ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Κρησάρισμα Μεταβλητών και Επιλογή Χαρακτηριστικών σε Δεδομένα Υψηλής Διάστασης

(Variables Screening and Feature Selection in High Dimensional Data)

Λοΐζου Μάριος

Επιβλέπων: Χρήστος Κουκουβίνος,
Καθηγητής ΕΜΠ

Αθήνα, 2016

Κρησάρισμα Μεταβλητών και Επιλογή Χαρακτηριστικών σε Δεδομένα Υψηλής Διάστασης

Κρησάρισμα Μεταβλητών και Επιλογή Χαρακτηριστικών σε Δεδομένα Υψηλής Διάστασης

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

APPLIED MATHEMATICAL AND PHYSICAL SCIENCE

DEPARTMENT OF MATHS

Dissertation

Variables Screening and Feature Selection in High Dimensional Data

Loizou Marios

Supervisor: Christos Koukouvinos,
Professor of NTUA

Athens, 2016

Κρησάρισμα Μεταβλητών και Επιλογή Χαρακτηριστικών σε Δεδομένα Υψηλής Διάστασης

Πρόλογος

Ζούμε σε μια εποχή όπου η τεχνολογία της πληροφορίας έχει αναπτυχθεί ραγδαία. Αυτό έχει ως αποτέλεσμα την συλλογή τεράστιων πολύπλοκων δεδομένων που μπορούν να φανούν πολύ χρήσιμα στην ανθρωπότητα. Για παράδειγμα η χρήση του διαδικτύου (*World Wide Web*) ως ένα παγκόσμιο πληροφοριακό σύστημα μας έχει πλημμυρήσει με τεράστιες ποσότητες δεδομένων και πληροφοριών. Τέτοια δεδομένα μπορεί να κρύβουν σημαντικές πληροφορίες αλλά η εξαγωγή τέτοιων πληροφοριών αποτελεί μεγάλη πρόκληση για τα άτομα που ασχολούνται με τέτοιου είδους προβλήματα, κυρίως τους στατιστικούς. Μέσω αυτών μπορούμε να προβλέψουμε καταστάσεις, όπως οικονομικούς κινδύνους, να συσχετίσουμε ασθένειες με συγκεκριμένα γονίδια, να κατανοήσουμε τις προτιμήσεις των καταναλωτών και άλλα πολλά. Για να επιτευχθούν όμως όλα αυτά χρειάζονται στατιστικές μέθοδοι η οποίες μπορούν να "δαμάσουν" δεδομένα υψηλής διάστασης. Ο όρος "υψηλή διάσταση" αναφέρεται στην κατάσταση όπου ο αριθμός των πιθανών άγνωστων παραμέτρων που επηρεάζουν το πρόβλημά μας είναι αρκετές τάξεις μεγαλύτερος από τον αριθμό του δείγματος. Για παράδειγμα η πρόβλεψη της κίνησης μιας μετοχής μπορεί να επηρεάζεται από τεράστιο αριθμό παραγόντων. Για να βρούμε ποιοι παράγοντες όντως επηρεάζουν την κίνηση μιας μετοχής πρέπει να γίνει κάποιο είδος στατιστικής ανάλυσης χρησιμοποιώντας ένα δείγμα από μετοχές. Προφανώς αν οι πιθανοί παράγοντες είναι πάρα πολλοί η συλλογή εξίσου μεγάλου δείγματος είναι αδύνατη. Τέτοιου είδους προβλήματα ονομάζονται προβλήματα επιλογής χαρακτηριστικών (*variable selection*). Η κλασσική στατιστική συμπερασματολογία αποτυγχάνει να λύσει προβλήματα με δεδομένα υψηλής διάστασης. Τα τελευταία χρόνια αναπτύχθηκαν, και αναπτύσσονται αρκετές μεθοδολογίες που επιτρέπουν την στατιστική συμπερασματολογία για υψηλής διάστασης δεδομένα βασιζόμενα στην παραδοχή κάποιων εννοιών όπως της σποραδικότητας (*sparsity*). Με τον όρο σποραδικότητα αναφερόμαστε στην παραδοχή ότι μόνο κάποιοι παράγοντες από το συνολικό αριθμό παραγόντων επιδρούν σε ένα μοντέλο, κάτι που εκ των πραγμάτων φαίνεται να λειτουργεί.

Περιεχόμενα

Πρόλογος **5**

Περιεχόμενα **7**

Περίληψη **11**

Abstract **13**

Ευχαριστίες **15**

ΚΕΦΑΛΑΙΟ 1: ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΠΟΙΝΙΚΟΠΟΙΗΣΗΣ **17**

1.1 Εισαγωγή 17

1.2 Το Γενικό Γραμμικό Μοντέλο 18

1.2.1 Μέθοδος Ελάχιστων Τετραγώνων (*Ordinary least square method*) 19

1.2.2 Μέθοδος Μέγιστης Πιθανοφάνειας (*Maximum Likelihood Method*) 20

1.3 Ποινικοποιημένα Ελάχιστα Τετράγωνα (*Penalized Least Squares*) 22

1.3.1 Εισαγωγή 22

1.3.2 Στάθμιση Μεροληψίας-Μεταβλητότητας (*Bias-Variance Tradeoff*) 25

1.3.3 Επιλογή μεταβλητών μέσω ποινικοποιημένων ελάχιστων τετραγώνων 25

1.3.4 Μέθοδος LASSO (*Least Absolute Shrinkage and Selection Operator*) 29

1.3.5 Μέθοδος προσαρμοσμένη (*Adaptive*) LASSO 31

1.3.6 SCAD (*Smoothly Clipped Absolute Deviation*) και MCP (*Minimax Concave Penalty*) 32

1.3.7 Επίδοση Ποινικοποιημένων Εκτιμητών 33

1.3.8 Επιλογή παραμέτρων για ποινικοποιημένες μεθόδους 34

ΚΕΦΑΛΑΙΟ 2: ΣΙΓΟΥΡΟ ΑΝΕΞΑΡΤΗΤΟ ΚΡΗΣΑΡΙΣΜΑ (*SURE INDEPENDECE SCREENING*) 35

2.1 Εισαγωγή -Δεδομένα υψηλής Διάστασης (*high dimensional data*) 35

2.1.1 Προβλήματα ανάλυσης δεδομένων υψηλής διάστασης 36

2.2 Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος (*Sure Independence Screening* (*SIS*)) 39

2.2.1 Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος 39

2.2.2 Iteratively Thresholded Ridge Regression Screener (ITRRS) 40

2.2.3 Ιδιότητα Σίγουρου Κρησαρίσματος της SIS 41

2.2.4 Τεχνικές επιλογές μοντέλου βασισμένες στη SIS 42

2.2.3 Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος για ταξινόμηση *(Sure Independent Screening (SIS) for classification)* 44

2.2.4 Προβλήματα της μεθόδου SIS 45

2.3 Επαναληπτική Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος (*Iterative Sure Independence Screening* (*ISIS*)) 46

ΚΕΦΑΛΑΙΟ 3: ΠΡΟΒΟΛΗ ΚΑΝΟΝΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ ΣΕ ΥΨΗΛΗΣ ΔΙΑΣΤΑΣΗΣ ΔΕΔΟΜΕΝΑ ΓΙΑ ΚΡΗΣΑΡΙΣΜΑ ΜΕΤΑΒΛΗΤΩΝ (*HIGH DIMENSIONAL ORDINARY LEAST SQUARES PROJECTION FOR SCREENING VARIABLES*) 49

3.1 Εισαγωγή 49

3.2 Προβολή Κανονικών Ελάχιστων σε Υψηλής Διάστασης Δεδομένα (*High Dimensional Ordinary Least Square Screening* (*HOLP*)) 50

3.2.1 Εισαγωγικά στοιχεία 50

3.2.2 Ορισμός Μεθόδου HOLP 52

3.2.3 Ιδιότητα Σίγουρου Κρησαρίσματος της HOLP 54

3.3 Τεχνικές επιλογές μοντέλου βασισμένες στη HOLP 56

3.4 Προβλήματα της μεθόδου HOLP 57

3.5 Προσομοιώσεις 58

ΚΕΦΑΛΑΙΟ 4: ΚΡΗΣΑΡΙΣΜΑ ΜΕΤΑΒΛΗΤΩΝ ΜΕΣΩ ΤΗΣ ΣΥΣΧΈΤΙΣΗ ΤΗΣ ΑΠΟΣΤΑΣΗΣ (*FEATURE SCREENING VIA DISTANCE CORRELATION*)... 61

4.1 Εισαγωγή 61

4.2 Θεωρητικό Υπόβαθρο-Συσχέτιση της Απόστασης (*Distance Correlation*) 62

4.3 Μέθοδος Συσχέτιση της Απόστασης βασισμένη στο Σίγουρο Ανεξάρτητο Κρησάρισμα (*Distance Correlation-based Sure Independence Screening* (*DC-SIS*)) . 64

4.3.1 Ορισμός Μεθόδου DC-SIS 64

4.3.2 Ιδιότητα Σίγουρου Κρησαρίσματος της DC-SIS 65
4.3.3 Προβλήματα της μεθόδου DC-SIS 66
4.4 Επαναληπτική Μέθοδος Συσχέτιση της Απόστασης βασισμένη στο Σίγουρο Ανεξάρτητο Κρησάρισμα (*Iterative Distance Correlation-based Sure Independence Screening* (*DC-ISIS*)). 67
4.4.1 Εύρεση σημαντικών αλληλεπιδράσεων με την βοήθεια της DC-ISIS..... 68

ΚΕΦΑΛΑΙΟ 5: ΕΦΑΡΜΟΓΕΣ ΜΕΘΟΔΩΝ ΚΑΙ ΣΥΓΚΡΙΣΕΙΣ 71
5.1 Εισαγωγή 71
5.2 Γραμμικά Μοντέλα 72
5.2.1 Ανεξάρτητες Επεξηγηματικές Μεταβλητές 72
5.2.2 Ψευδώς Αρνητικά (*False Negative*) 73
5.2.3 Ψευδώς Θετικά (*False Positive*) 74
5.2.4 Έντονη Συσχέτιση 75
5.3 Κατηγορική Μεταβλητή Απόκρισης 76
5.4 Μη γραμμικά Μοντέλα 77
5.5 Επίδραση Αλληλεπιδράσεων 78
5.6 Συμπεράσματα 79

ΒΙΒΛΙΟΓΡΑΦΙΑ (*REFERENCES*) 81

ΠΑΡΑΡΤΗΜΑΤΑ 83
ΠΑΡΑΡΤΗΜΑ 1: ΑΠΟΔΕΙΞΕΙΣ ΘΕΩΡΗΜΑΤΩΝ ΤΗΣ ΜΕΘΟΔΟΥ HOLP 83
ΠΑΡΑΡΤΗΜΑ 2: ΚΩΔΙΚΕΣ R 89
Μέθοδος SIS 89
Μέθοδος HOLP 90
Μέθοδος Divide HOLP 91
Μέθοδος DC-SIS 92
Μέθοδος DC-ISIS 92

