είναι να προβλεφθεί η μεταβλητή απόκρισης. Αυτή η έννοια είναι ευρύτερη από τη συσχέτιση κρησαρίσματος και εφαρμόζεται σε γενικευμένα γραμμικά μοντέλα, σε προβλήματα ταξινόμησης με διάφορες λειτουργίες απώλειας/ζημιάς .

Το υπολογιστικό κόστος της SIS ή μάθησης συσχέτισης εμφανίζεται όταν πολλαπλασιάζουμε ένα $p \times n$ πίνακα με ένα *n*-διάστασης διάνυσμα και παίρνουμε το μεγαλύτερο *d* περιεχόμενο του *p*-διανύσματος, έτσι η SIS παρουσιάζει υπολογιστική πολυπλοκότητα $O(n\ \ p)$.

Αξίζει να σημειώσουμε ότι η SIS χρησιμοποιεί μόνο την τάξη των χαρακτηριστικών των συνιστωσών του *ω*, έτσι ώστε να είναι αναλλοίωτη κάτω υπό κλίμακα. Ακόμα, η ιδέα της SIS είναι ιδανική στην επιλογή επεξηγηματικών μεταβλητών χρησιμοποιώντας τις συσχετίσεις τους με την απόκριση. Για να εφαρμοστεί η SIS, πρέπει τα γραμμικά μοντέλα με πάνω από *n* παραμέτρους, να μην είναι αναγνωρίσιμα με μόνο *n* στοιχεία δεδομένων. Έτσι, μπορούμε να επιλέξουμε $d = [\gamma_n]$ να είναι συντηρητικό, για π.χ. $n-1$ ή $n/\log(n)$, ανάλογα με την τάξη του μεγέθους του δείγματος *n*. Παρόλο που η SIS έχει προταθεί για να μειώσει τη διάσταση *p* από υψηλή σε χαμηλότερη από το μέγεθος του δείγματος *n*, τίποτα δεν μπορεί να μας σταματήσει από το να χρησιμοποιήσουμε ως τελικό μέγεθος του δείγματος μοντέλου $d \geq n$, έστω $\gamma \geq 1$. Είναι φανερό ότι μεγαλύτερο από *d* σημαίνει μεγαλύτερη πιθανότητα να περιέχεται το πραγματικό μοντέλο M_* στο τελικό μοντέλο M_γ.

Η SIS είναι μια δύσκολη μέθοδος. Για ορθογώνιους σχεδιασμούς πινάκων είναι κατανοητό. Όμως για γενικές μεθόδους σχεδιασμού πινάκων, δεν υπάρχει θεωρητικό υπόβαθρο γι' αυτό, παρόλο που συχνά χρησιμοποιείται στις εφαρμογές. Είναι σημαντικό να ορίσουμε τις συνθήκες κάτω από τις οποίες οι ιδιότητες του

σίγουρου κρησαρίσματος ισχύουν. Με άλλα λόγια, επιλέγουμε $[\gamma_n]$ μεταβλητές, που έχουν το μεγαλύτερο μέγεθος συσχέτισης με τη μεταβλητή απόκρισης, χωρίς να υπολογίσουμε τη συμβολή των άλλων συμμεταβλητών στην παλινδρόμηση. Αποδεικνύεται ότι μια τέτοια απλή διαδικασία έχει την ιδιότητα του σίγουρου κρησαρίσματος με την εξής έννοια:

$$\mathrm{P}(\mathrm{M}_* \subset \mathrm{M}_\gamma) \to 1 \text{ καθώς } n \to \infty$$

για δοθέν γ. Πιο συγκεκριμένα, έχουμε:

Θεώρημα 1. Υποθέτουμε ότι:

(α) $p > n$ και $\log(\mathrm{p}) = \mathrm{O}(\mathrm{n}^{\xi})$ για κάποιο $\xi \in (0, 1-2\kappa)$, όπου το κ δίνεται από την τρίτη συνθήκη

(β) $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ για κάποιο $\sigma > 0$ και $\Sigma^{-1/2} x$ έχει σφαιρική συμμετρική κατανομή που ικανοποιεί την ιδιότητα της συγκέντρωσης, όπου Σ ο πίνακας συνδιασποράς του x,

(γ) $\mathrm{var}(\mathrm{Y}) < \infty$ και για κάποιο $\kappa \geq 0$ και $c > 0$,

$$\min_{t \in M_*} |\beta_i| \geq \frac{c}{n^{\kappa}} \text{ και } \min_{t \in M_*} \left|\mathrm{cov}(\beta_i^{-1} \mathrm{Y}, \mathrm{X}_i)\right| \geq c$$

(δ) υπάρχει κάποιο $\tau \geq 0$ και $c^* > 0$ έτσι ώστε

$$\lambda_{\max}(\Sigma) \leq \mathrm{c}^* n^T .$$

Αν $2\kappa + \tau < 1$, τότε υπάρχει κάποιο $\theta < 1 - 2\kappa - \tau$ έτσι ώστε όταν $\gamma \sim cn^{-\theta}$ με $c > 0$, τότε έχουμε για κάποιο $C > 0$,

$$P(\mathrm{M}_* \subset \mathrm{M}_\gamma) = 1 - \mathrm{O}(\exp(-\mathrm{Cn}^{1-2\kappa} / \log \mathrm{n})) .$$

Από το θεώρημα συνεπάγεται ότι ο μη αραιός αριθμός $s < [\gamma_n]$. Αποδεικνύει έτσι ότι η SIS μπορεί να μειώσει τη διάσταση p από την εκθετικά αυξανόμενη σε μια σχετικά μεγάλη κλίμακα $d = [\gamma_n] = \mathrm{O}(\mathrm{n}^{1-\theta}) < \mathrm{n}$ για κάποιο $\theta > 0$ και το μειωμένο μοντέλο M_γ εξακολουθεί να περιέχει όλες τις μεταβλητές που υπάρχουν στο πραγματικό μοντέλο με συντριπτική πιθανότητα. Πιο συγκεκριμένα, μπορούμε να επιλέξουμε το μέγεθος του υπομοντέλου $d = n-1$ ή $n/\log n$ να είναι συντηρητικό. Παρά το γεγονός ότι η SIS προτείνεται για τη μείωση του μοντέλου από p σε d που βρίσκεται κάτω από το μέγεθος του δείγματος n, τίποτα δεν μας σταματά από το να το εφαρμόσουμε με τελικό μέγεθος μοντέλου $d \geq n$. Είναι προφανές ότι μεγαλύτερο d σημαίνει και μεγαλύτερη πιθανότητα να περιλαμβάνει το πραγματικό μοντέλο M_* στο τελικό μοντέλο M_γ.

Όταν υπάρχουν περισσότερες επεξηγηματικές μεταβλητές από τις παρατηρήσεις, ο θόρυβος κατά την εκτίμηση μπορεί να είναι πολύ μεγάλος, προκαλώντας over-fitting και συσσώρευση θορύβου. Για να μειωθεί ο θόρυβος, συχνά χρησιμοποιείται κανονικοποίηση.

Έστω $\omega^\lambda = (\omega_1^\lambda, ..., \omega_p^\lambda)^T$ ένα διάνυσμα p-διάστασης που προκύπτει από την παλινδρόμηση :

$$\omega^\lambda = (\mathrm{X}^T\mathrm{X} + \lambda \mathrm{I}_p)^{-1} X^T y,$$

όπου $\lambda > 0$ είναι μια παράμετρος κανονικοποίησης. Τότε, μπορεί κάποιος να κρησάρει τις μεταβλητές με βάση το μέγεθος του ω_i^λ. Παρατηρούμε ότι

$$\lambda\omega^\lambda \to \omega \text{ καθώς } \lambda \to \infty.$$

Καταλήγουμε ότι η κατά συνιστώσες παλινδρόμηση ω αντιστοιχεί στη ridge παλινδρόμηση με $\lambda = \infty$.

2.3. Σύνδεση με άλλες μεθόδους μείωσης διάστασης

Η SIS χρησιμοποιεί τις οριακές πληροφορίες της συσχέτισης για να μειώσει τη διάσταση. Θα μελετήσουμε τη SIS στα πλαίσια της ταξινόμησης, στην οποία η ιδέα του ανεξάρτητου κρησαρίσματος εμφανίζεται φυσικά και έχει χρησιμοποιηθεί ευρέως.

Το πρόβλημα της ταξινόμησης μπορούμε να το δούμε ως μια ειδική περίπτωση παλινδρόμησης με μεταβλητές απόκρισης με διακριτές τιμές, όπως ± 1. Για προβλήματα μεγάλης διάστασης, όπως η ταξινόμηση του όγκου με χρήση γονιδιακής έκφρασης ή πρωτεϊνικών δεδομένων είναι προτιμότερο να μην ταξινομήσουμε τα δεδομένα χρησιμοποιώντας το πλήρες διάστημα χαρακτηριστικών γνωρισμάτων, λόγω της συσσώρευσης του θορύβου και της επεξηγηματικότητας. Επίσης, πολλά από τα χαρακτηριστικά έρχονται στο προσκήνιο μέσω της σύνδεσης με τις σημαντικές επεξηγηματικές μεταβλητές. Παρόλα αυτά, η επιλογή χαρακτηριστικών είναι σημαντική για την ταξινόμηση υψηλής διάστασης.

Η SIS μπορεί να χρησιμοποιηθεί για να μειώσει το χώρο των χαρακτηριστικών. Υποθέτουμε ότι έχουμε n_1 δείγματα από την κλάση 1 και n_2 δείγματα από την κλάση -1. Τότε ο κατά συνιστώσες εκτιμητής παλινδρόμησης γίνεται:

$$\omega = \sum_{Y_i=1} x_i - \sum_{Y_i=-1} x_i$$

όπου κάθε μεταβλητή έχει ομαλοποιηθεί οριακά.

Αν το γράψουμε πιο αναλυτικά, το j-οστό στοιχείο του p-διανύσματος ω είναι:

$$\omega_j = \left(n_1 \bar{X}_{j,1} - n_2 \bar{X}_{j,2}\right) / (\textit{τυπική} \text{ απόκλιση του j-οστού χαρακτηριστικού})$$

όπου τα $\bar{X}_{j,1}$ είναι ο μέσος του δείγματος του j χαρακτηριστικού κλάσης 1, και $\bar{X}_{j,2}$ είναι ο μέσος του δείγματος του j χαρακτηριστικού κλάσης -1. Όταν $n_1 = n_2$, ω_j είναι απλά μια εκδοχή δύο δειγμάτων t-στατιστικής, με εξαίρεση μια κλιμακωτή σταθερά.

Με τη χρήση της SIS μπορούμε να ξεχωρίσουμε τα σημαντικά χαρακτηριστικά και να μειώσουμε σημαντικά το χώρο των χαρακτηριστικών σε ένα χώρο πολύ χαμηλότερης διάστασης.

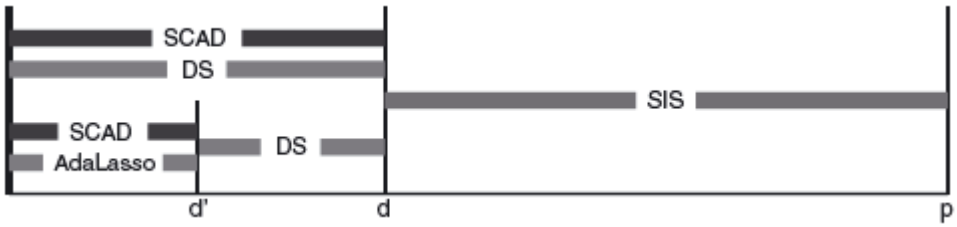
2.4. Τεχνικές επιλογής μοντέλου βασισμένες στη SIS

Με τη μάθηση συσχέτισης, μπορούμε να συρρικνώσουμε ολόκληρο το μοντέλο $\{1,....,\mathrm{p}\}$ σε ένα υποσύνολο του μοντέλου $\mathrm{M}=\mathrm{M}_{\gamma}$ με μέγεθος $d=[\gamma_n]$. Το αρχικό πρόβλημα της εκτίμησης του αραιού *p*-διανύσματος β στο αρχικό μοντέλο μειώνεται στο να εκτιμηθεί το αραιό *d*-διάνυσμα $\beta=(\beta_1,...,\beta_d)^T$ που βασίζεται τώρα στο πολύ μικρότερο υπομοντέλο

$$y=X_M\beta+\varepsilon$$

όπου $X_M=(\mathrm{x}_1,...,\mathrm{x}_n)^T$ υποδηλώνει ένα $n\times d$ υποπίνακα του *X* που προκύπτει από την αφαίρεση των στηλών που αντιστοιχούν στους δείκτες στο *M*. Η SIS μπορεί να επιταχύνει δραστικά την επιλογή μεταβλητών όταν η αρχική διάσταση *p* είναι πάρα πολύ μεγάλη.

Για τα προβλήματα επιλογής μεταβλητών μεγάλης διάστασης, αρχικά προτείνουμε να χρησιμοποιηθεί μια μέθοδος σίγουρου κρησαρίσματος όπως η SIS για τη μείωση της διάστασης από *p* σε σχετικά μεγάλη κλίμακα *d* κάτω από το μέγεθος του δείγματος *n*. Στη συνέχεια χρησιμοποιούμε μια μέθοδο επιλογής μοντέλου χαμηλότερης διάστασης όπως η SCAD, ο Dantzig επιλογέας, η LASSO ή η adaptive LASSO. Καλούμε τη SIS που ακολουθείται από τη SCAD και τον Dantzig επιλογέα SIS-SCAD και SIS-DS αντίστοιχα, για χάρη συντομίας. Σε ορισμένες περιπτώσεις, μπορεί να θέλουμε να μειώσουμε περισσότερο το μέγεθος του μοντέλου σε $\mathrm{d}'<\mathrm{d}$ χρησιμοποιώντας μια μέθοδο όπως ο Dantzig επιλογέας ή η LASSO με κατάλληλο συντονισμό, και τελικά να επιλέξουμε ένα μοντέλο με μια πιο εκλεπτυσμένη μέθοδο όπως η SCAD ή η adaptive LASSO, που θα καλούνται SIS-DS-SCAD και SIS-DS-AdaLASSO αντίστοιχα για ευκολία. Στο παρακάτω σχήμα έχουμε τη σχηματική απεικόνιση αυτών των προσεγγίσεων.



Σχήμα 2.4.1. Μέθοδοι της επιλογής μοντέλου Σε υψηλή διάσταση.

Η ιδέα του να χρησιμοποιήσουμε τη SIS, κάνει πιο εφικτή την επιλογή μοντέλου με πολύ μεγάλη διάσταση και επιταχύνει δραστικά την επιλογή μεταβλητών. Καθιστά επίσης το πρόβλημα της επιλογής μοντέλου πιο αποτελεσματικό. Η SIS μπορεί να χρησιμοποιηθεί με οποιαδήποτε άλλη μέθοδο επιλογής μοντέλου, όπως οι Bayesian μέθοδοι και LASSO.

2.5. Αριθμητικές μελέτες

Για τη μελέτη της απόδοσης των μεθόδων επιλογής μοντέλου που βασίζονται στη SIS και είδαμε παραπάνω, θα παρουσιάσουμε ένα παράδειγμα με πραγματικά δεδομένα.

2.5.1. Ανάλυση δεδομένων Λευχαιμίας

Εφαρμόσαμε τη SIS για να επιλέξουμε χαρακτηριστικά για την ταξινόμηση του συνόλου δεδομένων της λευχαιμίας. Τα δεδομένα λευχαιμίας από υψηλής πυκνότητας συστοιχίες ολιγονουκλεοτιδίων *Affymetrix* έχουν προηγουμένως αναλυθεί στο Golub et al. (1999). Υπάρχουν 7129 γονίδια και 72 δείγματα από τις δύο κατηγορίες: 47 στην *ALL* (*acute lymphocytic leukaemia*- οξεία λεμφοκυτταρική λευχαιμία) και 25 στην *AML* (*acute mylogenous leukaemia*- οξεία μυογενής λευχαιμία).

Μεταξύ αυτών των 72 δειγμάτων, 38 (27 από την ALL και 11 από την AML) από αυτά έχουν οριστεί ως δείγματα κατάρτισης και τα υπόλοιπα 34 (20 από την ALL και 14 από την AML) έχουν οριστεί να είναι το δείγμα δοκιμής.

Χρησιμοποιήσαμε τις δύο μεθόδους SIS-SCAD-LD και SIS-SCAD-NB που θα εισαχθούν παρακάτω για να πραγματοποιήσουν την κατάταξη. Για κάθε μέθοδο αρχικά εφαρμόσαμε τη SIS για να επιλέξουμε $d = [2n/\log(\mathrm{n})]$ γονίδια με $\mathrm{n} = 38$ το μέγεθος του δείγματος κατάρτισης που επιλέχθηκε παραπάνω και στη συνέχεια χρησιμοποιήσαμε τη SCAD για να πάρουμε μια οικογένεια μοντέλων που αναπροσαρμόζονται από την παράμετρο κανονικοποίησης λ. Εδώ, θα πρέπει να σημειώσουμε ότι τα αποτελέσματα ταξινόμησής μας δεν είναι πολύ ευαίσθητα στην επιλογή του d εφόσον δεν είναι πάρα πολύ μικρό. Σίγουρα υπάρχουν πολλοί τρόποι για να ρυθμίσουμε την παράμετρο κανονικοποίησης λ. Για λόγους απλότητας, επιλέξαμε ένα λ που παράγει ένα μοντέλο με μέγεθος που είναι ίσο με το βέλτιστο αριθμό των χαρακτηριστικών που προσδιορίστηκαν από τη διαδικασία χαρακτηριστικών ανόπτησης ανεξάρτητων κανόνων στους Fan & Fan (2008). 16 γονίδια επιλέχθηκαν από την προσέγγισή τους. Τώρα έχουμε επιλέξει 16 γονίδια και λάβαμε ένα γραμμικό μοντέλο με μέγεθος 16 με τη χρήση της μεθόδου SIS-SCAD. Τελικά, η SIS-SCAD-LD μέθοδος χρησιμοποίησε απευθείας τον παραπάνω γραμμικό κανόνα διάκρισης για να κάνει την ταξινόμηση, και η SIS-SCAD–NB μέθοδος εφάρμοσε τον απλό κανόνα Bayes στον 16-διάστατο χώρο των χαρακτηριστικών που προέκυψε.

Τα αποτελέσματα ταξινόμησης των SIS-SCAD-LD, SIS-SCAD-NB και της πλησιέστερης συρρικνωμένης κεντροειδής μεθόδου στο Tibshirani et al. (2002) εμφανίζονται στον Πίνακα 2.5.1.1. Και οι δύο μέθοδοι SIS-SCAD-LD και SIS-SCAD-NB επέλεξαν 16 γονίδια και έκανε ένα σφάλμα δοκιμής με σφάλματα κατάρτισης 0 και 4 αντίστοιχα, ενώ η πλησιέστερη συρρικνωμένη κεντροειδής μέθοδος επέλεξε 21 γονίδια και έκανε ένα σφάλμα κατάρτισης και δύο σφάλματα δοκιμής.

Method	*Training error*	*Test error*	*Number of genes*
SIS–SCAD–LD	0/38	1/34	16
SIS–SCAD–NB	4/38	1/34	16
Nearest shrunken centroids	1/38	2/34	21

Πίνακας 2.5.1.1. Σφάλματα ταξινόμησης στα σύνολα δεδομένων λευχαιμίας.

2.6. Κατάταξη Χαρακτηριστικού από τις Οριακές Χρησιμότητες.

Η σχέση μεταξύ των Y και $(\mathrm{X}_1,...,\mathrm{X}_p)^T$ συχνά μοντελοποιείται μέσω ενός διανύσματος παραμέτρου $\beta=(\beta_1,...,\beta_p)^T$ και η προσαρμογή του μοντέλου ισοδυναμεί με την ελαχιστοποίηση της αρνητικής συνάρτησης ψευδο-πιθανοφάνειας της μορφής

$$Q(\beta_0,\beta)=n^{-1}\sum_{i=1}^{n}L(\mathrm{Y}_i,\beta_0+x_i^T\beta)\,.$$

Εδώ το L μπορεί να θεωρηθεί ως η απώλεια της χρησιμοποίησης $\beta_0+x_i^T\beta$ για να προβλέψει το Y_i. Η οριακή χρησιμότητα του j-οστού χαρακτηριστικού είναι:

$$L_j=\min_{\beta_0,\beta_j}n^{-1}\sum_{i=1}^{n}L(\mathrm{Y}_i,\beta_0+X_{ij}\beta_j)\,,$$

η οποία ελαχιστοποιεί τη συνάρτηση απώλειας, όπου $x_i=(\mathrm{X}_{i1},...,\mathrm{X}_{ip})^T$. Η ιδέα της SIS σε αυτό το πλαίσιο είναι να υπολογίσουμε το διάνυσμα των οριακών χρησιμοτήτων $L=(\mathrm{L}_1,...,\mathrm{L}_p)^T$ και να τα κατατάξουμε σύμφωνα με τις οριακές χρησιμότητες: όσο μικρότερη είναι τόσο πιο σημαντική. Παρατηρούμε ότι προκειμένου να υπολογίσουμε τα L_j χρειαζόμαστε μόνο να ταιριάξουμε ένα μοντέλο με δύο παραμέτρους, β_0 και β_j, έτσι ώστε ο υπολογισμός του διανύσματος L να μπορεί να γίνει πολύ γρήγορα και σταθερά, ακόμη και για προβλήματα υπερυψηλής

διάστασης. Το χαρακτηριστικό X_j έχει επιλεγεί από τη SIS, αν η L_j είναι ένα από τα *d* μικρότερα συστατικά του *L*. Συνήθως, μπορούμε να πάρουμε $d = \lfloor \mathrm{n}/\log \mathrm{n} \rfloor$.

Η παραπάνω διαδικασία είναι μια μέθοδος ανεξάρτητου κρησαρίσματος. Χρησιμοποιεί μόνο μια οριακή σχέση μεταξύ των χαρακτηριστικών και της μεταβλητής απόκρισης για να κρησάρει τις μεταβλητές. Όταν το *d* είναι αρκετά μεγάλο, έχει υψηλή πιθανότητα επιλογής όλων των σημαντικών χαρακτηριστικών. Για αυτό το λόγο, καλούμε τη μέθοδο Sure Independence Screening (SIS). Για προβλήματα ταξινόμησης με τετραγωνική απώλεια *L*, οι Fan & Lv (2008) έδειξαν ότι η SIS μειώνεται σε κρησάρισμα μεταβλητής χρησιμοποιώντας δύο δείγματα t-στατιστικής.

2.7. Ποινικοποιημένη Ψευδοπιθανοφάνεια

Με χαρακτηριστικά πρόχειρα επιλεγμένα από τη SIS, η επιλογή μεταβλητών και η εκτίμηση παραμέτρων μπορεί ακόμη να διεξάγεται ταυτόχρονα, χρησιμοποιώντας μια πιο εκλεπτυσμένη ποινικοποιημένη μέθοδο ψευδο-πιθανοφάνειας. Η προσέγγιση αυτή λαμβάνει υπόψη πληροφορίες από κοινού. Με την αναδιάταξη, αν είναι απαραίτητο, μπορούμε να υποθέσουμε χωρίς βλάβη της γενικότητας ότι τα $X_1,...,\mathrm{X}_d$ είναι τα χαρακτηριστικά που προσλαμβάνονται από τη SIS. Έστω $x_{i,d} = (\mathrm{X}_{i1},...,\mathrm{X}_{id})^T$ και επαναπροσδιορίζουμε $\beta = (\beta_1,...,\beta_d)^T$. Στην ποινικοποιημένη προσέγγιση πιθανότητας θα επιδιώξουμε να ελαχιστοποιήσουμε:

$$l(\beta_0, \beta) = n^{-1}\sum_{i=1}^{n} L(\mathrm{Y}_i, \beta_0 + x_{i,d}^T\beta) + \sum_{j=1}^{d} p_\lambda\left(\left|\beta_j\right|\right) \qquad (2.7.1)$$

όπου $p_\lambda(\cdot)$ είναι μια συνάρτηση ποινής και $\lambda > 0$ είναι μια παράμετρος κανονικοποίησης, η οποία μπορεί να επιλεγεί από την πέντε φορές διασταυρωμένη επικύρωση, για παράδειγμα. Η συνάρτηση ποινής θα πρέπει να ικανοποιεί ορισμένες προϋποθέσεις ώστε οι προκύπτουσες εκτιμήσεις να έχουν επιθυμητές ιδιότητες και συγκεκριμένα να αποδίδει αραιές λύσεις στις οποίες κάποιοι από τους συντελεστές μπορεί να μηδενίζονται.

Τα πιο συχνά χρησιμοποιούμενα παραδείγματα συναρτήσεων ποινής περιλαμβάνουν την L_1 ποινή $p_\lambda\left(|\beta|\right)=\lambda|\beta|$ (Tibshirani, 1996; Park & Hastie, 2007), το ομαλό «ψαλίδισμα» της απόλυτης απόκλισης (SCAD) ποινής (Fan & Li, 2001), η οποία είναι μια τετραγωνική σφήνα με $p_\lambda(0)=0$ και $p_\lambda'\left(|\beta|\right)=\lambda\left\{1_{\{|\beta|\le\lambda\}}+\frac{(\alpha\lambda-|\beta|)_+}{(\alpha-1)\lambda}1_{\{|\beta|>\lambda\}}\right\}$, για κάποιο $\alpha>2$ και $|\beta|>0$, και την ελάχιστη κοίλη ποινή (MCP), $p_\lambda'\left(|\beta|\right)=(\lambda-|\beta|/\alpha)_+$ (Zhang, 2009). Σε αντίθεση με την L_1 ποινή, οι SCAD και MC συναρτήσεις ποινών που διαθέτουν επίπεδες ουρές είναι θεμελιώδεις για τη μείωση μεροληψίας λόγω ποινικοποίησης (Antoniadis & Fan (2001), Fan & Li (2001)). Οι Park & Hastie (2007) περιέγραψαν έναν επαναληπτικό αλγόριθμο για την ελαχιστοποίηση της αντικειμενικής συνάρτησης για την ποινή L_1, και ο Zhang (2009) πρότεινε ένα PLUS αλγόριθμο για την εύρεση διαδρομών λύσης στο ποινικοποιημένο πρόβλημα ελαχίστων τετραγώνων με μια γενική ποινή $p_\lambda(\cdot)$. Από την άλλη πλευρά, οι Fan & Li (2001) έδειξαν ότι ο SCAD τύπος της ποινικοποιημένης συνάρτησης απώλειας μπορεί να ελαχιστοποιηθεί επαναληπτικά χρησιμοποιώντας μια τοπική τετραγωνική προσέγγιση, ενώ οι Zou & Li (2008) πρότειναν μια τοπική γραμμική προσέγγιση λαμβάνοντας το πλεονέκτημα του αλγόριθμου που αναπτύχθηκε πρόσφατα για την ποινικοποιημένη L_1 βελτιστοποίηση (Efron et al., 2004). Ξεκινώντας από $\beta^{(0)}=0$, όπως προτάθηκε από τους Fan & Lv (2008), και χρησιμοποιώντας την τοπική γραμμική προσέγγιση $p_\lambda\left(|\beta|\right)\approx p_\lambda\left(\left|\beta^{(k)}\right|\right)+p_\lambda'\left(\left|\beta^{(k)}\right|\right)\left(|\beta|-\left|\beta^{(k)}\right|\right)$, στην (2.7.1), στην (k+1)-οστή επανάληψη, ελαχιστοποιούμε τη σταθμισμένη ποινή L_1

$$n^{-1}\sum_{i=1}^{n}L(\mathrm{Y}_i,\beta_0+x_{i,d}^T\beta)+\sum_{j=1}^{d}\mathrm{w}_j^{(k)}\left|\beta_j\right| \quad (2.7.2) \qquad \text{όπου } \mathrm{w}_j^{(k)}=p_\lambda'\left(\left|\beta_j^{(k)}\right|\right).$$

Πρέπει να σημειώσουμε ότι με την αρχική τιμή $\beta^{(0)}=0$ η $\beta^{(1)}$ είναι πράγματι μια εκτίμηση LASSO για τις SCAD και MC ποινές, δεδομένου ότι $p_\lambda'(0_+)=\lambda$. Με άλλα λόγια, το μηδέν δεν είναι μια απορροφητική κατάσταση. Αν και με λίγο διαφορετικά κίνητρα, μια σταθμισμένη L_1 ποινή είναι επίσης η βάση της προσαρμοστικής LASSO (Zou, 2006). Σε αυτή την περίπτωση $\mathrm{w}_j^{(k)}\equiv w_j=1\big/\left|\hat{\beta}_j\right|^\gamma$,

όπου $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_d)^{\mathrm{T}}$ μπορεί να θεωρηθεί ότι είναι ο εκτιμητής μέγιστης πιθανοφάνειας, και $\gamma > 0$ επιλέγεται από το χρήστη. Το μειονέκτημα μιας τέτοιας προσέγγισης είναι ότι το μηδέν είναι μια απορροφητική κατάσταση όταν η (2.7.2) χρησιμοποιείται επαναληπτικά- τα συστατικά που εκτιμώνται ως μηδέν σε μια επανάληψη δε θα ξεφύγουν ποτέ από το μηδέν.

Για μια κατηγορία συναρτήσεων ποινής που περιλαμβάνει την ποινή SCAD και όταν το d σταθεροποιείται καθώς το n αποκλίνει, οι Fan & Li (2001) καθιέρωσαν την ιδιότητα *Oracle*, δηλαδή οι ποινικοποιημένες εκτιμήσεις εκτελούνται ασυμπτωτικά καθώς και αν ένας χρησμός μας έχει πει εκ των προτέρων ποια συστατικά του β ήταν μη-μηδενικά. Οι Fan & Peng (2004) επέκτειναν αυτό το αποτέλεσμα για να καλυφθούν οι περιπτώσεις όπου το d μπορεί να αποκλίνει με $d = d_n = o(\mathrm{n}^{1/5})$. Ο Zou (2006) έδειξε ότι η προσαρμοστική LASSO κατέχει επίσης την ιδιότητα *Oracle*, όταν το d είναι πεπερασμένο.

Αναφερόμαστε στις διαδικασίες δύο σταδίων που περιγράφονται παρακάτω ως SIS-LASSO, SIS-SCAD και SIS-AdaLASSO.

2.8. Ομαδοποίηση και μετασχηματισμός των μεταβλητών εισόδου.

Η ομαδοποίηση των μεταβλητών εισόδου (*input variables*) χρησιμοποιείται συνήθως σε διάφορα προβλήματα. Για παράδειγμα, μπορούμε να χωρίσουμε τις p μεταβλητές σε ομάδες των πέντε μεταβλητών. Η ιδέα του κρησαρίσματος μεταβλητών μέσω της SIS μπορεί να εφαρμοστεί για την επιλογή ενός μικρότερου αριθμού ομάδων. Με αυτό τον τρόπο, υπάρχει μικρότερη πιθανότητα να χαθούν σημαντικές μεταβλητές εκμεταλλευόμενοι την κοινή βάση πληροφοριών μεταξύ των επεξηγηματικών μεταβλητών. Ως εκ τούτου, ένα πιο αξιόπιστο μοντέλο μπορεί να κατασκευαστεί.

Μια αξιοσημείωτη δυσκολία στην επιλογή μεταβλητών οφείλεται στη συγγραμμικότητα μεταξύ των συμμεταβλητών. Ένας αποτελεσματικός τρόπος να αποκλειστούν αυτές οι μη σημαντικές μεταβλητές, οι οποίες είναι αρκετά συσχετισμένες με τις σημαντικές, είναι να αποκλειστούν μετά. Μια καλή ιδέα είναι

να μετασχηματιστούν οι μεταβλητές εισόδου. Δύο πιθανοί τρόποι ξεχωρίζουν: Ο ένας είναι ο *subject-related* μετασχηματισμός και ο άλλος είναι ο στατιστικός μετασχηματισμός.

Ο μετασχηματισμός *subject-related* είναι ένα χρήσιμο εργαλείο . Σε μερικές περιπτώσεις, ένας απλός γραμμικός μετασχηματισμός των μεταβλητών εισόδου μπορεί να βοηθήσει στη μείωση της συσχέτισης μεταξύ των συμμεταβλητών. Για παράδειγμα, στις μελέτες σωματότυπου η κοινή λογική μας λέει ότι οι επεξηγηματικές μεταβλητές, όπως τα βάρη w_1, w_2, w_3, συσχετίζονται θετικά με τις ηλικίες 2, 9 και 18. Θα μπορούσαμε απευθείας να χρησιμοποιήσουμε τα w_1, w_2, w_3 ως τις μεταβλητές εισόδου σε ένα γραμμικό μοντέλο παλινδρόμησης, αλλά ένας καλύτερος τρόπος επιλογής μοντέλου σε αυτή την περίπτωση είναι να χρησιμοποιήσουμε λιγότερες συσχετισμένες επεξηγηματικές μεταβλητές όπως $\left(w_1, w_2 - w_1, w_3 - w_2\right)^T$, το οποίο είναι ένας γραμμικός μετασχηματισμός του $\left(w_1, w_2, w_3\right)^T$, που προσδιορίζει τις αλλαγές των βαρών αντί των ίδιων των βαρών. Ένα άλλο σημαντικό παράδειγμα είναι οι οικονομικές χρονοσειρές όπως οι τιμές των αποθεμάτων ή τα επιτόκια. Η διαφοροποίηση μπορεί να εξασθενίσει σημαντικά τη συσχέτιση ανάμεσα σε αυτές τις μεταβλητές.