Περίληψη

Στατιστικά προβλήματα υψηλής διάστασης προκύπτουν από διάφορα πεδία επιστημονικής έρευνας, και τεχνολογικής ανάπτυξης. Το κρησάρισμα μεταβλητών (*feature screening*) και η επιλογή χαρακτηριστικών (*feature selection*) παίζει σημαντικό ρόλο στην σύγχρονη στατιστική συμπερασματολογία και στις επιστημονικές ανακαλύψεις. Με τον όρο "κρησάρισμα μεταβλητών" αναφερόμαστε στην διαδικασία σκαρταρίσματος επεξηγηματικών μεταβλητών σε δεδομένα υψηλής διάστασης, έτσι ώστε να μειωθεί ο αριθμός τους αισθητά, χωρίς όμως να εκδιωχθούν από το μοντέλο οι πραγματικά σημαντικές μεταβλητές. Βλέποντας πιο προσεκτικά τον όρο αυτό βλέπουμε ότι ένα επιτυχές κρησάρισμα αποτελεί σπουδαίο επίτευγμα και διευκολύνει δραστικά την ανάλυση των δεδομένων αφού κλασσικές στατιστικές μέθοδοι που πριν αποτύγχαναν, τώρα μπορούν να χρησιμοποιηθούν. Έτσι η επιλογή χαρακτηριστικών γίνεται εφικτή, πιο εύκολη και πιο ακριβής. Την πρώτη μέθοδο κρησαρίσματος την πρότειναν οι Fan και Li (2008) και την ονόμασαν μέθοδος σίγουρου ανεξάρτητου κρησαρίσματος (*Sure Independence Screening* (*SIS*)) η οποία βασίζεται στη μάθηση συσχέτισης (*correlation learning*) . Στο ίδιο άρθρο πρότειναν μια επαναληπτική μορφή της SIS, την επαναληπτική μέθοδο σίγουρου ανεξάρτητου κρησαρίσματος (*Iterative Sure Independence Screening* (ISIS)) η οποία προτάθηκε για την επίλυση κάποιων προβλημάτων που αντιμετώπιζε η SIS. Οι Li, Zhong και Zhu (2012) πρότειναν μια άλλη μέθοδο κρησαρίσματος, την μέθοδο συσχέτισης της απόστασης βασισμένη στο σίγουρο ανεξάρτητο κρησάρισμα (*Distance Correlation-based Sure Independence Screening* (*DC-SIS*)), η οποία βασίζεται στην μάθηση συσχέτισης της απόστασης (*Distance Correlation*) που αποτελεί ένα νέος είδος συσχέτισης μεταβλητών που προτάθηκε από τους Szekely, Rizzo και Bakirov (2007). Λόγω του ότι η DC-SIS παρουσιάζει παρόμοια προβλήματα με την SIS, οι Zhong και Zhu (2015) πρότειναν μια επαναληπτική DC-SIS την *Iterative* DC-SIS που προσπαθεί να λύσει τέτοιου είδους προβλήματα. Οι Wang και Leng (2015) πρότεινα την μέθοδο προβολής κανονικών ελάχιστων τετραγώνων σε υψηλής διάστασης δεδομένα (*High Dimensional Ordinary Least Square Screening* (*HOLP*)) η οποία δεν ασχολείται με τη συσχέτιση μεταβλητών όπως οι προηγούμενες μέθοδοι. Στην παρούσα διπλωματική εργασία ασχολούμαστε κυρίως με μεθόδους κρησαρίσματος καθώς και με την επιλογή μεταβλητών χρησιμοποιώντας ποινικοποιημένες μεθόδους.

Πιο συγκεκριμένα, στο πρώτο κεφάλαιο γίνεται αναφορά στο γενικό γραμμικό μοντέλο και στις βασικές τεχνικές εκτίμησης παραμέτρων καθώς επίσης και στις μεθόδους επιλογής μεταβλητών και εκτίμησης παραμέτρων που βασίζονται στη εισαγωγή μιας συνάρτησης ποινής στην μέθοδο ελάχιστων τετραγώνων. Το κεφάλαιο 2 εισαγάγει τις έννοιες δεδομένων υψηλής διάστασης και σίγουρου κρησαρίσματος και εξετάζει εκτενώς

τις μεθόδους SIS και *Iterative*-SIS. Στο κεφάλαιο 3 παρουσιάζεται η μέθοδος σίγουρου κρησαρίσματος προβολή κανονικών ελάχιστων τετραγώνων σε υψηλής διάστασης δεδομένα (*High Dimensional Ordinary Least Square Screening* (*HOLP*)). Το τέταρτο κεφάλαιο παρουσιάζει συνοπτικά την συσχέτιση της απόσταση μεταβλητών (*Distance Correlation*) και ασχολείται με τις μεθόδους DC-SIS και DC-ISIS. Στο κεφάλαιο 5 γίνονται αριθμητικές συγκρίσεις μεταξύ των μεθόδων κρησαρίσματος, με την βοήθεια του στατιστικού πακέτου R.

Abstract

High dimensional statistical problems arise from diverse fields of scientific research and technological development. Feature screening and feature selection play a pivotal role in contemporary statistical learning and scientific discoveries. Feature screening refers to the procedure of screening potential explanatory variables in high dimension data so as to reduce their numbers significantly, without leaving the model's really important variables out. Paying a closer attention to this term, one can see that a successful feature screening is a great achievement since it drastically simplifies the data analysis making classical statistical methods that failed once before, to be used now. Therefore, the choice of characteristics becomes possible, easier and more accurate. The first method was introduced in an article by Fan and Li (2008), titled as *Sure Independence Screening* (*SIS*). This method is based on correlation learning. In the same article, an iterative form of SIS was suggested, which was called *Iterative Sure Independence Screening* (*ISIS*). This particular method was introduced in order to deal with some problems which derived from the first model (SIS). In another article, Li, Zhong and Zhu (2012) suggested an alternative method of feature screening that of *Distance Correlation-based Sure Independence Screening* (*DC-SIS*), which was based on distance correlation learning consisting of a new kind of distance correlation variables which was suggested by Szekely, Rizzo and Bakirov (2007). However, DC-SIS dealt with some similar problems as SIS, therefore Zhong and Zhu (2015) suggested *Iterative DC-SIS,* in order to solve these kinds of problems. Wang and Leng (2015) suggested another method called High Dimensional Ordinary Least Square Screening (HOLP), which does not deal with variable correlation like the previously mentioned methods. This particular paper mainly deals with various methods of feature screening but also with variable selection using penalised methods.

More specifically, the first chapter not only makes some references to the general linear model and basic techniques in parameter estimation, but it also refers to methods of choosing and estimates different variables which are based on the introduction of one penalty function to the least square method. Chapter two introduces the meaning of high dimensional data and sure screening και thoroughly examines the screening methods SIS και Iterative-SIS. Chapter three presents one new sure screening method, the High Dimensional Ordinary Least Square Screening (HOLP). The fourth chapter briefly presents Distance Correlation and deals with the DC-SIS και DC-ISIS methods. In the fifth chapter and final chapter of this paper numerical comparisons are presented between the screening methods with the help of the statistical package R.

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε υπό την επίβλεψη του καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κ. Χρήστου Κουκουβίνου, τον οποίο θα ήθελα να ευχαριστήσω γιατί μου έδωσε την δυνατότητα να ασχοληθώ με ένα θέμα πολύ σύγχρονο το οποίο στο μέλλον είμαι βέβαιος θα μου φανεί πολύ χρήσιμο τόσο από ερευνητικής πλευράς όσο και από επαγγελματικής αποκατάστασης.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτορα Κρυσταλλένια Δρόσου για την πολύτιμη βοήθεια της και το αμέριστο ενδιαφέρον που έδειξε κατά την διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου τόσο για την οικονομική στήριξη όσο και για την πνευματική στήριξη που μου παρείχαν κατά την διάρκεια των σπουδών.

Χωρίς την συμβολή των πιο πάνω ανθρώπων δεν θα ήταν δυνατή η ολοκλήρωση των σπουδών μου και η εκπόνηση της παρούσας εργασίας.

Μάριος Λοΐζου

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Αθήνα, 2016

ΚΕΦΑΛΑΙΟ 1

ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΠΟΙΝΙΚΟΠΟΙΗΣΗΣ

1.1 Εισαγωγή

Μεγάλο μέρος της επιστήμης της Στατιστικής ασχολείται με την κατασκευή και ανάλυση στατιστικών μοντέλων, δηλαδή τη μορφή της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών. Όταν αναφερόμαστε σε στατιστικά μοντέλα, αναφερόμαστε σε στοχαστικά μοντέλα, δηλαδή μελετάμε φαινόμενα που δεν είναι εντελώς προβλέψιμα. Για παράδειγμα η ανταπόκριση ενός ασθενή σε μια δόση φαρμάκου ή το φορτίο που χρειάζεται για να σπάσει μια δοκός, μπορεί να προβλεφθεί εν μέρη από μετρήσιμους παράγοντες αλλά ποτέ με απόλυτη ακρίβεια. Τα βασικά μοντέλα για την ανάλυση και την επεξεργασία δεδομένων που προέρχονται από στοχαστικά φαινόμενα είναι τα μοντέλα παλινδρόμησης.

Στη φυσική έχουμε σχέσεις ντετερμινιστικές, δηλαδή σχέσεις που μπορούν να καθορίζουν τη σχέση μεταξύ μεταβλητών με ακρίβεια μέσω μιας συνάρτησης $y = f(x)$, όπως η σχέση $u = u_o + at$ που δίνει την ταχύτητα u ως συνάρτηση του χρόνου t. Στην περίπτωση όμως που αναφερόμαστε σε στοχαστικά φαινόμενα είναι λογικό να εισάγουμε και μια τυχαία συνιστώσα. Έτσι θεωρούμε τη σχέση $y = f(x) + \varepsilon$, όπου ε είναι μια τυχαία μεταβλητή.

Έστω Y η μεταβλητή που μας ενδιαφέρει να εκτιμήσουμε, όπως για παράδειγμα η κίνηση μιας μετοχής, και $X_1, X_2, X_3, \dots . X_k$ οι μεταβλητές που πιθανόν επηρεάζουν την εκτιμώμενη μεταβλητή μας. Η μεταβλητή Y αναφέρεται ως μεταβλητή απόκρισης ενώ οι $X_1, X_2, X_3, \dots . X_k$, ως επεξηγηματικές μεταβλητές ή ανεξάρτητες μεταβλητές ή συμμεταβλητες ή χαρακτηριστικά. Σημειώνουμε ότι θα αναφέρουμε όλες τις ονομασίες στη συνέχεια. Αυτό που μας ενδιαφέρει είναι να εντοπίσουμε ποιες από τις επεξηγηματικές μεταβλητές όντως συνεισφέρουν στο μοντέλο, ποια είναι η επίδραση τους, καθώς και ποια είναι η σχέση που συνδέει αυτές με την μεταβλητή απόκρισης (π.χ γραμμική).

1.2 Το Γενικό Γραμμικό Μοντέλο

Είναι πολύ συχνό φαινόμενο οι τιμές κάποιας μεταβλητής να επηρεάζονται γραμμικά από περισσότερες από μία επεξηγηματικές μεταβλητές, όπως για παράδειγμα το ύψος ενός παιδιού ενδέχεται να εξαρτάται από την ηλικία και από το γένος του παιδιού. Το γενικό γραμμικό μοντέλο δίνεται από την σχέση:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \ldots + \beta_\kappa x_{ik} + \varepsilon_i, \qquad (1.2.1)$$

όπου,

- y_i, $i = 1,2,\ldots,n$ οι τιμές των παρατηρήσεων της μεταβλητής απόκρισης y.
- x_{ij}, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,k$ οι τιμές για την $i-$οστή παρατήρηση των επεξηγηματικών μεταβλητών x_j.
- $\beta_1, \beta_2, \beta_3 \ldots, \beta_k$ οι άγνωστες παράμετροι του μοντέλου.
- ε_i, $i = 1,2,\ldots,n$, τα τυχαία σφάλματα τα οποία υποθέτουμε ότι ικανοποιούν τις παρακάτω υποθέσεις.
    - $E(\varepsilon_i) = 0, \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ i.$
    - $V(\varepsilon_i) = \sigma^2, \gamma\iota\alpha\ \kappa\acute{\alpha}\theta\varepsilon\ i,$ δηλαδή τα τυχαία σφάλματα ικανοποιούν την υπόθεση της ομοσκεδαστικότητας.
    - $cov(\varepsilon_i, \varepsilon_j) = 0\ \gamma\iota\alpha\ i \neq j,$ δηλαδή τα ε_ι είναι ασυσχέτιστα μεταξύ τους.