Οι μέθοδοι στατιστικών μετασχηματισμών περιλαμβάνουν μια εφαρμογή ενός αλγόριθμου ομαδοποίησης, όπως είναι η ιεραρχική ομαδοποίηση ή ο αλγόριθμος *k*-μέσου, χρησιμοποιώντας μετρικές συσχέτισης πρώτα στην ομάδα μεταβλητών μέσα σε υψηλά συσχετισμένες ομάδες και ύστερα εφαρμόζοντας αραιή ανάλυση κύριων συνιστωσών για να κατασκευάσουμε ασθενώς συσχετισμένες επεξηγηματικές μεταβλητές. Τώρα, αυτές οι ασθενώς συσχετισμένες επεξηγηματικές μεταβλητές από κάθε ομάδα μπορεί να θεωρηθούν ως οι νέες συμμεταβλητές και μια μέθοδος επιλογής μεταβλητών βασισμένη στη SIS μπορεί να χρησιμοποιηθεί για την επιλογή τους.

Οι στατιστικές τεχνικές που αναφέραμε παραπάνω μπορούν να βοηθήσουν στο να προσδιοριστούν τα σημαντικά χαρακτηριστικά και επομένως να βελτιωθεί η αποτελεσματικότητα της στρατηγικής επιλογής μοντέλου βασισμένο στη SIS. Η εισαγωγή των μη γραμμικών όρων και η μετατροπή των μεταβλητών μπορεί επίσης

να χρησιμοποιηθεί για να μειώσει τις τάσεις μοντελοποίησης των γραμμικών μοντέλων. Ο Ravikumar et al. (2007) εισήγαγε αραιά προσθετικά μοντέλα για την αντιμετώπιση της επιλογής μη γραμμικών χαρακτηριστικών.

2.9. Μερικές επεκτάσεις της μάθησης συσχέτισης

Η βασική ιδέα της SIS είναι η εφαρμογή μιας ενιαίας παλινδρόμησης κατά συνιστώσες. Τρία πιθανά θέματα μπορεί να προκύψουν όμως με αυτή την προσέγγιση:

(α) κάποιες μη σημαντικές επεξηγηματικές μεταβλητές που συνδέονται πάρα πολύ με σημαντικές επεξηγηματικές μεταβλητές μπορεί να έχουν μεγαλύτερη προτεραιότητα να επιλεχθούν από τη SIS από οποιαδήποτε σημαντική επεξηγηματική μεταβλητή που συνδέεται ασθενά με την απόκριση.

(β) μια σημαντική επεξηγηματική μεταβλητή που είναι οριακά ασυσχέτιστη αλλά συσχετίζεται με την απόκριση, μπορεί να μην επιλεχθεί από τη SIS και έτσι να μην μπει στο μοντέλο.

(γ) το θέμα της συγγραμμικότητας μεταξύ δύο επεξηγηματικών μεταβλητών προσθέτει δυσκολία στο πρόβλημα της επιλογής μεταβλητών.

Όταν οι υποθέσεις του μοντέλου ικανοποιούνται, περιλαμβάνει δηλαδή τα τρία προαναφερθέντα προβλήματα, η SIS μπορεί να μειώσει με ακρίβεια τη διάσταση από υπερυψηλή σε μια μέτρια κλίμακα, κάτω από το μέγεθος του δείγματος. Όταν όμως οι προϋποθέσεις δεν ικανοποιούνται, μπορεί η SIS να χάσει κάποιες σημαντικές επεξηγηματικές μεταβλητές. Για να αντιμετωπιστεί αυτό το πρόβλημα μπορούμε να χρησιμοποιήσουμε άλλες μεθόδους όπως είναι η ISIS, η οποία είναι μια επαναληπτική εφαρμογή της προσέγγισης SIS στην επιλογή μεταβλητών.

2.9.1. Επαναληπτική SIS – Επαναληπτική μάθηση συσχέτισης

Θα δείξουμε τώρα ότι όταν οι υποθέσεις των μοντέλων ικανοποιούνται, οι οποίες αποκλείουν τα τρία προαναφερθέντα προβλήματα, η SIS μπορεί με ακρίβεια να μειώσει τη διάσταση από υψηλή σε μια μέτριας κλίμακας, κάτω από το μέγεθος του δείγματος. Αλλά όταν οι υποθέσεις αποτυγχάνουν, θα μπορούσε η SIS να χάσει κάποιες από τις σημαντικές επεξηγηματικές μεταβλητές. Για να ξεπεράσουμε αυτό το πρόβλημα, προτείνουμε την ISIS για να ενισχύσουμε την ισχύ της μεθοδολογίας. Η ISIS είναι μια επαναληπτική εφαρμογή της προσέγγισης SIS στην επιλογή μεταβλητών. Η ουσία είναι να εφαρμοστεί επαναληπτικά ένα κρησάρισμα μεταβλητών μεγάλης κλίμακας ακολουθούμενο από μιας μέτριας κλίμακας προσεκτική επιλογή μεταβλητών.

Η ISIS λειτουργεί ως εξής: Αρχικά, επιλέγουμε ένα υποσύνολο k_1 μεταβλητών $A_1 = \left\{ X_{i_1}, ..., X_{i_{k_1}} \right\}$ χρησιμοποιώντας μια μέθοδο επιλογής μοντέλου βασισμένο στη SIS, όπως είναι οι μέθοδοι SIS-SCAD ή SIS-LASSO. Αυτές οι μεταβλητές επιλέχθηκαν χρησιμοποιώντας τη SCAD ή τη LASSO, επί τη βάσει των κοινών πληροφοριών των $\left[n/\log(\mathrm{n}) \right]$ μεταβλητών, οι οποίες επιβίωσαν μετά το κρησάρισμα συσχέτισης. Κατόπιν, έχουμε ένα n-διάνυσμα υπολοίπων από την παλινδρόμηση της απόκρισης Y πάνω στα $X_{i_1}, ..., X_{i_{k_1}}$. Στο επόμενο βήμα, μεταχειριζόμαστε αυτά τα υπόλοιπα ως τις νέες αποκρίσεις και εφαρμόζουμε την ίδια μέθοδο όπως στο προηγούμενο βήμα στις εναπομείνασες $p - k_1$ μεταβλητές, το οποίο καταλήγει σε ένα υποσύνολο k_2 μεταβλητών $A_2 = \left\{ X_{j_1}, ..., X_{j_{k_2}} \right\}$. Σημειώνουμε ότι η προσαρμογή των υπολοίπων από το προηγούμενο βήμα στο $\left\{ X_1, ..., X_p \right\} / A_1$ μπορεί να εξασθενίσει σημαντικά την προτεραιότητα αυτών των μη σημαντικών μεταβλητών, που είναι αρκετά συσχετισμένες με την απόκριση μέσω των σχέσεων τους με τα $X_{i_1}, ..., X_{i_{k_1}}$, δεδομένου ότι τα υπόλοιπα είναι ασυσχέτιστα με αυτές τις επιλεγμένες μεταβλητές στο A_1. Αυτό βοηθάει στο να λυθεί το πρώτο ζήτημα. Επίσης, καθιστά δυνατό αυτές οι σημαντικές επεξηγηματικές μεταβλητές που έχουν χαθεί στο προηγούμενο βήμα να επιβιώσουν. Αυτό διευθύνει το δεύτερο ζήτημα.

Στην πραγματικότητα, μόλις οι μεταβλητές στο A_1 εισαχθούν στο μοντέλο, εκείνες που είναι οριακά ασθενώς συσχετισμένες με την Y λόγω της παρουσίας των μεταβλητών στο A_1 θα πρέπει να συσχετιστούν με τα υπόλοιπα. Μπορούμε να συνεχίσουμε να το κάνουμε αυτό μέχρι να αποκτήσουμε l υποσύνολα $A_1,...,A_l$ των οποίων η ένωση $A=\bigcup_{i=1}^{l} A_i$ έχει μέγεθος d, που είναι μικρότερο του n. Με πρακτική εφαρμογή, μπορούμε να επιλέξουμε, για παράδειγμα, το μεγαλύτερο l υποσύνολο, για το οποίο ισχύει $|A|<n$. Από τα επιλεγμένα χαρακτηριστικά στο A, μπορούμε να επιλέξουμε τα χαρακτηριστικά χρησιμοποιώντας μιας μέτριας κλίμακας μέθοδο, όπως είναι η SCAD, η LASSO ή ο επιλογέας Dantzig.

Για τα προβλήματα επιλογής μεταβλητών υψηλής διάστασης έχουμε τώρα τις μεθόδους επιλογής μοντέλου βασισμένες στην ISIS, οι οποίες είναι προεκτάσεις των μεθόδων επιλογής μοντέλου που βασίζονται στη SIS. Εφαρμόζοντας μια μέτριας διάστασης μέθοδο, όπως η SCAD, ο επιλογέας Dantzig, η LASSO ή η προσαρμοστική LASSO στο A θα παραχθεί ένα μοντέλο που είναι πολύ κοντά στο πραγματικό αραιό μοντέλο M_*. Η ιδέα της ISIS σχετίζεται με κάποιο τρόπο με τον αλγόριθμο ενίσχυσης (*boosting algorithm*) (Freund & Schapire, 1997). Ειδικότερα, αν η SIS χρησιμοποιείται για να διαλέγει μία μόνο μεταβλητή σε κάθε επανάληψη, π.χ. $|A_i|=1$, η ISIS είναι ισοδύναμη με μια μορφή ταιριάσματος ενός άπληστου αλγόριθμου για την επιλογή μεταβλητών. (Barron et al., 2008).

2.9.2. Επαναληπτική Επιλογή Χαρακτηριστικών

Η μεθοδολογία SIS μπορεί να καταρρεύσει αν ένα χαρακτηριστικό είναι οριακά ασύνδετο, αλλά σχετίζεται από κοινού με την απόκριση ή αν ένα χαρακτηριστικό είναι από κοινού ασυσχέτιστο με την απόκριση αλλά έχει υψηλότερη οριακή συσχέτιση με την απόκριση από κάποια σημαντικά χαρακτηριστικά. Στην πρώτη περίπτωση, το σημαντικό χαρακτηριστικό έχει ήδη κρησαριστεί στο πρώτο στάδιο, ενώ στη δεύτερη περίπτωση το ασήμαντο χαρακτηριστικό κατατάσσεται

πολύ ψηλά από την τεχνική ανεξάρτητου κρησαρίσματος. Η ISIS επιδιώκει να ξεπεράσει αυτές τις δυσκολίες χρησιμοποιώντας πληρέστερα την πληροφορία κοινής συμμεταβλητής διατηρώντας την υπολογιστική σκοπιμότητα και σταθερότητα στη SIS.

Στο πρώτο στάδιο, εφαρμόζουμε τη SIS για να πάρουμε ένα σύνολο $\hat{A}_1$ των δεικτών του μεγέθους k_1 και στη συνέχεια χρησιμοποιούμε μια ποινικοποιημένη μέθοδο (ψευδο-)πιθανοφάνειας όπως οι LASSO, SCAD, MCP ή η προσαρμοστική LASSO για να επιλέξουμε ένα υποσύνολο $\hat{M}_1$ αυτών των δεικτών. Αυτή είναι η αρχική μας εκτίμηση για το σύνολο των δεικτών των σημαντικών χαρακτηριστικών. Το στάδιο κρησαρίσματος λύνει μόνο τις δισδιάστατες βελτιστοποιήσεις και η προσαρμογή μέρους λύνει μόνο ένα πρόβλημα βελτιστοποίησης με μέτριο μέγεθος k_1. Αυτό είναι ένα ελκυστικό χαρακτηριστικό στην υψηλής διάστασης στατιστική μάθηση.

Αντί να υπολογίσουμε τα υπόλοιπα, όπως θα μπορούσε να γίνει στο γραμμικό μοντέλο, υπολογίζουμε

$$L_j^{(2)} = \min_{\beta_0, \beta_{\hat{M}_1}, \beta j} n^{-1} \sum_{i=1}^{n} L(\mathrm{Y}_i, \beta_0 + x_{i,\hat{M}_1}^T \beta_{\hat{M}_1} + \mathrm{X}_{ij}\, \beta_j) \qquad (2.9.2.1)$$

για $j \in \hat{M}_1^c = \{1, ..., \mathrm{p}\} \setminus \hat{M}_1$, όπου $x_{i,\hat{M}_1}$ είναι ένα υποδιάνυσμα του x_i που αποτελείται από αυτά τα στοιχεία στο $\hat{M}_1$. Αυτό είναι πάλι ένα χαμηλής διάστασης πρόβλημα βελτιστοποίησης που μπορεί εύκολα να λυθεί. Παρατηρούμε ότι $L_j^{(2)}$ (μετά την αφαίρεση του συνεχούς $\min_{\beta_0, \beta_{\hat{M}_1}} n^{-1} \sum_{i=1}^{n} L(\mathrm{Y}_i, \beta_0 + x_{i,\hat{M}_1}^T \beta_{\hat{M}_1})$ και αλλάζοντας το πρόσημο της διαφοράς) μπορούμε να το ερμηνεύσουμε ως την πρόσθετη συνεισφορά της μεταβλητής X_j δοσμένης της ύπαρξης των μεταβλητών στο $\hat{M}_1$. Αφού διατάξουμε τα $\{L_j^{(2)} : \mathrm{j} \in \hat{\mathrm{M}}_1^c\}$, σχηματίζουμε το σύνολο $\hat{A}_2$ που αποτελείται από τους δείκτες που αντιστοιχούν στα μικρότερα στοιχεία k_2, ας πούμε. Σε αυτό το στάδιο κρησαρίσματος, μια εναλλακτική προσέγγιση είναι να υποκαταστήσουμε την εφοδιασμένη μεταβλητή $\hat{\beta}_{\hat{M}_1}$ από το πρώτο στάδιο στο (2.9.2.1) και το πρόβλημα

βελτιστοποίησης (2.9.2.1) θα είναι δισδιάστατο. Η προσέγγιση αυτή είναι ακριβώς μια επέκταση των Fan & Lv (2008), όπου έχουμε

$$L(\mathrm{Y}_i, \beta_0 + x_{i,\hat{M}_1}^T \beta_{\hat{M}_1} + \mathrm{X}_{ij}\, \beta_j) = (\hat{\mathrm{r}}_i - \beta_0 - \mathrm{X}_{ij}\beta_j)^2$$

για την τετραγωνική απώλεια, όπου $\hat{\mathrm{r}}_i = \mathrm{Y}_i - x_{i,\hat{M}_1}^T \hat{\beta}_{\hat{M}_1}$ είναι το υπόλοιπο από το προηγούμενο στάδιο της εφαρμογής. Οι υπό όρους συνεισφορές των χαρακτηριστικών είναι πιο σχετικές στην πρόσληψη μεταβλητών στο δεύτερο στάδιο, αλλά ο υπολογισμός είναι πιο ακριβός. Τα αριθμητικά μας πειράματα παρακάτω δείχνουν τη βελτίωση της εν λόγω απόκλισης από τους Fan & Lv (2008).

Μετά το βήμα πριν το κρησάρισμα, χρησιμοποιούμε την ποινικοποιημένη πιθανότητα για να αποκτήσουμε

$$\hat{\beta}_2 = \underset{\beta_0, \beta_{\hat{M}_1}, \beta_{\hat{\mathrm{A}}_2}}{\arg\min}\, n^{-1} \sum_{i=1}^{n} L(\mathrm{Y}_i, \beta_0 + x_{i,\hat{M}_1}^T \beta_{\hat{M}_1} + x_{i,\hat{\mathrm{A}}_2}^T \beta_{\hat{\mathrm{A}}_2}) + \sum_{j \in \hat{M}_1 \cup \hat{\mathrm{A}}_2} p_\lambda(|\beta_j|).$$

Πάλι ο όρος ποινής ενθαρρύνει μια αραιή λύση. Οι δείκτες του $\hat{\beta}_2$ που είναι μη-μηδενικοί αποδίδουν ένα νέο εκτιμημένο σύνολο $\hat{M}_2$ των ενεργών δεικτών. Αυτό το βήμα αποκλίνει, επίσης, σημαντικά από την προσέγγιση των Fan & Lv (2008), ακόμη και στην περίπτωση των ελαχίστων τετραγώνων. Επιτρέπει στη διαδικασία να διαγράφει μεταβλητές από τα προηγούμενα επιλεγμένα χαρακτηριστικά με δείκτες στο $\hat{M}_1$.

Η διαδικασία, η οποία επαναληπτικά προσλαμβάνει και διαγράφει χαρακτηριστικά, μπορεί στη συνέχεια να επαναλαμβάνεται μέχρι να αποκτήσουμε ένα σύνολο δεικτών $\hat{M}_l$ το οποίο είτε έχει φτάσει στο προκαθορισμένο μέγεθος d ή ικανοποιεί $\hat{M}_l = \hat{M}_{l-1}$. Φυσικά μπορούμε επίσης να λάβουμε ένα τελικό εκτιμημένο διάνυσμα παραμέτρων $\hat{\beta}_l$. Η παραπάνω μέθοδος μπορεί να θεωρηθεί ανάλογη της ISIS διαδικασίας ελαχίστων τετραγώνων (Fan & Lv, 2008) χωρίς σαφή ορισμό των υπολοίπων. Στην πραγματικότητα, είναι μια βελτίωση ακόμη και για τα προβλήματα ελαχίστων τετραγώνων.

Σε γενικές γραμμές, επιλέγοντας μεγαλύτερες τιμές του κάθε k_r μειώνουμε το υπολογιστικό κόστος και την πιθανότητα η ISIS διαδικασία να τερματιστεί πρόωρα. Ωστόσο, κάνει επίσης τη διαδικασία περισσότερο σαν τη μη επαναλαμβανόμενη ομόλογή της, και έτσι μπορεί να προσφέρει λιγότερη βελτίωση στις δύσκολες περιπτώσεις για τη SIS που περιγράφτηκαν στην νωρίτερα. Στην εφαρμογή μας επιλέξαμε $k_1 = \lfloor 2d/3 \rfloor$ και στη συνέχεια στην r-οστή επανάληψη πήραμε $k_r = d - |\hat{M}_{r-1}|$. Αυτό εξασφαλίζει ότι οι επαναλαμβανόμενες εκδόσεις της SIS παίρνουν τουλάχιστον δύο επαναλήψεις για να τερματίσουν. Μια άλλη πιθανότητα θα ήταν να πάρουμε, για παράδειγμα, $\mathrm{k}_r = \min(5, \mathrm{d} - |\hat{\mathrm{M}}_{r-1}|)$.

Οι Fan & Lv (2008) έδειξαν εμπειρικά ότι το γραμμικό μοντέλο ISIS βελτιώνει σημαντικά την επίδοση της SIS και στις δύσκολες περιπτώσεις που περιγράφθηκαν παραπάνω. Ο λόγος είναι ότι η προσαρμογή των υπολοίπων από την (r-1) επανάληψη στα εναπομείναντα χαρακτηριστικά εξασθενίζει σημαντικά την προτεραιότητα αυτών των ασήμαντων χαρακτηριστικών που συσχετίζονται σε μεγάλο βαθμό με την απόκριση μέσω των σχέσεών του με $\{\mathrm{X}_j : \mathrm{j} \in \hat{M}_{r-1}\}$. Αυτό οφείλεται στο γεγονός ότι τα χαρακτηριστικά $\{\mathrm{X}_j : \mathrm{j} \in \hat{M}_{r-1}\}$ έχουν μικρότερη συσχέτιση με τα υπόλοιπα από ότι με τις αρχικές αποκρίσεις. Δίνει επίσης σε εκείνα τα σημαντικά χαρακτηριστικά που έχουν χαθεί στο προηγούμενο βήμα, την ευκαιρία να επιβιώσουν.

2.10. Μείωση των Ποσοστών Εσφαλμένης Επιλογής

Οι προσεγγίσεις του σίγουρου ανεξάρτητου κρησαρίσματος είναι απλές και γρήγορες μέθοδοι για να κρησάρουν τα «άσχετα» χαρακτηριστικά. Συνήθως είναι συντηρητικές και περιλαμβάνουν πολλά ασήμαντα χαρακτηριστικά. Σε αυτή την ενότητα, περιγράφουμε δύο πιθανές παραλλαγές της SIS και ISIS που έχουν ελκυστικές θεωρητικές ιδιότητες ώστε να μειωθούν τα FSR. Η πρώτη είναι μια επιθετική μέθοδος επιλογής μεταβλητών η οποία είναι ιδιαίτερα χρήσιμη όταν η

διάσταση είναι πολύ μεγάλη σε σχέση με το μέγεθος του δείγματος, ενώ η δεύτερη είναι μια πιο συντηρητική διαδικασία.

2.10.1. Η πρώτη παραλλαγή της ISIS

Είναι βολικό να εισάγουμε ένα νέο συμβολισμό. Σημειώνουμε με A το σύνολο των ενεργών δεικτών, που είναι το σύνολο που περιέχει αυτούς τους δείκτες j, για τους οποίους $\beta_j \neq 0$, στο πραγματικό μοντέλο. Σημειώνουμε $X_A = \{X_j : \mathrm{j} \in \mathrm{A}\}$ και $X_{A^c} = \{X_j : \mathrm{j} \in \mathrm{A}^c\}$ για τα αντίστοιχα σύνολα των ενεργών και ανενεργών μεταβλητών αντίστοιχα.

Υποθέτουμε για απλότητα ότι το n είναι άρτιο και χωρίζουμε το δείγμα τυχαία σε δύο μέρη. Εφαρμόζουμε τη SIS ή την ISIS ξεχωριστά στα δεδομένα σε κάθε διαμέριση (με $d = \lfloor \mathrm{n}/\log\mathrm{n} \rfloor$ ή μεγαλύτερο) αποδίδοντας δύο εκτιμήσεις $\hat{A}^{(1)}$ και $\hat{A}^{(2)}$ του συνόλου των ενεργών δεικτών A. Και τα δύο θα πρέπει να έχουν μεγάλα FSR, καθώς είναι κατασκευασμένα από μια ακατέργαστη μέθοδο κρησαρίσματος. Υποθέτουμε ότι και τα δύο σύνολα ικανοποιούν

$$P(\mathrm{A} \subset \hat{\mathrm{A}}^{(\mathrm{j})}) \to 1, \text{ για } j = 1 \text{ και } 2 .$$

Τότε τα ενεργά χαρακτηριστικά θα πρέπει να εμφανίζονται και στα δύο σύνολα με πιθανότητα που τείνει στο 1. Έτσι, κατασκευάζουμε $\hat{\mathrm{A}} = \hat{\mathrm{A}}^{(1)} \bigcap \hat{\mathrm{A}}^{(2)}$ ως μια εκτίμηση του A. Αυτή η εκτίμηση ικανοποιεί επίσης $P(\mathrm{A} \subset \hat{\mathrm{A}}) \to 1$. Ωστόσο, αυτή η εκτίμηση περιέχει λιγότερους δείκτες που αντιστοιχούν σε ανενεργά χαρακτηριστικά, όπως αυτοί οι δείκτες πρέπει να εμφανίζονται τυχαία δύο φορές στα σύνολα $\hat{A}^{(1)}$ και $\hat{A}^{(2)}$. Αυτό πράγματι φαίνεται παρακάτω στο Θεώρημα 1.

Ακριβώς όπως και στην αρχική διατύπωση της SIS, μπορούμε να χρησιμοποιήσουμε μια ποινικοποιημένη μέθοδο (ψευδο-)πιθανοφάνειας, όπως η SCAD, για να εκτελέσει την τελική επιλογή χαρακτηριστικών από το $\hat{A}$ και την εκτίμηση παραμέτρων. Μπορούμε ακόμη να προχωρήσουμε χωρίς την ποινικοποίηση δεδομένου ότι το ποσοστό εσφαλμένης επιλογής είναι μικρό.

Στη θεωρητική μας υποστήριξη για αυτή την παραλλαγή της SIS θα κάνουμε χρήση της ακόλουθης συνθήκης:

(A_1). Έστω $r \in \mathbb{N}$, το σύνολο των φυσικών αριθμών. Λέμε ότι το μοντέλο ικανοποιεί τη συνθήκη ανταλλαγής στο επίπεδο r αν το σύνολο των τυχαίων διανυσμάτων $\{(\mathrm{Y}, \mathrm{X}_A, \mathrm{X}_{j_1}, ..., \mathrm{X}_{j_r}) : \mathrm{j}_1, ..., \mathrm{j}_r$ είναι διακριτά στοιχεία της $\mathrm{A}^c\}$ εναλλάσσεται.

Αυτή η συνθήκη εξασφαλίζει ότι κάθε ανενεργό χαρακτηριστικό είναι εξίσου δυνατό να προσληφθεί από τη SIS. Παρατηρούμε ότι η (A_1) σίγουρα επιτρέπει στα ανενεργά χαρακτηριστικά να συσχετίζονται με την απόκριση αλλά δε συνεπάγεται ότι κάθε ανενεργό χαρακτηριστικό έχει την ίδια οριακή κατανομή. Στο Θεώρημα 1 παρακάτω, η περίπτωση $r = 1$ είναι ιδιαίτερα σημαντική καθώς δίνει ένα άνω φράγμα για την πιθανότητα της πρόσληψης οποιωνδήποτε ανενεργών χαρακτηριστικών στο μοντέλο. Παρατηρούμε ότι αυτό το ανώτερο όριο απαιτεί μόνο την ασθενέστερη έκδοση (επίπεδο 1) της συνθήκης εναλλαγής.

Θεώρημα 1. Έστω $r \in \mathbb{N}$ και υποθέτουμε ότι το μοντέλο ικανοποιεί τη συνθήκη εναλλαγής (A_1) στο επίπεδο r. Αν το $\hat{\mathrm{A}}$ δηλώνει τον εκτιμητή της A από την παραπάνω παραλλαγή της SIS, τότε

$$P\left(\left|\hat{\mathrm{A}} \cap \mathrm{A}^c\right| \geq \mathrm{r}\right) \leq \frac{\binom{d}{r}^2}{\binom{p-|A|}{r}} \leq \frac{1}{r!}\left(\frac{d^2}{p-|\mathrm{A}|}\right)^r,$$

όπου για τη δεύτερη ανισότητα απαιτούμε $d^2 \leq p - |\mathrm{A}|$ και d είναι ο προκαθορισμένος αριθμός των επιλεγμένων χαρακτηριστικών στο $\hat{A}^{(1)}$ και $\hat{A}^{(2)}$.

<u>Απόδειξη</u>

Ορίζουμε $r \in \mathbb{N}$ και έστω $J = \{(\mathrm{j}_1, ..., \mathrm{j}_r) : \mathrm{j}_1, ..., \mathrm{j}_r$ είναι διακριτά στοιχεία της $\mathrm{A}^c\}$

Τότε

$$P\left(\left|\hat{A}\bigcap A^c\right|\geq r\right)\leq \sum_{(j_1,\ldots,j_r)\in J} P(j_1\in\hat{A},\ldots,j_r\in\hat{A})$$

$$=\sum_{(j_1,\ldots,j_r)\in J} P(j_1\in\hat{A}^{(1)},\ldots,j_r\in\hat{A}^{(1)})^2\,,$$

στην οποία χρησιμοποιούμε τυχαίο διαχωρισμό στην τελευταία ισότητα. Προφανώς, η τελευταία πιθανότητα φράσσεται από το

$$\max_{(j_1,\ldots,j_r)\in J} P(j_1\in\hat{A}^{(1)},\ldots,j_r\in\hat{A}^{(1)})\sum_{(j_1,\ldots,j_r)\in J} P(j_1\in\hat{A}^{(1)},\ldots,j_r\in\hat{A}^{(1)}) \qquad (2.10.1.1).$$

Αφού υπάρχουν το πολύ *d* ανενεργά χαρακτηριστικά από το A^c στο σύνολο $\hat{A}^{(1)}$, ο αριθμός των *r*-πλειάδων από το *J* εμπίπτει στο σύνολο $\hat{A}^{(1)}$ και δεν μπορεί να υπερβαίνει το συνολικό αριθμό των τέτοιων r-πλειάδων στο $\hat{A}^{(1)}$, δηλαδή στο $\sum_{(j_1,\ldots,j_r)\in J} 1_{\{j_1\in\hat{A}^{(1)},\ldots,j_r\in\hat{A}^{(1)}\}}\leq\binom{d}{r}$. Έτσι έχουμε

$$\sum_{(j_1,\ldots,j_r)\in J} P(j_1\in\hat{A}^{(1)},\ldots,j_r\in\hat{A}^{(1)})\leq\binom{d}{r} \qquad (2.10.1.2).$$

Αντικαθιστώντας το στην (2.10.1.1) παίρνουμε

$$P\left(\left|\hat{A}\bigcap A^c\right|\geq r\right)\leq\binom{d}{r}\max_{(j_1,\ldots,j_r)\in J} P(j_1\in\hat{A}^{(1)},\ldots,j_r\in\hat{A}^{(1)})\,.$$

Τώρα, υπό τη συνθήκη της εναλλαγής (*A₁*), κάθε *r*-πλειάδα των διακριτών δεικτών στο A^c είναι ισοδύναμα πιθανό να προσληφθούν στο $\hat{A}^{(1)}$. Ως εκ τούτου, προκύπτει από την (2.10.1.2) ότι

$$\max_{(j_1,\ldots,j_r)\in J} P(j_1\in\hat{A}^{(1)},\ldots,j_r\in\hat{A}^{(1)})\leq\frac{\binom{d}{r}}{\binom{p-|A|}{r}}\,,$$

και το πρώτο αποτέλεσμα έπεται. Το δεύτερο αποτέλεσμα προκύπτει από το απλό γεγονός ότι

$$\frac{(\mathrm{d}-\mathrm{i})^2}{p^*-i} \le \frac{d^2}{p^*}, \qquad \text{για όλα τα} \quad 0 \le i \le d,$$

όπου $p^* = p - |A|$, και ο απλός υπολογισμός

$$\frac{\binom{d}{r}^2}{\binom{p^*}{r}} = \frac{1}{r!}\frac{d^2(d-1)^2\cdots(d-r+1)^2}{p^*(p^*-1)\cdots(p^*-r+1)} \le \frac{1}{r!}\binom{d}{p^*}^r$$

Και αυτό ολοκληρώνει την απόδειξη.

Το Θεώρημα 1 δίνει ένα ασυμπτωτικό όριο, χρησιμοποιώντας μόνο τα επιχειρήματα συμμετρίας και αυτό το όριο αναμένεται να είναι λογικά στενό, ειδικά όταν το p είναι μεγάλο. Από το Θεώρημα 1 παρατηρούμε ότι αν η συνθήκη εναλλαγής στο Επίπεδο 1 ικανοποιείται και αν το p είναι μεγάλο σε σύγκριση με το n^2 τότε όταν ο αριθμός των επιλεγμένων χαρακτηριστικών d είναι μικρότερος από το n έχουμε μεγάλη πιθανότητα αυτή η παραλλαγή της SIS να μην αναφέρει «εσφαλμένες θετικές», δηλαδή είναι πολύ πιθανό οποιοσδήποτε δείκτης στο εκτιμημένο ενεργό σύνολο να ανήκει επίσης στο ενεργό σύνολο στο αληθινό μοντέλο. Διαισθητικά, αν το p είναι μεγάλο τότε κάθε ανενεργό χαρακτηριστικό έχει μικρή πιθανότητα στο να συμπεριληφθεί στο εκτιμημένο ενεργό σύνολο έτσι είναι απίθανο, πράγματι, να εμφανιστεί στο εκτιμημένο ενεργό σύνολο και των δύο διαμερίσεων. Η φύση αυτού του αποτελέσματος είναι λίγο ασυνήθιστη δεδομένου ότι προτείνει μια «ευλογία της διάστασης»- το όριο στην πιθανότητα ψευδών θετικών μειώνεται με το p. Ωστόσο αυτό είναι μόνο ένα μέρος της πλήρους ιστορίας επειδή η πιθανότητα των στοιχείων που λείπουν του πραγματικού ενεργού συνόλου αναμένεται να αυξηθεί με το p.

Φυσικά, είναι δυνατό να χωρίσουμε τα δεδομένα σε $K > 2$ ομάδες κάθε μέγεθος n/K και εκτιμούμε A από $\hat{A}^{(1)} \cap \hat{A}^{(2)} \cap \ldots \cap \hat{A}^{(K)}$, όπου το $\hat{A}^{(k)}$ παριστάνει το εκτιμώμενο σύνολο ενεργών δεικτών από την k-οστή διαμέριση. Τέτοια διαδικασία επιλογής μεταβλητών θα ήταν ακόμη πιο επιθετική από την εκδοχή με $K = 2$ και βελτιωμένα όρια στο Θεώρημα 1 μπορούν να επιτευχθούν, αλλά η πιθανότητα να λείπουν πραγματικοί ενεργοί δείκτες θα αυξηθεί. Καθώς η διαδικασία

με $K = 2$ είναι ήδη αρκετά επιθετική, το θεωρούμε να είναι η πιο φυσική επιλογή στην πράξη.

Στην επαναληπτική εκδοχή αυτής της πρώτης παραλλαγής της SIS, εφαρμόζουμε τη SIS σε κάθε διαμέριση ξεχωριστά για να πάρουμε δύο σύνολα από δείκτες $\hat{A}_1^{(1)}$ και $\hat{A}_1^{(2)}$, που το καθένα έχει k_1 στοιχεία. Μετά το σχηματισμό της τομής $\hat{A}_1 = \hat{A}_1^{(1)} \bigcap \hat{A}_1^{(2)}$ πραγματοποιούμε την εκτίμηση της ποινικοποιημένης πιθανοφάνειας όπως και πριν για να δώσουμε μια πρώτη προσέγγιση $\hat{M}_1$ προς το πραγματικό ενεργό σύνολο των χαρακτηριστικών. Στη συνέχεια, πραγματοποιούμε ένα δεύτερο στάδιο της διαδικασίας ISIS για κάθε διαμέριση ξεχωριστά για να αποκτήσουμε ομάδες δεικτών $\hat{M}_1 \bigcup \hat{A}_2^{(1)}$ και $\hat{M}_1 \bigcup \hat{A}_2^{(2)}$. Λαμβάνοντας την τομή αυτών των συνόλων και εκτιμώντας εκ νέου τις παραμέτρους χρησιμοποιώντας την ποινικοποιημένη πιθανοφάνεια παίρνουμε μια δεύτερη προσέγγιση $\hat{M}_2$ στο πραγματικό ενεργό σύνολο. Αυτή η διαδικασία μπορεί να συνεχιστεί μέχρι να επιτευχθεί μια επανάληψη l με $\hat{M}_l = \hat{M}_{l-1}$, ή να έχουμε προσλάβει d δείκτες.

2.10.2. Δεύτερη παραλλαγή της ISIS

Η δεύτερη παραλλαγή της SIS είναι μια πιο συντηρητική διαδικασία επιλογής χαρακτηριστικών και βασίζεται επίσης σε τυχαία διαμέριση των δεδομένων σε $K = 2$ ομάδες όπως πριν. Και πάλι εφαρμόζουμε τη SIS σε κάθε διαμέριση ξεχωριστά αλλά τώρα προσλαμβάνουμε όσα περισσότερα χαρακτηριστικά σε ισομεγέθη σύνολα ενεργών δεικτών $\tilde{A}^{(1)}$ και $\tilde{A}^{(2)}$ που απαιτούνται για να εξασφαλίσουμε ότι η τομή τους $\tilde{A} = \tilde{A}^{(1)} \bigcap \tilde{A}^{(2)}$ έχει d στοιχεία. Στη συνέχεια, εφαρμόζουμε μια ποινικοποιημένη μέθοδο ψευδο-πιθανοφάνειας στα χαρακτηριστικά $X_{\tilde{A}} = \{X_j : j \in \tilde{A}\}$ για την τελική επιλογή χαρακτηριστικών και εκτίμηση παραμέτρου.

Θεωρητική στήριξη για αυτή τη μέθοδο μπορεί να παρέχεται στην περίπτωση του γραμμικού μοντέλου, δηλαδή κάτω από ορισμένες συνθήκες κανονικότητας, αυτή η παραλλαγή της SIS κατέχει την ιδιότητα του σίγουρου κρησαρίσματος. Πιο

συγκεκριμένα, αν οι συνθήκες (1)-(4) των Fan & Lv (2008) ισχύουν με $2\kappa+\tau<1$, και επιλέγουμε $d=\lfloor n/\log n\rfloor$, τότε υπάρχει $C>0$ έτσι ώστε

$$P(A\subseteq \tilde{A})=1-O\{\exp(-Cn^{1-2\kappa}/\log n+\log p)\}.$$

Η παράμετρος $\kappa\geq 0$ ελέγχει το ρυθμό με τον οποίο το ελάχιστο σήμα $\min_{j\in A}|\beta_j|$ επιτρέπεται να συγκλίνει στο μηδέν, ενώ το $\tau\geq 0$ ελέγχει το ρυθμό με τον οποίο η μέγιστη ιδιοτιμή του πίνακα συνδιακύμανσης $\Sigma=\mathrm{cov}\left(X_1,...,X_p\right)$ επιτρέπεται να αποκλίνει στο άπειρο. Στην πραγματικότητα, επιμένουμε ότι $\min_{j\in A}\left|\beta_j\right|\geq n^{-\kappa}$ και $\lambda_{\max}(\Sigma)\leq n^{\tau}$ για μεγάλο *n*, όπου $\lambda_{\max}(\Sigma)$ συμβολίζει τη μέγιστη ιδιοτιμή του *Σ*. Έτσι, αυτές οι τεχνικές συνθήκες εξασφαλίζουν ότι οποιοδήποτε μη-μηδενικό σήμα δεν είναι πολύ μικρό, και ότι τα χαρακτηριστικά δεν είναι πάρα πολύ κοντά στο να είναι συγγραμικά και η διάσταση ελέγχεται επίσης μέσω $\log p=o(n^{1-2\kappa}/\log n)$, το οποίο είναι ακόμα εκθετικής τάξης.

Πρόσφατα, οι Fan & Song (2009) επέκτειναν το αποτέλεσμα των Fan & Lv (2008) στα γενικευμένα γραμμικά μοντέλα. Έστω $\hat{L}_0=\min_{\beta_0}n^{-1}\sum_{i=1}^{n}L(Y_i,\beta_0)$ είναι η τιμή αναφοράς στην ταξινόμηση χαρακτηριστικών. Η διαδικασία κατάταξης χαρακτηριστικών είναι ισοδύναμη με τη μέθοδο κατωφλιού (*thresholding)*: $\hat{M}_{v_n}=\{j:L_j-L_0\geq v_n\}$, στο οποίο το v_n είναι η δοσμένη τιμή κατωφλιού. Κάτω από συγκεκριμένες συνθήκες κανονικότητας, αν $\min_{j\in A}\left|\mathrm{cov}(X_j,Y)\right|\geq c_1 n^{-\kappa}$, για κάποιο $c_1>0$ και $\kappa<1/2$ και $v_n=c_0 n^{-2\kappa}$ για αρκετά μικρό c_0 τότε έχουμε $P(A\subset\hat{M}_{v_n})\to 1$, εκθετικά γρήγορα, με την προϋπόθεση ότι $\log p_n=o(n^{1-2\kappa})$. Η ιδιότητα σίγουρου κρησαρίσματος δεν εξαρτάται από τη συσχέτιση των χαρακτηριστικών, όπως αναμενόταν. Ωστόσο, το επιλεγμένο μέγεθος του μοντέλου βασίζεται στη δομή συσχέτισης: Όσο πιο πολύ συσχετίζονται τα χαρακτηριστικά, τόσο μεγαλύτερο είναι το επιλεγμένο μέγεθος του μοντέλου. Στην πραγματικότητα, οι Fan & Song (2009) έδειξαν περαιτέρω ότι με πιθανότητα που τείνει εκθετικά γρήγορα στο 1, $\left|\hat{M}_{v_n}\right|=O(v_n^{-2}\lambda_{\max}(\Sigma))$. Όταν $\lambda_{\max}(\Sigma)=O(n^{\tau})$ και $\lambda_{\max}(\Sigma)=O(n^{-2\kappa})$

, το επιλεγμένο μέγεθος μοντέλου είναι $\left|\hat{\mathrm{M}}_{v_n}\right| = \mathrm{O}(\mathrm{n}^{2\kappa+\tau})$. Πιο συγκεκριμένα, αν η συνθήκη $2\kappa+\tau<1$ επιβάλλεται όπως και στους Fan & Lv (2008) μπορούμε να μειώσουμε με ασφάλεια το μέγεθος του δείγματος σε $o(\mathrm{n})$ μέσω της ανεξάρτητης μάθησης.

Μια επαναληπτική εκδοχή της δεύτερης παραλλαγής της SIS είναι επίσης διαθέσιμη. Κατά το πρώτο στάδιο εφαρμόζουμε τη SIS λαμβάνοντας αρκετά χαρακτηριστικά στα ισομεγέθη σύνολα των ενεργών δεικτών $\tilde{A}_1^{(1)}$ και $\tilde{A}_1^{(2)}$ για να εξασφαλίσει ότι η τομή τους $\tilde{A} = \tilde{A}_1^{(1)} \bigcap \tilde{A}_1^{(2)}$ έχει k_1 στοιχεία. Εφαρμόζοντας την ποινικοποιημένη πιθανοφάνεια στα χαρακτηριστικά με δείκτες στο $\tilde{A}_1$ δίνει μια πρώτη προσέγγιση $\tilde{\mathrm{M}}_1$ στο πραγματικό σύνολο των ενεργών δεικτών. Στη συνέχεια, θα προβούμε σε ένα δεύτερο στάδιο της διαδικασίας ISIS για κάθε διαμέριση ξεχωριστά για να αποκτήσουμε ισομεγέθη νέα σύνολα δεικτών $\tilde{A}_2^{(1)}$ και $\tilde{A}_2^{(2)}$, λαμβάνοντας αρκετά χαρακτηριστικά για να εξασφαλίσουμε ότι η τομή τους $\tilde{A}_2 = \tilde{A}_2^{(1)} \bigcap \tilde{A}_2^{(2)}$ έχει k_2 στοιχεία. Η ποινικοποιημένη πιθανοφάνεια που εφαρμόζεται στο $\tilde{\mathrm{M}}_1 \bigcap \tilde{A}_2$ δίνει μια δεύτερη προσέγγιση $\tilde{\mathrm{M}}_2$ στο πραγματικό σύνολο των ενεργών δεικτών. Όπως και με την πρώτη παραλλαγή, συνεχίζουμε μέχρι να φτάσουμε σε μια επανάληψη l με $\tilde{\mathrm{M}}_l = \tilde{\mathrm{M}}_{l-1}$, ή έχουμε προσλάβει d δείκτες.

2.11. Αριθμητικά στοιχεία

Για να μελετηθεί η απόδοση της μεθόδου ISIS που προτάθηκε παραπάνω, παρουσιάζουμε ένα παράδειγμα προσομοίωσης. Ο στόχος είναι να εξετάσουμε το βαθμό στον οποίο η ISIS μπορεί να βελτιώσει τη SIS στην κατάσταση όπου οι συνθήκες της SIS αποτυγχάνουν. Αξιολογούμε τη μέθοδο μετρώντας τις συχνότητες όπου τα επιλεγμένα μοντέλα περιλαμβάνουν όλες τις μεταβλητές στο πραγματικό μοντέλο, δηλαδή η ικανότητα του σωστού κρησαρίσματος μη-σημαντικών μεταβλητών.

2.11.1. Προσομοιωμένο παράδειγμα

Για το προσομοιωμένο παράδειγμα, χρησιμοποιήσαμε ένα γραμμικό μοντέλο

$$Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon$$

όπου τα $X_1,...,X_p$ είναι p επεξηγηματικές μεταβλητές και $\varepsilon \sim N(0,1)$ είναι ο θόρυβος που είναι ανεξάρτητος από τις επεξηγηματικές μεταβλητές. Στην προσομοίωση, ένα δείγμα $\left(X_1,...,X_p\right)$ με μέγεθος n συντάχθηκε από μια πολυμεταβλητή κανονική κατανομή $N(0,\Sigma)$ του οποίου ο πίνακας συνδιακύμανσης $\Sigma = \left(\sigma_{ij}\right)_{p\times p}$ έχει καταχωρήσεις $\sigma_{ii} = 1,\ i = 1,...,p$ και $\sigma_{ij} = \rho,\ i \neq j$. Θεωρήσαμε 20 τέτοια μοντέλα που χαρακτηρίζονται από $\left(p,n,\rho\right)$ με $p = 100,1000$, n=20,50,70 και ρ=0,0.1,0.5,0.9 και για κάθε μοντέλο προσομοιώνουμε 200 σύνολα δεδομένων.

Για κάθε μοντέλο, εφαρμόσαμε τη SIS και την ISIS για να επιλέξουμε n μεταβλητές και δοκιμάσαμε την ακρίβειά τους στο να περιλαμβάνουν το πραγματικό μοντέλο $\left\{X_1, X_2, X_3\right\}$. Για την ISIS, η SIS-SCAD μέθοδος με $d = \left[n/\log(\mathrm{n})\right]$ χρησιμοποιήθηκε σε κάθε βήμα και συνεχίσαμε να συλλέγουμε μεταβλητές σε αυτές τις ασυνεχείς A_j μέχρι να έχουμε n μεταβλητές (αν υπήρχαν περισσότερες μεταβλητές από αυτές που χρειάζονται στο τελικό στάδιο, θα περιλαμβάναμε μόνο εκείνες με τους μεγαλύτερους απόλυτους συντελεστές). Στον Πίνακα 2.11.1.1, αναφέρουμε τα ποσοστά της SIS, LASSO και ISIS που περιλαμβάνει το πραγματικό μοντέλο. Όλες αυτές οι τρεις μέθοδοι επιλέγουν $\mathrm{n}-1$ μεταβλητές, για να κάνουν δίκαιες συγκρίσεις. Είναι σαφές ότι η συγγραμμικότητα (μεγάλη τιμή του ρ) και η υψηλή διάσταση επιδεινώνει την απόδοση της SIS και της LASSO, και η LASSO υπερτερεί της SIS κατά κάποιο τρόπο. Ωστόσο, όταν το μέγεθος του δείγματος είναι 50 ή μεγαλύτερο, η διαφορά στην απόδοση είναι πολύ μικρή, αλλά η SIS έχει πολύ λιγότερο υπολογιστικό κόστος. Σε αντίθεση, η ISIS βελτιώνει δραματικά την απόδοση αυτής της απλής SIS και LASSO. Πράγματι, σε αυτή την προσομοίωση, η ISIS επιλέγει πάντα όλες τις πραγματικές μεταβλητές. Μπορεί ακόμη και να έχει

λιγότερο υπολογιστικό κόστος από ότι η LASSO, όταν η LASSO χρησιμοποιείται κατά την εφαρμογή της ISIS.

p	*n*	*Method*	*Results for the following values of ρ:*			
			$\rho=0$	$\rho=0.1$	$\rho=0.5$	$\rho=0.9$
100	20	SIS	0.755	0.855	0.690	0.670
		Lasso	0.970	0.990	0.985	0.870
		ISIS	1	1	1	1
	50	SIS	1	1	1	1
		Lasso	1	1	1	1
		ISIS	1	1	1	1
1000	20	SIS	0.205	0.255	0.145	0.085
		Lasso	0.340	0.555	0.556	0.220
		ISIS	1	1	1	1
	50	SIS	0.990	0.960	0.870	0.860
		Lasso	1	1	1	1
		ISIS	1	1	1	1
	70	SIS	1	0.995	0.97	0.97
		Lasso	1	1	1	1
		ISIS	1	1	1	1

<u>Πίνακας 2.11.1.1.</u> Αποτελέσματα του παραδείγματος προσομοίωσης: ακρίβεια της SIS, της LASSO και της ISIS που περιλαμβάνονται στο πραγματικό μοντέλο $\{X_1, X_2, X_3\}$.

2.12. Συμπερασματικά σχόλια

Η ιδέα του σίγουρου κρησαρίσματος εισάγεται και μια μέθοδος σίγουρου κρησαρίσματος, που βασίζεται στη μάθηση συσχέτισης, προτάθηκε και ονομάζεται SIS. Η SIS είναι ικανή να μειώσει τη διάσταση από εκθετικά αυξανόμενη σε κάτω από το μέγεθος του δείγματος, με ακρίβεια. Επιταχύνει δραστικά την επιλογή μεταβλητών και μπορεί να βελτιώσει την ακρίβεια της εκτίμησης όταν η διάσταση είναι υψηλή.

ΚΕΦΑΛΑΙΟ 3ο

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

3.1. Εισαγωγή

Τα γενικευμένα γραμμικά μοντέλα είναι μια αξιοσημείωτη σύνθεση και επέκταση των γνωστών μοντέλων παλινδρόμησης όπως τα γραμμικά μοντέλα και τα μοντέλα logit και probit (Nelder και Wedderburn (1972)). Είναι σημαντικά στην αποτελεσματική ανάλυση στατιστικών δεδομένων, για αυτό αξίζει την προσπάθεια που απαιτείται για την κατανόησή τους.

Τα **γενικευμένα γραμμικά μοντέλα** ή *Generalized Linear Model-GLM* αναπτύχθηκαν για να μας βοηθήσουν να προσαρμόσουμε μοντέλα παλινδρόμησης για μονοπαραμετρικά δεδομένα απόκρισης που ακολουθούν μια πολύ γενική κατανομή που καλείται **εκθετική οικογένεια**. Η εκθετική οικογένεια περιλαμβάνει την κανονική, τη διωνυμική, την Poisson, τη γεωμετρική, την αρνητική διωνυμική, την εκθετική, τη Γάμμα και την αντίστροφη κανονική κατανομή. Μεταγενέστερες εργασίες, όμως, έχουν επεκτείνει τα GLM σε πολυμεταβλητές εκθετικές οικογένειες (όπως η πολυωνυμική κατανομή) σε ορισμένες μη εκθετικές οικογένειες (όπως η δύο παραμέτρων αρνητική διωνυμική κατανομή), και σε ορισμένες περιστώσεις στις οποίες η κατανομή του y_i δεν καθορίζεται πλήρως.

Έστω y_i, $i = 1, 2, ..., n$ τα οποία παριστάνουν τις τιμές απόκρισης, τότε τα GLM δίνονται από τον παρακάτω τύπο:

$$g(\mu_i) = \mathrm{g}\left[\mathrm{E}(y_i)\right] = \mathrm{x}_i' \beta$$

όπου x_i είναι το διάνυσμα των μεταβλητών απόκρισης ή οι συμμεταβλητές για την i-οστή παρατήρηση και β είναι το διάνυσμα των παραμέτρων ή των συντελεστών παλινδρόμησης.

3.2. Τυπική δομή για τα γενικευμένα γραμμικά μοντέλα

Κάθε γενικευμένο γραμμικό μοντέλο αποτελείται από τρεις συνιστώσες:

1. Μια τυχαία συνιστώσα, που προσδιορίζει τη δεσμευμένη κατανομή της μεταβλητής απόκρισης y_i (για το *i*-οστό στοιχείο των *n* ανεξάρτητων παρατηρήσεων δείγματος), και δίνει τις τιμές των επεξηγηματικών μεταβλητών στο μοντέλο (μερικές φορές καλείται και δομή σφάλματος-*error structure*).

2. Μια γραμμική επεξηγηματική μεταβλητή που περιλαμβάνει τις μεταβλητές παλινδρόμησης ή τις συμμεταβλητές.

   Το μοντέλο κατασκευάζεται γύρω από αυτή τη μεταβλητή $\eta = \mathrm{x}'\beta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$. Η συμμετοχή αυτής της γραμμικής επεξηγηματικής μεταβλητής προτείνει την ορολογία των γενικευμένων γραμμικών μοντέλων.

   Όπως και στο γραμμικό μοντέλο και στα logit και probit, οι μεταβλητές παλινδρόμηση x_i είναι προκαθορισμένες συναρτήσεις των επεξηγηματικών μεταβλητών και ως εκ τούτου μπορεί να περιλαμβάνουν ποσοτικές επεξηγηματικές μεταβλητές, μετασχηματισμούς ποσοτικών επεξηγηματικών μεταβλητών, πολυωνυμικές μεταβλητές παλινδρόμησης, dummy μεταβλητές παλινδρόμησης, αλληλεπιδράσεις και ούτω καθ' εξής. Πράγματι, ένα από τα πλεονεκτήματα των GLM είναι ότι η δομή της γραμμικής επεξηγηματικής μεταβλητής είναι η γνωστή δομή ενός γραμμικού μοντέλου.

3. Μια ομαλή και αντιστρέψιμη γραμμική συνάρτηση σύνδεσης (*link function*) *g* που συνδέει τη γραμμική επεξηγηματική μεταβλητή με τη μέση τιμή της μεταβλητής απόκρισης. Η *g* μετασχηματίζει τη μέση τιμή της μεταβλητής απόκρισης $\mu_i \equiv \mathrm{E}(\mathrm{y}_i)$ στη γραμμική επεξηγηματική μεταβλητή:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

   Επειδή η συνάρτηση σύνδεσης είναι αντιστρέψιμη, μπορούμε να τη γράψουμε

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

και έτσι τα GLM μπορούν να θεωρηθούν ως ένα γραμμικό μοντέλο για το μετασχηματισμό της αναμενόμενης απόκρισης ή ως μη γραμμικό μοντέλο παλινδρόμησης για την απόκριση. Ο αντίστροφος σύνδεσμος $g^{-1}(\cdot)$ καλείται και *συνάρτηση μέσου*.

Το μοντέλο βρίσκεται μέσω της χρήσης της συνάρτησης σύνδεσης.

$$\eta_i = g(\mu_i)\,,\ i = 1, 2, ..., n$$

Ο όρος σύνδεση προέρχεται από το γεγονός ότι η συνάρτηση είναι η σύνδεση μεταξύ της μέσης τιμής και της γραμμικής επεξηγηματικής μεταβλητής. Παρατηρούμε ότι η αναμενόμενη απόκριση είναι

$$\mathrm{E}(y_i) = g^{-1}(\eta_i) = g^{-1}(\mathrm{x}_i'\beta)$$

Στην πραγματικότητα, στην πολλαπλή γραμμική παλινδρόμηση το μοντέλο

$$\mu_i = \eta_i = \mathrm{x}_i'\beta\,,\ i = 1, 2, ..., n$$

προτείνει μια ειδική περίπτωση στην οποία $g(\mu_i) = \mu_i$, και για αυτό η συνάρτηση σύνδεσης που χρησιμοποιείται είναι η **ταυτοτική σύνδεση**.

Η συνάρτηση σύνδεσης είναι μια μονότονη διαφορίσιμη συνάρτηση.

Υπάρχουν πολλές πιθανές επιλογές για τη συνάρτηση σύνδεσης. Αν επιλέξουμε $\eta_i = \theta_i$ τότε λέμε ότι η η_i είναι η **κανονική σύνδεση**.

Άλλες συναρτήσεις σύνδεσης που μπορούν να χρησιμοποιηθούν με τα γενικευμένα γραμμικά μοντέλα είναι οι εξής:

1. Η σύνδεση probit

$$\eta_i = \Phi^{-1}\left[\mathrm{E}(y_i)\right]$$

όπου το Φ παριστάνει την αθροιστική τυπική κανονική συνάρτηση κατανομής.

2. Η log-log σύνδεση

$$\eta_i = \ln\{\ln(\mu_i)\}$$

3. Η complementary log-log

$$\eta_i = \ln\{-\ln(1-\mu_i)\}$$

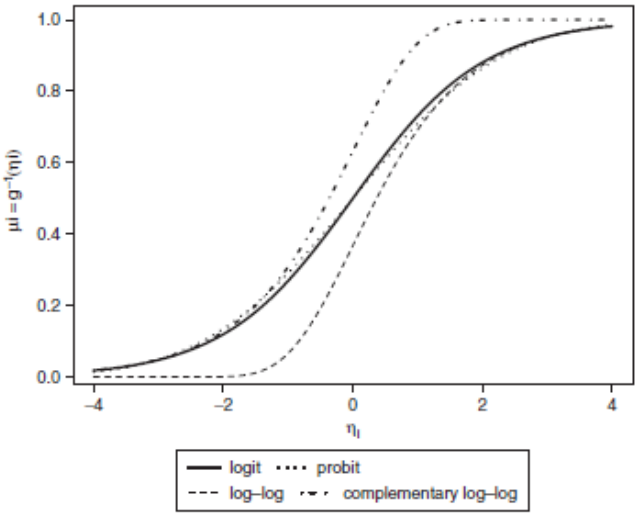
Οι πιο συνηθισμένες συναρτήσεις σύνδεσης και οι αντίστροφές τους φαίνονται στον παρακάτω Πίνακα 3.2.1. Παρατηρούμε ότι η ταυτοτική σύνδεση απλά επιστρέφει το όρισμα της αναλλοίωτο, $\eta_i = g(\mu_i) = \mu_i$, και έτσι $\mu_i = g^{-1}(\eta_i) = \eta_i$.

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

<u>Πίνακας 3.2.1</u>. Μερικές συνηθισμένες συναρτήσεις σύνδεσης και οι αντίστροφές τους, όπου μ_i είναι η μέση τιμή της απόκρισης, η_i η γραμμική επεξηγηματική μεταβλητή, Φ_i η αθροιστική συνάρτηση κατανομής της κανονικής κατανομής.

Οι τέσσερις τελευταίες συναρτήσεις σύνδεσης στον Πίνακα 3.2.1 είναι για διωνυμικά δεδομένα, όπου το y_i αντιπροσωπεύει την παρατηρούμενη αναλογία των επιτυχιών σε n_i ανεξάρτητες δυαδικές δοκιμές. Έτσι, το y_i μπορεί να λάβει οποιεσδήποτε από τις τιμές $0, 1/n_i, 2/n_i, \ldots, (n_i-1)/n_i, 1$. Θυμόμαστε ακόμη ότι τα

διωνυμικά δεδομένα συμπεριλαμβάνουν δυαδικά δεδομένα, όπου όλες οι παρατηρήσεις αντιπροσωπεύουν $n_i = 1$ δοκιμές, και συνεπώς το y_i είναι είτε 0 είτε 1. Η μέση τιμή της απόκρισης $\mu_i = E(y_i)$ είναι τότε η πιθανότητα επιτυχίας, την οποία συμβολίζουμε με π_i. Οι logit, probit, log-log και συμπληρωματικές log-log συνδέσεις απεικονίζονται στο Σχήμα 3.2.1. Σε αντίθεση με τους συνδέσμους logit και probit (οι οποίες είναι σχεδόν δυσδιάκριτες όταν οι διακυμάνσεις των κανονικών και λογιστικών κατανομών εξισώνονται), οι log-log και οι συμπληρωματικοί log-log σύνδεσμοι προσεγγίζουν τις ασύμπτωτες του 0 και 1 ασύμμετρα.



<u>Σχήμα 3.2.1</u>. Οι logit, probit, log-log και συμπληρωματικές log-log συνδέσεις για διωνυμικά δεδομένα. Η διασπορά των κανονικών και λογιστικών κατανομών έχει εξισωθεί για να διευκολύνουν τη σύγκριση των logit και probit συνδέσεων.

Πέρα από τη γενική επιθυμία να επιλέξουμε μια συνάρτηση σύνδεσης που καθιστά την παλινδρόμηση του y_i πάνω στο x γραμμική, μια πολλά υποσχόμενη σύνδεση θα άρει τους περιορισμούς σχετικά με το εύρος της αναμενόμενης απόκρισης. Πρόκειται για κάτι γνώριμο από τα logit και probit μοντέλα, όπου σκοπός ήταν να διαμορφωθεί η πιθανότητα επιτυχίας, εκπροσωπούμενη από το μ_i. Ως πιθανότητα το μ_i περιορίζεται στο διάστημα [0,1]. Οι σύνδεσμοι logit και probit απεικονίζουν αυτό το διάστημα σε ολόκληρη την πραγματική ευθεία από το $-\infty$ έως

το $+\infty$. Ομοίως, αν η απόκριση y είναι υπολογίσιμη, λαμβάνοντας μόνο τις μη αρνητικές ακέραιες τιμές $0,1,2,\ldots$, και κατά συνέπεια και μ_i είναι μια αναμενόμενη καταμέτρηση, η οποία (αν και όχι απαραίτητα ένας ακέραιος) είναι επίσης μη αρνητική, η log σύνδεση απεικονίζει το μ_i σε ολόκληρη την πραγματική ευθεία. Αυτό δε σημαίνει ότι η συνάρτηση σύνδεσης είναι εντελώς καθορισμένη από το εύρος της μεταβλητής απόκρισης.

Μια βολική ιδιότητα των κατανομών των εκθετικών οικογενειών είναι ότι η δεσμευμένη διακύμανση του y_i είναι μια συνάρτηση του μέσου της μ_i (έστω $v(\mu_i)$) και, ενδεχομένως, μιας παραμέτρου διασποράς φ. Οι συναρτήσεις διακύμανσης για τις συνηθισμένες εκθετικές οικογένειες φαίνονται στον Πίνακα 3.2.2. Η δεσμευμένη διακύμανση της απόκρισης στη κανονική οικογένεια είναι μια σταθερά φ, η οποία είναι απλά αυτό που ονομάζουμε και *σφάλμα διακύμανσης*. Στις διωνυμικές και Poisson οικογένειες, η παράμετρος διασποράς έχει οριστεί στη σταθερή τιμή $\varphi = 1$.

Ο Πίνακας 3.2.2 δείχνει ακόμα το εύρος της διακύμανσης της μεταβλητής απόκρισης σε κάθε οικογένεια, και τη λεγόμενη κανονική (ή «φυσική») συνάρτηση σύνδεσης που σχετίζεται με κάθε οικογένεια. Η κανονική σύνδεση απλοποιεί τα GLM, μπορούν όμως να χρησιμοποιηθούν και άλλες συναρτήσεις σύνδεσης. Πράγματι, ένα από τα πλεονεκτήματα του παραδείγματος GLM- σε αντίθεση με τους μετασχηματισμούς της μεταβλητής απόκρισης στη γραμμική παλινδρόμηση- είναι ότι η επιλογή του μετασχηματισμού γραμμικοποίησης εν μέρει διαχωρίζεται από την κατανομή της απόκρισης, και ο ίδιος μετασχηματισμός δεν έχει να ομαλοποιήσει την κατανομή του y και να κάνει και την παλινδρόμησή της πάνω στα x γραμμική. Οι ειδικές συνδέσεις που μπορούν να χρησιμοποιηθούν διαφέρουν από τη μία οικογένεια στην άλλη και επίσης- σε ορισμένο βαθμό- από το ένα λογισμικό εφαρμογής των GLM σε ένα άλλο. Για παράδειγμα, δεν είναι πολλά υποσχόμενο να χρησιμοποιήσουμε την ταυτοτική, τη log, την αντίστροφη, και το αντίστροφο τετράγωνο ή το σύνδεσμο τετραγωνικής ρίζας με διωνυμικά δεδομένα, ούτε θα ήταν λογικό να χρησιμοποιήσουμε τα logit, probit, log-log ή τη συμπληρωματική log-log σύνδεση με μη διωνυμικά δεδομένα.

Family	Canonical Link	Range of Y_i	$V(Y_i\|\eta_i)$
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0,1,\ldots,n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	$0,1,2,\ldots$	μ_i
Gamma	Inverse	$(0,\infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	$(0,\infty)$	$\phi\mu_i^3$

<u>Πίνακας 3.2.2</u>. Κανονική σύνδεση, Εύρος Απόκρισης και Συνεχής Συνάρτηση Διασποράς για τις Εκθετικές Οικογένειες.

3.3. Στατιστική θεωρία για τα Γενικευμένα Γραμμικά Μοντέλα

3.3.1. Εκθετικές Οικογένειες

Μια σημαντική αρχή που διέπει τα GLM είναι η εκθετική οικογένεια κατανομών. Στη σύγχρονη στατιστική, οι περισσότερες σημαντικές κατανομές μπορούν να εκφραστούν από τη «γραμμική-εκθετική» μορφή του Fisher, που είναι η συνάρτηση πυκνότητας πιθανότητας μιας παρατηρούμενης απόκρισης y:

$$\mathrm{f}(\mathrm{y};\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(\mathrm{y},\phi)\right] \tag{3.3.1}$$

όπου

- $f(\mathrm{y};\theta,\phi)$ είναι η συνάρτηση πιθανότητας για τη διακριτή τυχαία μεταβλητή y, ή η συνάρτηση πυκνότητας πιθανότητας για συνεχή y.
- $a(\cdot)$, $b(\cdot)$ και $c(\cdot)$ είναι γνωστές συναρτήσεις που διαφέρουν από μια εκθετική οικογένεια σε μια άλλη
- $\theta = \mathrm{g}_c(\mu)$, η κανονική παράμετρος για την εκθετική οικογένεια είναι μια συνάρτηση της μέσης τιμής της y, $\mu \equiv \mathrm{E}(\mathrm{y})$. Επιπλέον η κανονική συνάρτηση παραμέτρου $\mathrm{g}_c(\cdot)$ δεν εξαρτάται από το ϕ .

- $\phi > 0$ είναι μια *παράμετρος διακύμανσης*, η οποία, σε μερικές οικογένειες, παίρνει μια σταθερή, γνωστή τιμή, ενώ για ορισμένες άλλες οικογένειες είναι μια άγνωστη παράμετρος που πρέπει να εκτιμηθεί από τα δεδομένα μαζί με το θ.

Η συνάρτηση $a(\phi)$ είναι συνήθως της μορφής $a(\phi) = \phi \cdot \omega$, όπου ω είναι μια γνωστή σταθερά. Θυμίζουμε ότι σε κάποιες περιπτώσεις, όπως η διωνυμική και Poisson ισχύει $\phi = 1.0$.

Το πιο διακεκριμένο μέλος της εκθετικής οικογένειας είναι η γνωστή κανονική κατανομή. Η συνάρτηση πυκνότητας πιθανότητας για μια κανονική τυχαία μεταβλητή y με παραμέτρους μ και σ δίνονται από τον τύπο:

$$\mathrm{f}(\mathrm{y};\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\mathrm{y}-\mu)^2}{2\sigma^2}\right]$$

$$= \exp\left\{(\mathrm{y}\,\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}[y^2/\sigma^2 + \ln(2\pi\sigma^2)]\right\} \qquad (3.3.2)$$

Αυτή η συνάρτηση πυκνότητας είναι της μορφής της εξίσωσης (3.3.1) με $\theta = \mu$, $b(\theta) = \mu^2/2$, $a(\phi) = \phi$, $\phi = \sigma^2$, και $c(\mathrm{y},\phi) = -\frac{1}{2}[y^2/\sigma^2 + \ln(2\pi\sigma^2)]$, όπου μ είναι η παράμετρος θέσης και σ^2 είναι η φυσική παράμετρος κλίμακας.