Η σχέσης (1.2.1) μπορεί να γραφτεί και υπό τη μορφή πινάκων ως εξής:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \qquad (1.2.2)$$

όπου

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)' \quad (n \times 1).$
- $\boldsymbol{X} = \begin{pmatrix} \mathbf{1} & x_{\mathbf{11}} & x_{\mathbf{12}} & \cdots & x_{\mathbf{1k}} \\ \mathbf{1} & x_{\mathbf{21}} & x_{\mathbf{22}} & \cdots & x_{\mathbf{2k}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1} & x_{\boldsymbol{n1}} & x_{\boldsymbol{n2}} & \cdots & x_{\boldsymbol{nk}} \end{pmatrix} \quad (n \times p)$ με $p = k + 1.$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_\kappa)' \quad (p \times 1).$
- $\boldsymbol{\varepsilon} \sim \boldsymbol{N_n}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, όπου $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ η αναμενόμενη τιμή, και $V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$ ο πίνακας διασποράς-συνδιασποράς, με τον $\boldsymbol{I}$ να είναι $n \times n$ μοναδιαίος πίνακας.

Έχουν προταθεί αρκετές μέθοδοι για την εκτίμηση των παραμέτρων του μοντέλου, δηλαδή των β_j, με , $j = 1,2, \dots, k$. Παρακάτω παραθέτουμε τις μεθόδους:

- Ελαχίστων Τετραγώνων (least squares method).
- Μέγιστης Πιθανοφάνειας (maximum likelihood method).

1.2.1 Μέθοδος Ελάχιστων Τετραγώνων (*Ordinary least square method*)

Η μέθοδος ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων $\boldsymbol{\beta}$, βασίζεται στην ελαχιστοποίηση του αθροίσματος των τετραγώνων των αποκλίσεων μεταξύ των y_i και $\hat{y}_i$, δηλαδή της παράστασης:

$$RSS(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \varepsilon_{\boldsymbol{i}}^{\boldsymbol{2}} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^{\boldsymbol{2}}$$

$$= \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta})^{\boldsymbol{2}}, \qquad (1.2.3)$$

με $\boldsymbol{x}_{\boldsymbol{i}}' = (1, x_{i1}, x_{i2}, \dots, x_{ik})$.

Παραγωγίζοντας ως προς $\boldsymbol{\beta}$ παίρνουμε:

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta}). \qquad (1.2.4)$$

Θέτοντας την παραπάνω παράσταση ίση με το μηδέν έχουμε:

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}. \qquad (1.2.5)$$

Αν ο πίνακας $\boldsymbol{X}'\boldsymbol{X}$ αντιστρέφεται τότε η εκτιμήτρια ελαχίστων τετραγώνων (*ordinary least square estimator-OLS*) του $\boldsymbol{\beta}$ είναι η εξής:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{y}. \qquad (1.2.6)$$

Επομένως το μοντέλο που εκτιμούμε δίνεται από τη σχέση $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ και τα υπόλοιπα (*residuals*) δίνονται από την σχέση $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$. Παρατηρούμε ότι κατά τη διαδικασία εύρεσης των εκτιμητριών ελαχίστων τετραγώνων δεν απαιτείται η γνώση της κατανομής των τυχαίων σφαλμάτων.

Η εκτιμήτρια ελάχιστων τετραγώνων έχει την επιθυμητή ιδιότητα της αμεροληψίας (*unbiased estimator*).

Πράγματι:

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) = E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right) \\ &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})) \\ &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})) = \boldsymbol{\beta}, \end{aligned}$$

αφού $E(\boldsymbol{\varepsilon}) = \mathbf{0}$.

Επίσης ισχύει:

$$\begin{aligned} V(\widehat{\boldsymbol{\beta}}) &= V((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= ((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\,\boldsymbol{V}(\boldsymbol{y})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \end{aligned}$$

αφού $V(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$.

1.2.2 Μέθοδος Μέγιστης Πιθανοφάνειας (*Maximum Likelihood Method*)

Μια διαφορετική μέθοδος κατασκευής εκτιμητριών είναι η μέθοδος μέγιστης πιθανοφάνειας (*Maximum likelihood method*). Έστω $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \ldots, \theta_m)' \in \boldsymbol{\Theta}$ μια άγνωστη παράμετρος ενός πληθυσμού. Η συνάρτηση πιθανότητας ενός τυχαίου δείγματος $\boldsymbol{Y} = (Y_1, Y_2, Y_3, \ldots, Y_n)$ δίνεται από την σχέση:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}), \tag{1.2.7}$$

όπου $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \ldots, \theta_m)'$ οι παράμετροι προς εκτίμηση.

Όταν είναι γνωστές οι τιμές $y_i, i = 1, \ldots, n$ των τυχαίων μεταβλητών $Y_i, i = 1, \ldots, n$, η συνάρτηση πιθανότητας $f(\boldsymbol{x}|\boldsymbol{\theta})$ μπορεί να θεωρηθεί ως συνάρτηση της παραμέτρου $\boldsymbol{\theta}$. Επομένως έχουμε:

$$L(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}). \tag{1.2.8}$$

Η $L(\boldsymbol{\theta})$ ονομάζεται **συνάρτηση πιθανοφάνειας** (*likelihood function*), αφού εκφράζει πόσο "πιθανοφανείς", δηλαδή πόσο σύμφωνες με το συγκεκριμένο δείγμα είναι οι διάφορες τιμές της παραμέτρου $\boldsymbol{\theta}$. Μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας $L(\boldsymbol{\theta})$ ως προς $\boldsymbol{\theta}$ προκύπτει η εκτιμήτρια μέγιστης πιθανοφάνειας (ΕΜΠ) της παραμέτρου $\boldsymbol{\theta}$, η οποία είναι η $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)'$.

Επομένως έχουμε να μεγιστοποιήσουμε την συνάρτηση:

$$L(\widehat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}), \tag{1.2.9}$$

ή ισοδύναμα

$$l(\widehat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta}} l(\boldsymbol{\theta}), \tag{1.2.10}$$

όπου $l(\boldsymbol{\theta}) = lnL(\boldsymbol{\theta})$ αύξουσα συνάρτηση της πιθανοφάνειας $L(\boldsymbol{\theta})$.

Θα πρέπει λοιπόν οι μερικές παράγωγοι της $l(\boldsymbol{\theta})$ ως προς $\theta_j,\ j = 1, \ldots, m$ να μηδενίζονται στο σημείο $\widehat{\boldsymbol{\theta}}$ αφού $\widehat{\boldsymbol{\theta}}$ ακρότατο. Έτσι έχουμε το σύστημα:

$$\frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}) = 0, \qquad j = 1, \ldots, m \tag{1.2.11}$$

Λύνοντας το παραπάνω σύστημα προκύπτει η ΕΜΠ $\widehat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)'$. Για να είναι η λύση αυτή σημείο μεγίστου προφανώς θα πρέπει ο Εσσιανός πίνακας $\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]_{m \times m}$ να είναι γνήσια αρνητικός για $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

Στο γενικό γραμμικό μοντέλο που περιγράψαμε πιο πάνω γνωρίζουμε ότι:

$$\boldsymbol{\varepsilon} \sim \boldsymbol{N_n}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

και άρα:

$$\boldsymbol{y} \sim \boldsymbol{N_n}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}).$$

Επομένως:

$$y_{\boldsymbol{i}} \sim N_{\boldsymbol{n}}(\boldsymbol{x_i'\beta}, \sigma^2),$$

όπου $\boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$.

Άρα για την εύρεση των εκτιμητριών μέγιστης πιθανοφάνειας μεγιστοποιούμε την συνάρτηση πιθανοφάνειας:

$$L(\sigma^2, \boldsymbol{\beta}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{\boldsymbol{1}}{(2\pi \boldsymbol{\sigma^2})^{\boldsymbol{1/2}}} exp\left\{-\frac{(y_i - \boldsymbol{x_i'\beta})^{\boldsymbol{2}}}{2\sigma^2}\right\}, \tag{1.2.12}$$

ως προς σ^2 και $\boldsymbol{\beta}$.

Εφαρμόζοντας την μέθοδο όπως την περιγράψαμε πιο πάνω, εύκολα μπορούμε να δείξουμε ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι οι εξής:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-\boldsymbol{1}} \boldsymbol{X'y}. \tag{1.2.13}$$

$$\widehat{\boldsymbol{\sigma}}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{\boldsymbol{n}} = \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{n}}. \tag{1.2.14}$$

1.3 Ποινικοποιημένα Ελάχιστα Τετράγωνα (*Penalized Least Squares*)

1.3.1 Εισαγωγή

Όπως είδαμε σε προηγούμενο κεφάλαιο, η εκτίμηση των παραμέτρων ενός γραμμικού μοντέλου μπορεί να γίνει με τη μέθοδο ελαχίστων τετραγώνων (*OLS)* και την μέθοδο μέγιστης πιθανοφάνειας. Η εκτιμήτρια που προκύπτει από την μέθοδο ελαχίστων τετραγώνων, ελαχιστοποιώντας το άθροισμα των τετραγώνων των υπολοίπων (*Residual Sum of Squares* (*RSS*)) είναι σε αρκετές περιπτώσεις μια καλή εκτιμήτρια. Υπάρχουν όμως περιπτώσεις όπου η εκτιμήτρια αυτή δεν είναι αρκετά καλή. Όταν ο αριθμός των παραμέτρων είναι μεγαλύτερος από τις παρατηρήσεις $(p > n)$ ή όταν στήλες του πίνακα σχεδιασμού $\boldsymbol{X}$ έχουν υψηλή συσχέτιση, δηλαδή αντιμετωπίζουμε το πρόβλημα της πολυσυγγραμμικότητας, τότε η εκτιμήτρια που προκύπτει είναι ασταθής με μεγάλη μεταβλητότητα. Στην περίπτωση "απόλυτης πολυσυγραμμικότητας" ο $\boldsymbol{X}'\boldsymbol{X}$ δεν είναι αντιστρέψιμος, αφού $\det(\boldsymbol{X}'\boldsymbol{X}) = 0$, οπότε η εκτιμήτρια ελαχίστων τετραγώνων δεν μπορεί να χρησιμοποιηθεί.