Θεωρούμε τώρα την Poisson κατανομή. Γνωρίζουμε ότι η συνάρτηση πιθανότητας δίνεται από τη σχέση

$$\mathrm{f}(\mathrm{y};\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

$$= \exp[y\ln\mu - \mu - \ln(\mathrm{y}!)]$$

Ως αποτέλεσμα, $\theta = \ln\mu$, $b(\theta) = e^{\theta}$ και $c(\mathrm{y},\phi) = -\ln(\mathrm{y}!)$. Έτσι, η παράμετρος θέσης είναι μ και η παράμετρος κλίμακας είναι $\phi = 1.0$. Οι κατανομές Poisson για τις διάφορες τιμές της παραμέτρου μ φαίνονται γραφικά στο Σχήμα 3.3.1.2. Καθώς το

μ αυξάνει, η κατανομή Poisson μεγαλώνει πιο συμμετρικά και τελικά προσεγγίζεται καλά από την κανονική κατανομή.

Στην περίπτωση της διωνυμικής κατανομής, θεωρούμε ότι η y ακολουθεί την κανονική κατανομή με παράμετρο μ και P το ποσοστό των επιτυχιών (δηλαδή η πιθανότητα σε μια δοκιμή να προκύψει το υπό μελέτη γεγονός) οπότε για τη συνάρτηση πυκνότητας πιθανότητας έχουμε

$$f(\mathrm{y}) = \binom{n}{P} P^{y} (1-\mathrm{P})^{n-y}, \; y = 0,1,2,....,n, \; 0 \le P \le 1$$

$$f(\mathrm{y}) = \exp\left\{ y \ln P + (\mathrm{n}-\mathrm{y}) \ln(1-\mathrm{P}) + \ln\binom{n}{y} \right\}$$

$$= \exp\left\{ y \ln \frac{P}{1-P} + n \ln(1-\mathrm{P}) + \ln\binom{n}{y} \right\}$$

οπότε, $\theta = \ln\left(\frac{P}{1-P}\right)$, $b(\theta) = n\ln(1+e^{\theta})$, $\phi = 1.0$, $a(\phi) = 1$ και $c(\cdot) = \ln\binom{n}{y}$.

Η κατανομή Γάμμα είναι μια συνεχής οικογένεια με συνάρτηση πυκνότητας πιθανότητας προσαρμοσμένη από την παράμετρο κλίμακας $\omega > 0$ και παράμετρο σχήματος $\psi > 0$:

$$f(\mathrm{y}) = \left(\frac{y}{\omega}\right)^{\psi-1} \times \frac{\exp\left(\frac{-y}{\omega}\right)}{\omega\Gamma(\psi)} \quad \text{για} \quad y > 0$$

όπου $\Gamma(\cdot)$ είναι η Γάμμα συνάρτηση. Η μέση τιμή και η διασπορά της Γάμμα κατανομής είναι αντίστοιχα $E(y) = \omega\psi$ και $V(y) = \omega^2\psi$. Στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, όπου για τη Γάμμα οικογένεια, $(y) = \phi\mu^2$, η

παράμετρος διασποράς είναι απλά η αντίστροφη της παραμέτρου ϕ, δηλαδή $\phi = 1/\psi$. Η παράμετρος κλίμακας στη Γάμμα οικογένεια επηρεάζει την έκταση (και παρεμπιπτόντως τη θέση) αλλά όχι το σχήμα της κατανομής, ενώ η παράμετρος σχήματος ελέγχει την ασυμμετρία της κατανομής. Το Σχήμα 3.3.1.3 δείχνει τις Γάμμα κατανομές για κλίμακα $\omega = 1$ και τις διάφορες τιμές της παραμέτρου σχήματος ψ. (Η αλλαγή της παραμέτρου κλίμακας θα αλλάξει μόνο την επισήμανση του οριζόντιου άξονα στο γράφημα.) Καθώς η παράμετρος σχήματος μεγαλώνει, η κατανομή γίνεται όλο και πιο συμμετρική. Η Γάμμα κατανομή είναι χρήσιμη για τη μοντελοποίηση μιας θετικής συνεχούς μεταβλητής απόκρισης, όπου η σχετική διακύμανση της απόκρισης μεγαλώνει με τη μέση τιμή της αλλά όπου ο συντελεστής διακύμανσης της απόκρισης, $SD(y)/\mu$, είναι σταθερά.

Η αντίστροφη-κανονική κατανομή είναι άλλη μια συνεχής οικογένεια προσαρμοσμένη από δύο παραμέτρους, μ και λ, με συνάρτηση πυκνότητας

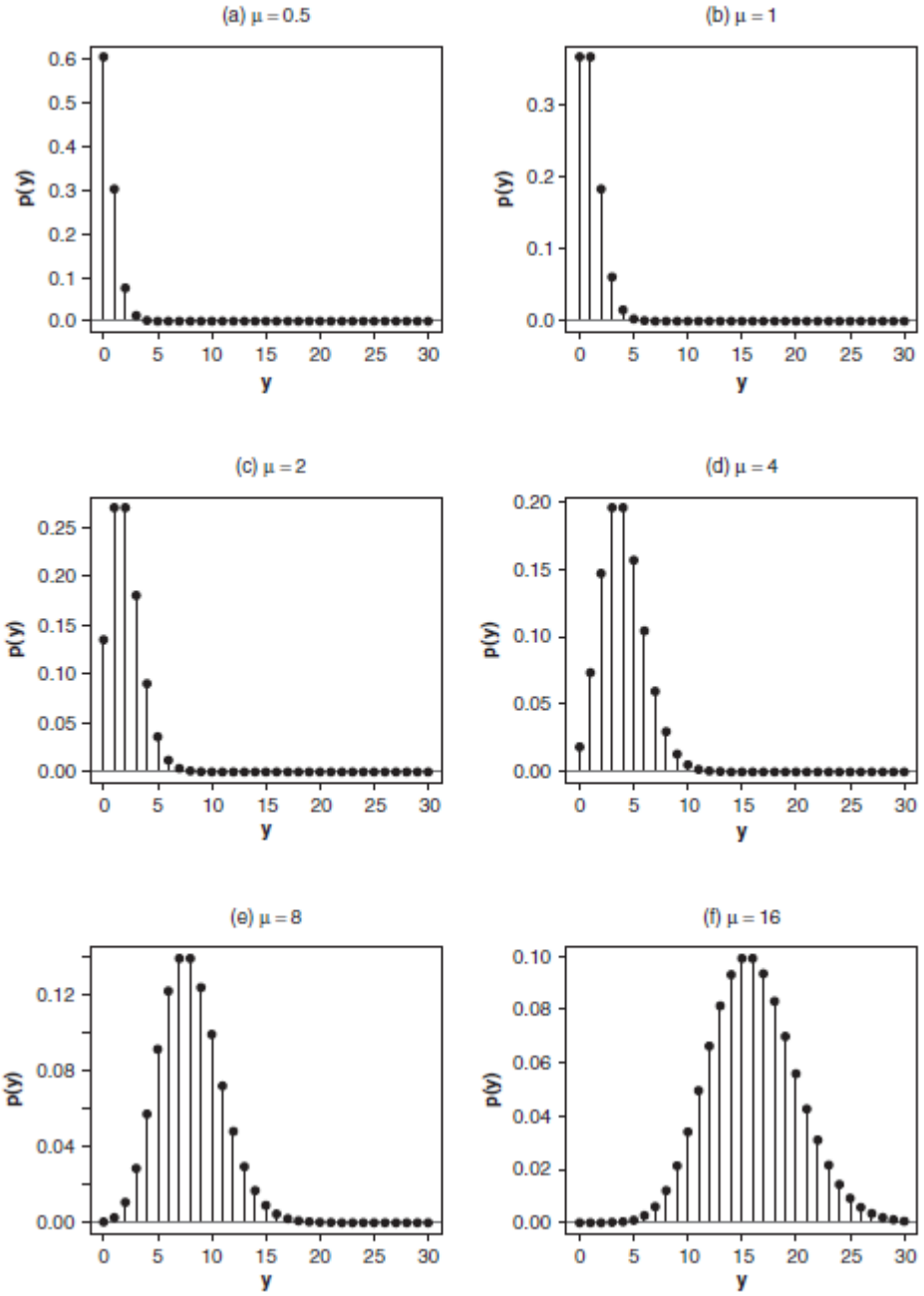
$$f(\mathrm{y}) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(\mathrm{y}-\mu)^2}{2y\mu^2} \right] \text{ για } y > 0$$

Η μέση τιμή και η διασπορά του y είναι $E(y) = \mu$ και $V(y) = \mu^3/\lambda$. Στο πλαίσιο των γενικευμένων γραμμικών μοντέλων, για την αντίστροφη-κανονική οικογένεια, $V(y) = \phi\mu^3$, λ είναι η αντίστροφη της παραμέτρου διασποράς ϕ. Όπως και με τη Γάμμα κατανομή, η διασπορά της αντίστροφης-κανονικής κατανομής αυξάνει με το μέσο της, αλλά σε ένα πιο γρήγορο ρυθμό. Η ασυμμετρία επίσης αυξάνει με την τιμή μ και μειώνεται με την τιμή του λ. Η αντίστροφη-κανονική κατανομή φαίνονται στο Σχήμα 3.3.1.4 .

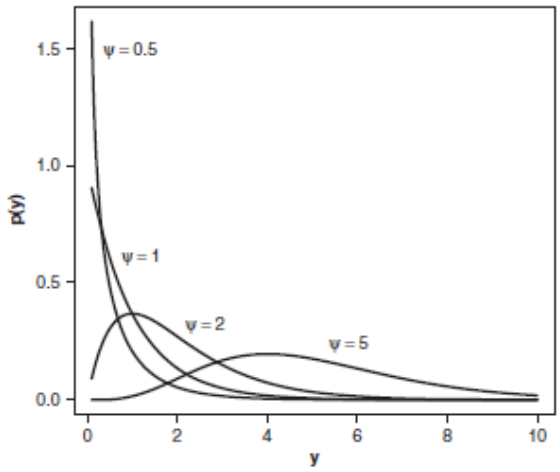
Τα αποτελέσματα για την κανονική, τη διωνυμική, την Poisson, τη Γάμμα και την αντίστροφη-κανονική οικογένεια φαίνονται συγκεντρωτικά στον παρακάτω πίνακα 3.3.1.1.

Family	$a(\phi)$	$b(\theta)$	$c(y,\phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2}\left[y^2/\phi+\log_e(2\pi\phi)\right]$
Binomial	$1/n$	$\log_e(1+e^\theta)$	$\log_e\binom{n}{ny}$
Poisson	1	e^θ	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2}\log_e(y/\phi)-\log_e y-\log_e\Gamma(\phi^{-1})$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left[\log_e(\pi\phi y^3)+1/(\phi y)\right]$

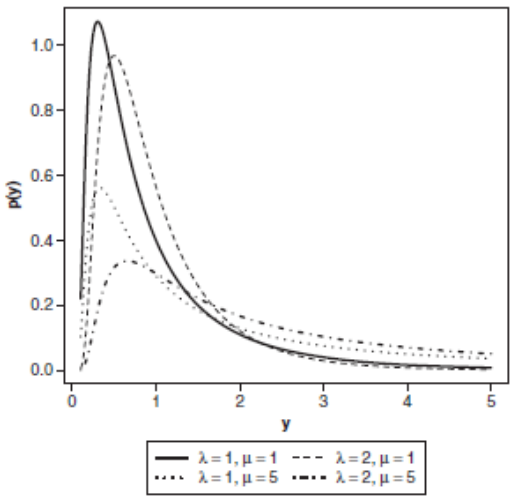
<u>Πίνακας 3.3.1.1</u>. Οι συναρτήσεις $a(\cdot)$, $b(\cdot)$ και $c(\cdot)$ για την κατασκευή των εκθετικών οικογενειών. (Στον πίνακα έχουμε θεωρήσει *n* τον αριθμό των διωνυμικών παρατηρήσεων)



<u>Σχήμα 3.3.1.2</u>. Poisson κατανομές για τις διάφορες τιμές της παραμέτρου ποσοστού μ .



<u>Σχήμα 3.3.1.3.</u> Διάφορες Γάμμα κατανομές για κλίμακα $\omega = 1$ και διάφορες τιμές της παραμέτρου σχήματος ψ .



<u>Σχήμα 3.3.1.4</u>. Αντίστροφες-Κανονικές κατανομές για διάφορους συνδυασμούς των τιμών του μέσου μ και την αντίστροφη-διασπορά λ .

3.4. SIS στα γενικευμένα γραμμικά μοντέλα

Θεωρούμε την ανεξάρτητη μάθηση από την κατάταξη του εκτιμητή μέγιστης οριακής πιθανοφάνειας (*MMLE*) ή την ίδια τη μέγιστη οριακή πιθανοφάνεια για τα γενικευμένα γραμμικά μοντέλα. Δηλαδή, ταιριάζουμε *p* οριακές παλινδρομήσεις από τη μεγιστοποίηση της οριακής πιθανοφάνειας με απόκριση *y* και οριακή συμμεταβλητή $x_i, \mathrm{i} = 1, ..., \mathrm{p}$ (και το σημείο τομής) κάθε φορά. Το μέγεθος των απόλυτων τιμών του MMLE μπορεί να διατηρήσει τις πληροφορίες της μη-σποραδικότητας του μοντέλου της κοινής παλινδρόμησης, με την προϋπόθεση ότι οι πραγματικές τιμές της οριακής πιθανοφάνειας διατηρούν τη μη-σποραδικότητα των κοινών μοντέλων παλινδρόμησης και ότι η MMLE εκτιμάει τις πραγματικές τιμές της οριακής πιθανοφάνειας ομοιόμορφα καλά. Οι Hall, Titterington & Xue (2009) χρησιμοποίησαν μια διαφορετική οριακή χρησιμότητα, που προέρχεται από ένα σημείο εμπειρικής πιθανοφάνειας. Οι Hall & Miller (2009) πρότειναν μια γενικευμένη κατάταξη συσχέτισης, η οποία επιτρέπει τη μη-γραμμική παλινδρόμηση. Οι δύο αυτές εργασίες πρότειναν μια ενδιαφέρουσα μέθοδο bootstrap για να εκτιμήσουν το κύρος των επιλεγμένων χαρακτηριστικών.

Καθώς η MMLE ή η κατάταξη μέγιστης πιθανοφάνειας είναι ισοδύναμη με την κατάταξη οριακής συσχέτισης στα συνηθισμένα γραμμικά μοντέλα, το έργο μας μπορεί επομένως να θεωρηθεί ως μια σημαντική επέκταση της SIS στους Fan & Lv (2008), όπου η κοινή κανονικότητα της απόκρισης και οι συμμεταβλητές επιβάλλονται. Επιπλέον, τα αποτελέσματά μας είναι βελτιωμένα έναντι εκείνων που υπάρχουν στους Fan & Lv (2008) σε τουλάχιστον 3 στοιχεία. Πρώτον, δημιουργούμε ένα νέο πλαίσιο για τις SIS ιδιότητες, οι οποίες δεν βασίζονται στην υπόθεση της κανονικότητας ακόμη και στον καθορισμό του γραμμικού μοντέλου. Δεύτερον, ενώ δεν είναι προφανές (και θα μπορούσε να είναι δύσκολο) να γενικεύσουμε την απόδειξη των Fan & Lv (2008) σε πιο πολύπλοκα μοντέλα, στο τρέχον πλαίσιο, η διαδικασία SIS μπορεί να εφαρμοστεί στα γενικευμένα γραμμικά μοντέλα και ενδεχομένως και σε άλλα μοντέλα. Τρίτον, τα αποτελέσματά μας μπορούν εύκολα να εφαρμοστούν στην κατάταξη γενικευμένης συσχέτισης (Hall & Miller (2009)) και σε άλλες κατατάξεις που βασίζονται σε μια ομάδα οριακών μεταβλητών.

Η εφαρμογή των οριακών μοντέλων σε μια κοινή παλινδρόμηση είναι ένα είδος ατελούς μοντέλου (White, 1982) από τη στιγμή που αφαιρούμε τις περισσότερες συμμεταβλητές από το προσαρμοσμένο μοντέλο. Δημιουργούμε ένα όριο μη–ασυμπτωτικής ουράς πιθανότητας για τη MMLE κάτω από το μοντέλο ατελειών, το οποίο είναι πέρα από το παραδοσιακό ασυμπτωτικό πλαίσιο του ατελούς μοντέλου. Ως μια πρακτική μέθοδος κρησαρίσματος, το ανεξάρτητο κρησάρισμα μπορεί να χάσει μεταβλητές που είναι οριακά ασθενώς ασυσχέτιστες με τις μεταβλητές απόκρισης, αλλά είναι από κοινού άκρως σημαντικές για τις μεταβλητές απόκρισης, και ταξινομεί ακόμα κάποιες κοινές μη σημαντικές υψηλές μεταβλητές με τη χρήση οριακών μεθόδων. Οι Fan & Lv (2008) και οι Fan, Samworth & Wu (2009) ανέπτυξαν μεθόδους επαναληπτικού συμβατικού κρησαρίσματος και επιλογής για να κάνει τις διαδικασίες ισχυρές (*robust*) και πρακτικές. Η πρώτη επικεντρώνεται στα συνήθη γραμμικά μοντέλα ενώ η τελευταία βελτιώνει την ιδέα της πρώτης και διευρύνει σημαντικά το πεδίο εφαρμογής, συμπεριλαμβανομένων των γενικευμένων γραμμικών μοντέλων.

Η SIS ιδιότητα μπορεί να επιτευχθεί όσο η οριακή χρησιμότητα μπορεί να διατηρήσει τη μη-σποραδικότητα των τιμών της πραγματικής παραμέτρου. Με μια παρόμοια ιδέα, οι Fan, Samworth & Wu (2009) πρότειναν μια διαδικασία SIS για τα γενικευμένα γραμμικά μοντέλα, με τη διαλογή των συναρτήσεων μέγιστης πιθανοφάνειας, η οποία είναι μια μορφή κατάταξης του «λόγου οριακής πιθανότητας», ενώ η MMLE μπορεί να θεωρηθεί ως ένα είδος *Wald* της στατιστικής. Οι δύο μέθοδοι είναι ισοδύναμες με όρους των ιδιοτήτων του σίγουρου κρησαρίσματος στο προτεινόμενο πλαίσιό μας. Η βασική τεχνική πρόκληση στην κατάταξη της μέγιστης οριακής πιθανοφάνειας είναι ότι το σήμα μπορεί να είναι πιο αδύναμο από το θόρυβο. Ξεπερνάμε αυτή την τεχνική δυσκολία χρησιμοποιώντας την ιδιότητα συμμετρίας της κατάταξης με μονότονους μετασχηματισμούς.

Στη συνέχεια θα αναφερθούμε στις ρυθμίσεις των γενικευμένων γραμμικών μοντέλων. Θα αναφερθούμε επίσης στο εκθετικό όριο για τον εκτιμητή ημι-μέγιστης πιθανοφάνειας, στις SIS ιδιότητες της MMLE μάθησης στο κρησάρισμα οριακής πιθανοφάνειας και θα δείξουμε τη SIS ιδιότητα. Επίσης, θα παρουσιαστούν και μερικά αποτελέσματα προσομοίωσης.

3.4.1. Γενικευμένα Γραμμικά Μοντέλα

Η συνάρτηση πυκνότητας πιθανότητας στην κανονική μορφή δίνεται από τη σχέση (3.1). Θεωρούμε το πρόβλημα της εκτίμησης ενός (*p+1*)-διανύσματος της παραμέτρου $\beta = (\beta_0, \beta_1, ..., \beta_p)$ από το ακόλουθο γενικευμένο γραμμικό μοντέλο:

$$\mathrm{E}(\mathrm{y} \mid \mathrm{x}) = \mathrm{b}'(\theta(\mathrm{x})) = \mathrm{g}^{-1}(\sum_{j=0}^{p} \beta_j x_j) \qquad (3.4.1.2)$$

όπου $x = \left\{x_0, x_1, ..., x_p\right\}^T$ είναι ένα (*p+1*)-διάστασης συμμεταβλητή και $x_0 = 1$ το σημείο τομής. Αν g είναι η κανονική σύνδεση, δηλαδή, $g = (\mathrm{b}')^{-1}$, τότε $\theta(\mathrm{x}) = \sum_{j=0}^{p} \beta_j x_j$. Επικεντρωνόμαστε στη συνάρτηση κανονικής σύνδεσης για την απλότητα της παρουσίασης.

Υποθέτουμε ότι τα παρατηρούμενα δεδομένα $\left\{(\mathrm{x}_i, \mathrm{y}_i), \mathrm{i} = 1, ... \mathrm{n}\right\}$ είναι τα i.i.d. αντίγραφα του (x, y). Αφήνουμε το *p* να αυξάνεται με το *n* και το δηλώνουμε ως p_n όπου χρειάζεται.

Παρατηρούμε ότι το συνηθισμένο γραμμικό μοντέλο $Y = X^T \beta + \varepsilon$, με $\varepsilon \sim \mathrm{N}(0,1)$ είναι μια ειδική περίπτωση του μοντέλου (3.4.1.2), παίρνοντας $g(\mu) = \mu$ και $b(\theta) = \theta^2/2$. Όταν ο πίνακας σχεδιασμού X είναι κανονικοποιημένος, η κατάταξη του μεγέθους της οριακής συσχέτισης είναι στην πραγματικότητα η ίδια με την κατάταξη του μεγέθους οριακής πιθανοφάνειας του εκτιμητή (*MMLE*). Στη συνέχεια, προτείνουμε μια μέθοδο ανεξάρτητου κρησαρίσματος για τα γενικευμένα γραμμικά μοντέλα (*GLM*) που βασίζεται στα MMLE. Επίσης, υποθέτουμε ότι οι συμμεταβλητές είναι κανονικοποιημένες ώστε να έχουν μέση τιμή μηδέν και τυπική απόκλιση:

$$\mathrm{E}\ \mathrm{X}_j = 0 \quad \text{και} \quad \mathrm{E}\ \mathrm{X}_j^{\ 2} = 1, \qquad j = 1, ..., p_n$$

3.4.2. Ανεξάρτητο κρησάρισμα με MMLE

Έστω $M_* = \left\{1 \le j \le p_n : \beta_j^* \ne 0\right\}$ να είναι το πραγματικό αραιό μοντέλο με μέγεθος μη σποραδικότητας $s_n = \left|\mathrm{M}_*\right|$, όπου το $\beta = (\beta_0^*, \beta_1^*, ..., \beta_{p_n}^*)$ δηλώνει την πραγματική αξία. Ο εκτιμητής μέγιστης οριακής πιθανοφάνειας (*MMLE*) $\hat{\beta}_j^M$, για $j = 1, ..., p_n$ ορίζεται ως ο ελαχιστοποιητής της κατά συνιστώσες παλινδρόμησης:

$$\hat{\beta}_j^M = (\hat{\beta}_{j,0}^M, \hat{\beta}_j^M) = \underset{\beta_0, \beta_j}{\arg\min} \mathrm{P}_n \, l(\beta_0 + \beta_j \, \mathrm{X}_j, \mathrm{Y})$$

όπου $\mathrm{l}(\mathrm{Y}; \theta) = -\left[\theta Y - b(\theta) - \log c(\mathrm{Y})\right]$ και $P_n f(\mathrm{X}, \mathrm{Y}) = \mathrm{n}^{-1} \sum_{i=1}^{n} f(\mathrm{X}_i, \mathrm{Y}_i)$ είναι το εμπειρικό μέτρο. Αυτό μπορεί να υπολογιστεί γρήγορα και η εφαρμογή του είναι ισχυρή. Ορίζουμε, αντίστοιχα, την εκδοχή του πληθυσμού του ελαχιστοποιητή της παλινδρόμησης κατά συνιστώσες,

$$\beta_j^M = (\beta_{j,0}^M, \beta_j^M) = \underset{\beta_0, \beta_j}{\arg\min} \mathrm{E} \, l(\beta_0 + \beta_j \, \mathrm{X}_j, \mathrm{Y}) \quad \text{για} \quad j = 1, ..., p_n$$

όπου το E ορίζει τη μέση τιμή κάτω από το πραγματικό μοντέλο.

Επιλέγουμε ένα σύνολο μεταβλητών

$$\hat{M}_{\gamma_n} = \left\{1 \le j \le p_n : \left|\hat{\beta}_j^M\right| \ge \gamma_n\right\}, \qquad (3.4.2.1)$$

όπου το γ_n είναι μια προκαθορισμένη τιμή. Μια τέτοια ανεξάρτητη μάθηση ταξινομεί τη σημαντικότητα των χαρακτηριστικών σύμφωνα με το μέγεθος των συντελεστών οριακής παλινδρόμησης. Με την ανεξάρτητη μάθηση, μειώνουμε ραγδαία τη διάσταση του χώρου παραμέτρων από p_n (πιθανόν εκατοντάδων χιλιάδων) σε ένα πολύ μικρότερο αριθμό, επιλέγοντας ένα μεγάλο γ_n, και ως εκ τούτου, ο υπολογισμός είναι πολύ πιο εφικτός. Παρά το γεγονός ότι οι ερμηνείες και οι επιπτώσεις των οριακών μοντέλων είναι προκατειλημμένες από το κοινό μοντέλο, η μη-αραιή πληροφορία σχετικά με το κοινό μοντέλο μπορεί να περάσει μαζί με το

οριακό μοντέλο κάτω από ήπιες συνθήκες. Ως εκ τούτου, είναι κατάλληλο για το σκοπό του κρησαρίσματος μεταβλητών. Στη συνέχεια, θα δείξουμε ότι κάτω από ορισμένες συνθήκες η ιδιότητα του σίγουρου κρησαρίσματος διατηρείται, δηλαδή, το σύνολο M_* ανήκει στο $\hat{M}_{\gamma_n}$ με πιθανότητα 1 ασυμπτωτικά, για κατάλληλη επιλογή του γ_n. Για να επιτευχθεί αυτό, χρειαζόμαστε τις ακόλουθες τεχνικές κατασκευής.

3.5. Ένα εκθετικό φράγμα της QMLE

Σε αυτή την ενότητα, παίρνουμε ένα εκθετικό φράγμα της ψευδο-MLE (*QMLE*). Δεδομένου ότι το αποτέλεσμα αυτό διατηρείται κάτω από γενικές συνθήκες και είναι ίδιου ενδιαφέροντος, στη συνέχεια κάνουμε μια πιο γενική περιγραφή του μοντέλου και των συνθηκών του. Θεωρούμε τα δεδομένα $\{X_i, Y_i\}, i = 1,...,n$ να είναι τα n i.i.d. δείγματα του $(\mathrm{X}, \mathrm{Y}) \in X \times Y$ για κάποιο χώρο X και Y. Θεωρούμε ένα μοντέλο παλινδρόμησης για X και Y με συνάρτηση μερικής-πιθανοφάνειας $-\mathrm{l}(\mathrm{X}^T \beta, \mathrm{Y})$. Εδώ, τα Y και $\mathrm{X} = (\mathrm{X}_1,...,\mathrm{X}_q)^T$ παριστάνουν την απόκριση και το q-διάστασης διάνυσμα συμμεταβλητών, το οποίο περιλαμβάνει τόσο διακριτά όσο και συνεχή στοιχεία και η διάσταση μπορεί επίσης να εξαρτάται από το n. Έστω

$$\beta_0 = \arg\min_{\beta} El(\mathrm{X}^T \beta, \mathrm{Y})$$

είναι η παράμετρος πληθυσμού. Θεωρούμε ότι το β_0 είναι ένα εσωτερικό σημείο ενός επαρκώς μεγάλου, συμπαγούς και κυρτού συνόλου $\mathrm{B} \in R^q$. Οι επόμενες συνθήκες στο μοντέλο είναι αναγκαίες.

(A) Η Fisher πληροφορία,

$$\mathrm{I}(\beta) = \mathrm{E}\left\{\left[\frac{\partial}{\partial\beta} l(\mathrm{X}^T \beta, \mathrm{Y})\right]\left[\frac{\partial}{\partial\beta} l(\mathrm{X}^T \beta, \mathrm{Y})\right]^T\right\},$$

είναι πεπερασμένη και θετικά ορισμένη σε $\beta = \beta_0$. Επιπλέον, υπάρχει το $\|\mathrm{I}(\beta)\|_B = \sup_{\beta\in\mathrm{B}, \|x\|=1} \left\|\mathrm{I}(\beta)^{\frac{1}{2}} x\right\|$, όπου $\|\cdot\|$ είναι η Ευκλείδεια νόρμα.

(B) Η συνάρτηση $l(\mathrm{x}^T\beta,\mathrm{y})$ ικανοποιεί τη συνθήκη Lipschitz με θετική σταθερά k_n

$$\left|l(\mathrm{x}^T\beta,\mathrm{y})-l(\mathrm{x}^T\beta',\mathrm{y})\right|\mathrm{I}_n(\mathrm{x},\mathrm{y})\le \mathrm{k}_n\,|\,\mathrm{x}^T\beta-\mathrm{x}^T\beta'\,|\,\mathrm{I}_n(\mathrm{x},\mathrm{y})$$

για $\beta,\beta'\in\mathrm{B}$, όπου $I_n(\mathrm{x},\mathrm{y})=\mathrm{I}((\mathrm{x},\mathrm{y})\in\Omega_n)$ με $\Omega_n=\left\{(\mathrm{x},\mathrm{y}):\|\,\mathrm{x}\,\|_\infty\le K_n,|\,\mathrm{y}\,|\le K_n^*\right\}$

για κάποιες αρκετά μεγάλες θετικές σταθερές K_n και K_n^* και $\|\cdot\|_\infty$ να είναι η *supremum* νόρμα. Επιπλέον, υπάρχει μια αρκετά μεγάλη σταθερά C έτσι ώστε $b_n=Ck_nV_n^{-1}(q/n)^{1/2}$ και V_n σε συνάρτηση με το C

$$\sup_{\beta\in\mathrm{B},\|\beta-\beta_0\|\le \mathrm{b}_n}\left|\mathrm{E}[\mathrm{l}(\mathrm{X}^T\beta,\mathrm{Y})-\mathrm{l}(\mathrm{X}^T\beta_0,\mathrm{Y})](1-I_n(\mathrm{X},\mathrm{Y}))\le \mathrm{o}(\mathrm{q}/\mathrm{n})\right|,$$

όπου το V_n είναι σταθερά συναρτήσει του C.

(Γ) Η συνάρτηση $l(\mathrm{x}^T\beta,\mathrm{y})$ είναι κυρτή στο β, ικανοποιώντας

$$\mathrm{E}(\mathrm{l}(\mathrm{X}^T\beta,\mathrm{Y})-\mathrm{l}(\mathrm{X}^T\beta_0,\mathrm{Y}))\ge V_n\,\|\,\beta-\beta_0\,\|^2$$

για όλα τα $\|\beta-\beta_0\|\le b_n$ και κάποιες θετικές σταθερές V_n.

Η συνθήκη Α διασφαλίζει την ταυτότητα και την ύπαρξη της QMLE και ικανοποιείται για πολλά παραδείγματα των γενικευμένων γραμμικών μοντέλων. Οι συνθήκες Α και Γ επικαλύπτονται αλλά δεν είναι οι ίδιες.

Δημιουργούμε ένα εκθετικό φράγμα για την πιθανότητα ουράς της QMLE

$$\hat\beta=\arg\min_\beta P_n l(\mathrm{X}^T\beta,\mathrm{Y})\,.$$

Η ιδέα της απόδειξης είναι να συνδέσουμε το $\sqrt{n}\,\|\,\hat\beta-\beta_0\,\|$ με την ουρά ορισμένων εμπειρικών διαδικασιών και να χρησιμοποιήσουμε την κυρτότητα και τις συνθήκες Lipschitz.

Θεώρημα 1. Κάτω από τις συνθήκες Α-Γ, ισχύει, για κάποιο $t > 0$,

$$P(\sqrt{n}\left\|\hat{\beta} - \beta_0\right\| \geq 16\,\mathrm{k}_n(1+\mathrm{t})/V_n) \leq \exp(-2\,\mathrm{t}^2/\mathrm{K}_n{}^2) + \mathrm{nP}(\Omega_n^c).$$

3.6. Ιδιότητες Σίγουρου Κρησαρίσματος με MMLE

Σε αυτή την ενότητα, εισάγουμε ένα νέο πλαίσιο για τη θέσπιση της ιδιότητας του σίγουρου κρησαρίσματος με MMLE στην κανονική εκθετική οικογένεια (3.3.2). Χωρίζουμε σε τρεις ενότητες για να παρουσιάσουμε τα ευρήματά μας.

Στοιχεία πληθυσμού. Καθώς προσαρμόζουμε τις οριακές παλινδρομήσεις στην κοινή παλινδρόμηση, είναι ένα είδος ατελούς μοντέλου, ένα σημαντικό ερώτημα θα ήταν σε ποιο επίπεδο οι πληροφορίες του μοντέλου διατηρούνται. Ειδικά, για τους σκοπούς του κρησαρίσματος, ενδιαφερόμαστε για τη διατήρηση της μη-σποραδικότητας από την κοινή παλινδρόμηση στην οριακή παλινδρόμηση. Αυτό μπορούμε να το συνοψίσουμε στα ακόλουθα δύο ερωτήματα. Πρώτον, για το σκοπό του σίγουρου κρησαρίσματος, αν μια μεταβλητή X_j είναι από κοινού σημαντική ($\beta_j^* \neq 0$), (και κάτω από ποιες συνθήκες) θα είναι οριακά σημαντική ($\beta_j^{\mathrm{M}} \neq 0$). Δεύτερον, για το σκοπό της συνοχής της επιλογής μοντέλου, αν μια μεταβλητή X_j είναι από κοινού ασήμαντη ($\beta_j^* = 0$) θα εξακολουθεί να είναι οριακά ασήμαντη ($\beta_j^{\mathrm{M}} = 0$).