Ένας τρόπος για να αντιμετωπιστεί αυτό το πρόβλημα είναι αντί να ελαχιστοποιήσουμε μόνο το άθροισμα των τετραγώνων των υπολοίπων (RSS), ελαχιστοποιούμε την παράσταση:

$$RSS(\boldsymbol{\beta}) + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^{\boldsymbol{2}} + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta}), \tag{1.3.1}$$

όπου $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ είναι μια συνάρτηση ποινής που εξαρτάται από μια παράμετρο κανονικοποίησης $\lambda \geq 0$ και $RRS(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^{\boldsymbol{2}}$, $\boldsymbol{x}_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$ **.**

Οπότε προκύπτει ο εκτιμητής:

$$\widehat{\boldsymbol{\beta}} = arg\min_{\beta}\big(RSS(\boldsymbol{\beta}) + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\big) = arg\min_{\beta}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^{\boldsymbol{2}} + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\right). \tag{1.3.2}$$

Η πιο πάνω μέθοδος καλείται μέθοδος ποινικοποιημένων ελαχίστων τετραγώνων (*Penalized least squares* (*PLS*)) με την οποία θα ασχοληθούμε εκτενέστερα σε επόμενο κεφάλαιο της παρούσας εργασίας.

Σημειώνουμε ότι η παράσταση (1.3.1) είναι η Λαγκραντζιανή μορφή (*Lagrangian form*) για την ελαχιστοποίηση της εξίσωσης:

$$RSS(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^{\boldsymbol{2}}, \tag{1.3.3}$$

υπό τον περιορισμό $P(\boldsymbol{\beta}) < t, t \geq 0$.

Ισοδύναμα ο εκτιμητής που προκύπτει είναι:

$$\widehat{\boldsymbol{\beta}} = arg\min_{\beta}\big(RSS(\boldsymbol{\beta})\big) = arg\min_{\beta}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(y_i - \boldsymbol{x_i'\beta})^{\boldsymbol{2}}\right), \tag{1.3.4}$$

υπό τον περιορισμό:

$$P(\boldsymbol{\beta}) \leq t,$$

όπου $t \geq 0$ μια ρυθμιστική παράμετρος και $\boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$.

Για καλύτερη κατανόηση παραθέτουμε το πιο κάτω παράδειγμα.

<u>Παράδειγμα</u>

Έστω ότι η μεταβλητή απόκρισης Y εξαρτάται γραμμικά από δύο επεξηγηματικές μεταβλητές X_1, X_2 και έστω ότι οι επεξηγηματικές μεταβλητές έχουν υψηλή συσχέτιση μεταξύ τους:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_1 X_2.$$

Χρησιμοποιώντας το στατιστικό πακέτο R προσομοιάσαμε ένα τέτοιο παράδειγμα δύο επεξηγηματικών μεταβλητών με υψηλή συσχέτιση οι οποίες έχουν την παρακάτω γραμμική σχέση με την μεταβλητή απόκρισης :

$$E(Y) = 3 + X_1 + X_2.$$

Εφαρμόσαμε γραμμική ανάλυση παλινδρόμησης με δείγμα 1000 παρατηρήσεων. Τα αποτελέσματα που λάβαμε ήταν $\hat{\beta}_1 \approx -8.57$ και $\hat{\beta}_2 \approx 10.54$ που προφανώς είναι αρκετά κακές εκτιμήσεις και αποκλίνουν από τις πραγματικές τιμές $\beta_1 = 1$ και $\beta_2 = 1$. Επειδή οι δύο επεξηγηματικές μεταβλητές έχουν υψηλή συσχέτιση, η τιμή που προκύπτει από το $RSS(-8.57\,,10.54) = 953.04$ είναι παραπλήσια με την τιμή του $RSS(1,1) = 955.29$, με την τιμή $RRS(-8.57\,,10.54)$ να είναι λίγο μικρότερη αφού επιλέχθηκε κατά την ελαχιστοποίηση.

Ένας πολύ αποτελεσματικός τρόπος να αντιμετωπιστεί το πιο πάνω πρόβλημα είναι η χρήση των ποινικοποιημένων ελαχίστων τετραγώνων, χρησιμοποιώντας ως συνάρτηση ποινικοποίησης τη συνάρτηση $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2$. Αυτή η προσέγγιση είναι γνωστή ως παλινδρόμηση κορυφογραμμής (*ridge regression).*

Γιατί αυτό βοηθά στην καλύτερη εκτίμηση των παραμέτρων;

Παρόλο που ισχύει ότι $RRS(-8.57\,,10.54) \approx RRS(1,1)$, προφανώς $1^2 + 1^2 = 2$ είναι μικρότερο του $10.54^2 + (-8.57)^2 = 184.53$, και άρα η τιμή της συνάρτησης ποινής που αθροίζεται στην παράσταση (1.3.1) είναι μικρότερη, οπότε κατά την ελαχιστοποίηση της παράστασης (1.3.1) θα προτιμηθεί μια λύση κοντά στο $\widehat{\boldsymbol{\beta}} \approx (1,1)$**.** Για επαλήθευση των όσων αναφέραμε, εφαρμόσαμε την παλινδρόμηση κορυφογραμμής για το συγκεκριμένο παράδειγμα με την βοήθεια του στατιστικού πακέτου R και λάβαμε τις εκτιμήσεις

$\hat{\beta}_1 \approx 0.98$ και $\hat{\beta}_2 \approx 0.99$, που προφανώς είναι αρκετά καλές εκτιμήσεις για τα β_1 και β_2 αντίστοιχα.

Πέρα των όσων αναφέραμε πιο πάνω, οι μέθοδοι ποινικοποίησης χρησιμοποιούνται και για ακόμα ένα σημαντικό κομμάτι στην ανάλυση παλινδρόμησης, την επιλογή σημαντικών μεταβλητών (*variable selection*), δηλαδή αυτών που όντως συνεισφέρουν στη μεταβλητή απόκρισης. Στην πράξη, στα αρχικά στάδια επιλογής ενός μοντέλου, εισάγεται ένας μεγάλος αριθμός επεξηγηματικών μεταβλητών για να μετριαστεί πιθανή μεροληψία κατά την μοντελοποίηση, χωρίς αυτές οι μεταβλητές να είναι σίγουρα σημαντικές για το μοντέλο. Για την ενίσχυση της προβλεψιμότητας ενός μοντέλου αλλά και για την επιλογή σημαντικών μεταβλητών οι στατιστικοί χρησιμοποιούν μεθόδους όπως η κατά βήματα επιλογή (*stepwise selection*), η επιλογή καλύτερου υποσυνόλου (*best subset selection*) χρησιμοποιώντας διάφορα κριτήρια όπως είναι το *AIC, BIC* και το *Cp-Mallows.* Το πρόβλημα που παρουσιάζεται στην περίπτωση επιλογής καλύτερου υποσυνόλου έγκειται στο γεγονός ότι ο αριθμός όλων των πιθανών υποσυνόλων που πρέπει να εξεταστεί, όταν ο αριθμός των μεταβλητών p είναι μεγάλος, μεγαλώνει εκθετικά με το p και έτσι το υπολογιστικό κόστος είναι μεγάλο, ενώ παράλληλα αγνοούνται στοχαστικά σφάλματα. Οι μέθοδοι ποινικοποίησης, αυτό που επιτυγχάνουν είναι την συρρίκνωση (*shrinkage*) των εκτιμητών προς το μηδέν που προέκυψαν από μια μέθοδο χωρίς ποινικοποίηση. Αν η συρρίκνωση αυτή είναι πολύ μεγάλη τότε επιτυγχάνεται ο μηδενισμός κάποιων μεταβλητών, με αποτέλεσμα να γίνεται επιλογή σημαντικών μεταβλητών (*variable selection*).

Η διαδικασία ποινικοποίησης χρησιμοποιείται συχνά στην επιλογή μεταβλητών και συνιστάται στην εισαγωγή κάποιων συναρτήσεων ποινής (*penalty functions*) οι οποίες πρέπει να ικανοποιούν τις ακόλουθες ιδιότητες (τις οποίες θα εξετάσουμε στη συνέχεια):

1. Να είναι ιδιάζουσες (*singular)* στην αρχή ώστε να παράγουν σποραδικές λύσεις (*sparse solutions),* δηλαδή πολλοί εκ των εκτιμώμενων συντελεστών να έχουν τιμή μηδέν.

2. Να ικανοποιούν συγκεκριμένες απαιτήσεις ώστε ο παραγόμενος εκτιμητής να είναι συνεχής, και η επιλογή του μοντέλου να χαρακτηρίζεται από σταθερότητα (*stability*).

3. Να παράγουν αμερόληπτους εκτιμητές για μεγάλους συντελεστές.

Σημειώνουμε ότι η μέθοδος ποινικοποίησης, παλινδρόμηση κορυφογραμμής (*ridge regression),* δεν μπορεί να χρησιμοποιηθεί για επιλογή μεταβλητών αφού η συνάρτηση ποινής δεν ικανοποιεί την ιδιότητα (1.), με αποτέλεσμα ακόμα και αν η μέθοδος συρρικνώσει αρκετά τους ασήμαντους εκτιμώμενους συντελεστές, αυτοί παραμένουν στο μοντέλο.

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 (*ridge regression)* αλλά και της μεθόδου επιλογής καλύτερου υποσυνόλου, η οποία επεκτείνεται και σε μοντέλα βασισμένα στην πιθανοφάνεια.

1.3.2 Στάθμιση Μεροληψίας-Μεταβλητότητας (*Bias-Variance Tradeoff*)

Η κατανόηση της λειτουργίας "Στάθμιση Μεροληψίας-Μεταβλητότητας" (*Bias-Variance Tradeoff*) παίζει σημαντικό ρόλο στην κατανόηση της ποινικοποιημένης παλινδρόμησης. Η "Στάθμιση Μεροληψίας-Μεταβλητότητας" (*Bias-Variance Tradeoff*) εξηγείται καλύτερα μελετώντας το μέσο τετραγωνικό σφάλμα (*Mean Square Error* (*MSE*)) ενός μοντέλου, που ουσιαστικά δίνει το αναμενόμενο σφάλμα πρόβλεψης. Για ένα μοντέλο με εκτιμώμενους συντελεστές $\hat{\beta}_j$, το μέσο τετραγωνικό σφάλμα (*MSE)* ορίζεται να είναι:

$$\begin{aligned} MSE &= E(y_{new} - (\hat{\beta}_0 + \textstyle\sum_{j=1}^{p-1} \hat{\beta}_j x_{new,j})^2 \\ &= Var\left[y_{new} - (\hat{\beta}_0 + \textstyle\sum_{j=1}^{p-1} \hat{\beta}_j x_{new,j})\right] + Bias(\widehat{\boldsymbol{\beta}})^2, \end{aligned} \tag{1.3.5}$$

όπου (y_{new}, x_{new}) είναι νέες παρατηρήσεις που δεν χρησιμοποιήθηκαν αρχικά από το μοντέλο για την εκτίμηση των συντελεστών $\hat{\beta}_j$.

Οι ποινικοποιημένες μέθοδοι παλινδρόμησης προσθέτουν μεροληψία κατά την εκτίμηση των συντελεστών παλινδρόμησης, αφού όπως αναφέραμε, συνεχώς συρρικνώνουν (*shrinkage)* τους εκτιμητές παλινδρόμησης $\hat{\beta}_j$. Παρόλα αυτά η συρρίκνωση αυτή προκαλεί και μείωση της διακύμανσης (μεταβλητότητας). Αυτό ονομάζεται "Στάθμιση Μεροληψίας-Μεταβλητότητας" (*Bias-Variance Tradeoff*). Συχνά η αύξηση της μεροληψίας είναι μικρότερη από την μείωση της διακύμανσης, οπότε το μέσο τετραγωνικό σφάλμα (*MSE*) που προκύπτει από τη χρησιμοποίηση μιας μεθόδου ποινικοποίησης για την εκτίμηση των συντελεστών του μοντέλου είναι μικρότερο από αυτό που θα προέκυπτε αν χρησιμοποιούσαμε τον αμερόληπτο εκτιμητή ελάχιστων τετραγώνων (*OLS*). Επομένως οι μέθοδοι ποινικοποιημένης παλινδρόμησης παράγουν μοντέλα με καλύτερη προβλεπτική ικανότητα για νέες παρατηρήσεις.

1.3.3 Επιλογή μεταβλητών μέσω ποινικοποιημένων ελάχιστων τετραγώνων

Έστω το γραμμικό μοντέλο παλινδρόμησης:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{1.3.6}$$

όπου

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)' \quad (n \times 1)$.
- $\boldsymbol{X} = \begin{pmatrix} \mathbf{1} & x_{\mathbf{11}} & x_{\mathbf{12}} & \cdots & x_{\mathbf{1k}} \\ \mathbf{1} & x_{\mathbf{21}} & x_{\mathbf{22}} & \cdots & x_{\mathbf{2k}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1} & x_{\mathbf{n1}} & x_{\mathbf{n2}} & \cdots & x_{\mathbf{nk}} \end{pmatrix} \quad (n \times p)$ με $p = k + 1$.
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_\kappa)' \quad (p \times 1)$.

- $\boldsymbol{\varepsilon} \sim \boldsymbol{N_n}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, όπου $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ η αναμενόμενη τιμή, και $V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$ ο πίνακας διασποράς-συνδιασποράς, με τον $\boldsymbol{I}$ να είναι $n \times n$ μοναδιαίος πίνακας.

Θεωρούμε ότι τα y_i είναι υπό συνθήκη ανεξάρτητα δοθέντων των $x_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, k$ και ότι οι στήλες του πίνακα σχεδιασμού $\boldsymbol{X}$ είναι ορθοκανονικές (*orthonormal*), δηλαδή $\boldsymbol{X'X} = \boldsymbol{I}$. Τα αποτελέσματα που θα προκύψουν στην πιο κάτω ανάλυση γενικεύονται και σε μη ορθοκανονικούς πίνακες σχεδιασμού. Όπως είδαμε ο εκτιμητής ελαχίστων τετραγώνων δίνεται από την ελαχιστοποίηση της ποσότητας:

$$RSS(\boldsymbol{\beta}) = \sum_{\boldsymbol{\iota=1}}^{\boldsymbol{n}} \varepsilon_{\boldsymbol{\iota}}^{\mathbf{2}} = \boldsymbol{\varepsilon'\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^{\mathbf{2}}, \qquad (1.3.7)$$

και ο εκτιμητής προκύπτει να είναι:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = (\boldsymbol{X'X})^{-\mathbf{1}}\boldsymbol{X'y} = \boldsymbol{I}^{-\mathbf{1}}\boldsymbol{X'y} = \boldsymbol{X'y}, \qquad (1.3.8)$$

αφού οι στήλες του πίνακα $\boldsymbol{X}$ είναι ορθοκανονικές. Επίσης ισχύει ότι $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{XX'y}$.

Μια μορφή ποινικοποιημένων ελάχιστων τετραγώνων κατάλληλη για επιλογή μεταβλητών είναι η εξής:

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^{\mathbf{2}} + \boldsymbol{\lambda}\sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} p_j\left(|\beta_j|\right) =$$

$$= \frac{1}{2}\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^{\mathbf{2}} + \frac{1}{2}\sum_{\boldsymbol{j=1}}^{\boldsymbol{p}}(\hat{\beta}_{j,OLS} - \beta_{\boldsymbol{j}})^{\mathbf{2}} + \lambda\sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} p_{\boldsymbol{j}}\left(|\beta_{\boldsymbol{j}}|\right), \qquad (1.3.9)$$

Η ισότητα της (1.3.9) προκύπτει εύκολα έχοντας υπόψιν το εξής :

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^{\mathbf{2}} = \|\boldsymbol{y} - \boldsymbol{X\beta} + \boldsymbol{XX'y} - \boldsymbol{XX'y}\|^{\mathbf{2}} = \left\|(\boldsymbol{y} - \widehat{\boldsymbol{y}}) + \left(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X\beta}\right)\right\|^{\mathbf{2}}. \qquad (1.3.10)$$

Οι συναρτήσεις ποινής (*penalty functions*) $p_j(\,.\,)$ στην (1.3.9) δεν είναι αναγκαίο να είναι ίδιες για όλα τα j. Για παράδειγμα, μπορεί να επιθυμούμε τη διατήρηση κάποιων σημαντικών επεξηγηματικών μεταβλητών σε ένα παραμετρικό μοντέλο και έτσι να μην επιθυμούμε την ποινικοποίηση των αντίστοιχων παραμέτρους τους. Για χάρη απλότητας, θεωρούμε ότι οι συναρτήσεις ποινής είναι ίδια για όλους τους συντελεστές, και θα τη συμβολίζουμε με $p(|\,.\,|)$. Η χρησιμοποίηση διαφορετικών συναρτήσεων δεν προσθέτει κάποια δυσκολία στην παρακάτω ανάλυση. Επίσης θα συμβολίζουμε το $\lambda p(|\,.\,|)$, με $p_\lambda(|\,.\,|)$, εννοώντας ότι η συνάρτηση $p(|\,.\,|)$ εξαρτάται από το λ.

Για να βρούμε τον εκτιμητή του $\boldsymbol{\beta}$ πρέπει να ελαχιστοποιήσουμε την παράσταση (1.3.2) ως προς $\boldsymbol{\beta}$. Το πρόβλημα ελαχιστοποίησης είναι ισοδύναμο με την ελαχιστοποίηση των συνιστωσών. Επομένως έχουμε να ελαχιστοποιήσουμε την παράσταση:

$$\frac{\mathbf{1}}{\mathbf{2}}(\hat{\beta}_{j,OLS} - \beta_{\boldsymbol{j}})^{\mathbf{2}} + p_\lambda\left(|\beta_{\boldsymbol{j}}|\right), \quad j = 1, \ldots, p. \qquad (1.3.11)$$

Ισοδύναμα έχουμε το πρόβλημα ποινικοποιημένων ελάχιστων τετραγώνων (*PLS):*

$$\widehat{\beta_j}(\hat{\beta}_{j,OLS}) = \underset{\Theta}{\operatorname{argmin}}\{\tfrac{\mathbf{1}}{\mathbf{2}}(\hat{\beta}_{j,OLS} - \beta_j)^{\mathbf{2}} + p_\lambda(|\beta_j|)\}, j = 1, \ldots, p. \qquad (1.3.12)$$

Οι Antoniadis και Fan (2001) έδειξαν ότι ο PLS εκτιμητής (1.3.12) έχει τις εξής τρεις ιδιότητες:

1. Σχεδόν Αμερόληπτος (*Approximate Unbiasedness*) όταν $p'_\lambda(|\beta_j|) = 0$ για μεγάλο $|\beta_j|$.

2. Σποραδικότητας (*Sparsity*), όταν $\min_{\beta_j \neq 0}\{|\beta_j| + p_\lambda{}'(|\beta_j|)\} > 0$ (δηλαδή ο εκτιμητής που προκύπτει αποτελεί ένα κανόνα περιορισμού (*thresholding rule*), έτσι ώστε να θέτει αυτόματα τις μικρές τιμές των εκτιμώμενων συντελεστών ίσες με το μηδέν, με σκοπό να μειωθεί η πολυπλοκότητα του μοντέλου).

3. Συνέχειας (Continuity), αν και μόνο αν $\underset{\beta_j}{\operatorname{argmin}}\{|\beta_j| + p_\lambda{}'(|\beta_j|)\} = 0$.

Πράγματι, παρατηρούμε ότι η πρώτη παράγωγος ως προς β_j της (1.3.11) είναι:

$$sgn(\beta_j)\{|\beta_j| + p_\lambda{}'(|\beta_j|)\} - \hat{\beta}_{j,OLS}. \qquad (1.3.13)$$

Όταν $p'_\lambda(|\beta_j|) = 0$ για μεγάλο $|\beta_j|$ είναι εύκολο να δείξουμε ότι ο εκτιμητής που προκύπτει είναι το $\hat{\beta}_{j,OLS}$ όταν το $|\hat{\beta}_{j,OLS}|$ είναι επαρκώς μεγάλο. Έτσι όταν η παράμετρος $|\beta_j|$ είναι μεγάλη, η παρατηρούμενη τιμή $|\hat{\beta}_{j,OLS}|$ είναι μεγάλη με μεγάλη πιθανότητα. Επομένως ο εκτιμητής ποινικοποιημένων ελάχιστων τετραγώνων (*PLS)* είναι ο $\hat{\beta}_{j,PLS} = \hat{\beta}_{j,OLS}$, ο οποίος είναι σχεδόν αμερόληπτος. Έτσι η συνθήκη $p'_\lambda(|\beta_j|) = 0$ για μεγάλο $|\beta_j|$ είναι επαρκής συνθήκη για την αμεροληψία μιας μεγάλη πραγματικής παραμέτρου.

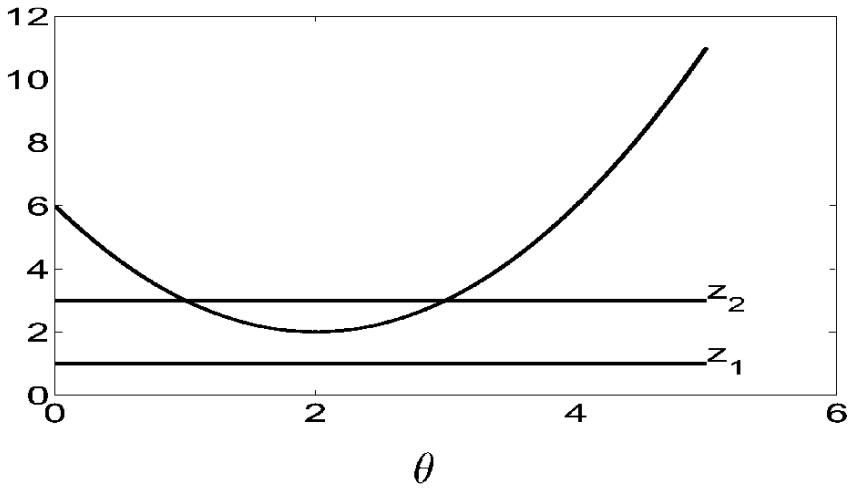
Σχετικά με την ιδιότητα της σποραδικότητας μία επαρκής συνθήκη για ένα εκτιμητή να αποτελεί κανόνα περιορισμού (*thresholding rule)* είναι το ελάχιστο της συνάρτησης $\{|\beta_j| + p_\lambda{}'(|\beta_j|)\}$ να είναι θετικό δηλαδή:

$$\min_{\beta_j \neq 0}\{|\beta_j| + p_\lambda{}'(|\beta_j|)\} > 0. \qquad (1.3.14)$$

Το πιο πάνω ισχύει γιατί όταν $|\hat{\beta}_{j,OLS}| < \min_{\beta_j \neq 0}\{|\beta_j| + p_\lambda{}'(|\beta_j|)\}$, η παράγωγος της (1.3.2) $sgn(\beta_j)\{|\beta_j| + p_\lambda{}'(|\beta_j|)\} - \hat{\beta}_{j,OLS}$, είναι θετική για όλα τα θετικά β_j (και αρνητική για όλα τα αρνητικά). Επομένως ο εκτιμητής ποινικοποιημένων ελάχιστων τετραγώνων (*PLS)* προκύπτει να είναι ίσος με μηδέν ($\hat{\beta}_{j,PLS} = 0$). Όταν τώρα $|\hat{\beta}_{j,OLS}| > \min_{\theta \neq 0}\{|\beta_j| + p_\lambda{}'(|\beta_j|)\}$, μπορούμε να έχουμε δύο διασταυρώσεις (*crossings)*. Η μεγαλύτερη είναι ο PLS εκτιμητής. Αυτό υπονοεί ότι ικανή και αναγκαία συνθήκη για να είναι ένας εκτιμητής συνεχής (ιδιότητα (3.)), είναι το ελάχιστο της συνάρτησης

$|\beta_j| + p_\lambda'(|\beta_j|)$ να επιτυγχάνεται στο μηδέν ($\underset{\beta_j}{\operatorname{argmin}}\{|\beta_j| + p_\lambda'(|\beta_j|)\} = 0$).