Τα παρακάτω θεωρήματα αποκαλύπτουν ότι η παράμετρος οριακής παλινδρόμησης είναι στην πραγματικότητα ένα μέτρο συσχέτισης μεταξύ της οριακής συμμεταβλητής και της μέσης συνάρτησης απόκρισης.

Θεώρημα 2. Για $j = 1,\ldots,p_n$, οι παράμετροι οριακής παλινδρόμησης $\beta_j^{\mathrm{M}} = 0$ αν και μόνο αν $\mathrm{cov}(\mathrm{b}'(\mathrm{X}^T\beta^*), \mathrm{X}_j) = 0$.

Χρησιμοποιώντας το γεγονός ότι $X^T \beta^* = \beta_0^* + \sum_{j \in M_*} X_j \beta_j^*$, μπορούμε εύκολα να δείξουμε το ακόλουθο πόρισμα:

Πόρισμα 1. Αν η συνθήκη μερικής ορθογωνιότητας ισχύει, δηλαδή $\{X_j, j \notin M_*\}$ είναι ανεξάρτητη του $\{X_i, i \in M_*\}$, τότε $\beta_j^M = 0$, για $j \notin M_*$.

Η συνθήκη μερικής ορθογωνιότητας είναι ουσιαστικά η υπόθεση που έγινε στους Huang, Horowitz & Ma (2008), οι οποίοι έδειξαν τη συμβατότητα της επιλογής μοντέλου στην ειδική περίπτωση με το σύνηθες γραμμικό μοντέλο και τη bridge παλινδρόμηση. Παρατηρούμε ότι $\mathrm{cov}(b'(X^T \beta^*), X_j) = \mathrm{cov}(Y, X_j)$. Απαραίτητη προϋπόθεση για το σίγουρο κρησάρισμα είναι ότι οι σημαντικές μεταβλητές X_j με $\beta_j^* \neq 0$ να σχετίζονται με την απόκριση, κάτι το οποίο συνήθως ισχύει. Όταν συσχετίζονται με την απόκριση, από το Θεώρημα 2, $\beta_j^M \neq 0$, για $j \in M_*$. Με άλλα λόγια, το οριακό μοντέλο σχετίζεται με τις πληροφορίες σχετικά με τις σημαντικές μεταβλητές στο κοινό μοντέλο. Αυτή είναι η θεωρητική βάση για τη SIS. Από την άλλη πλευρά, αν η συνθήκη μερικής ορθογωνιότητας του Πορίσματος 1 ισχύει, τότε $\beta_j^M = 0$ για $j \notin M_*$. Σε αυτή την περίπτωση, υπάρχει ένα κατώτατο όριο γ_n έτσι ώστε το οριακά επιλεγμένο μοντέλο να είναι συμβατό με την επιλογή μοντέλου

$$\min_{j \in M_*} \left| \beta_j^M \right| \geq \gamma_n \qquad \max_{j \notin M_*} \left| \beta_j^M \right| = 0 .$$

Για να έχουμε την ιδιότητα του σίγουρου κρησαρίσματος του δείγματος, χρειαζόμαστε

$$\min_{j \in M_*} \left| \beta_j^M \right| \geq O(n^{-k})$$

για κάποιο $\kappa < 1/2$, έτσι ώστε τα οριακά σήματα να είναι ισχυρότερα από το στοχαστικό θόρυβο. Το παρακάτω θεώρημα δείχνει ότι αυτό είναι δυνατό.

Θεώρημα 3. Αν $\left|\mathrm{cov}(\mathrm{b}'(\mathrm{X}^T \beta^*), \mathrm{X}_j)\right| \geq \mathrm{c}_1 n^{-\kappa}$ για $j \in M_*$ και μια θετική σταθερά $\mathrm{c}_1 > 0$, τότε υπάρχει μια θετική σταθερά c_2, έτσι ώστε

$$\min_{j \in M_*} \left|\beta_j^M\right| \geq c_2 \, \mathrm{n}^{-\kappa}$$

με την προϋπόθεση ότι $b''(\cdot)$ είναι φραγμένο ή $EG\left(a\left|\mathrm{X}_j\right|\right)\left|\mathrm{X}_j\right|\mathrm{I}\left(\left|\mathrm{X}_j\right| \geq n^{\eta}\right) \leq dn^{-\kappa}$ για κάποιο $0 \leq \eta \leq \kappa$, και κάποιες επαρκώς μικρές θετικές σταθερές α και d όπου $G(|x|) = \sup_{|u| \leq |x|} |\mathrm{b}'(u)|$.

Παρατηρούμε ότι για την κανονική κατανομή και την κατανομή Bernoulli το $b''(\cdot)$ είναι φραγμένο, ενώ για την Poisson κατανομή $G(|x|) = \exp\left(|x|\right)$ και το Θεώρημα 3 απαιτεί οι ουρές του X_j να είναι «ελαφριές». Κάτω από ορισμένες πρόσθετες προϋποθέσεις ισχύει

$$\sum_{j=1}^{p} \left|\beta_j^M\right|^2 = O\left(\left\|\sum \beta^*\right\|^2\right) = O(\lambda_{\max}(\Sigma))$$

όπου $\Sigma = \mathrm{var}\left(\mathrm{X}\right)$ και $\lambda_{\max}(\Sigma)$ είναι η μέγιστη ιδιοτιμή του. Η πρώτη ισότητα απαιτεί κάποιες προσπάθειες για να αποδειχθεί, ενώ η δεύτερη ισότητα προκύπτει εύκολα από την υπόθεση ότι

$$\mathrm{var}\left(\mathrm{X}^T \beta^*\right) = \beta^{*\mathrm{T}} \Sigma \beta^* = \mathrm{O}(1).$$

Η συνέπεια αυτού του αποτελέσματος είναι ότι δεν μπορούν να υπάρχουν πάρα πολλές μεταβλητές που να έχουν οριακό συντελεστή $|\beta_j^M|$ που να υπερβαίνει ορισμένο επίπεδο «κατωφλιού». Αυτό επιτυγχάνει την σποραδικότητα στο τελικό επιλεγμένο μοντέλο.

Όταν οι συμμεταβλητές είναι από κοινού κανονικά κατανεμημένες, οι συνθήκες του Θεωρήματος 3 μπορούν να απλουστευθούν περαιτέρω.

Πρόταση 1. Έστω ότι τα X και Z είναι από κοινού κανονικά με μέση τιμή μηδέν και τυπική απόκλιση 1. Για μια αυστηρά μονότονη συνάρτηση $f(\cdot)$, $\mathrm{cov}(\mathrm{X},\mathrm{Z})=0$ αν και μόνο αν $\mathrm{cov}(\mathrm{X},\mathrm{f(Z)})=0$ μας δίνει ότι η τελευταία συνδιακύμανση υπάρχει. Επιπλέον,

$$\left|\mathrm{cov}(\mathrm{X},\mathrm{f(Z)})\right| \geq |\rho| \inf_{|\mathrm{x}|\leq c|\rho|} |\mathrm{g'(x)}| \mathrm{EX}^2 I\left(|\mathrm{X}|\leq c\right)$$

για οποιοδήποτε $c>0$, όπου $\rho = EXZ$, $\mathrm{g(x)}=\mathrm{Ef}\left(\mathrm{x}+\varepsilon\right)$ με $\varepsilon \sim \mathrm{N}(0,1-\rho^2)$.

Η παραπάνω πρόταση δείχνει ότι η συνδιακύμανση των X και $\mathrm{f}\left(\mathrm{Z}\right)$ μπορεί να φραχθεί από κάτω από τη συνδιακύμανση μεταξύ των X και Z, δηλαδή,

$$\left|\mathrm{cov}(\mathrm{X},\mathrm{f}\left(\mathrm{Z}\right))\right| \geq \mathrm{d}|\rho| \quad , \quad d = \inf_{|\mathrm{x}|\leq \mathrm{c}} |\mathrm{g'(x)}| \mathrm{EX}^2 I\left(|\mathrm{X}|\leq \mathrm{c}\right),$$

στα οποία $d>0$ για αρκετά μικρό c. Το πρώτο μέρος της πρότασης πραγματικά ισχύει όταν η σχετική πυκνότητα $\mathrm{f}\left(\mathrm{z}|\mathrm{x}\right)$ του Z δοσμένου X είναι μια οικογένεια μονότονων πιθανοφανειών (Bickel και Doksum (2001)), όταν το x θεωρείται ως παράμετρος. Παίρνοντας $\mathrm{Z}=\mathrm{X}^T\beta^*$, μια άμεση εφαρμογή του θεωρήματος είναι $\beta_j^M=0$ αν και μόνο αν $\mathrm{cov}(\mathrm{X}^T\beta^*,\mathrm{X}_j)=0$, εφόσον το X είναι από κοινού κανονικό, δεδομένου ότι το $b'(\cdot)$ είναι μια αύξουσα συνάρτηση. Επιπλέον, αν

$$\left|\mathrm{cov}(\mathrm{X}^T\beta^*,\mathrm{X}_j)\right| \geq c_0 n^{-\kappa}, \qquad \kappa<1/2 \tag{3.6.1}$$

για κάποια θετική σταθερά c_0, μια ελάχιστη προϋπόθεση που απαιτείται ακόμη και για τα μοντέλα ελαχίστων τετραγώνων (Fan & Lv (2008)), τότε από το δεύτερο μέρος της Πρότασης 1, έχουμε

$$\left|\mathrm{cov}(\mathrm{b'}(\mathrm{X}^T\beta^*),\mathrm{X}_j)\right| \geq \mathrm{c}_1 n^{-\kappa}$$

για κάποια σταθερά c_1. Ως εκ τούτου, από το Θεώρημα 2, υπάρχει μια θετική σταθερά c_2 τέτοια ώστε $\left|\beta_j^M\right| \geq c_2\, n^{-\kappa}$. Με άλλα λόγια, η (3.6.1) αρκεί να έχει οριακά σήματα που είναι πάνω από το επίπεδο μέγιστου θορύβου.

3.7. Ομοιόμορφη σύγκλιση και σίγουρο κρησάρισμα

Για να εδραιώσουμε την ιδιότητα της SIS της MMLE, ένα βασικό σημείο είναι να καθιερώσουμε την ομοιόμορφη σύγκλιση των MMLEs. Αυτό γίνεται για να ελέγξουμε το επίπεδο μέγιστου θορύβου σε σχέση με το σήμα. Στη συνέχεια, καθιερώνουμε την κατάταξη της ομοιόμορφης σύγκλισης για τα MMLEs και την ιδιότητα του σίγουρου κρησαρίσματος. Το πρώτο θα είναι χρήσιμο στον έλεγχο του μεγέθους του επιλεγμένου συνόλου.

Έστω $\beta_j = (\beta_{j,0}, \beta_j)^T$ η δισδιάστατη παράμετρος και $X_j = (1, X_j)^T$. Λόγω του κοίλου της log-πιθανοφάνειας στα GLM με την κανονική σύνδεση, $El(X_j^{\mathrm{T}} \beta_j, Y)$ να έχει ένα μοναδικό ελάχιστο πάνω στο $\beta_j \in B$ σε ένα εσωτερικό σημείο $\beta_j^M = (\beta_{j,0}^M, \beta_j^M)^T$, όπου $B = \{\left|\beta_{j,0}^M\right| \leq B, \left|\beta_j^M\right| \leq B\}$ είναι ένα τετράγωνο με πλάτος B πάνω από το οποίο η οριακή πιθανοφάνεια μεγιστοποιείται. Το παρακάτω, είναι μια ενημερωμένη εκδοχή των συνθηκών Α-Γ για κάθε οριακή παλινδρόμηση και δύο πρόσθετες συνθήκες για τις συμμεταβλητές και τις παραμέτρους πληθυσμού.

Α΄. Η οριακή πληροφορία Fisher: $I_j(\beta_j) = E\{b''(X_j^T \beta_j) X_j X_j^T\}$ είναι πεπερασμένη και θετικά ορισμένη στο $\beta_j = \beta_j^M$ για $j = 1, \ldots, p_n$. Επιπλέον, $\left\|I_j(\beta_j)\right\|_B$ είναι φραγμένη από πάνω.

Β΄. Η δεύτερη παράγωγος του $b(\theta)$ είναι συνεχής και θετική. Υπάρχει ένα $\varepsilon_1 > 0$ έτσι ώστε για όλα τα $j = 1, \ldots, p_n$,

$$\sup_{\beta \in B, \left\|\beta - \beta_j^M\right\| \leq \varepsilon_1} \left| E\, b(X_j^T \beta)\, I\left(\left|X_j\right| > K_n\right) \right| \leq o(n^{-1})$$

Γ΄. Για όλα τα $\beta_j \in B$, έχουμε $E(l(\mathrm{X}_j^{\mathrm{T}}\beta_j, \mathrm{Y}) - l(\mathrm{X}_j^{\mathrm{T}}\beta_j^M, \mathrm{Y})) \geq \mathrm{V}\left\|\beta_j - \beta_j^M\right\|^2$, για κάποια θετικά V, φραγμένα από κάτω ομοιόμορφα πάνω από $j = 1, \ldots, p_n$.

Δ΄. Υπάρχουν κάποιες θετικές σταθερές m_0, m_1, s_0, s_1 και α, έτσι ώστε για επαρκώς μεγάλα t,

$$P\left(\left|\mathrm{X}_j\right| > \mathrm{t}\right) \leq (\mathrm{m}_1 - \mathrm{s}_1)\exp\{-\mathrm{m}_0\mathrm{t}^a\} \quad \text{για} \quad j = 1, \ldots, p_n$$

και $\mathrm{E}\exp(\mathrm{b}(\mathrm{X}^T\beta^* + s_0) - \mathrm{b}(\mathrm{X}^T\beta^*)) + \mathrm{E}\exp(\mathrm{b}(\mathrm{X}^T\beta^* - s_0) - b(\mathrm{X}^T\beta^*)) \leq \mathrm{s}_1$

Ε΄. Οι συνθήκες στο θεώρημα 3 ισχύουν.

Οι συνθήκες Α΄-Γ΄ ικανοποιούνται σε πολλά παραδείγματα των γενικευμένων γραμμικών μοντέλων, όπως στη γραμμική παλινδρόμηση, στη λογιστική παλινδρόμηση και στην παλινδρόμηση Poisson. Παρατηρούμε ότι το δεύτερο μέρος της συνθήκης Δ διασφαλίζει την ουρά της μεταβλητής απόκρισης Y να είναι εκθετικά «φωτεινή», όπως φαίνεται στο παρακάτω λήμμα.

Λήμμα 1. Αν η συνθήκη Δ ικανοποιείται, για οποιοδήποτε $t > 0$,

$$P\left(\left|\mathrm{Y}\right| \geq \mathrm{m}_0\mathrm{t}^a / \mathrm{s}_o\right) \leq \mathrm{s}_1 \exp(-\mathrm{m}_0\mathrm{t}^a)$$

Έστω $k_n = b'(\mathrm{K}_n\mathrm{B} + \mathrm{B}) + \mathrm{m}_0\mathrm{K}_n{}^a / s_0$. Τότε η συνθήκη Β ισχύει για την εκθετική οικογένεια με $K_n^* = m_0 K_n^a / s_0$. Η σταθερά Lipschitz k_n είναι φραγμένη για τη λογιστική παλινδρόμηση, αφού τα Y και $b'(\cdot)$ είναι φραγμένα. Το ακόλουθο θεώρημα δίνει ένα αποτέλεσμα ομοιόμορφης σύγκλισης των MMLEs και μια ιδιότητα σίγουρου κρησαρίσματος. Κάτι ενδιαφέρον είναι ότι η ιδιότητα του σίγουρου κρησαρίσματος δεν εξαρτάται απευθείας από την ιδιότητα του πίνακα συνδιακύμανσης των συμμεταβλητών, όπως με την αύξηση του τελεστή norm. Αυτό είναι ένα πλεονέκτημα σε σχέση με τη χρήση ολόκληρης της πιθανοφάνειας.

Θεώρημα 4. Υποθέτουμε ότι οι συνθήκες Α΄, Β΄, Γ΄ και Δ΄ ισχύουν. Τότε:

(i) Αν $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$, τότε για οποιοδήποτε $c_3 > 0$, υπάρχει μια θετική σταθερά c_4, τέτοια ώστε

$$P\left(\max_{1\le j\le p_n}\left|\hat{\beta}_j^M - \beta_j^M\right| \ge c_3\, n^{-k}\right) \le p_n\{\exp(-c_4 n^{1-2\kappa}/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^a)\}$$

(ii) Αν επιπλέον, ισχύει η συνθήκη Ε, τότε παίρνοντας $\gamma_n = c_5 n^{-k}$ με $c_5 \le c_2/2$, έχουμε:

$$P(M_* \subset \hat{M}_{\gamma_n}) \ge 1 - s_n\{\exp(-c_4 n^{1-2\kappa})/(k_n K_n)^2) + n m_1 \exp(-m_0 K_n^a)\}$$

όπου $s_n = |M_*|$, το μέγεθος των μη-αραιών στοιχείων.

ΠΑΡΑΤΗΡΗΣΗ 1. Αν υποθέσουμε ότι $\min_{j\in M_*}\left|\mathrm{cov}(b'(X^T\beta^*), X_j)\right| \ge c_1 n^{-\kappa+\delta}$ για οποιοδήποτε $\delta > 0$, τότε μπορούμε να πάρουμε $\gamma_n = c\cdot n^{-\kappa+\delta/2}$ για οποιοδήποτε $c > 0$ στο Θεώρημα 4.

Παρατηρούμε ότι όταν $b'(\cdot)$ είναι φραγμένο, όπως στο μοντέλο Bernoulli, το k_n είναι μια πεπερασμένη σταθερά. Σε αυτή την περίπτωση, εξισώνοντας τους δύο όρους στο άνω φράγμα του Θεωρήματος 4 (i), η βέλτιστη τάξη της K_n δίνεται από $K_n = n^{(1-2\kappa)/(\alpha+2)}$ και $P(\max_{1\le j\le p_n}\left|\hat{\beta}_j^M - \beta_j^M\right| \ge c_3 n^{-\kappa}) = O\{p_n \exp(-c_4 n^{(1-2k)\alpha/(\alpha+2)})\}$, για μια θετική σταθερά c_4. Όταν οι συμμεταβλητές X_j είναι φραγμένες, τότε οι k_n και K_n μπορούν να θεωρηθούν ως πεπερασμένες σταθερές. Σε αυτή την περίπτωση,

$$P(\max_{1\le j\le p_n}\left|\hat{\beta}_j^M - \beta_j^M\right| \ge c_3 n^{-\kappa}) \le O\{p_n \exp(-c_4 n^{(1-2k)})\}.$$

Και στις δύο περιπτώσεις που προαναφερθήκαν, η ουρά πιθανότητας στο Θεώρημα 4 είναι εκθετικά μικρή. Με άλλα λόγια, μπορούμε να χειριστούμε την NP-διάσταση,

$$\log p_n = o(n^{(1-2k)\alpha/(\alpha+2)}),$$

όπου $\alpha = \infty$, στην περίπτωση των φραγμένων συμμεταβλητών.

Για το κλασσικό γραμμικό μοντέλο, $k_n = B(K_n+1) + K_n^a/(2s_0)$ και παίρνοντας τη βέλτιστη τάξη του $K_n = n^{(1-2\kappa)/A}$ με $A = \max(\alpha+4, 3\alpha+2)$, έχουμε

$$P(\max_{1\le j\le p_n} \left|\hat{\beta}_j^M - \beta_j^M\right| > c_3 n^{-\kappa}) = O\{p_n \exp(-c_4 n^{(1-2k)\alpha/A})\}.$$

Όταν οι συμμεταβλητές είναι κανονικές, $\alpha = 2$ και τα αποτελέσματά μας είναι πιο αδύναμα από αυτά που δίνονται στους Fan & Lv (2008), οι οποίοι επιτρέπουν $\log p_n = o(n^{1-2k})$, ενώ στο Θεώρημα 4 μπορούμε να χειριστούμε μόνο το $\log p_n = o(n^{(1-2k)/4})$. Παρόλα αυτά, επιτρέπουμε μη φυσιολογικές συμμεταβλητές και άλλα σφάλματα κατανομών.

Τα παραπάνω εφαρμόζονται στην ιδιότητα του σίγουρου κρησαρίσματος που δόθηκε στο Θεώρημα 4 (ii). Μόνο το μέγεθος των μη αραιών στοιχείων s_n μετράει για το σκοπό του σίγουρου κρησαρίσματος, και όχι η διάσταση p_n.

3.8. Έλεγχος κατάταξης επιλογής σφαλμάτων.

Αφού εφαρμόσουμε τη διαδικασία του κρησαρίσματος μεταβλητών, η ερώτηση που τίθεται φυσικά είναι πόσο μεγάλο είναι το σύνολο $\hat{M}_{\gamma_n}$. Με άλλα λόγια, έχει μειωθεί πραγματικά ο αριθμός των μεταβλητών από τη μάθηση ανεξαρτησίας;

Μια απλή απάντηση σε αυτή την ερώτηση είναι η ιδανική περίπτωση στην οποία $\mathrm{cov}(b'(X^T\beta^*), X_j) = o(n^{-\kappa})$ για $j \notin M_*$.

Σε αυτή την περίπτωση, κάτω από ήπιες συνθήκες, μπορούμε να δείξουμε ότι

$$\max_{j\notin M_*} \left|\beta_j^M\right| = o(n^{-\kappa}).$$

Αυτό, μαζί με το Θεώρημα 4(i) μας δείχνει ότι

$$\max_{j \notin M_*} \left| \hat{\beta}_j^M \right| \le c_3 \, \mathrm{n}^{-\kappa} \text{ για οποιοδήποτε } c_3 > 0 ,$$

με πιθανότητα που τείνει στο 1, αν η πιθανότητα του Θεωρήματος 4(i) τείνει στο 0. Ως εκ τούτου, από την επιλογή του γ_n, όπως στο Θεώρημα 4(ii), μπορούμε να πετύχουμε την επιλογή μοντέλου σταθερότητας

$$P(\hat{M}_{\gamma_n} = M_*) = 1 - \mathrm{o}(1) .$$

Αυτό το είδος της συνέχειας ήταν πράγματι συνέπεια από τη συνθήκη Huang, Horowitz & Ma (2008) στην ειδική περίπτωση με το κλασσικό γραμμικό μοντέλο, χρησιμοποιώντας τη bridge παλινδρόμηση που οδηγεί σε παρόμοιο συμπέρασμα.

Τώρα ερχόμαστε αντιμέτωποι με την πιο γενική περίπτωση. Η ιδέα είναι να φράξουμε το μέγεθος του επιλεγμένου συνόλου χρησιμοποιώντας το γεγονός ότι η $\mathrm{var}(Y)$ είναι φραγμένη.Αυτό συνήθως έχει ως συνέπεια $\mathrm{var}(\mathrm{X}^T \beta^*) = \beta^{*\mathrm{T}} \Sigma \beta^* = \mathrm{O}(1)$. Χρειαζόμαστε τις ακόλουθες επιπρόσθετες συνθήκες:

ΣΤ. Η διασπορά $\mathrm{var}(\mathrm{X}^T \beta^*)$ είναι φραγμένη από πάνω και από κάτω.

Ζ. Είτε το $\mathrm{b}''(\cdot)$ είναι φραγμένο ή η $\mathrm{X}_M = (\mathrm{X}_1, ..., \mathrm{X}_{p_n})^T$ ακολουθεί μια ελλειπτική κατανομή, η οποία είναι $X_M = \Sigma_1^{1/2} RU$, και $\left| Eb'(\mathrm{X}^T \beta^*)(\mathrm{X}^T \beta^* - \beta_0^*) \right|$ είναι φραγμένο, όπου U είναι ομοιόμορφα κατανεμημένο πάνω στη μοναδιαία σφαίρα στον p-διάστασης Ευκλείδειο χώρο, ανεξάρτητα από τη μη-αρνητική τυχαία μεταβλητή R και $\Sigma_1 = \mathrm{var}(\mathrm{X}_M)$.

Παρατηρούμε ότι $\Sigma = diag(0, \Sigma_1)$ με τη συνθήκη ΣΤ΄, αφού οι πίνακες συνδιακύμανσης διαφέρουν μόνο στους όρους της τομής. Ως εκ τούτου, $\lambda_{\max}(\Sigma) = \lambda_{\max}(\Sigma_1)$. Το ακόλουθο αποτέλεσμα είναι σχετικό με το μέγεθος του $\hat{\mathrm{M}}_{\gamma_n}$.

ΘΕΩΡΗΜΑ 5. Κάτω από τις συνθήκες Α΄, Β΄, Γ΄, Δ, Ε και ΣΤ, έχουμε για οποιοδήποτε $\gamma_n = c_5 n^{-2\kappa}$ ότι υπάρχει ένα c_4 τέτοιο ώστε

$$P\left[\left|\hat{M}_{\gamma_n}\right| \le \mathrm{O}\{\mathrm{n}^{2\kappa}\,\lambda_{\max}(\Sigma)\}\right] \ge 1 - p_n\{\exp(-\mathrm{c}_4\mathrm{n}^{1-2\kappa}/(\mathrm{k}_n\mathrm{K}_n)^2) + \mathrm{nm}_1\exp(-\mathrm{m}_0\mathrm{K}_n^{\alpha})\}$$

Δείξαμε λοιπόν ότι ο αριθμός των επιλεγμένων μεταβλητών είναι τάξης $\left\|\Sigma\beta^*\right\|^2 / \gamma_n^2$, το οποίο είναι επιπλέον φραγμένο από $O\{\mathrm{n}^{2\kappa}\,\lambda_{\max}(\Sigma)\}$ χρησιμοποιώντας $\mathrm{var}(\mathrm{X}^T\beta^*) = \mathrm{O}(1)$. Κάτι ενδιαφέρον είναι ότι ενώ η ιδιότητα του σίγουρου κρησαρίσματος δεν εξαρτάται από τη συμπεριφορά του *Σ*, ο αριθμός των επιλεγμένων μεταβλητών επηρεάζεται από το πόσο συσχετισμένες είναι οι συμμεταβλητές. Όταν $\mathrm{n}^{2\kappa}\,\lambda_{\max}(\Sigma)/\mathrm{p} \to 0$, ο αριθμός των επιλεγμένων μεταβλητών είναι πράγματι αμελητέος συγκριτικά με το αρχικό. Σε αυτή την περίπτωση, το ποσοστό των ψευδομεταβλητών που έχουν ανακαλυφθεί είναι φυσικά αμελητέο. Πιο συγκεκριμένα, όταν $\lambda_{\max}(\Sigma) = O(\mathrm{n}^{\tau})$, το μέγεθος της επιλεγμένης μεταβλητής είναι τάξης $O(\mathrm{n}^{2\kappa+\tau})$. Αυτό είναι της ίδιας τάξης όπως και των Fan & Lv (2008) για το μοντέλο πολλαπλής παλινδρόμησης με τα κανονικά δεδομένα, που χρειάζονται επιπλέον τη συνθήκη 2κ + τ < 1. Επιπρόσθετα, τα αποτελέσματά μας είναι πιο διαισθητικά, ο αριθμός των επιλεγμένων μεταβλητών σχετίζεται με το $\lambda_{\max}(\Sigma)$ ή με περισσότερη ακρίβεια $\left\|\Sigma\beta^*\right\|^2$ και την παράμετρο κατωφλιού γ_n.

3.9. Κρησάρισμα λόγου πιθανοφάνειας

Σε ένα παρόμοιο πρόβλημα κρησαρίσματος μεταβλητών με τα γενικευμένα γραμμικά μοντέλα, οι Fan, Samworth & Wu (2009) πρότειναν το κρησάρισμα των μεταβλητών με ταξινόμηση της οριακής πιθανοφάνειας. Αυτή τη μέθοδο μπορούμε να τη δούμε ως ένα *κρησάρισμα του λόγου της οριακής πιθανοφάνειας*, καθώς χτίστηκε πάνω στις προσαυξήσεις της log-πιθανοφάνειας. Σε αυτή την ενότητα δείχνουμε ότι το κρησάρισμα του λόγου πιθανοφάνειας είναι ισοδύναμο με το MMLE κρησάρισμα, με την έννοια ότι και οι δύο κατέχουν την ιδιότητα του σίγουρου κρησαρίσματος και ότι ο αριθμός των επιλεγμένων μεταβλητών και των δύο μεθόδων είναι της ίδιας τάξης μεγέθους.

Αρχικά, τυποποιούμε τη διαδικασία της οριακής πιθανοφάνειας. Έστω

$L_{j,n} = P_n\{\mathrm{l}(\hat{\beta}_0^{\mathrm{M}}, \mathrm{Y}) - \mathrm{l}(\mathrm{X}_j^{\mathrm{T}}\hat{\beta}_j^M, Y)\}$, $\mathrm{j} = 1, ..., \mathrm{p}_n$ και $L_n = (\mathrm{L}_{1,n}, ..., \mathrm{L}_{p_n,n})^T$ όπου $\hat{\beta}_0^M = \arg\min_{\beta_0} P_n l(\beta_0, \mathrm{Y})$. Αντίστοιχα, έστω $L_j^* = E\{\mathrm{l}(\beta_0^{\mathrm{M}}, \mathrm{Y}) - \mathrm{l}(\mathrm{X}_j^{\mathrm{T}}\beta_j^M, Y)\}$, $\mathrm{j} = 1, ..., \mathrm{p}_n$, και $\mathrm{L}^* = (\mathrm{L}_1^*, ..., \mathrm{L}_{p_n}^*)^T$ όπου $\beta_0^M = \arg\min_{\beta_0} El(\beta_0, \mathrm{Y})$.

Μπορεί να δειχθεί ότι $\mathrm{EY} = \mathrm{b}'(\beta_0^M)$ και ότι $\bar{\mathrm{Y}} = b'(\hat{\beta}_0^M)$, όπου $\bar{\mathrm{Y}}$ είναι το μέσο του δείγματος. Ταξινομούμε το διάνυσμα L_n σε αύξουσα σειρά και επιλέγουμε το σύνολο των μεταβλητών

$$\hat{\mathrm{N}}_{v_n} = \{1 \le \mathrm{j} \le \mathrm{p}_n : \mathrm{L}_{j,n} \ge \mathrm{v}_n\},$$

όπου το v_n είναι μια προκαθορισμένη τιμή κατωφλιού. Τέτοια μάθηση ανεξαρτησίας ταξινομεί τη σημαντικότητα των χαρακτηριστικών σύμφωνα με την οριακή τους συνεισφορά στο μέγεθος της συνάρτησης πιθανοφάνειας. Το κρησάρισμα της οριακής πιθανοφάνειας και το MMLE κρησάρισμα μοιράζονται μια κοινή υπολογιστική διαδικασία όπως την επίλυση των p_n προβλημάτων βελτιστοποίησης πάνω σε δισδιάστατο παραμετρικό χώρο. Ως εκ τούτου, ο υπολογισμός είναι πολύ πιο εφικτός από τις παραδοσιακές μεθόδους επιλογής μεταβλητών.

Συγκρίνοντας με το MMLE κρησάρισμα, όπου οι πληροφορίες που χρησιμοποιούνται είναι μόνο τα μεγέθη των εκτιμητών, το κρησάρισμα της οριακής πιθανοφάνειας ενσωματώνει όλες τις συνεισφορές των χαρακτηριστικών των προσαυξήσεων πιθανοφάνειας και τα μεγέθη των εκτιμητών και τη σχετισμένη παραλλαγή. Υπό αυτές τις συνθήκες (συνθήκη Γ΄), η διακύμανση των MMLE είναι ένα συγκρίσιμο επίπεδο (μέσω του μεγέθους του V, μια επίπτωση της κυρτότητας της αντικειμενικής συνάρτησης), και των δύο μεθόδων κρησαρίσματος είναι ότι είναι ισοδύναμες. Διαφορετικά, αν το V εξαρτάται από το *n*, το κρησάρισμα της οριακής πιθανοφάνειας μπορεί να διατηρεί τη δομή της μη-σποραδικότητας, ενώ το MMLE κρησάρισμα μπορεί να χρειάζεται κάποιες ανάλογες προσαρμογές.

Στη συνέχεια θα δείξουμε ότι η ιδιότητα του σίγουρου κρησαρίσματος ισχύει κάτω από συγκεκριμένες συνθήκες. Ομοίως, με το MMLE κρησάρισμα, αρχικά «χτίζουμε» το θεωρητικό θεμέλιο του κρησαρίσματος οριακής πιθανοφάνειας. Αυτό

σημαίνει ότι η προσαύξηση της οριακής πιθανοφάνειας είναι ακόμη ένα μέτρο της συσχέτισης μεταξύ των οριακών συμμεταβλητών και της συνάρτησης μέσης απόκρισης.

ΘΕΩΡΗΜΑ 6. Για $j = 1, ..., p_n$, η προσαύξηση της οριακής πιθανοφάνειας $L_j^* = 0$ αν και μόνο αν $\mathrm{cov}(b'(X^T \beta^*), X_j) = 0$.

Ένα απευθείας πόρισμα του Θεωρήματος 1 είναι ότι εύκολα μπορούμε να δείξουμε το επόμενο πόρισμα για το σκοπό της συνοχής της επιλογής μοντέλου.

Πόρισμα 2. Αν η συνθήκη μερικής ορθογωνιότητας του Πορίσματος 1 ισχύει, τότε $L_j^* = 0$, για $j \notin M_*$.