Για καλύτερη κατανόηση των προαναφερθέντων βλέπε σχήμα (1.3.1). Από την πιο πάνω ανάλυση προκύπτει ότι μια συνάρτηση ποινής (*penalty function*) ικανοποιεί τις ιδιότητες της σποραδικότητας (*sparsity*) και της συνέχειας (*continuity*) όταν είναι ιδιάζουσα (*singular*) στην αρχή.



Σχήμα 1.3.1: Γραφική παράσταση της συνάρτησης $|\beta_j| + p_\lambda'(|\beta_j|)$ ως προς $\theta = \beta_j$.

Χρησιμοποιώντας για συνάρτηση ποινής την hard οριακή συνάρτηση ποινής (*hard thresholding penalty function*) (βλέπε σχήμα (1.3.3(α))):

$$p_\lambda(\beta_j) = \lambda^2 - (|\beta_j| - \lambda)^2 I(|\beta_j| < \lambda) \quad j = 1, \ldots, p, \tag{1.3.15}$$

όπου $I(|\beta_j| < \lambda) = \begin{cases} 1, \alpha\nu\ |\beta_j| < \lambda \\ 0, \alpha\nu |\beta_j| \geq \lambda \end{cases}$, οι Antoniadis και Fan (1997) έδειξαν ότι προκύπτει ο hard κανόνας περιορισμού (*hard thresholding rule)* (βλέπε σχήμα 1.3.4 (α)):

$$\hat{\beta}_{j,Hard} = \hat{\beta}_{j,OLS} I(|\hat{\beta}_{j,OLS}| > \lambda), \quad j = 1, \ldots, p, \tag{1.3.16}$$

ο οποίος αποτελεί τον εκτιμητή που προκύπτει από την χρήση των ποινικοποιημένων ελαχίστων τετραγώνων με συνάρτηση ποινής την (1.3.15), και συμπίπτει με την επιλογή καλύτερου υποσυνόλου (*best subset selection*) και με τη κατά βήματα πρόσθεση και αφαίρεση (*stepwise deletion and addition*) για ορθοκανονικούς σχεδιασμούς.

Επίσης η L_2 συνάρτηση ποινής: $p_\lambda(|\boldsymbol{\beta}|) = \lambda\|\boldsymbol{\beta}\|_2$ έχει ως αποτέλεσμα την παλινδρόμηση κορυφογραμμής. Η μέθοδος LASSO (*Least Absolute Shrinkage and Selection Operator)* που προτάθηκε από τον Tibshirani (1996,1997) είναι η εκτίμηση των ποινικοποιημένων ελάχιστων τετραγώνων με την L_1 συνάρτηση ποινής $p_\lambda(|\boldsymbol{\beta}|) = \lambda|\boldsymbol{\beta}|$, και οδηγεί στο Soft κανόνα περιορισμού (*Soft thresholding rule):*

$$\hat{\beta}_{j,Lasso} = sgn(\hat{\beta}_{j,OLS})(|\hat{\beta}_{j,OLS}| - \lambda)_+. \tag{1.3.17}$$

Για την μέθοδο LASSO θα αναφερθούμε εκτενώς πιο κάτω.

Η L_q συνάρτηση ποινής: $p_\lambda(|\boldsymbol{\beta}|) = \lambda\|\boldsymbol{\beta}\|_q$, έχει ως αποτέλεσμα την παλινδρόμηση bridge. Η λύση είναι συνεχής μόνο αν $q \geq 1$. Όταν όμως $q > 1$ το ελάχιστο της συνάρτησης $|\beta_j| + p_\lambda{}'(|\beta_j|)$ ισούται με μηδέν ($\min_{\beta_j \neq 0}\{|\beta_j| + p_\lambda{}'(|\beta_j|)\} = 0$), οπότε δεν παράγεται σποραδική λύση σύμφωνα με τα όσα προαναφέραμε. Η μόνη συνεχής λύση που αποτελεί ένα κανόνα περιορισμού (*thresholding rule*) στη οικογένεια αυτή είναι η L_1 συνάρτησης ποινής, αλλά αυτό έχει ως τίμημα την μετακίνηση του εκτιμητή που προκύπτει κατά μια σταθερά λ.

1.3.4 Μέθοδος LASSO (*Least Absolute Shrinkage and Selection Operator*)

Η μέθοδος LASSO (Least Absolute Shrinkage and Selection Operator) που προτάθηκε από τον Tibshirani (1996,1997) είναι η εκτίμηση των ποινικοποιημένων ελαχίστων τετραγώνων με την L_1 συνάρτηση ποινής: $p_\lambda(|\boldsymbol{\beta}|) = \lambda\|\boldsymbol{\beta}\|_1$. Ένα πρόβλημα που προκύπτει στην μέθοδο αυτή είναι ότι η συνάρτηση ποινής δεν είναι διαφορίσιμη, οπότε ούτε και η συνάρτηση προς βελτιστοποίηση, ως εκ τούτου απαιτούνται ειδικές τεχνικές βελτιστοποίησης. Παρόλα αυτά είδαμε ότι όταν οι στήλες του πίνακα σχεδιασμού είναι ορθοκανονικές, ο εκτιμητής προκύπτει να είναι ο (1.3.17).

Επίσης όπως είδαμε το πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων:

$$\widehat{\boldsymbol{\beta}} = arg\min_{\beta}\big(RSS(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})\big) = arg\min_{\beta}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(y_i - \boldsymbol{x_i'\beta})^{\boldsymbol{2}} + \lambda P(\boldsymbol{\beta})\right), \quad (1.3.18)$$

είναι η Λαγκρασιανή μορφή της εξίσωσης:

$$\widehat{\boldsymbol{\beta}} = arg\min_{\beta}\big(RSS(\boldsymbol{\beta})\big) = arg\min_{\beta}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(y_i - \boldsymbol{x_i'\beta})^{\boldsymbol{2}}\right), \quad (1.3.19)$$

υπό τον περιορισμό $P(\boldsymbol{\beta}) < t$, όπου $t \geq 0$ είναι μια ρυθμιστική παράμετρος. Αυτή η μορφή κάνει πιο κατανοητό γιατί οι εκτιμώμενοι συντελεστές παλινδρόμησης συρρικνώνονται, αφού παρατηρούμε ότι προσθέτοντας αυτό τον περιορισμό μειώνεται η περιοχή των αποδεκτών λύσεων (βλέπε σχήμα (1.3.4)). Σημειώνουμε ότι για να διαμεριστεί η συρρίκνωση (*shrinkage*) ισότιμα μεταξύ των συντελεστών παλινδρόμησης, κάθε επεξηγηματική μεταβλητή (*predictor*) πρέπει να είναι κανονικοποιηνένη πριν εφαρμοστεί η μέθοδος για την επιλογή μεταβλητών.

Για την μέθοδο LASSO η συνάρτηση ποινής είναι η $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$, οπότε έχουμε:

$$\widehat{\boldsymbol{\beta}} = arg\min_{\beta}\big(RSS(\boldsymbol{\beta})\big) = arg\min_{\beta}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(y_i - \boldsymbol{x_i'\beta})^{\boldsymbol{2}}\right), \quad (1.3.20)$$

υπό τον περιορισμό $\sum_{j=1}^{p}|\beta_j| \leq t\,, t \geq 0$.

Αν θεωρήσουμε ότι οι στήλες του πίνακα σχεδιασμού είναι ορθοκανονικές, οι εκτιμητές προκύπτουν να είναι οι:

$$\hat{\beta}_{j,lasso} = sgn(\hat{\beta}_{j,OLS})(|\hat{\beta}_{j,OLS}| - \lambda)_{+}. \tag{1.3.21}$$

Επομένως για κάποιο λ προκύπτει ότι :

$$\hat{\beta}_{j,lasso} = \begin{cases} \hat{\beta}_{j,OLS} - \lambda, & \gamma\iota\alpha\ \hat{\beta}_{j,OLS} > \lambda \\ 0, & \gamma\iota\alpha\ |\hat{\beta}_{j,OLS}| \leq \lambda. \\ \hat{\beta}_{j,OLS} + \lambda, & \gamma\iota\alpha\ \hat{\beta}_{j,OLS} < -\lambda \end{cases} \tag{1.3.22}$$

Λόγω της l_1-γεωμετρίας, η μέθοδος LASSO, εκτελεί επιλογή μεταβλητών λόγω του ότι μια εκτιμώμενη συνιστώσα μπορεί να είναι ακριβώς μηδέν. Αυτό το φαινόμενο δε συμβαίνει με την παλινδρόμηση κορυφογραμμής (*ridge regression*) όπου η συνάρτηση ποινής είναι $P(\boldsymbol{\beta}) = \| \beta \|_2 = \sum_{j=1}^{p} |\beta_j|^2$ και επομένως έχουμε το πρόβλημα:

$$\widehat{\boldsymbol{\beta}} = arg\min_{\beta}\big(RSS(\boldsymbol{\beta})\big) = arg\min_{\beta}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - x_{\boldsymbol{i}}'\boldsymbol{\beta})^{\boldsymbol{2}}\right), \tag{1.3.23}$$

υπό τον περιορισμό $\sum_{j=1}^{p} |\beta_j|^2 \leq t\ , t \geq 0$.

Η λύση του προβλήματος παλινδρόμησης κορυφογραμμής μπορούμε να δείξουμε ότι είναι:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{y}, \tag{1.3.24}$$

και στην ειδική περίπτωση όπου οι στήλες του πίνακα σχεδιασμού είναι ορθοκανονικές $\boldsymbol{X}$ τότε:

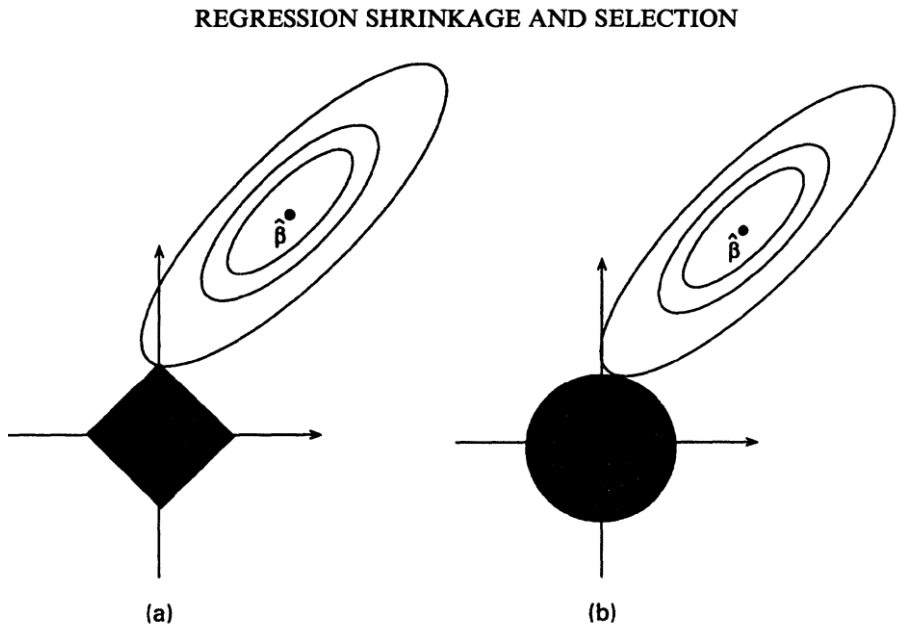
$$\widehat{\boldsymbol{\beta}}_{ridge} = \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}}{1+\lambda} \iff \hat{\beta}_{j,ridge} = \frac{\hat{\beta}_{j,OLS}}{1+\lambda}. \tag{1.3.25}$$

Βλέποντας την γεωμετρική ερμηνεία της μεθόδου LASSO γίνεται πιο κατανοητό γιατί η μέθοδος αυτή μηδενίζει κάποιους συντελεστές κατά την συρρίκνωση των εκτιμώμενων παραμέτρων, αντίθετα με την παλινδρόμηση κορυφογραμμής (*ridge regression*).

<u>Γεωμετρική Ερμηνεία</u>

Βλέποντας το σχήμα (1.3.2) το οποίο απεικονίζει τις μεθόδους LASSO και παλινδρόμηση κορυφογραμμής (*ridge regression*) για ένα γραμμικό μοντέλο με δύο επεξηγηματικές μεταβλητές, μπορούμε να συμπεράνουμε το λόγο για τον οποίο η μέθοδος LASSO μπορεί να μηδενίσει συντελεστές αντίθετα με την παλινδρόμηση κορυφογραμμής (*ridge regression*). Οι ελλείψεις που φαίνονται στο σχήμα απεικονίζουν τα ελλειπτικά περιγράμματα της τετραγωνικής συνάρτησης $\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - x_{\boldsymbol{i}}'\boldsymbol{\beta})^{\boldsymbol{2}}$ τα οποία είναι κεντραρισμένα στους *OLS* εκτιμητές. Η λύση που προκύπτει τόσο από την LASSO όσο και από την παλινδρόμηση κορυφογραμμής (*ridge regression*) συμπίπτει με το πρώτο

σημείο που θα ακουμπήσει τον ρόμβο $\| \boldsymbol{\beta} \|_1 < t$ και τον κύκλο $\| \boldsymbol{\beta} \|_2 < t$ αντίστοιχα. Λόγω του ότι ο ρόμβος έχει γωνίες είναι πολύ πιθανό μια λύση να ακουμπήσει σε μια γωνιά έτσι κάποια συντεταγμένη του $\widehat{\boldsymbol{\beta}}$ να μηδενιστεί, κάτι που δεν είναι πολύ πιθανόν για τον κύκλο αφού δεν υπάρχουν γωνιές για να συναντήσουν οι ελλείψεις.



Σχήμα 1.3.2: (a)Γεωμετρία της LASSO για $p = 2$, (b) γεωμετρία της παλινδρόμησης κορυφογραμμής για $p = 2$.

1.3.5 Μέθοδος προσαρμοσμένη (*Adaptive*) LASSO

Ένα αρνητικό που ίσως χαρακτηρίζει την LASSO είναι ότι δημιουργεί μεροληψία εκτίμησης. Το πρόβλημα αυτό επισημάνθηκε πρώτα από τους Fan και Li (2001) και αργότερα από τον Zou (2006). Για να ξεπεραστεί το πρόβλημα αυτό ο Zou (2006) πρότεινε μια προσαρμοσμένη LASSO (Adaptive LASSO), η οποία διαφέρει από την κλασσική LASSO στην συνάρτηση ποινής που χρησιμοποιεί, αφού χρησιμοποιεί μια προσαρμοσμένη σταθμισμένη L_1 −ποινή, την:

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \omega_j |\beta_j|, \tag{1.3.26}$$

όπου $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)$ είναι ένα γνωστό σταθμισμένο διάνυσμα. Ως $\widehat{\omega}_j$ ο Zou πρότεινε το $\widehat{\omega}_j = \frac{1}{|\widehat{\beta}_j|^{\gamma}}$ με $\gamma \geq 0$ και έδειξε ότι η προσαρμοσμένη LASSO μπορεί να λυθεί από τον αλγόριθμο LARS, που προτάθηκε από τους Efron, Hastie, Johnstone και Tibshirani (2004).

1.3.6 SCAD (*Smoothly Clipped Absolute Deviation*) και MCP (*Minimax Concave Penalty*)

Είναι γνωστό ότι οι κυρτές L_q συναρτήσεις ποινής με $q > 1$ δεν ικανοποιούν την ιδιότητα της σποραδικότητας (*sparsity*) , ενώ η L_1 συνάρτηση ποινής δεν ικανοποιεί την ιδιότητα της αμεροληψίας (*unbiasedness*) , και οι κοίλες συναρτήσεις ποινής L_q με $0 \leq q < 1$ δεν ικανοποιούν την ιδιότητα της συνέχειας (*continuity*). Επομένως οι L_q συναρτήσεις ποινής δεν ικανοποιούν και τις τρεις ιδιότητες ταυτόχρονα. Για το λόγο αυτό ο Fan (1997), και οι Fan και Li (2001) πρότειναν την μέθοδο SCAD (*Smoothly Clipped Absolute Deviation*), όπου η παράγωγος της δίνεται από την σχέση:

$$P'_\lambda(\beta_j) = \lambda\left\{I(\beta_j \leq \lambda) + \frac{(\alpha\lambda - \beta_j)_+}{\alpha} I(\beta_j \leq \lambda)\right\}, \ \ \gamma\iota\alpha\ \alpha > 2, \tag{1.3.27}$$

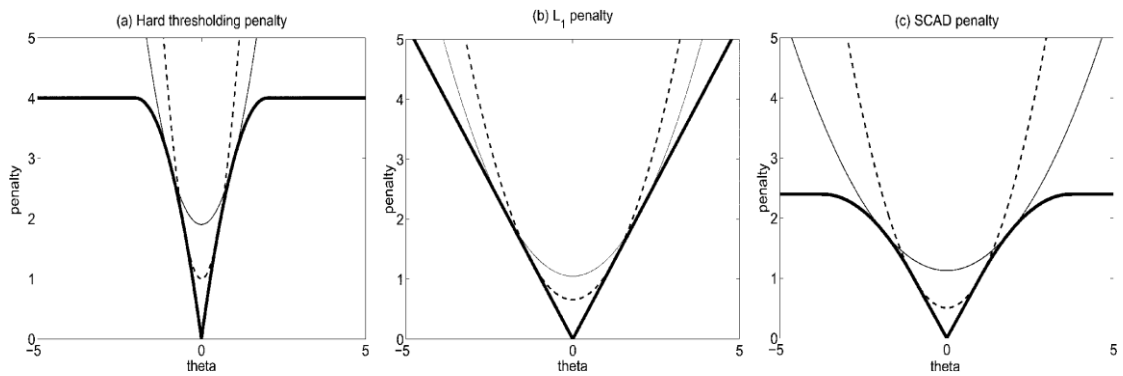
όπου $p_\lambda(0) = 0$ και $\alpha = 3.7$ όπως προέκυψε από την Μπεϋζιανή ανάλυση ρίσκου που εφάρμοσαν οι Fan και Li (2001).

Αυτή η συνάρτηση ποινής δίνει ένα εκτιμητή που ικανοποιεί και τις τρεις προαναφερθείσες ιδιότητες. Πιο συγκεκριμένα ο εκτιμητής δίνεται από την εξίσωση:

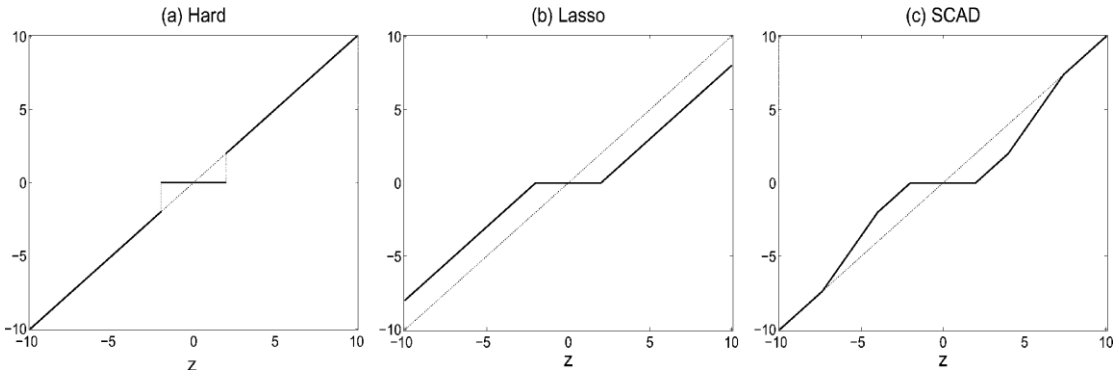
$$\hat{\beta}_{j,SCAD} = \begin{cases} sgn(\hat{\beta}_{j,OLS})(|\hat{\beta}_{j,OLS}| - \lambda)_+, & \gamma\iota\alpha\ |\hat{\beta}_{j,OLS}| \leq 2\lambda \\ \dfrac{\{(\alpha-1)\hat{\beta}_{j,OLS} - sgn(\hat{\beta}_{j,OLS})a\lambda\}}{a-2}, & \gamma\iota\alpha\ 2\lambda < |\hat{\beta}_{j,OLS}| < \alpha\lambda \\ \hat{\beta}_{j,OLS}, & \gamma\iota\alpha\ |\hat{\beta}_{j,OLS}| > \alpha\lambda \end{cases}. \tag{1.3.28}$$

Μια συνάρτηση ποινής που ακολουθεί το ίδιο πνεύμα με την SCAD είναι minimax κοίλη συνάρτηση ποινής (*Minimax Concave penalty* (*MCP*)) την οποία πρότεινε ο Zhang (2009), και δίνεται από τη σχέση:

$$p_\lambda(\beta_j) = \lambda \int_0^{\beta_j} \left(1 - \frac{x}{a\lambda}\right) dx \Rightarrow P'_\lambda(\beta_j) = \frac{(\alpha\lambda - \beta_j)_+}{\alpha}. \tag{1.3.29}$$



Σχήμα 1.3.3: Τρεις συναρτήσεις ποινών $p_\lambda(\boldsymbol{\theta})$ και οι τετραγωνικές τους προσεγγίσεις.