Μπορούμε να δυναμώσουμε το αποτέλεσμα των μικρότερων σημάτων όπως ακολουθεί. Από την άλλη πλευρά, θα δείξουμε επίσης ότι όλα τα σήματα δεν μπορούν να είναι πολύ μεγάλα. Αυτό σημαίνει ότι δεν μπορούν να υπάρχουν πολλά σήματα.

ΘΕΩΡΗΜΑ 7. Κάτω από τις συνθήκες του Θεωρήματος 3 και της συνθήκης Γ΄, έχουμε $\min_{j \in M_*} \left| L_j^* \right| \geq c_6 \, n^{-2\kappa}$ για κάποια θετική σταθερά c_6, με τον όρο $\left| \mathrm{cov}(b'(X^T \beta^*), X_j) \right| \geq c_1 n^{-\kappa}$ για $j \in M_*$. Αν επιπλέον ισχύουν οι συνθήκες ΣΤ και Z τότε

$$\left\| L^* \right\| = O\left(\left\| \beta^M \right\|^2 \right) = O\left(\left\| \Sigma \beta^* \right\|^2 \right) = O(\lambda_{\max}(\Sigma)).$$

Η τεχνική πρόκληση είναι ότι ο στοχαστικός θόρυβος $\left\|L_n - L^*\right\|_\infty$ είναι συνήθως τάξης $O(\mathrm{n}^{-2\kappa} + \mathrm{n}^{-1/2}\log \mathrm{p}_n)$, η οποία μπορεί να είναι τάξης μεγέθους μεγαλύτερη από τα σήματα που δίνονται στο Θεώρημα 7, εκτός αν $\kappa < 1/4$. Παρόλα αυτά, με ένα διαφορετικό τέχνασμα που χρησιμοποιεί το γεγονός ότι η κατάταξη είναι αναλλοίωτη στα πλαίσια ενός αυστηρού μονότονου σίγουρου κρησαρίσματος ανεξαρτησίας για $\kappa < 1/2$.

ΘΕΩΡΗΜΑ 8. Υποθέτουμε ότι οι συνθήκες Α΄, Β΄, Γ΄ και Δ, Ε και ΣΤ ισχύουν. Τότε λαμβάνοντας $v_n = c_7 n^{-2\kappa}$ για ένα επαρκώς μικρό $c_7 > 0$, υπάρχει $c_8 > 0$ έτσι ώστε $P(\mathrm{M}_* \subset \hat{\mathrm{N}}_{v_n}) \geq 1 - \mathrm{s}_n\{\exp(-\mathrm{c}_8 \mathrm{n}^{1-2\kappa}/(\mathrm{k}_n \mathrm{K}_n)^2) + \mathrm{nm}_1 \exp(-\mathrm{m}_0 \mathrm{K}_n^\alpha)\}$.

Ομοίως με το MMLE κρησάρισμα μπορούμε να ελέγξουμε το μέγεθος του N_{v_n} ως εξής. Για την απλότητα του τεχνικού επιχειρήματος θα επικεντρωθούμε μόνο στην περίπτωση όπου $b''(\cdot)$ είναι φραγμένο.

ΘΕΩΡΗΜΑ 9. Κάτω από τις συνθήκες Α΄, Β΄, Γ΄, Δ, Ε και ΣΤ, αν $b''(\cdot)$ είναι φραγμένο τότε έχουμε

$$P\left[\left|\hat{\mathrm{N}}_{v_n}\right| \leq O\{\mathrm{n}^{2\kappa}\, \lambda_{\max}(\Sigma)\}\right] \geq 1 - \mathrm{p}_n\{\exp(-\mathrm{c}_8 \mathrm{n}^{1-2\kappa}/(\mathrm{k}_n \mathrm{K}_n)^2) + \mathrm{nm}_1 \exp(-\mathrm{m}_0 \mathrm{K}_n^\alpha)\}.$$

3.10. Τελικές παρατηρήσεις

Προτείναμε δύο μεθόδους ανεξάρτητου κρησαρίσματος με την κατάταξη των εκτιμητών μέγιστης οριακής πιθανοφάνειας και της μέγιστης οριακής πιθανοφάνειας στα γενικευμένα γραμμικά μοντέλα. Με τους Fan & Lv (2008) ως ειδική περίπτωση, η προτεινόμενη μέθοδος αποδεικνύεται να κατέχει την ιδιότητα του σίγουρου ανεξάρτητου κρησαρίσματος. Η επιτυχία του οριακού κρησαρίσματος δημιουργεί την

ιδέα ότι οποιοδήποτε υποκατάστατο κρησαρίσματος εφόσον μπορεί να διατηρήσει τη δομή μη-σποραδικότητας του πραγματικού μοντέλου και είναι εφικτή στον υπολογισμό, μπορεί να είναι μια καλή επιλογή του κρησαρίσματος μεταβλητών πληθυσμού. Επίσης, ανοίγει το δρόμο για το κρησάρισμα δείγματος μεταβλητών, καθώς τα υποκατάστατα σήματα είναι ομοιόμορφα διακριτά από το στοχαστικό θόρυβο. Κατά μήκος αυτής της γραμμής, όπως το *R* τετράγωνο στατιστικών, η οριακή ψευδο-πιθανοφάνεια (ελάχιστη τετραγωνική εκτίμηση, για παράδειγμα) μπορεί να είναι μια δυνητική βάση για την ανεξάρτητη μάθηση.

Εν τω μεταξύ, οι προτεινόμενες ιδιότητες του σίγουρου κρησαρίσματος και η εξαφάνιση του ποσοστού λάθους επιλογής θα είναι καλά κριτήρια για την αξιολόγηση μεθόδων υψηλής διάστασης επιλογής μεταβλητών.

Καθώς τα τρέχοντα αποτελέσματά μας ισχύουν μόνο όταν η συνάρτηση λογαριθμικής πιθανοφάνειας είναι κοίλη στις παραμέτρους παλινδρόμησης, η προτεινόμενη διαδικασία δεν καλύπτει όλα τα γενικευμένα γραμμικά μοντέλα, όπως σε κάποιες μη κανονικές συνδετικές περιπτώσεις. Αυτό αφήνει χώρο για μελλοντική έρευνα.

Σε αντίθεση με τους Fan & Lv (2008), η βασική ιδέα των τεχνικών μας αποδείξεων είναι σε γενικές γραμμές εφαρμόσιμες. Υποθέτουμε ότι τα αποτελέσματά μας θα ισχύουν όταν η δεσμευμένη κατανομή των αποτελεσμάτων *Y*, δοσμένων των συμμεταβλητών *X*, εξαρτάται μόνο από το $X^T\beta^*$ και είναι αυθαίρετες και άγνωστες διαφορετικά. Επομένως, η μέθοδος SIS μπορεί να εφαρμοστεί σε μια πλούσια τάξη γενικών μοντέλων παλινδρόμησης, συμπεριλαμβανομένων των μοντέλων μετατροπής (Box & Cox (1964), Bickel & Doksum(1981)), αποκομμένων μοντέλων παλινδρόμησης (Cox(1972), Kosorok, Lee & Fine (2004), Zeng & Lin (2007)) και αναζήτηση προβολής παλινδρόμησης (Friedman & Stuetzle(1981)).

Μια άλλη σημαντική επέκταση είναι να γενικευθεί η έννοια της οριακής παλινδρόμησης στην οριακή ομάδα παλινδρόμησης, όπου ο αριθμός των συμμεταβλητών *m* σε κάθε οριακή παλινδρόμηση είναι μεγαλύτερη ή ίση του 1. Αυτό οδηγεί σε μια νέα διαδικασία που καλείται κρησάρισμα ομαδοποιημένων μεταβλητών. Αναμένεται να βελτιώσει την κατάσταση όταν οι μεταβλητές συσχετίζονται σε μεγάλο βαθμό και είναι από κοινού σημαντικές, αλλά οριακά η

συσχέτιση μεταξύ κάθε επιμέρους μεταβλητής και απόκρισης είναι ασθενής. Οι τρέχουσες θεωρητικές μελέτες για την κατά συνιστώσες οριακή παλινδρόμηση μπορεί άμεσα να επεκταθεί στο κρησάρισμα ομαδοποιημένων μεταβλητών, με τις κατάλληλες συνθήκες και προσαρμογές.

Στην πράξη, ο τρόπος με τον οποίο θα επιλέξουμε την παράμετρο συντονισμού γ_n είναι ένα ενδιαφέρον και σημαντικό πρόβλημα. Σύμφωνα με τους Fan & Lv (2008), για το πρώτο στάδιο της επαναληπτικής SIS διαδικασίας, επιλέγουμε πολλά χαρακτηριστικά, όπως $\left|M_{\gamma_n}\right| = n$ ή $n|\log(\mathrm{n})$. Οι FDR μέθοδοι που βασίζονται σε πολλαπλές συγκρίσεις μπορούν ενδεχομένως να χρησιμοποιηθούν. Στο δεύτερο και τελικό στάδιο, μπορεί να εφαρμοστεί το είδος πληροφοριών Bayes του κριτηρίου. Στην πράξη, κάποιες μέθοδοι με γνώμονα τα δεδομένα μπορούν επίσης να είναι ευπρόσδεκτες για την επιλογή της παραμέτρου συντονισμού γ_n.

3.11. Αριθμητικά αποτελέσματα.

Περιγράφουμε τώρα το εύρος της εφαρμογής της (I)SIS και των παραλλαγών τους μελετώντας τις επιδόσεις τους πάνω στα προσομοιωμένα δεδομένα σε τέσσερα διαφορετικά πλαίσια: λογιστική παλινδρόμηση, παλινδρόμηση Poisson, ισχυρή παλινδρόμηση (με το ελάχιστο απόλυτο κριτήριο απόκλισης) και πολύ-κατηγορική κατάταξη με μηχανές υποστήριξης διανυσμάτων. Θα εξετάσουμε τρεις διαφορετικές διαμορφώσεις του $p = 1000$ χαρακτηριστικών $X_1, ..., X_p$:

Περίπτωση 1: $X_1, ..., X_p$ είναι ανεξάρτητες και ταυτόσημα κατανεμημένες $N(0,1)$ τυχαίες μεταβλητές.

Περίπτωση 2: $X_1, ..., X_p$ είναι από κοινού κανονική, οριακά $N(0,1)$ και με $\mathrm{corr}(\mathrm{X}_i, \mathrm{X}_4) = \frac{1}{\sqrt{2}}$ για όλα τα $i \neq 4$ και $\mathrm{corr}(\mathrm{X}_i, \mathrm{X}_j) = 1/2$ αν τα i και j είναι διακριτά στοιχεία του $\{1, ..., \mathrm{p}\} \setminus \{4\}$.

Περίπτωση 3: $X_1,...., X_p$ είναι από κοινού κανονική, οριακά $N(0,1)$ και με $\mathrm{corr}(\mathrm{X}_i, \mathrm{X}_5) = 0$ για όλα τα $i \neq 5$, $\mathrm{corr}(\mathrm{X}_i, \mathrm{X}_4) = \frac{1}{\sqrt{2}}$ για όλα τα $i \notin \{4,5\}$ και $\mathrm{corr}(\mathrm{X}_i, \mathrm{X}_j) = 1/2$ αν τα i και j είναι διακριτά στοιχεία του $\{1,...,\mathrm{p}\} \setminus \{4,5\}$.

Η Περίπτωση 1, με ανεξάρτητα χαρακτηριστικά, είναι η πιο απλή για την επιλογή μεταβλητών. Στις Περιπτώσεις 2 και 3, όμως, έχουμε διαδοχική συσχέτιση έτσι ώστε το $\mathrm{corr}(\mathrm{X}_i, \mathrm{X}_j)$ να μην «φθείρεται» καθώς το $|\mathrm{i} - \mathrm{j}|$ αυξάνεται. Θα δούμε στη συνέχεια ότι και για την Περίπτωση 2 και 3 οι πραγματικοί συντελεστές έχουν επιλεγεί έτσι ώστε η απόκριση να είναι οριακά ασυσχέτιστη με τη X_4. Συνεπώς, περιμένουμε η επιλογή χαρακτηριστικών σε αυτές τις περιπτώσεις να είναι πιο δύσκολη, ειδικά για τις μη-επαναλαμβανόμενες εκδοχές της SIS. Παρατηρούμε ότι στην ασυμπτωτική θεωρία της SIS στους Fan & Lv (2008), αυτό το είδος της εξάρτησης αποκλείεται από τη συνθήκη (4).

Όσον αφορά την επιλογή του *d*, η ασυμπτωτική θεωρία των Fan & Lv (2008) δείχνει ότι στο γραμμικό μοντέλο υπάρχει $\theta^* > 0$ τέτοιο ώστε να μπορούμε να λάβουμε την ιδιότητα του σίγουρου κρησαρίσματος με $\lfloor \mathrm{n}^{1-\theta^*} \rfloor < d < n$. Ωστόσο το θ^* είναι άγνωστο στην πράξη, και ως εκ τούτου οι Fan & Lv πρότειναν $d = \lfloor \mathrm{n}/\log \mathrm{n} \rfloor$ ως μια λογική επιλογή. Φυσικά, επιλέγοντας μια μεγαλύτερη τιμή της *d*, αυξάνεται η πιθανότητα η SIS να περιλαμβάνει όλες τις σωστές μεταβλητές, αλλά συμπεριλαμβάνοντας περισσότερες ανενεργές μεταβλητές θα τείνει να έχει μια ελαφριά αρνητική συνέπεια στην απόδοση της τελικής επιλογής μεταβλητών και στη μέθοδο εκτίμησης παραμέτρων. Έχουμε βρει ότι αυτό το τελευταίο αποτέλεσμα είναι πιο εμφανές στα μοντέλα όπου η απόκριση παρέχει λιγότερες πληροφορίες. Ειδικότερα, η δυαδική απόκριση ενός μοντέλου λογιστικής παλινδρόμησης και σε ένα μικρότερο βαθμό η απόκριση με ακέραια τιμή στο μοντέλο παλινδρόμησης Poisson παρέχει λιγότερες πληροφορίες από ότι η απόκριση με πραγματικές τιμές σε ένα γραμμικό μοντέλο. Ως εκ τούτου, χρησιμοποιώντας $d = \left\lfloor \frac{n}{4\log n} \right\rfloor$ στη λογιστική

παλινδρόμηση, $d = \left\lfloor \frac{n}{2\log n} \right\rfloor$ στην παλινδρόμηση Poisson και $\mathrm{d} = \left\lfloor \frac{n}{2} \right\rfloor$ στη γραμμική παλινδρόμηση. Αυτές οι επιλογές του d με βάση το μοντέλο παρά τα προσαρμοστικά δεδομένα φαίνεται να είναι ικανοποιητικές, καθώς η απόδοση των διαδικασιών είναι αρκετά ισχυρή σε διαφορετικές επιλογές του d (στην πραγματικότητα χρησιμοποιώντας $d = \left\lfloor \frac{n}{\log n} \right\rfloor$ για όλα τα μοντέλα θα εξακολουθούν να δίνουν καλή απόδοση).

3.11.1 Λογιστική Παλινδρόμηση

Σε αυτό το παράδειγμα, τα δεδομένα $(\mathrm{x}_1^T, \mathrm{Y}_1), \ldots, (\mathrm{x}_n^T, \mathrm{Y}_n)$ είναι ανεξάρτητα αντίγραφα του ζεύγους $(\mathrm{x}^T, \mathrm{Y})$, όπου το Y κατανέμεται με την προϋπόθεση ότι $\mathrm{X} = \mathrm{x}$, όπως $\mathrm{Bin}(1, \mathrm{p}(\mathrm{x}))$ με $\log\left(\frac{p(\mathrm{x})}{1 - p(\mathrm{x})}\right) = \beta_0 + x^T \beta$. Επιλέγουμε $n = 400$.

Επιλέξαμε $d = \left\lfloor \frac{n}{4\log n} \right\rfloor = 16$ και στις δύο vanilla εκδοχές της SIS που περιγράφονται παραπάνω (Van-SIS) και στη δεύτερη παραλλαγή (Var2-SIS). Για την πρώτη παραλλαγή (Var1-SIS) ωστόσο χρησιμοποιούμε $d = \left\lfloor \frac{n}{\log n} \right\rfloor = 66$. Παρατηρούμε ότι αφού αυτό σημαίνει ότι τα επιλεγμένα χαρακτηριστικά ανήκουν στην τομή των δύο συνόλων μεγέθους *d*, καταλήγουμε με πολύ λιγότερα από τα *d* χαρακτηριστικά που επιλέγονται από αυτή τη μέθοδο.

Για το παράδειγμα της λογιστικής παλινδρόμησης, η επιλογή της τελικής παραμέτρου κανονικοποιήσης λ για τη SCAD ποινή (μετά από όλα τα (I)SIS βήματα) έγινε με τη βοήθεια ενός ανεξάρτητου συνόλου δεδομένων επικύρωσης του μεγέθους *n* (που παράγεται από το ίδιο μοντέλο όπως τα αρχικά δεδομένα, χρησιμοποιείται μόνο για την προσαρμογή των παραμέτρων, παρά με διασταυρωμένη επικύρωση.

Αυτό ισχύει επίσης για τη LASSO και τις μεθόδους Nearest Shrunken Centroid (NSC, Tibshirani et al., 2003) τις οποίες περιλαμβάνουμε για σύγκριση, αντί να χρησιμοποιήσουμε τη SIS, η μέθοδος αυτή κανονικοποιεί τη λογαριθμική πιθανοφάνεια με L_1-ποινή. Ο λόγος για τον οποίο χρησιμοποιούμε το σύνολο των ανεξάρτητων δεδομένων συντονισμού είναι ότι η έλλειψη πληροφοριών στη δυαδική απόκριση σημαίνει ότι η διασταυρωμένη επικύρωση είναι ιδιαίτερα επιρρεπής στην προσαρμογή στη λογιστική παλινδρόμηση, και ως εκ τούτου εκτελείται ανεπαρκώς για όλες τις μεθόδους.

Οι συντελεστές που χρησιμοποιούνται σε κάθε μια από τις τρεις περιπτώσεις είναι οι εξής:

Περίπτωση 1: $\beta_0 = 0$, $\beta_1 = 1.2439$, $\beta_2 = -1.3416$, $\beta_3 = -1.3500$, $\beta_4 = -1.7971$, $\beta_5 = -1.5810$, $\beta_6 = -1.5967$ και $\beta_j = 0$ για $j > 6$. Το αντίστοιχο Bayes σφάλμα δοκιμής είναι 0.1368.

Περίπτωση 2: $\beta_0 = 0$, $\beta_1 = 4$, $\beta_2 = 4$, $\beta_3 = 4$, $\beta_4 = -6\sqrt{2}$ και $\beta_j = 0$ για $j > 4$. Το Bayes σφάλμα δοκιμής είναι 0.1074.

Περίπτωση 3: $\beta_0 = 0$, $\beta_1 = 4$, $\beta_2 = 4$, $\beta_3 = 4$, $\beta_4 = -6\sqrt{2}$ και $\beta_j = 0$ για $j > 5$. Το Bayes σφάλμα δοκιμής είναι 0.1040.

Στην Περίπτωση 1, οι συντελεστές που επιλέχθηκαν τυχαία και δημιουργήθηκαν ως $(4\log n/\sqrt{n} + |\mathrm{Z}|/4)\,\mathrm{U}$ με $\mathrm{Z} \sim \mathrm{N}(0,1)$ και $\mathrm{U} = 1$ με πιθανότητα 0.5 και $U = -1$ με πιθανότητα -0.5 ανεξαρτήτως του *Z*. Για τις Περιπτώσεις 2 και 3, οι επιλογές διασφαλίζουν ότι ακόμα και αν $\beta_4 \neq 0$ έχουμε ότι τα X_4 και Y είναι ανεξάρτητα. Το γεγονός ότι το X_4 είναι οριακά ανεξάρτητο με την απόκριση είναι σχεδιασμένο να το κάνει δύσκολο για μια δημοφιλή μέθοδο όπως το t-test δύο δειγμάτων ή άλλων μεθόδων ανεξάρτητης μάθησης για να αναγνωρίσει αυτό το σημαντικό χαρακτηριστικό. Επιπλέον, για την Περίπτωση 3, προσθέτουμε μια άλλη σημαντική μεταβλητή X_5 με ένα μικρό συντελεστή για να το κάνει ακόμα πιο δύσκολο να προσδιορίσει το πραγματικό μοντέλο. Για την Περίπτωση 2, οι ιδανικές μεταβλητές που πήραμε από το test δύο δειγμάτων ή την τεχνική ανεξάρτητου

κρησαρίσματος είναι τα X_1, X_2 και X_3. Χρησιμοποιώντας αυτές τις μεταβλητές για να κατασκευάσουμε τον ιδανικό ταξινομητή, ο κίνδυνος Bayes είναι 0.3443, που είναι πολύ μεγαλύτερος από το σφάλμα Bayes 0.1074 του πραγματικού μοντέλου με X_1, X_2, X_3, X_4. Στην πραγματικότητα, μπορεί κανείς να υπερβάλλει στην Περίπτωση 2 για να κάνει το σφάλμα Bayes χρησιμοποιώντας την τεχνική του ανεξάρτητου κρησαρίσματος κοντά στο 0.5, το οποίο αντιστοιχεί σε τυχαία εικασία, θέτοντας $\beta_0 = 0$, $\beta_1 = \beta_2 = \beta_3 = \alpha$, $\beta_m = \alpha$ για $m = 5, 6, ..., j$, $\beta_4 = -a(\mathrm{j}-1)\sqrt{2}/2$ και $\beta_m = 0$ για $m > j$. Για παράδειγμα, το σφάλμα Bayes χρησιμοποιώντας την τεχνική ανεξάρτητου κρησαρίσματος το οποίο διαγράφει το X_4 είναι 0.4290 όταν $j = 20$ και $a = 4$ ενώ το αντίστοιχο σφάλμα Bayes χρησιμοποιώντας τα $X_m, m = 1, 2, ..., 20$ είναι 0.0445.

Στους παρακάτω πίνακες εκθέτουμε διάφορα μέτρα απόδοσης, τα οποία βασίζονται στις 100 Monte Carlo επαναλήψεις. Οι δύο πρώτες σειρές δίνουν το μέσο εκτίμησης L_1 και τα τετραγωνικά σφάλματα εκτίμησης L_2 $\|\beta - \hat{\beta}\|_1 = \sum_{j=0}^{p} |\beta_j - \hat{\beta}_j|$ και $\|\beta - \hat{\beta}\|_2^2 = \sum_{j=0}^{p} (\beta_j - \hat{\beta}_j)^2$. Η τρίτη γραμμή δίνει την αναλογία των φορών που η υπό εξέταση (I)SIS διαδικασία περιλαμβάνει όλα τα σημαντικά χαρακτηριστικά στο μοντέλο, ενώ η τέταρτη γραμμή αναφέρει την αντίστοιχη αναλογία των φορών που έχουν επιλεγεί τα τελικά χαρακτηριστικά μετά την εφαρμογή της SCAD ή της ποινής LASSO ανάλογα με την περίπτωση και περιλαμβάνουν όλα τα σημαντικά. Η πέμπτη σειρά δίνει το μέσο τελικό αριθμό των χαρακτηριστικών που έχουν επιλεγεί. Στην έκτη, έβδομη και όγδοη σειρά παρέχονται μέτρα προσαρμογής στα δεδομένα εκπαίδευσης, δηλαδή οι μέσες τιμές του $2Q(\hat{\beta}_0, \hat{\beta})$ που ορίστηκαν νωρίτερα, οι πληροφορίες του κριτηρίου Akaike (Akaike, 1974) το οποίο προσθέτει το διπλάσιο αριθμό χαρακτηριστικών στο τελικό μοντέλο και το κριτήριο Bayesian πληροφοριών (Schwarz, 1978), η οποία προσθέτει το γινόμενο των $\log n$ και τον αριθμό των χαρακτηριστικών στο τελικό μοντέλο. Τέλος, ένα ανεξάρτητο σύνολο πειραματικών δεδομένων μεγέθους $100n$ χρησιμοποιήθηκε για να αξιολογήσει τη μέση τιμή της $2Q(\hat{\beta}_0, \hat{\beta})$ στα πειραματικά δεδομένα (Σειρά 9), καθώς και να εκθέσει τη διάμεσο 0-

1 σφάλμα δοκιμής (Σειρά 10), όπου παρατηρούμε ένα σφάλμα αν η απόκριση δοκιμής διαφέρει από την προσαρμοσμένη απόκριση κατά περισσότερο από $1/2$. Ο Πίνακας 1 συγκρίνει πέντε μεθόδους: Van-SIS, Var1-SIS, Var2-SIS, LASSO και NSC. Η πιο αξιοσημείωτη παρατήρηση είναι ότι ενώ η LASSO πάντα περιλαμβάνει όλα τα σημαντικά χαρακτηριστικά, το κάνει επιλέγοντας πολύ μεγάλα μοντέλα- μια διάμεσο 94 μεταβλητών- σε αντίθεση με το σωστό αριθμό, 6, ο οποίος είναι το μέσο μέγεθος μοντέλου που αναφέρθηκαν από όλες τις τρεις μεθόδους που βασίζονται στη SIS. Αυτό οφείλεται στην προκατάληψη της LASSO, όπως τόνισαν οι Fan & Li (2001) και Zou (2006), το οποίο ενθαρρύνει την επιλογή μιας μικρής παραμέτρου κανονικοποίησης για να καταστήσει το συνολικό μέσο τετραγωνικό σφάλμα μικρό. Κατά συνέπεια πολλά ανεπιθύμητα χαρακτηριστικά προσλαμβάνονται επίσης. Αυτό αποδεικνύεται επίσης από τη σύγκριση των διαφορών μεταξύ των L_1 και L_2 απωλειών στις δύο πρώτες σειρές. Έτσι, η μέθοδος LASSO έχει μεγάλο σφάλμα εκτίμησης και ενώ το $2Q(\hat{\beta}_0, \hat{\beta})$ είναι σχετικά μικρό με το σύνολο δεδομένων εκπαίδευσης αυτό είναι ένα αποτέλεσμα υπερπροσαρμογής όπως φαίνεται από τις μεγάλες τιμές των AIC/BIC, $2Q(\hat{\beta}_0, \hat{\beta})$ πάνω στα πειραματικά δεδομένα και του 0-1 σφάλματος δοκιμής.

Δεδομένου ότι τα χαρακτηριστικά είναι ανεξάρτητα στην Περίπτωση 1, δεν είναι έκπληξη να δούμε ότι η Van-SIS έχει την καλύτερη επίδοση από τις τρεις μεθόδους που βασίζονται στη SIS. Ακόμη και με τη μεγαλύτερη τιμή του d που χρησιμοποιείται για τη Var1-SIS τείνει να χάσει σημαντικά χαρακτηριστικά πιο συχνά από τις άλλες μεθόδους. Αν και η μέθοδος φαίνεται να έχει αξία ως μέσο απόκτησης ενός ελάχιστου συνόλου χαρακτηριστικών που θα πρέπει να συμπεριληφθούν σε ένα τελικό μοντέλο δε θα λάβουμε υπόψη τη Var1-SIS περαιτέρω στη μελέτη προσομοίωσής μας.

Ο Πίνακας 3.11.2 παρουσιάζει τα αποτελέσματα της επανάληψης των προσομοιώσεων της Περίπτωσης 1 για τις Van-SIS, Var1-SIS και Var2-SIS κάτω από τις ίδιες συνθήκες αλλά χρησιμοποιώντας τη LASSO ποινικοποιημένη συνάρτηση παρά τη SCAD ποινικοποιημένη συνάρτηση μετά το στάδιο της SIS. Αυτές οι εκδοχές καλούνται Van-SIS-LASSO, Var1-SIS-LASSO και Var2-SIS-LASSO αντίστοιχα. Παρατηρούμε ότι καθώς μειώνεται το υπολογιστικό κόστος

χρησιμοποιώντας οποιαδήποτε από τις τρείς εκδοχές της SIS πριν τη LASSO βελτιώνεται σημαντικά και η απόδοση σε σχέση με την εφαρμογή της LASSO στο ολοκληρωμένο σύνολο χαρακτηριστικών. Από την άλλη πλευρά, τα αποτελέσματα είναι λιγότερο επιτυχή από το να εφαρμόζαμε τη SIS και τις παραλλαγές της σε συνδυασμό με τη SCAD ποινή για την τελική επιλογή χαρακτηριστικών και την εκτίμηση παραμέτρου.

Στις Περιπτώσεις 2 και 3 θεωρούμε επίσης τις επαναληπτικές εκδοχές των Van-SIS και Var2-SIS τις οποίες δηλώνουμε Van-ISIS και Var2-ISIS αντίστοιχα. Σε κάθε ενδιάμεσο στάδιο των ISIS διαδικασιών, το κριτήριο των Bayesian πληροφοριών χρησιμοποιήθηκε ως ένας γρήγορος τρόπος επιλογής της SCAD παραμέτρου κανονικοποίησης.

Από τους Πίνακες 3.11.3 και 3.11.4, παρατηρούμε ότι οι μη-επαναληπτικές SIS μέθοδοι αποτυγχάνουν άσχημα σε αυτές τις δύσκολες περιπτώσεις. Οι επιδόσεις τους είναι παρόμοιες με εκείνες της μεθόδου LASSO. Από την άλλη πλευρά, και οι δύο επαναληπτικές μέθοδοι Van-ISIS και Var2-ISIS εκτελούνται εξαιρετικά καλά .

	Van-SIS	Var1-SIS	Var2-SIS	LASSO	NSC
$\|\beta - \widehat{\beta}\|_1$	1.1093	1.2495	1.2134	8.4821	N/A
$\|\beta - \widehat{\beta}\|_2^2$	0.4861	0.5237	0.5204	1.7029	N/A
Prop. incl. (I)SIS models	0.99	0.84	0.91	N/A	N/A
Prop. incl. final models	0.99	0.84	0.91	1.00	0.34
Median final model size	6	6	6	94	3
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	237.21	247.00	242.85	163.64	N/A
AIC	250.43	259.87	256.26	352.54	N/A
BIC	277.77	284.90	282.04	724.70	N/A
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	271.81	273.08	272.91	318.52	N/A
0-1 test error	0.1421	0.1425	0.1426	0.1720	0.3595

Πίνακας 3.11. 1. Λογιστική παλινδρόμηση, Περίπτωση 1

	Van-SIS-LASSO	Var1-SIS-LASSO	Var2-SIS-LASSO
$\|\beta - \widehat{\beta}\|_1$	3.8500	2.1050	3.0055
$\|\beta - \widehat{\beta}\|_2^2$	1.0762	0.7536	0.9227
Prop. incl. (I)SIS models	0.99	0.84	0.91
Prop. incl. final models	0.99	0.84	0.91
Median final model size	16.0	9.0	14.5
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	207.86	240.44	226.95
AIC	239.69	260.49	255.99
BIC	302.98	295.40	316.36
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	304.79	280.95	291.79
0-1 test error	0.1621	0.1476	0.1552

Πίνακας 3.11.2. Λογιστική παλινδρόμηση, Περίπτωση 1

	Van-SIS	Van-ISIS	Var2-SIS	Var2-ISIS	LASSO	NSC
$\|\beta - \widehat{\beta}\|_1$	20.0504	1.9445	20.1100	1.8450	21.6437	N/A
$\|\beta - \widehat{\beta}\|_2^2$	9.4101	1.0523	9.3347	0.9801	9.1123	N/A
Prop. incl. (I)SIS models	0.00	1.00	0.00	1.00	N/A	N/A
Prop. incl. final models	0.00	1.00	0.00	1.00	0.00	0.21
Median final model size	16	4	16	4	91	16.5
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	307.15	187.58	309.63	187.42	127.05	N/A
AIC	333.79	195.58	340.77	195.58	311.10	N/A
BIC	386.07	211.92	402.79	211.55	672.34	N/A
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	344.25	204.23	335.21	204.28	258.65	N/A
0-1 test error	0.1925	0.1092	0.1899	0.1092	0.1409	0.3765

Πίνακας 3.11.3. Λογιστική παλινδρόμηση, Περίπτωση 2

	Van-SIS	Van-ISIS	Var2-SIS	Var2-ISIS	LASSO	NSC
$\|\beta - \widehat{\beta}\|_1$	20.5774	2.6938	20.6967	3.2461	23.1661	N/A
$\|\beta - \widehat{\beta}\|_2^2$	9.4568	1.3615	9.3821	1.5852	9.1057	N/A
Prop. incl. (I)SIS models	0.00	1.00	0.00	1.00	N/A	N/A
Prop. incl. final models	0.00	0.90	0.00	0.98	0.00	0.17
Median final model size	16	5	16	5	101.5	10
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	269.20	187.89	296.18	187.89	109.32	N/A
AIC	289.20	197.59	327.66	198.65	310.68	N/A
BIC	337.05	218.10	389.17	219.18	713.78	N/A
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	360.89	225.15	358.13	226.25	275.55	N/A
0-1 test error	0.1933	0.1120	0.1946	0.1119	0.1461	0.3866

Πίνακας 3.11.4. Λογιστική παλινδρόμηση, Περίπτωση 3

3.11.2. Παλινδρόμηση Poisson

Στο δεύτερο παράδειγμά μας, η γενική απόκριση Y κατανέμεται με την προϋπόθεση ότι $\mathrm{X} = x$ κατά Poisson($\mu(\mathrm{x})$), όπου $log\,\mu(\mathrm{x}) = \beta_0 + x^T\beta$.

Λόγω της επιπλέον πληροφορίας στην απόκριση καταμέτρησης επιλέγουμε $n = 200$ και εφαρμόζουμε όλες τις εκδοχές της (I)SIS με $d = \lfloor n/2\log n \rfloor = 37$. Χρησιμοποιούμε επίσης 10πλάσια διασταυρωμένη επικύρωση για να επιλέξουμε την τελική παράμετρο κανονικοποίησης για τις SCAD και LASSO ποινές. Οι συντελεστές που χρησιμοποιήθηκαν ήταν οι εξής:

Περίπτωση 1: $\beta_0 = 5$, $\beta_1 = -0.5423$, $\beta_2 = 0.5314$, $\beta_3 = -0.5012$, $\beta_4 = -0.4850$, $\beta_5 = -0.4133$, $\beta_6 = 0.5234$ και $\beta_j = 0$ για $j > 6$.

Περίπτωση 2: $\beta_0 = 5$, $\beta_1 = 0.6$, $\beta_2 = 0.6$, $\beta_3 = 0.6$, $\beta_4 = -0.9\sqrt{2}$ και $\beta_j = 0$ για $j > 4$.

Περίπτωση 3: $\beta_0 = 5$, $\beta_1 = 0.6$, $\beta_2 = 0.6$, $\beta_3 = 0.6$, $\beta_4 = -0.9\sqrt{2}$, $\beta_5 = 0.15$ και $\beta_j = 0$ για $j > 5$.

Στην Περίπτωση 1 τα μεγέθη των συντελεστών $\beta_1,....,\beta_6$ παράχθηκαν από $(\frac{\log n}{\sqrt{n}} + |\mathrm{Z}|/8)\,\mathrm{U}$ με $\mathrm{Z} \sim \mathrm{N}(0,1)$ και $U = 1$ με πιθανότητα 1 και $U = -1$ με πιθανότητα 0.5, ανεξάρτητα του Z. Και πάλι, οι επιλογές στις Περιπτώσεις 2 και 3 εξασφαλίζουν ότι ακόμη και αν $\beta_4 \neq 0$, έχουμε $corr(\mathrm{X}_4, \mathrm{Y}) = 0$. Οι συντελεστές είναι μια αναπροσαρμοσμένη εκδοχή εκείνων στο μοντέλο λογιστικής παλινδρόμησης, εκτός από $\beta_0 = 5$ που χρησιμοποιείται για τον έλεγχο του κατάλληλου σήματος-προς-το-θόρυβο.

Τα αποτελέσματα εμφανίζονται στους Πίνακες 3.11.5, 3.11.6 και 3.11.7. Ακόμη και στην Περίπτωση 1, με ανεξάρτητα χαρακτηριστικά, οι ISIS μέθοδοι ξεπερνούν τη SIS. Και πάλι τόσο η Van-ISIS όσο και η Var2-ISIS εκτελούνται

αρκετά καλά, σχεδόν πάντα, συμπεριλαμβανομένων όλων των σημαντικών χαρακτηριστικών σε σχετικά μικρά τελικά μοντέλα. Η LASSO μέθοδος συνεχίζει να πάσχει από την υπερπροσαρμογή, ιδιαίτερα στις δύσκολες Περιπτώσεις 2 και 3.

	Van-SIS	Van-ISIS	Var2-SIS	Var2-ISIS	LASSO
$\|\beta - \widehat{\beta}\|_1$	0.0695	0.1239	1.1773	0.1222	0.1969
$\|\beta - \widehat{\beta}\|_2^2$	0.0225	0.0320	0.4775	0.0330	0.0537
Prop. incl. (I)SIS models	0.76	1.00	0.45	1.00	N/A
Prop. incl. final models	0.76	1.00	0.45	1.00	1.00
Median final model size	12	18	13	17	27
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	1560.85	1501.80	7735.51	1510.38	1534.19
AIC	1586.32	1537.80	7764.51	1542.14	1587.23
BIC	1627.06	1597.17	7812.34	1595.30	1674.49
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	1557.74	1594.10	14340.26	1589.51	1644.63

Πίνακας 3.11.5. Poisson παλινδρόμηση, Περίπτωση 1.

	Van-ISIS	Var2-ISIS	LASSO
$\|\beta - \widehat{\beta}\|_1$	0.2705	0.2252	3.0710
$\|\beta - \widehat{\beta}\|_2^2$	0.0719	0.0667	1.2856
Prop. incl. (I)SIS models	1.00	0.97	N/A
Prop. incl. final models	1.00	0.97	0.00
Median final model size	18	16	174
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	1494.53	1509.40	1369.96
AIC	1530.53	1541.17	1717.91
BIC	1589.90	1595.74	2293.29
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	1629.49	1614.57	2213.10

Πίνακας 3.11.6. Poisson παλινδρόμηση, Περίπτωση 2.

	Van-ISIS	Var2-ISIS	LASSO
$\|\beta - \widehat{\beta}\|_1$	0.2541	0.2319	3.0942
$\|\beta - \widehat{\beta}\|_2^2$	0.0682	0.0697	1.2856
Prop. incl. (I)SIS models	0.97	0.91	0.00
Prop. incl. final models	0.97	0.91	0.00
Median final model size	18	16	174
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	1500.03	1516.14	1366.63
AIC	1536.03	1546.79	1715.35
BIC	1595.40	1600.17	2293.60
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	1640.27	1630.58	2389.09

Πίνακας 3.11.7. Poisson παλινδρόμηση, Περίπτωση 3.

3.11.3. Ανθεκτική Παλινδρόμηση

Έχουν διεξαχθεί παρόμοια αριθμητικά πειράματα χρησιμοποιώντας την L_1-παλινδρόμηση για τις τρεις περιπτώσεις με ένα ανάλογο τρόπο με τα προηγούμενα δύο παραδείγματα. Παίρνουμε παρόμοια αποτελέσματα. Και οι δύο εκδοχές της ISIS είναι αποτελεσματικές στην επιλογή σημαντικών χαρακτηριστικών σε σχετικά χαμηλά εσφαλμένα θετικά ποσοστά. Ως εκ τούτου, τα σφάλματα πρόβλεψης είναι επίσης μικρά. Από την άλλη πλευρά, η LASSO έχασε τις δύσκολες μεταβλητές στις Περιπτώσεις 2 και 3 και επίσης επιλεγμένα μοντέλα με ένα μεγάλο αριθμό χαρακτηριστικών για να μετριάσει την προκατάληψη της διαδικασίας της επιλογής μεταβλητών. Ως αποτέλεσμα, τα σφάλματα πρόβλεψής τους είναι πολύ μεγαλύτερα.

3.11.4. Γραμμική Παλινδρόμηση

Παρατηρούμε ότι η νέα διαδικασία ISIS μας επιτρέπει τη διαγραφή χαρακτηριστικών σε κάθε βήμα. Είναι μια σημαντική βελτίωση πάνω από την αρχική πρόταση των Fan & Lv (2008) ακόμη και στη συνήθη ρύθμιση ελαχίστων

τετραγώνων. Για να το δείξουμε αυτό επιλέγουμε την Περίπτωση 3, την πιο δύσκολη, με συντελεστές που δίνονται ως εξής:

Περίπτωση 3: $\beta_0 = 0$, $\beta_1 = 5$, $\beta_2 = 5$, $\beta_3 = 5$, $\beta_4 = -15\sqrt{2}/2$, $\beta_5 = 1$ και $\beta_j = 0$ για $j > 5$.

Η απόκριση Y ορίζεται ως $Y = x^T\beta + \varepsilon$ με ανεξάρτητη $\varepsilon \sim \mathrm{N}(0,1)$. Χρησιμοποιώντας $\mathrm{n} = 70$ και $d = n/2$, η νέα ISIS μέθοδος περιλαμβάνει τις πέντε σημαντικές μεταβλητές για τις 91 από τις 100 επαναλήψεις, ενώ η αρχική ISIS χωρίς τη δυνατότητα διαγραφής χαρακτηριστικών περιλαμβάνει όλα τα σημαντικά χαρακτηριστικά για 36 μόνο από τις 100 επαναλήψεις. Το μεσαίο μέγεθος μοντέλου της νέας διαδικασίας επιλογής μεταβλητών με διαγραφή μεταβλητής είναι 21, ενώ το διάμεσο μέγεθος μοντέλου που αντιστοιχεί στην αρχική ISIS είναι 19.

Έχει διεξαχθεί ακόμα και το αριθμητικό πείραμα με διαφορετικό μέγεθος δείγματος $\mathrm{n} = 100$ και $d = n/2 = 50$. Για 97 από τις 100 επαναλήψεις, η νέα ISIS περιλαμβάνει όλα τα σημαντικά χαρακτηριστικά ενώ η ISIS χωρίς τη διαγραφή μεταβλητών περιλαμβάνει όλα τα σημαντικά χαρακτηριστικά για μόνο 72 επαναλήψεις. Τα ενδιάμεσα μεγέθη του μοντέλου είναι 26. Αυτό καταδεικνύει σαφώς τη βελτίωση της δυνατότητας διαγραφής χαρακτηριστικών σε αυτό το παράδειγμα.

3.11.5 Πολυκατηγορική Ταξινόμηση

Το τελευταίο μας παράδειγμα σε αυτή την ενότητα είναι ένα πρόβλημα ταξινόμησης τεσσάρων κλάσεων. Θα μελετήσουμε δύο διαφορετικές διαμορφώσεις χαρακτηριστικών τα οποία και τα δύο εξαρτώνται από την πρώτη ανεξάρτητη παραγωγή $\tilde{X}_1,...,\tilde{X}_p$ έτσι ώστε τα $\tilde{X}_1,...,\tilde{X}_4$ να είναι ομοιόμορφα κατανεμημένα στο $\left[-\sqrt{3},\sqrt{3}\right]$ και $\tilde{X}_5,...,\tilde{X}_p$ να είναι κατανεμημένα στο $\mathrm{N}(0,1)$. Χρησιμοποιούμε αυτές τις τυχαίες μεταβλητές για να δημιουργήσουμε τις ακόλουθες περιπτώσεις:

Περίπτωση 1: $X_j = \tilde{X}_j$ για $j = 1,...., p$.

Περίπτωση 2: $X_1 = \tilde{X}_1 - \sqrt{2}\tilde{X}_5$, $X_2 = \tilde{X}_2 + \sqrt{2}\tilde{X}_5$, $X_3 = \tilde{X}_3 - \sqrt{2}\tilde{X}_5$, $X_4 = \tilde{X}_4 + \sqrt{2}\tilde{X}_5$ και $X_j = \sqrt{3}\tilde{X}_j$ για $j = 5, ..., p$.

Με την προϋπόθεση ότι $X = x$, η απόκριση Y που παράγεται σύμφωνα με την $P(\mathrm{Y} = \mathrm{k} \mid \tilde{\mathrm{X}} = \tilde{\mathrm{x}}) \propto \exp\{\mathrm{f}_k(\tilde{\mathrm{x}})\}$, για $\mathrm{k} = 1, ..., 4$ όπου $f_1(\tilde{\mathrm{x}}) = -\alpha\, \tilde{\mathrm{x}}_1 + a\, \tilde{\mathrm{x}}_4$, $f_2(\tilde{\mathrm{x}}) = \alpha\, \tilde{\mathrm{x}}_1 - a\, \tilde{\mathrm{x}}_2$, $f_3(\tilde{\mathrm{x}}) = \alpha\, \tilde{\mathrm{x}}_2 - a\, \tilde{\mathrm{x}}_3$ και $f_4(\tilde{\mathrm{x}}) = \alpha\, \tilde{\mathrm{x}}_3 - a\, \tilde{\mathrm{x}}_4$ για $a = 5/\sqrt{3}$.

Και στις δύο Περιπτώσεις 1 και 2 όλα τα χαρακτηριστικά έχουν την ίδια τυπική απόκλιση αφού $sd(\mathrm{X}_j) = 1$ για $j = 1, 2, ..., \mathrm{p}$ στην Περίπτωση 1 και $sd(\mathrm{X}_j) = \sqrt{3}$ για $j = 1, 2, ..., \mathrm{p}$ στην Περίπτωση 2. Επιπλέον για αυτή την Περίπτωση η μεταβλητή X_5 είναι οριακά ασήμαντη αλλά από κοινού σημαντική, έτσι αντιπροσωπεύει μια πρόκληση για να την ταυτοποιήσουμε ως μια σημαντική μεταβλητή. Και για τις δύο Περιπτώσεις 1 και 2, το Bayes σφάλαμα είναι 0.1373.

Για την πολυκατηγορική κατάταξη χρησιμοποιούμε τη συνάρτηση απώλειας που προτείνουν οι Lee, Lin & Wahba (2004). Δείχνουμε τους συντελεστές για την k-οστή κλάση από τα β_{0k} και β_k για $k = 1, 2, 3, 4$ και έστω $B = (\beta_{01}, \beta_1^T)^T$, $(\beta_{02}, \beta_2^T)^T$, $(\beta_{03}, \beta_3^T)^T$, $(\beta_{04}, \beta_4^T)^T$. Έστω $f_k(\mathrm{x}) \equiv f_k(\mathrm{x}, \beta_{0k}, \beta_k) = \beta_{0k} + x^T \beta_k$, $k = 1, 2, 3, 4$ και

$f(\mathrm{x}) \equiv \mathrm{f}(\mathrm{x}, \mathrm{B}) = (\mathrm{f}_1(\mathrm{x}), \mathrm{f}_2(\mathrm{x}), \mathrm{f}_3(\mathrm{x}), \mathrm{f}_4(\mathrm{x}))^T$.

Η συνάρτηση απώλειας δίνεται από $L(\mathrm{Y}, \mathrm{f}(\mathrm{x})) = \sum_{j \neq Y} [1 + \mathrm{f}_j(\mathrm{x})]_+$ όπου $[\psi]_+ = \psi$ αν $\psi \geq 0$ και 0 διαφορετικά. Άμα αποκλίνουμε ελαφρά από την τυπική μας διαδικασία η οριακή χρησιμότητα του j-οστού χαρακτηριστικού ορίζεται από $L_j = \min_B \sum_{i=1}^{n} L(\mathrm{Y}, \mathrm{f}(\mathrm{X}_{ij}, \mathrm{B})) + \frac{1}{2} \sum_{k=1}^{4} \beta_{jk}^2$ για να αποφευχθούν πιθανά θέματα μη αναγνωρισιμότητας λόγω της άρθρωσης συνάρτησης απώλειας. Με εκτιμημένους συντελεστές $\hat{\beta}_{0k}$ και $\hat{\beta}_k$ και $\hat{f}_k(\mathrm{x}) = \hat{\beta}_{0k} + x^T \hat{\beta}_k$ για $k = 1, 2, 3, 4$, ο κανόνας εκτιμημένης ταξινόμησης δίνεται από $\mathrm{argmax}_k\, \hat{f}_k(\mathrm{x})$.

Όπως και με τη λογιστική παλινδρόμηση στην ενότητα 3.10.1, χρησιμοποιούμε $n = 400$, $d = \left\lfloor \frac{n}{4\log n} \right\rfloor = 16$ και ένα ανεξάρτητο σύνολο επικυρωμένων δεδομένων του μεγέθους n για να πάρουμε την τελική παράμετρο κανονικοποίησης για την SCAD ποινή.

Τα αποτελέσματα δίνονται στον Πίνακα 3.11.8. Το μέσο εκτιμώμενο σφάλμα δοκιμής βασίστηκε σε περαιτέρω δοκιμές του συνόλου δεδομένων του μεγέθους $200n$. Αναφέρουμε επίσης το τυπικό σφάλμα της μέσης εκτίμησης. Στην περίπτωση των ανεξάρτητων χαρακτηριστικών, όλες οι (I)SIS μέθοδοι έχουν παρόμοια απόδοση. Τα οφέλη από τη χρήση επαναληπτικών εκδοχών της ISIS μεθοδολογίας είναι και πάλι σαφή για την Περίπτωση 2, με εξαρτώμενα χαρακτηριστικά.

	Van-SIS	Van-ISIS	Var2-SIS	Var2-ISIS	LASSO	NSC
			Case 1			
Prop. incl. (I)SIS models	1.00	1.00	0.99	1.00	N/A	N/A
Prop. incl. final model	1.00	1.00	0.99	1.00	0.00	0.68
Median modal size	2.5	4	10	5	19	4
0-1 test error	0.3060	0.3010	0.2968	0.2924	0.3296	0.4524
Test error standard error	0.0067	0.0063	0.0067	0.0061	0.0078	0.0214
			Case 2			
Prop. incl. (I)SIS models	0.10	1.00	0.03	1.00	N/A	N/A
Prop. incl. final models	0.10	1.00	0.03	1.00	0.33	0.30
Median modal size	4	11	5	9	54	9
0-1 test error	0.4362	0.3037	0.4801	0.2983	0.4296	0.6242
Test error standard error	0.0073	0.0065	0.0083	0.0063	0.0043	0.0084

Πίνακας 3.11.8. Πολυκατηγορική ταξινόμηση.

3.11.6. Παραδείγματα Πραγματικών Δεδομένων.

Σε αυτή την ενότητα εφαρμόζουμε τις προτεινόμενες μεθόδους σε δύο πραγματικά σύνολα δεδομένων. Το πρώτο μέρος έχει μια δυαδική απόκριση ενώ το δεύτερο είναι πολυ-κατηγορικό. Αντιμετωπίζουμε και τα δύο ως προβλήματα ταξινόμησης και χρησιμοποιούμε την απώλεια άρθρωσης που συζητήθηκε στην

πολυκατηγορική ταξινόμηση. Συγκρίνουμε τις μεθόδους μας με δύο εναλλακτικές: τη LASSO και τη NSC.

3.11.6.1. Δεδομένα Νευροβλαστώματος.

Αρχικά θεωρούμε τα δεδομένα νευροβλαστώματος που χρησιμοποιήθηκαν στο Oberthuer et al. (2006). Το νευροβλάστωμα είναι ένας εξωκρανιακός στερεός καρκίνος. Είναι πιο συχνό στην παιδική και νηπιακή ηλικία. Ο ετήσιος αριθμός περιστατικών είναι περίπου αρκετές εκατοντάδες στις Ηνωμένες Πολιτείες. Το νευροβλάστωμα είναι ένας κακοήθης όγκος της παιδιατρικής που προέρχεται από νευρικά ακρολοφία στοιχεία του σπλαχνικού νευρικού συστήματος. Η μελέτη αποτελείται από 251 ασθενείς των γερμανικών δοκιμών Νευροβλαστώματος NB90-NB2004, που είχαν διαγνωστεί μεταξύ 1989 και 2004. Κατά τη διάγνωση, οι ηλικίες των ασθενών κυμαίνονται από 0 έως 296 μήνες με μέση ηλικία τους 15 μήνες. Αναλύθηκαν 251 δείγματα νευροβλαστώματος χρησιμοποιώντας μια προσαρμοσμένη μικροσυστοιχία ολιγονουκλεοτιδίων με στόχο την ανάπτυξη μιας γονιδιακής έκφρασης με βάση τον κανόνα ταξινόμησης για ασθενείς με νευροβλάστωμα για να προβλέψει με αξιοπιστία στοιχεία της νόσου. Αυτό παρέχει επίσης μια ολοκληρωμένη άποψη για το ποια σύνολα γονιδίων είναι υπεύθυνα για το νευροβλάστωμα.

Το πλήρες σύνολο δεδομένων το οποίο λαμβάνουμε μέσω του Ποιοτικού Ελέγχου MicroArray II (MAQC-II) περιλαμβάνει την έκφραση γονιδίων πάνω από 10.707 θέσεις καθετήρα. Ιδιαίτερου ενδιαφέροντος είναι η πρόβλεψη της απόκρισης με την ένδειξη «3-χρόνια επιβίωσης χωρίς συμβάντα» (3-χρόνια EFS), η οποία είναι μια δυαδική μεταβλητή που δείχνει αν κάθε ασθενής επέζησε 3 χρόνια μετά τη διάγνωση του νευροβλαστώματος.

Αποκλείοντας 5 ακραίες συστοιχίες, υπάρχουν 246 άτομα από τα οποία 239 άτομα έχουν 3-χρόνια EFS πληροφορίες διαθέσιμα με 49 θετικά και 190 αρνητικά. Εφαρμόζουμε τις SIS και ISIS για να μειώσουμε τη διάσταση από $p = 10707$ σε $d = 50$. Από την άλλη πλευρά, οι ανταγωνιστικές μέθοδοι LASSO και NSC εφαρμόζονται απευθείας στα $p = 10707$ γονίδια. Όταν είναι απαραίτητο, πενταπλάσια διασταυρωμένη επικύρωση χρησιμοποιείται για την επιλογή

παραμέτρων συντονισμού. Επιλέγουμε τυχαία 125 άτομα (25 θετικά και 100 αρνητικά) να είναι το σύνολο εκπαίδευσης και τα υπόλοιπα χρησιμοποιούνται ως το σύνολο δοκιμής. Τα αποτελέσματα αναφέρονται στο πάνω μισό του Πίνακα 3.11.9.

Στο MAQC-II, ένα ειδικά σχεδιασμένο τελικό σημείο είναι το φύλο του κάθε ασθενούς το οποίο θα πρέπει να αποτελεί εύκολη ταξινόμηση. Ο στόχος αυτού του ειδικά σχεδιασμένου τελικού σημείου είναι να συγκρίνουμε την απόδοση των διαφορετικών ταξινομητών για απλές εργασίες ταξινόμησης. Οι πληροφορίες για το φύλο είναι διαθέσιμες για όλες τις ακραίες τιμές των 246 συστοιχιών με 145 άντρες και 101 γυναίκες. Επιλέγουμε τυχαία 70 άντρες και 50 γυναίκες να είναι στο σύνολο εκπαίδευσης και χρησιμοποιούμε τους υπόλοιπους ως το σύνολο δοκιμής. Θέτουμε $d = 50$ για τις SIS και ISIS όπως στην περίπτωση των 3-χρόνων EFS τελικού σημείου. Τα αποτελέσματα δίνονται στο δεύτερο μισό του Πίνακα 3.11.9.

Μπορούμε να δούμε από τον Πίνακα 3.11.9 ότι οι (I)SIS μέθοδοί μας συγκρίνονται ευνοϊκά με τις LASSO και NSC. Ειδικά για το τελικό σημείο 3-χρόνια EFS, οι μέθοδοί μας χρησιμοποιούν λιγότερα χαρακτηριστικά ενώ δίνουν μικρότερα δοκιμαστικά σφάλματα.

End point		SIS	ISIS	var2-SIS	var2-ISIS	LASSO	NSC
3-year EFS	No. of features	5	23	10	12	57	9413
	Testing error	19/114	22/114	22/114	21/114	22/114	24/114
Gender	No. of features	6	2	4	2	42	3
	Testing error	4/126	4/126	4/126	4/126	5/126	4/126

<u>Πίνακας 3.11.9.</u> Αποτελέσματα ανάλυσης δύο τελικών σημείων για τα δεδομένα νευροβλαστώματος.

3.11.6.2. Δεδομένα SRBCT

Σε αυτή την ενότητα, εφαρμόζουμε τη μέθοδό μας στα δεδομένα παιδικού καρκίνου που αναφέρθηκαν στο Khan et al. (2001). Ο Khan et al. (2001) χρησιμοποίησε τεχνητά νευρωνικά δίκτυα για την ανάπτυξη μιας μεθόδου ταξινόμησης των μικρών στρογγυλών μπλε όγκων των κυττάρων (SRBCTs) της παιδικής ηλικίας σε μια από τις τέσσερις κατηγορίες: νευροβλάστωμα (NB),

ραβδομυοσάρκωμα (RMS), non-Hodgkin λέμφωμα (NHL) και την οικογένεια όγκων Ewing (EWS) χρησιμοποιώντας cDNA προφίλ γονιδιακής έκφρασης. Η ακριβής διάγνωση των SRBCTs σε αυτές τις τέσσερις διαφορετικές διαγνωστικές κατηγορίες είναι σημαντικές στην επιλογή θεραπείας και οι αποκρίσεις στη θεραπεία είναι διαφορετικές από τη μία κατηγορία στην άλλη.

Μετά το φιλτράρισμα 2308 προφίλ γονιδίων από τα 6567 γονίδια δίνονται στα SRBCTs σύνολα δεδομένων. Περιλαμβάνει ένα σύνολο εκπαίδευσης του μεγέθους 63 (12NB, 20RMS, 8NHL & 23EWS) και ένα ανεξάρτητο σύνολο δοκιμών του μεγέθους 20 (6NB, 5RMS, 3NHL & 6EWS).

Πριν την εφαρμογή της κατάταξης τυποποιούμε σύνολα δεδομένων εφαρμόζοντας ένα απλό γραμμικό μετασχηματισμό τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής. Ο γραμμικός μετασχηματισμός βασίζεται στα δεδομένα εκπαίδευσης έτσι ώστε μετά την τυποποίηση τα δεδομένα εκπαίδευσης να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Η (I)SIS μειώνει τη διάσταση από $p = 2308$ σε $d = \left\lfloor \frac{63}{\log 63} \right\rfloor = 15$ πρώτα, ενώ οι εναλλακτικές μέθοδοι LASSO και NSC εφαρμόζονται απευθείας σε $p = 2308$ γονίδια. Ανάλογα με την περίπτωση μια τετραπλάσια διασταυρωμένη επικύρωση (*cross validation*) χρησιμοποιείται για την επιλογή παραμέτρων συντονισμού.

Οι ISIS, Var2-ISIS, LASSO και NSC όλες επιτυγχάνουν το μηδενικό σφάλμα δοκιμής για τα 20 δείγματα στο σύνολο δοκιμής. Η NSC χρησιμοποιεί 343 γονίδια και η LASSO απαιτεί 71 γονίδια. Ωστόσο οι ISIS και Var2-ISIS χρησιμοποιούν 15 και 14 γονίδια, αντίστοιχα.

Αυτή η πραγματική εφαρμογή δεδομένων παρέχει το ίδιο μήνυμα, ότι οι νέες ISIS και Var2-ISIS μέθοδοι μπορούν να επιτύχουν ανταγωνιστικές επιδόσεις της κατάταξης χρησιμοποιώντας λιγότερα χαρακτηριστικά.

ΠΑΡΑΡΤΗΜΑ

Σε αυτή την ενότητα παρουσιάζουμε διάφορα παραδείγματα προσομοίωσης για την αξιολόγηση της απόδοσης της SIS διαδικασίας με γενικευμένα γραμμικά μοντέλα. Οι Fan & Lv (2008) και οι Fan, Samworth & Wu (2009) απέδειξαν ότι το ανεξάρτητο κρησάρισμα είναι μια γρήγορη αλλά ακατέργαστη μέθοδος μείωσης της διάστασης σε πιο μέτριο μέγεθος. Ορισμένες μεθοδολογικές επεκτάσεις περιλαμβάνουν την επαναληπτική SIS (ISIS) και διαδικασίες πολλών σταδίων, όπως η SIS-SCAD και SIS-LASSO, μπορούν να εφαρμοστούν για να εκτελέσουν την τελική επιλογή μεταβλητών και εκτίμηση παραμέτρων ταυτόχρονα.

Εκτεταμένες προσομοιώσεις σε αυτές τις διαδικασίες παρουσιάστηκαν επίσης στους Fan, Samworth & Wu (2009). Για να αποφευχθεί η επανάληψη επικεντρωνόμαστε στη vanilla SIS που στοχεύει στην αξιολόγηση της ιδιότητας του σίγουρου κρησαρίσματος και για να διαπιστώσει την επιρροή κάποιων παραγόντων στο ποσοστό εσφαλμένης επιλογής. Το μέγεθος του δείγματος ποικίλλει από 80 ως 600 για διαφορετικά σενάρια για να μετρήσουμε τις δυσκολίες των μοντέλων προσομοίωσης. Οι ακόλουθες τρεις διαμορφώσεις με $p = 2000,\ 5000$ και 40000 επεξηγηματικές μεταβλητές θεωρούνται για τη δημιουργία των συμμεταβλητών $\mathrm{X} = (\mathrm{X}_1, ..., \mathrm{X}_p)^T$:

S1. Οι συμμεταβλητές παράγονται σύμφωνα με

$$X_j = \frac{\varepsilon_j + \alpha_j \varepsilon}{\sqrt{1+\alpha_j^2}} \tag{1}$$

όπου ε και $\{\varepsilon_j\}_{j=1}^{[p/3]}$ είναι i.i.d. τυπικές κανονικές τυχαίες μεταβλητές, $\{\varepsilon_j\}_{j=[p/3]+1}^{[2p/3]}$ είναι i.i.d. και ακολουθούν μια διπλή εκθετική κατανομή με θέση παραμέτρου Ο και παράμετρο κλίμακας 1 και $\{\varepsilon_j\}_{j=[2p/3]+1}^{p}$ είναι i.i.d. και ακολουθούν μια μεικτή κανονική κατανομή με δύο «συστατικά» $N(-1,1)$, $N(1,0.5)$ και ίση αναλογία μίγματος. Οι συμμεταβλητές είναι κανονικοποιημένες να έχουν μέση τιμή 0 και διακύμανση 1. Οι σταθερές $\{a\}_{j=1}^{q}$ είναι οι ίδιες και έχουν επιλεγεί ώστε η συσχέτιση $\rho = corr(\mathrm{X}_i, \mathrm{X}_j) = 0, 0.2, 0.4, 0.6,$ και 0.8 μεταξύ των πρώτων q μεταβλητών και $\alpha_j = 0$ για $j > q$. Η παράμετρος q σχετίζεται επίσης με τη συνολική συσχέτιση στη μήτρα συνδιακύμασης. Θα παρουσιάσουμε τα αριθμητικά αποτελέσματα για $q = 15$ για αυτή τη ρύθμιση.

S2. Οι συμμεταβλητές παράγονται επίσης από την (1), εκτός αν οι $\{a_j\}_{j=1}^{q}$ είναι i.i.d. κανονικές τυχαίες μεταβλητές με μέση τιμή α και διασπορά 1 και $a_j = 0$ για $j > q$. Η μέση τιμή του α λαμβάνεται τέτοια ώστε $Ecorr(\mathrm{X}_i, \mathrm{X}_j) = 0, 0.2,$ $0.4, 0.6,$ και 0.8 μεταξύ των πρώτων q μεταβλητών. Τα αποτελέσματα προσομοίωσης που θα χρησιμοποιηθούν για αυτή τη ρύθμιση χρησιμοποιούν $q = 50$.

S3. Έστω $\{\mathrm{X}_j\}_{j=1}^{p-50}$ οι i.i.d. τυπικές κανονικές μεταβλητές και $X_k = \sum_{j=1}^{s} X_j (-1)^{j+1} / 5 + \sqrt{25-s} / 5\varepsilon_k$, $k = p-49, ..., p$ όπου $\{\varepsilon_k\}_{k=p-49}^{p}$ είναι τυπική κανονική κατανομή.

		ρ				
(p, n)	Setting	0	0.2	0.4	0.6	0.8
(40,000, 80)	S1 ($q = 15$)	549.9 (1.4)	550.1 (1.4)	550.1 (1.3)	550.1 (1.3)	550.1 (1.4)
(40,000, 80)	S2 ($q = 50$)	550.0 (1.4)	550.1 (1.4)	550.4 (1.5)	552.9 (1.8)	558.5 (2.4)
(40,000, 300)	S1 ($q = 15$)	157.3 (0.4)	157.4 (0.4)	157.4 (0.4)	157.4 (0.3)	157.7 (0.4)
(40,000, 300)	S2 ($q = 50$)	157.4 (0.4)	157.5 (0.4)	160.9 (1.2)	168.2 (1.0)	176.9 (1.0)
(5000, 300)	S1 ($q = 15$)	25.68 (0.2)	25.68 (0.2)	26.18 (0.2)	27.99 (0.4)	30.28 (0.4)
(5000, 300)	S1 ($q = 50$)	25.69 (0.1)	29.06 (0.5)	37.98 (0.7)	47.49 (0.7)	57.17 (0.5)
(2000, 600)	S1 ($q = 15$)	7.92 (0.07)	8.32 (0.15)	10.5 (0.3)	13.09 (0.3)	15.79 (0.2)
(2000, 600)	S1 ($q = 50$)	7.93 (0.07)	14.62 (0.40)	23.95 (0.7)	33.90 (0.6)	43.56 (0.5)
(2000, 600)	S2 ($q = 50$)	7.93 (0.07)	14.62 (0.40)	23.95 (0.7)	33.90 (0.6)	43.56 (0.5)

<u>Πίνακας 1</u>. Η διάμεσος των 200 εμπειρικών μέγιστων ιδιοτιμών με την ισχυρή εκτίμηση της SD σε παρένθεση.

Ο Πίνακας 1 συνοψίζει τη διάμεσο των εμπειρικών μέγιστων ιδιοτιμών του πίνακα συνδιασποράς και την ανθεκτική εκτίμηση της τυπικής απόκλισης (*RSD*) που βασίζεται σε 200 προσομοιώσεις στις δύο πρώτες ρυθμίσεις (S1 και S2) με μερικούς συνδυασμούς του μεγέθους του δείγματος $n = 80, 300, 600$, και $p = 2000$, 5000, 4000 και $q = 15, 50$. RSD είναι το ενδοτεταρτημοριακό εύρος (*IQR: interquartile range*) διαιρούμενο με 1.34. Οι εμπειρικές μέγιστες ιδιοτιμές είναι πάντα μεγαλύτερες από την έκδοση του πληθυσμού τους, ανάλογα με τα επιτεύγματα της μήτρας σχεδιασμού. Η εμπειρική μικρότερη ιδιοτιμή είναι πάντα μηδέν και οι αριθμοί εμπειρικών συνθηκών για τη μήτρα συνδιακύμασης του δείγματος είναι άπειροι, δεδομένου ότι $p > n$. Σε γενικές γραμμές, οι εμπειρικές μέγιστες ιδιοτιμές αυξάνονται καθώς οι παράμετροι συσχετισμού *p, q,* και οι αριθμοί των συμμεταβλητών p αυξάνονται και τα μεγέθη *n* του δείγματος μειώνονται.

Με αυτές τις τρεις προσομοιώσεις, έχουμε ως στόχο να τονίσουμε τις συμπεριφορές των δύο SIS διαδικασιών κάτω από διαφορετικές δομές συσχέτισης. Για κάθε προσομοίωση και για κάθε μοντέλο, εφαρμόζουμε τις δύο SIS διαδικασίες, την οριακή MLE και τις μεθόδους αναλογίας οριακής πιθανοφάνειας, για να κρησάρουμε τις μεταβλητές. Το ελάχιστο μέγεθος μοντέλου (MMS) που απαιτείται για κάθε μέθοδο για να έχει ένα σίγουρο κρησάρισμα, είναι ότι για να περιέχει αληθινό μοντέλο M_* είναι ενδοτεταρτημοριακό που χρησιμοποιείται ως ένα μέτρο αποτελεσματικότητας της μεθόδου κρησαρίσματος. Αυτό αποτρέπει τα θέματα επιλογής της παραμέτρου κατωφλιού. Για να μετρήσουμε τη δυσκολία του προβλήματος, περιλαμβάνουμε επίσης τη LASSO και τη SCAD ως σημεία αναφοράς για τη σύγκριση όταν $p = 2000$ και 5000. Το μικρότερο *p* χρησιμοποιείται λόγω της υπολογιστικής επιβάρυνσης της LASSO και της SCAD. Επιπλέον, όπως φαίνεται στα αποτελέσματα της προσομοίωσής μας, δεν αποδίδουν καλά όταν το *p* είναι μεγάλο. Αρχική μας πρόθεση είναι να αποδείξουμε ότι η απλή SIS δεν αποδίδει πολύ χειρότερα από τις πιο περίπλοκες διαδικασίες όπως η LASSO και η SCAD. Προς έκπληξή μας, η SIS μπορεί ακόμη και να ξεπεράσει αυτές τις πιο περίπλοκες μεθόδους από την πλευρά του κρησαρίσματος μεταβλητών. Και πάλι καταγράφουμε τις MMS για τη LASSO και τη SCAD για κάθε προσομοίωση και κάθε μοντέλο, οι

οποίες δεν εξαρτώνται από την επιλογή των παραμέτρων κανονικοποίησης. Όταν η LASSO ή η SCAD δεν μπορεί να επαναφέρει το πραγματικό μοντέλο ακόμα και για τη μικρότερη παράμετρο κανονικοποίησης, παίρνουμε κατά μέσο όρο το μέγεθος του μοντέλου με τη μικρότερη παράμετρο κανονικοποίησης και *p*. Αυτά τα παρεμβαλλόμενα MMS παρουσιάζονται με πλάγια γράμματα στους παρακάτω πίνακες για να τα διακρίνουμε από τα πραγματικά MMS. Τα αποτελέσματα για τις λογιστικές παλινδρομήσεις και τις γραμμικές παλινδρομήσεις παρουσιάζονται στις ακόλουθες υποενότητες.

Λογιστικές παλινδρομήσεις.

Τα δεδομένα που δημιουργούνται $(\mathrm{X}_1^T, \mathrm{Y}_1), \ldots, (\mathrm{X}_n^T, \mathrm{Y}_n)$ είναι i.i.d. αντίγραφα του ζεύγους $(\mathrm{X}^T, \mathrm{Y})$, στα οποία η δεσμευμένη κατανομή της απόκρισης *Y*, δεδομένου $X = x$, είναι διωνυμική κατανομή με πιθανότητα επιτυχίας $p(\mathrm{x}) = \exp(\mathrm{x}^T \beta^*) / [1 + \exp(\mathrm{x}^T \beta^*)]$. Παίρνουμε το μέγεθος του μη αραιού συνόλου συντελεστών $s = 3, 6, 12, 15,$ και 24. Για κάθε προσομοίωση, αξιολογούμε κάθε μέθοδο συνοψίζοντας τη διάμεσο ελάχιστου μεγέθους μοντέλου (MMMS) των επιλεγμένων μοντέλων καθώς και των συνδεδεμένων RSD. Τα αποτελέσματα, που βασίζονται στις 200 προσομοιώσεις για κάθε σενάριο καταγράφονται στο δεύτερο και τρίτο πάνελ. Πιο συγκεκριμένα, ο Πίνακας 1 καταγράφει τα MMMS και τη σχετική RSD για τη SIS σύμφωνα με τις δύο πρώτες ρυθμίσεις όταν $p = 40000$, ενώ οι Πίνακες 2-4 καταγράφουν αυτά τα αποτελέσματα για τη SIS, τη LASSO και τη SCAD όταν $p = 2000$ και 5000 κάτω από τις ρυθμίσεις 1, 2 και 3 αντίστοιχα. Οι πραγματικές παράμετροι καταγράφονται επίσης σε κάθε αντίστοιχο πίνακα.

Για να αποδειχθεί η δυσκολία των μοντέλων προσομοίωσή μας, απεικονίζουμε την κατανομή, μεταξύ των 200 προσομοιώσεων, των ελάχιστων |t|-στατιστικών των s εκτιμημένων συντελεστών παλινδρόμησης στο μοντέλο *Oracle*, στο οποίο ο στατιστικολόγος δε γνωρίζει ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές. Αυτό δείχνει τη δυσκολία στην ανάκτηση όλων των σημαντικών μεταβλητών ακόμα και στο μοντέλο *Oracle* με το ελάχιστο μέγεθος μοντέλου *s*. Στην πραγματικότητα, οι κατανομές κάτω από τις Ρυθμίσεις 1 είναι παρόμοιες με αυτές

κάτω από τις Ρυθμίσεις 2 όταν λαμβάνεται η ίδια τιμή q. Μπορεί να δειχθεί ότι το μέγεθος των ελάχιστων |t|-στατιστικών είναι αρκετά μικρό και γίνεται μικρότερο καθώς η συσχέτιση στις συμμεταβλητές (μετρούμενη με p και q) αυξάνεται, και μερικές φορές επιτυγχάνει τα τρία δεκαδικά ψηφία. Δεδομένων τόσο μικρών λόγων σημάτων προς θόρυβο στα μοντέλα *Oracle*, η δυσκολία των μοντέλων προσομοίωσής μας είναι προφανής ακόμα και αν τα σήματα δεν φαίνονται τόσο μικρά.

ρ	n	SIS-MLR	SIS-MMLE	n	SIS-MLR	SIS-MMLE
Setting 1, $q = 15$						
		$s = 3, \boldsymbol{\beta}^* = (1, 1.3, 1)^T$			$s = 6, \boldsymbol{\beta}^* = (1, 1, 3, 1, \ldots)^T$	
0	300	87.5 (381)	89 (375)	300	47 (164)	50 (170)
0.2	200	3 (0)	3 (0)	300	6 (0)	6 (0)
0.4	200	3 (0)	3 (0)	300	7 (1)	7 (1)
0.6	200	3 (1)	3 (1)	300	8 (1)	8 (2)
0.8	200	4 (1)	4 (1)	300	9 (3)	9 (3)
		$s = 12, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$			$s = 15, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$	
0	500	297 (589)	302.5 (597)	600	350 (607)	359.5 (612)
0.2	300	13 (1)	13 (1)	300	15 (0)	15 (0)
0.4	300	14 (1)	14 (1)	300	15 (0)	15 (0)
0.6	300	14 (1)	14 (1)	300	15 (0)	15 (0)
0.8	300	14 (1)	14 (1)	300	15 (0)	15 (0)
Setting 2, $q = 50$						
		$s = 3, \boldsymbol{\beta}^* = (1, 1.3, 1)^T$			$s = 6, \boldsymbol{\beta}^* = (1, 1.3, 1, \ldots)^T$	
0	300	84.5 (376)	88.5 (383)	500	6 (1)	6 (1)
0.2	300	3 (0)	3 (0)	500	6 (0)	6 (0)
0.4	300	3 (0)	3 (0)	500	6 (1)	6 (1)
0.6	300	3 (1)	3 (1)	500	8.5 (4)	9 (5)
0.8	300	5 (4)	5 (4)	500	13.5 (8)	14 (8)
		$s = 12, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$			$s = 15, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$	
0	600	77 (114)	78.5 (118)	800	46 (82)	47 (83)
0.2	500	18 (7)	18 (7)	500	26 (6)	26 (6)
0.4	500	25 (8)	25 (10)	500	34 (7)	33 (8)
0.6	500	32 (9)	31 (8)	500	39 (7)	38 (7)
0.8	500	36 (8)	35 (9)	500	40 (6)	42 (7)

<u>Πίνακας 1.</u> Οι MMMS και RSD (σε παρένθεση) των προσομοιωμένων παραδειγμάτων για τις λογιστικές παλινδρομήσεις για τις δύο πρώτες ρυθμίσεις όταν $p = 40000$.

Οι MMMS και RSD με σταθερή συσχέτιση (S1) και τυχαία συσχέτιση (S2) είναι συγκρίσιμες με την ίδια q. Καθώς η συσχέτιση αυξάνει και το μη αραιό μέγεθος αυξάνει, τα MMMS και το σχετικό RSD συνήθως αυξάνει για όλες τις SIS, LASSO και SCAD. Ανάμεσα σε όλα τα σχεδιασμένα σενάρια των Ρυθμίσεων 1 και 2, η SIS εκτελείται καλά, ενώ οι LASSO και SCAD περιστασιακά αποτυγχάνει κάτω από υψηλές συσχετίσεις και σχετικά μεγάλο μη αραιό μέγεθος ($s = 12, 15$ και 24). Η

συσχέτιση στις συμμεταβλητές μπορεί μερικές φορές να βοηθήσει τη SIS να μειώσει το εσφαλμένο ποσοστό επιλογής, καθώς μπορεί να αυξήσει τα οριακά σήματα. Αξιοσημείωτο είναι ότι οι LASSO και SCAD συνήθως δεν μπορούν να επιλέξουν τις σημαντικές μεταβλητές στην τρίτη ρύθμιση, λόγω της παραβίασης της συνθήκης που δεν μπορεί να παρουσιαστεί για $s = 6,12$ και 24, ενώ η SIS αποδίδει ικανοποιητικά.

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
$s = 3, \boldsymbol{\beta}^* = (1, 1.3, 1)^T$					
0	300	3 (0)	3 (0)	3 (1)	3 (1)
0.2	300	3 (0)	3 (0)	3 (0)	3 (0)
0.4	300	3 (0)	3 (0)	3 (0)	3 (0)
0.6	300	3 (0)	3 (0)	3 (0)	3 (1)
0.8	300	3 (1)	3 (1)	4 (1)	4 (1)
$s = 6, \boldsymbol{\beta}^* = (1, 1.3, 1, 1.3, 1, 1.3)^T$					
0	200	8 (6)	9 (7)	7 (1)	7 (1)
0.2	200	18 (38)	20 (39)	9 (4)	9 (2)
0.4	200	51 (77)	64.5 (76)	20 (10)	16.5 (6)
0.6	300	77.5 (139)	77.5 (132)	20 (13)	19 (9)
0.8	400	306.5 (347)	313 (336)	86 (40)	70.5 (35)
$s = 12, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$					
0	300	297.5 (359)	300 (361)	72.5 (3704)	12 (0)
0.2	300	13 (1)	13 (1)	12 (1)	12 (0)
0.4	300	14 (1)	14 (1)	14 (1861)	13 (1865)
0.6	300	14 (1)	14 (1)	*2552 (85)*	12 (3721)
0.8	300	14 (1)	14 (1)	*2556 (10)*	12 (3722)
$s = 15, \boldsymbol{\beta}^* = (3, 4, \ldots)^T$					
0	300	479 (622)	482 (615)	69.5 (68)	15 (0)
0.2	300	15 (0)	15 (0)	16 (13)	15 (0)
0.4	300	15 (0)	15 (0)	38 (3719)	15 (3720)
0.6	300	15 (0)	15 (0)	*2555 (87)*	15 (1472)
0.8	300	15 (0)	15 (0)	*2552 (8)*	15 (1322)

<u>Πίνακας 2.</u> Οι MMMS και RSD (σε παρένθεση) των προσομοιωμένων παραδειγμάτων για τις λογιστικές παλινδρομήσεις στην πρώτη ρύθμιση όταν $p = 5000$ και $q = 15$.

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
$s=3, \boldsymbol{\beta}^\star=(3,4,3)^T$					
0	200	3 (0)	3 (0)	3 (0)	3 (0)
0.2	200	3 (0)	3 (0)	3 (0)	3 (0)
0.4	200	3 (0)	3 (0)	3 (0)	3 (1)
0.6	200	3 (1)	3 (1)	3 (1)	3 (1)
0.8	200	5 (5)	5.5 (5)	6 (4)	6 (4)
$s=6, \boldsymbol{\beta}^\star=(3,-3,3,-3,3,-3)^T$					
0	200	8 (6)	9 (7)	7 (1)	7 (1)
0.2	200	18 (38)	20 (39)	9 (4)	9 (2)
0.4	200	51 (77)	64.5 (76)	20 (10)	16.5 (6)
0.6	300	77.5 (139)	77.5 (132)	20 (13)	19 (9)
0.8	400	306.5 (347)	313 (336)	86 (40)	70.5 (35)
$s=12, \boldsymbol{\beta}^\star=(3,4,\ldots)^T$					
0	600	13 (6)	13 (7)	12 (0)	12 (0)
0.2	600	19 (6)	19 (6)	13 (1)	13 (2)
0.4	600	32 (10)	30 (10)	18 (3)	17 (4)
0.6	600	38 (9)	38 (10)	22 (3)	22 (4)
0.8	600	38 (7)	39 (8)	*1071 (6)*	*1042 (34)*
$s=24, \boldsymbol{\beta}^\star=(3,4,\ldots)^T$					
0	600	180 (240)	182 (238)	35 (9)	31 (10)
0.2	600	45 (4)	45 (4)	35 (27)	32 (24)
0.4	600	46 (3)	47 (2)	*1099 (17)*	1093 (1456)
0.6	600	48 (2)	48 (2)	*1078 (5)*	*1065 (23)*
0.8	600	48 (1)	48 (1)	*1072 (4)*	*1067 (13)*

<u>Πίνακας 3.</u> Οι MMMS και RSD (σε παρένθεση) των προσομοιωμένων παραδειγμάτων για τις λογιστικές παλινδρομήσεις στη ρύθμιση 2 όταν $p=2000$ και $q=50$.

s	M-$\lambda_{\max}$(RSD)	SIS-MLR	SIS-MMLE	LASSO	SCAD
3	8.47 (0.17)	3 (0)	3 (0)	3 (1)	3 (0)
6	10.36 (0.26)	56 (0)	56 (0)	*1227 (7)*	*1142 (64)*
12	14.69 (0.39)	63 (6)	63 (6)	*1148 (8)*	*1093 (59)*
24	23.70 (0.14)	214.5 (93)	208.5 (82)	*1120 (5)*	*1087 (24)*

<u>Πίνακας 4.</u> Οι MMMS RSD (σε παρένθεση) για τα προσομοιωμένα παραδείγματα των λογιστικών παλινδρομήσεων στη ρύθμιση 3 όταν $p=2000$ και $n=600$.

Γραμμικά μοντέλα

Τα δεδομένα που δημιουργούνται $(X_1^T, Y_1), \ldots, (X_n^T, Y_n)$ είναι i.i.d. αντίγραφα του ζεύγους (X^T, Y), στο οποίο η απόκριση Y ακολουθεί ένα γραμμικό μοντέλο με $Y = X^T \beta^* + \varepsilon$, όπου το σφάλμα ε είναι τυπικά κανονικά κατανεμημένο. Οι συμμεταβλητές παράγονται κατά τον ίδιο τρόπο όπως οι λογιστικές ρυθμίσεις παλινδρόμησης. Λαμβάνουμε τους ίδιους πραγματικούς συντελεστές και δομές συσχέτισης για μέρος των σεναρίων ($p = 40000$) όπως τα παραδείγματα λογιστικής παλινδρόμησης ενώ διαφέρουν οι πραγματικοί συντελεστές για τα άλλα σενάρια, έτσι ώστε να μετρηθεί η δυσκολία του προβλήματος. Το μέγεθος του δείγματος για κάθε σενάριο είναι αντίστοιχα μειωμένο για να απεικονίζει το γεγονός ότι τα γραμμικά μοντέλα είναι πιο κατατοπιστικά. Τα αποτελέσματα καταγράφονται στους παρακάτω πίνακες. Η τάση των MMMS και της σχετικής RSD της SIS, της LASSO και SCAD να ποικίλλουν με τη συσχέτιση και το μη-αραιό μέγεθος να είναι παρόμοιο με τα παραδείγματα λογιστικής παλινδρόμησης, αλλά τα μεγέθη τους είναι συνήθως μικρότερα στα παραδείγματα γραμμικής παλινδρόμησης, καθώς το μοντέλο είναι πιο κατατοπιστικό. Συνολικά, η SIS κάνει μια πολύ λογική δουλειά στο κρησάρισμα άσχετων μεταβλητών και μερικές φορές ξεπερνά τις LASSO και SCAD.

ρ	n	SIS-MLR	SIS-MMLE	n	SIS-MLR	SIS-MMLE
Setting 1, $q = 15$						
		$s = 3, \boldsymbol{\beta}^* = (1, 1.3, 1)^T$			$s = 6, \boldsymbol{\beta}^* = (1, 1, 3, 1, \ldots)^T$	
0	80	12 (18)	12 (18)	150	42 (157)	42 (157)
0.2	80	3 (0)	3 (0)	150	6 (0)	6 (0)
0.4	80	3 (0)	3 (0)	150	6.5 (1)	6.5 (1)
0.6	80	3 (0)	3 (0)	150	6 (1)	6 (1)
0.8	80	3 (0)	3 (0)	150	7 (1)	7 (1)
		$s = 12, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$			$s = 15, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$	
0	300	143 (282)	143 (282)	400	135.5 (167)	135.5 (167)
0.2	200	13 (1)	13 (1)	200	15 (0)	15 (0)
0.4	200	13 (1)	13 (1)	200	15 (0)	15 (0)
0.6	200	13 (1)	13 (1)	200	15 (0)	15 (0)
0.8	200	13 (1)	13 (1)	200	15 (0)	15 (0)
Setting 2, $q = 50$						
		$s = 3, \boldsymbol{\beta}^* = (1, 1.3, 1)^T$			$s = 6, \boldsymbol{\beta}^* = (1, 1.3, 1, \ldots)^T$	
0	100	3 (2)	3 (2)	200	7.5 (7)	7.5 (7)
0.2	100	3 (0)	3 (0)	200	6 (1)	6 (1)
0.4	100	3 (0)	3 (0)	200	7 (1)	7 (1)
0.6	100	3 (0)	3 (0)	200	7 (2)	7 (2)
0.8	100	3 (1)	3 (1)	200	8 (4)	8 (4)
		$s = 12, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$			$s = 15, \boldsymbol{\beta}^* = (1, 1.3, \ldots)^T$	
0	400	22 (27)	22 (27)	500	35 (52)	35 (52)
0.2	300	16 (5)	16 (5)	300	24 (7)	24 (7)
0.4	300	19 (8)	19 (8)	300	30 (10)	30 (10)
0.6	300	25 (8)	25 (8)	300	33.5 (7)	33.5 (7)
0.8	300	24 (7)	24 (7)	300	35 (8)	35 (8)

Οι MMMS και RSD (σε παρένθεση) των προσομοιωμένων παραδειγμάτων για τις δύο πρώτες ρυθμίσεις για τις γραμμικές παλινδρομήσεις όταν $p = 40000$**.**

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
$s = 3, \boldsymbol{\beta}^* = (0.5, 0.67, 0.5)^T$					
0	100	12 (40)	12 (40)	3 (1)	3 (1)
0.2	100	3 (1)	3 (1)	3 (0)	3 (0)
0.4	100	3 (0)	3 (0)	3 (0)	3 (0)
0.6	100	3 (1)	3 (1)	5 (7)	5 (5)
0.8	100	4 (2)	4 (2)	4 (1)	4 (1)
$s = 6, \boldsymbol{\beta}^* = (0.5, 0.67, 0.5, 0.67, 0.5, 0.67)^T$					
0	100	210.5 (422)	210.5 (422)	33.5 (651)	25 (22)
0.2	100	7 (2)	7 (2)	6 (1)	6 (1)
0.4	100	7 (2)	7 (2)	6 (1)	6 (1)
0.6	100	8 (2)	8 (2)	7 (1)	7 (1)
0.8	100	9 (3)	9 (3)	7 (2)	8 (1)
$s = 12, \boldsymbol{\beta}^* = (0.5, 0.67, \ldots)^T$					
0	300	49 (76)	49 (76)	12 (1)	12 (0)
0.2	100	14 (2)	14 (2)	12 (1)	12 (1)
0.4	100	14 (1)	14 (1)	12 (1)	12 (1)
0.6	100	14 (1)	14 (1)	13 (1)	13 (1)
0.8	100	14 (1)	14 (1)	13 (1)	13 (1)
$s = 15, \boldsymbol{\beta}^* = (0.5, 0.67, \ldots)^T$					
0	300	199 (251)	199 (251)	17 (2)	15 (0)
0.2	100	17 (5)	17 (5)	15 (1)	15 (0)
0.4	100	15 (0)	15 (0)	15 (0)	15 (0)
0.6	100	15 (0)	15 (0)	15 (0)	15 (0)
0.8	100	15 (0)	15 (0)	15 (0)	15 (1)

Οι MMMS και RSD (σε παρένθεση) των προσομοιωμένων παραδειγμάτων για τις γραμμικές παλινδρομήσεις στη ρύθμιση 1 όταν $p = 5000$ και $q = 15$.

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
$s = 3, \boldsymbol{\beta}^* = (0.6, 0.8, 0.6)^T$					
0	100	5 (14)	6 (16)	4 (4)	4 (2)
0.2	100	3 (1)	3 (1)	3 (0)	3 (0)
0.4	100	3 (1)	4 (1)	3 (1)	3 (1)
0.6	100	5 (3)	7 (5)	4 (1)	4 (1)
0.8	100	7 (7)	14 (12)	5 (57)	7 (4)
$s = 6, \boldsymbol{\beta}^* = (3, -3, 3, -3, 3, -3)^T$					
0	100	15 (43)	18 (47)	6 (0)	6 (1)
0.2	100	42 (116)	47 (99)	7 (1)	7 (1)
0.4	100	143 (207)	129 (226)	12 (4)	12 (5)
0.6	200	47 (93)	49 (110)	7 (1)	7 (1)
0.8	200	360 (470)	376.5 (486)	54 (32)	51 (25)
$s = 12, \boldsymbol{\beta}^* = (0.6, 0.8, \ldots)^T$					
0	200	151 (212)	140 (207)	15 (4)	15 (4)
0.2	100	37.5 (10)	36 (12)	16 (3)	16 (4)
0.4	100	39 (7)	40.5 (8)	18 (3)	17 (2)
0.6	100	41 (7)	42 (6)	19 (3)	18 (3)
0.8	100	44 (5)	46 (6)	23 (1478)	24 (50)
$s = 24, \boldsymbol{\beta}^* = (3, 4, \ldots)^T$					
0	400	229 (283)	227 (279)	24 (0)	25 (0)
0.2	100	61 (43)	67 (46)	30 (2)	30 (2)
0.4	100	48 (2)	47 (2)	31 (2)	30 (1)
0.6	100	48 (2)	49 (2)	32 (2)	32 (3)
0.8	100	49 (2)	49 (1)	32 (2)	32 (2)

Οι MMMS και RSD (σε παρένθεση) των προσομοιωμένων παραδειγμάτων για τις γραμμικές παλινδρομήσεις στη ρύθμιση 2 όταν $p = 2000$ **και** $q = 50$

s	M-$\lambda_{\max}$(RSD)	SIS-MLR	SIS-MMLE	LASSO	SCAD
3	8.47 (0.17)	3 (0)	3 (0)	3 (0)	3 (0)
6	10.36 (0.26)	56 (0)	56 (0)	47 (4)	45 (3)
12	14.69 (0.39)	62 (0)	62 (0)	*1610 (10)*	*1304 (2)*
24	23.70 (0.14)	81 (19)	81 (23)	*1637 (14)*	*1303 (1)*

Οι MMMS RSD (σε παρένθεση) των προσομοιωμένων παραδειγμάτων για τις γραμμικές παλινδρομήσεις στη ρύθμιση 3 όταν $p = 2000$ **και** $n = 600$**.**

ΒΙΒΛΙΟΓΡΑΦΙΑ

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Statist. Math*., **21**, pp. 243-247.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, pp. 716-723.

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, **22**, pp. 162-176.

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables. *Technical Report No. 23*. Department of Statistics. University of Kentuky.

Antoniadis, A. and Fan, J. (2001) Regularization of wavelets approximation (with discussion). *J. Am. Statist. Ass*, **96**, 939-967.

Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion). *J. Italian Statist. Assoc*., **6**, pp. 97-144.

Barron, A., Cohen, A., Dahmen, W. and DeVore, R. (2008) Approximation and learning by greedy algorithms. *Ann. Statist*., **36**, 64-94.

Barron, D., Wakin, M. B., Duarte, M. F., Sarvotham, S. and Baraniuk, R. G. (2005) Distributed compressed sensing. *Manuscript*.

Bickel, P. J. (1975), One-Step Huber Estimates in Linear Models, *Journal of the American Statistical Association*, **70**, pp. 428–433.

Bickel, P. J. and Doksum K. A. (1981) An analysis of transformations revisited. *J. Amer. Statist. Assoc.*, **76**, 296-311.

Bickel, P. J. and Doksum K. A. (2001) *Mathematical statistics: Basic Ideas and Selected Topics*, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2008) Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist., **36**, in the press.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *J. R. Stat. Soc. Ser. B*, **26**, 211-246.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression. *Journal of the American Statistical Association*, **87**, pp. 738-754.

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote, *Technometrics*, **37**, pp.373–384.

Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection, *The Annals of Statistics*, **24**, pp. 2350–2383.

Bryant, P.G. and Cordero-Brana, O.I. (2000). Model selection using the minimum description length principle. *The Amer. Statist*., **54**, pp. 257-268.

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when *p* is much larger than *n* (with discussion). *Ann. Statist*., **35**, 2313-2404.

Cox, D. R. (1974) Regression models and life-tables (with discussion). *J. R. Stat Soc. Ser. B Stat. Methodol*, **34**, 187-220.

Donoho, D. L. (2000) High-dimensional data analysis: the curses and blessings of dimensionality. *American Mathematical Society Conf. Math Challenges of the 21st Century*.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaption by wavelet shrinkage. *Biometrika*, **81**, 425-455.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann Statist*., **32**, 407-499.

Eksioglu, B., Demirer, R., Capar, I., (2005). Subset selection in multiple linear regression: a new mathematical programming approach. *Computer and Industrial engineering*, **49**, 155-167

Fan, J. (1997) Comments on “Wavelets in statistics: a review,” by Antoniadis, *J. Ital. Statist. Ass.*, **6**, 131-138.

Fan, J. and Fan, Y. (2008) High dimensional classification using features annealed independence rules. *Ann. Statist.*, to be published.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Statist. Ass*., **96**, 1348-1360.

Fan, J. and Li, R. (2002) Variable selection for Cox’s proportional hazards model and frailty model. *Ann. Statist*., **30**, 74-99.

Fan, J. and Li, R. (2006) Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proc. Int. Congr. Mathematicians* (eds M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), vol.III, pp. 595-622. Freiburg: European Mathematical Society.

Fan, J. and Peng, H. (2004) Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928-961.

Fan, J. and Ren, Y. (2006) Statistical analysis of DNA microarray data. *Clinical Cancer Research*, **12**, 4469-4473.

Fan, J., Samworth, R., Wu, Y. (2009) Ultra-dimensional variable selection via independent learning: Beyond the linear model. *J. Mach. Learn. Res.*, **10**, 1829-1853.

Frank, I. E., and Friedman, J. H. (1993), A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, pp. 109–148.

Freund, Y. and Schapire, R. E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119-139.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Amer. Starist. Assoc.*, **76**, 817-823.

Fu, W. J. (1998). Penalized Regression: The Bridge Versus the LASSO. *Journal of Computational and Graphical Statistics*, **7**, pp. 397–416.

Greenshtein, E. (2006) Best subset selection, persistence in high dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.*, **34**, 2367-2386.

Gribonval, R., Mailhe, B., Rauhut, H., Schnass, K. and Vandergheynst, P. (2007) Average case analysis of multi-channel thresholding. In *Proc. Int. Conf. Acoustic and Speech Signal Processing*. New York: Institute of Electrical and Electronics Engineers.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339-373.

Hall, P. and Miller, H. (2009) Using generalized correlation to effect variable selection in very high dimensional problems, *J. Comput. Graph. Statist.*, **18**, 533

Hall, P., Titterington, D. M. and Xue, J.-H. (2009) Tilting methods for assessing the influence of components in a classifier, *J. R. Stat. Soc. Ser. B Stat. Methodol.* , **71**, 783-803.

Hannan, E.J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc*., **41**, pp. 190-195.

Hardin, J.W. and Hilbe, J.M. (2007) *Generalized Linear Models and Extensions*, 2nd Ed., College Station, TX: Stata Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed., New York: Springer-Verlag.

Hoerl, A. E. and Kennard, R.W. (1970a). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*. **12**, pp. 55-67.

Hoerl, A. E. and Kennard, R.W. (1970b). Ridge regression: Applications to non-orthogonal problems. *Technometrics*. **12**, pp. 69-82.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics*, **4**, pp. 105-123.

Huber P. (1981). *Robust Estimation*. Wiley, New York.

Huang, J., Horowitz, J. and Ma, S. (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.*, **36**, 587-613.

Hunter, D. and Li, R. (2005) Variable selection using MM algorithms. *Ann. Statist*., **33**, 1617-1642

Hurvich, C. M. and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, pp. 297-307.

Khan, J., Wei, J., Ringér, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C. and Meltzer, P. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**, 673-679.

Konishi, S. and Kitagama, G. (2008). *Information Criteria and Statistical Modeling*, Springer, New York.

Kosorok, M. R. , Lee, B. L., and Fine, J. P. (2004) Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.*, **32**, 1448-1491

Koukouvinos, C., Mylona, K. and Vonta, F. (2008). A Comparative study of variable selection procedures applied in high dimensional medical problems. *Journal of Applied Probability & Statistics*, **3**, pp. 195-209.

Lee, Y., Lin, Y. and Wahba, G. (2004) Multicategory Support Vector Machines, Theory and Application to the Classification of Microarray Data and Satellite Radiance Data, *J. Amer. Statist. Assoc.*, **99 (465)**, 67-81

Linhart, H. and Zucchini, W. (1986). *Model selection*, John Wiley, New York.

Mallows, C. L. (1973). Some comments on C_p. Technometrics, **15**, pp. 661-675.

Mandel J. (1989). Some thoughts on variable selection in multiple regression. *Journal of Quality Technology*, **21**(1), 2-6

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H., and Patil, P. (1998). Exact Risk Analysis of Wavelet Regression. *Journal Computational and Graphical Statistics*, **7**, pp. 278–309.

Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in Practice. *The American Statistician*, **29**, pp. 3-20.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.

McDonald**,** G. C. and Galarneau**,** D. I., (**1975).** A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, **70**, pp. 407-416.

Meier, L., Van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *J. R. Statist. Soc*. B, **70**, 53-71.

Meinshausen, N. (2007) Relaxed Lasso. *Computnl Statist. Data Anal.*, **52**, 374-393.

Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436-1462.

Meinshausen, N., Rocha, G. and Yu, B. (2007) Discussion of "The Dantzig selector: statistical estimation when *p* is larger than *n*". *Ann. Statist.*, **35**, 2373-2384.

Narula, SC., Andre, CDS. , Elian, SN., Tavares, RA. (2003) An overview of the variable selection methods for the minimum sum of absolute error regression. *Statistics in Medicine*, **22**, 2101-2111.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K. König, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006) Customized Oligonucleotide Microarray Gene Expression Based Classification of Neuroblastoma Patients Outperforms Current Clinical Risk Stratification, *Journal of Clinical Oncology*, **24**, 5070-5078.

Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008) "Pre-conditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.*, to be published.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2007) Sparse additive models. *Manuscript*.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica*, **14**, pp. 465-471.

Robinson, P. M. (1988). The Stochastic Difference Between Econometrics

and Statistics. *Econometrica*, **56**, pp. 531–547.

Schwartz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, pp. 461-464.

Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natn. Acad. Sci. USA*, 100, 9940-9445.

Tibshirani, R. J. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, pp. 385-395.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, **18**, 104-117

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natn. Acad. Sci. USA*, **99**, 6567-6572.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-26.

Zeng, D. and Lin, D. Y. (2007) Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Stat. Soc. Ser B Stat. Methodol.*, **69**, 507-564.

Zhang, C.-H. (2007) Penalized linear unbiased selection. *Technical Report 2007-003*. Department of Statistics, Rutgers University, Piscataway.

Zhang, C.-H. and Huang, J. (2008) The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567-1594.

Zhao, P. and Yu, B. (2006) on model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541-2567.

Zou, H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418-1429.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. Ann. Statist., to be published.