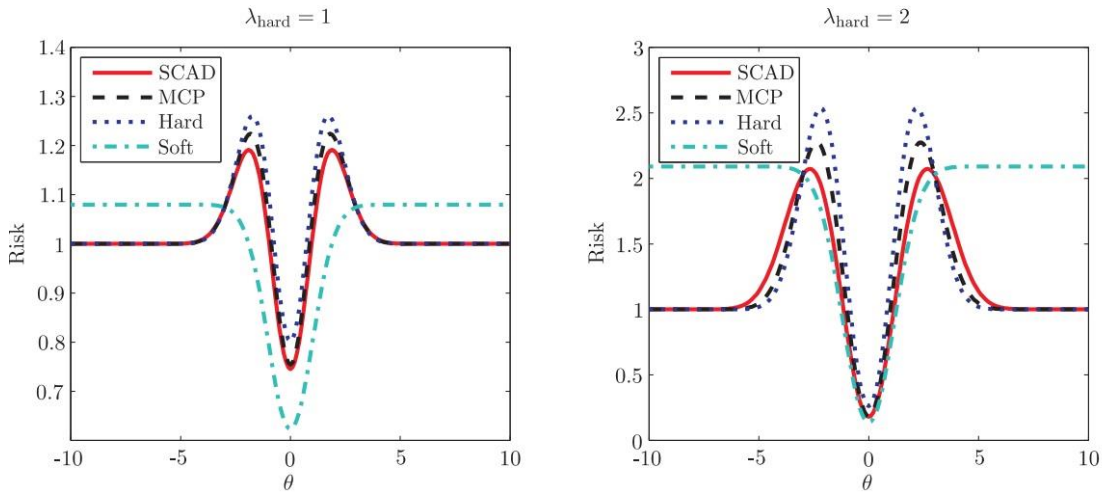
Σχήμα 1.3.4: Γραφικές παραστάσεις των thresholding συναρτήσεων για (α) τη Hard, (β) τη Soft, (γ)και τη SCAD για λ=2 και a=3.7 για την SCAD.

1.3.7 Επίδοση Ποινικοποιημένων Εκτιμητών

Για να μελετήσουν την επίδοση των εκτιμητών οι Jia και Lv (2010) μελετήσαν τις L_2 συναρτήσεις κινδύνου (*risk functions*):

$$R(\theta) = E(\hat{\theta}(\hat{\beta}_{OLS}) - \theta)^2, \tag{1.3.30}$$

για το κλασσικό γραμμικό μοντέλο με $\hat{\beta}_{OLS} \sim N(\theta, 1)$ να είναι η εκτιμήτρια ελαχίστων τετραγώνων (OLS). Πιο συγκεκριμένα θα δούμε την επίδοση των εκτιμητών των μεθόδων LASSO, SCAD, MCP, και hard κανόνα περιορισμού. Για να μετατρέψουν την κλίμακα των κανόνων περιορισμού (thresholding rule) συγκρίσιμη, θεωρήσαν λ=1 και λ=2 για τον hard κανόνα περιορισμού και προσάρμοσαν τις τιμές του λ για τους άλλους κανόνες περιορισμού έτσι ώστε οι εκτιμώμενες τιμές να είναι ίδιες όταν $\theta = 3$. Όπως μπορούμε να διακρίνουμε από το σχήμα (1.3.5), όλοι οι ποινικοποιημένοι εκτιμητές βελτιώνουν τον εκτιμητή ελαχίστων τετραγώνων OLS όταν το θ είναι κοντά στο μηδέν και έχουν την ίδια απόδοση με αυτόν όταν το θ απομακρύνεται από το μηδέν, εκτός από τον εκτιμητή LASSO. Όταν το θ είναι μεγάλο, ο εκτιμητής LASSO έχει μεροληψία ανάλογη με την τιμή του λ και αυτό προκαλεί μεγαλύτερο ρίσκο όπως φαίνεται στο σχήμα (1.3.5 (β)). Για αυτό είναι προτιμότερα να επιλέγονται μικρότερα λ για την μέθοδο αυτή. Όταν $\lambda_{hard} = 2$, ο LASSO εκτιμητής έχει μεγαλύτερο ρίσκο από αυτόν της μεθόδου SCAD, εκτός από κάποιες μικρές περιοχές. Για $\lambda_{hard} = 1$ το πλεονέκτημα της μεθόδου LASSO για θ κοντά στο μηδέν είναι εμφανές (βλέπε σχήμα (1.3.5 (α))). Κατά την επιλογή μεταβλητών σε ένα μοντέλο, όταν το λ επιλέγεται αυτόματα έτσι ώστε να μειωθεί το πρόβλημα της μεροληψίας, η LASSO πρέπει να διαλέξει ένα μικρότερο λ έτσι ώστε να έχει επιθυμητό μέσο τετραγωνικό σφάλμα (βλέπε παρ. (1.3.2)). Μικρότερη τιμή λ συνεπάγεται πιο πολύπλοκο μοντέλο αφού επιλέγονται περισσότερες μεταβλητές. Για αυτό πολλές φορές ο εκτιμητής LASSO τείνει να διατηρεί αρκετές μη σημαντικές μεταβλητές στο επιλεγόμενο μοντέλο.



Σχήμα 1.3.5: Συναρτήσεις κινδύνου (risk functions) ως προς θ για της μεθόδους SCAD, MCP, LASSO (Soft), και Hard.

1.3.8 Επιλογή παραμέτρων για ποινικοποιημένες μεθόδους

Στην ανάλυση που έγινε για τους μεθόδους ποινικοποίησης δεν ορίσαμε ποια πρέπει να είναι η τιμή της παραμέτρου λ και της παραμέτρου α στην περίπτωση των μεθόδων SCAD και MCP. Μια μέθοδος για τον υπολογισμό αυτών των παραμέτρων είναι (k)fold-*cross validation* η οποία είναι γενίκευση του *cross validation.* Η μέθοδος *cross validation* ουσιαστικά χωρίζει τα δεδομένα σε ένα σύνολο εκπαίδευσης (*training set*) και σε ένα σύνολο ελέγχου (*test set*). Εκτιμά το $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, όπου $\boldsymbol{\theta}$ οι παράμετροι προς εκτίμηση, από το σύνολο εκπαίδευσης και στην συνέχεια ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολοίπων ως προς $\boldsymbol{\theta}$ χρησιμοποιώντας τα δεδομένα του συνόλου ελέγχου και τις εκτιμήσεις του συνόλου εκπαίδευσης. Από την ελαχιστοποίηση προκύπτει το $\widehat{\boldsymbol{\theta}}$. Στη συνέχεια περιγράφουμε την μέθοδο (k)*fold-cross validation.*

Έστω $\boldsymbol{\theta}$ οι παράμετροι που αναζητούμε. Έστω όλα τα δεδομένα συμβολίζονται με T. Παίρνουμε από τα δεδομένα k υποσύνολα $T^{(v)}, v = 1, \ldots, k$ τα οποία θεωρούμε ως σύνολα ελέγχου (*test sets*) και ως σύνολα εκπαίδευσης (*training sets*) τα $T - T^{(v)},\ v = 1, \ldots, k$. Για κάθε $\boldsymbol{\theta}$ και v βρίσκουμε τους εκτιμητές $\widehat{\boldsymbol{\beta}}^{(v)}(\theta)$ του $\boldsymbol{\beta}$ χρησιμοποιώντας τα σύνολα εκπαίδευσης. Ορίζουμε ως (k)*fold-cross validation* την ελαχιστοποίηση του αθροίσματος των τετραγώνων των υπολοίπων αθροιζόμενο και τις (k) φορές ως προς την παράμετρο $\boldsymbol{\theta}$. Δηλαδή ελαχιστοποιούμε την παράσταση:

$$CV(\boldsymbol{\Theta}) = \sum_{\boldsymbol{v=1}}^{\boldsymbol{k}} \sum_{(\boldsymbol{x_l}, \boldsymbol{y_l}) \in T^{(v)}} \{y_l - \boldsymbol{x}_l' \widehat{\boldsymbol{\beta}}^{(v)}(\theta)\}^{\boldsymbol{2}}. \qquad (1.3.31)$$

Επομένως επιλέγουμε εκείνο το $\boldsymbol{\theta}$ που ελαχιστοποιεί την πιο πάνω παράσταση.

ΚΕΦΑΛΑΙΟ 2

ΣΙΓΟΥΡΟ ΑΝΕΞΑΡΤΗΤΟ ΚΡΗΣΑΡΙΣΜΑ (*SURE INDEPENDECE SCREENING*)

2.1 Εισαγωγή -Δεδομένα υψηλής Διάστασης (*high dimensional data*)

Ζούμε σε μια εποχή όπου οι δυνατότητες συλλογής και αναπαραγωγής δεδομένων έχουν αυξηθεί ραγδαία. Για παράδειγμα η χρήση του διαδικτύου (*World Wide Web*) ως ένα παγκόσμιο πληροφοριακό σύστημα μας έχει πλημμυρήσει με τεράστιες ποσότητες δεδομένων και πληροφοριών. Λόγω αυτού του γεγονότος υπάρχει ανάγκη εύρεσης νέων τεχνικών ανάλυσης δεδομένων έτσι ώστε τα δεδομένα αυτά να μετατραπούν σε χρήσιμες πληροφορίες.

Ο όρος δεδομένα υψηλής διάστασης (*high dimensional data*) προκύπτει από καταστάσεις όπου ο αριθμός των υποψήφιων μεταβλητών που μπορούν να χρησιμοποιηθούν σε ένα στατιστικό μοντέλο είναι πολύ μεγάλος, και συγκεκριμένα πολύ μεγαλύτερος από το δείγμα που χρησιμοποιείται στην στατιστική ανάλυση. Για παράδειγμα, δεδομένα που αφορούν την κατάσταση υγείας ενός ασθενή μπορεί να είναι υψηλής διάστασης επειδή οι μεταβλητές που πιθανόν να επηρεάζουν την υγεία του ασθενή είναι πάρα πολλές (π.χ. πίεση αίματος, γονίδια, κάπνισμα, αλκοόλ, ασθένειες κοκ) και συνήθως ο αριθμός των ασθενών που επιλέγεται στην ανάλυση δεν μπορεί να ξεπερνά ένα συγκεκριμένο αριθμό n γιατί το υπολογιστικό κόστος θα είναι μεγάλο.

Πολλές παραδοσιακές στατιστικές μέθοδοι επιλογής μεταβλητών έχουν ισχυρό θεωρητικό υπόβαθρο υπό τις υποθέσεις ότι ο αριθμός του δείγματος είναι μεγαλύτερος από τον αριθμό των επεξηγηματικών μεταβλητών ($n > p$) και ότι $n \to \infty$. Όμως η υπόθεση ότι $n \to \infty$ με p σταθερό δεν αντικατοπτρίζει την πραγματικότητα αφού η πιο ρεαλιστική περίπτωση είναι $p \to \infty$ με n σταθερό. Όταν $p > n$ οι παραδοσιακές μέθοδοι μέτριας κλίμακας τείνουν να αποτυγχάνουν. Ακόμα και οι μέθοδοι ποινικοποίησης που αναφέραμε στο κεφάλαιο 1 αποτυγχάνουν όταν η διάσταση p είναι πολύ μεγάλη.

Για την αντιμετώπιση προβλημάτων υψηλής διάστασης οι Fan και Li (2008) πρότειναν την μέθοδο σίγουρου ανεξάρτητου κρησαρίσματος (SIS) καθώς και μια επαναληπτική της μορφή την *Iterative SIS*. Με το όρο κρησάρισμα μεταβλητών εννοούμε το σκαρτάρισμα ασήμαντων επεξηγηματικών μεταβλητών και με τον όρο σίγουρο κρησάρισμα

αναφερόμαστε στην ιδιότητα όπου μετά το κρησάρισμα οι πραγματικά σημαντικές μεταβλητές παραμένουν στο μοντέλο με πιθανότητα να τείνει στο ένα.

2.1.1 Προβλήματα ανάλυσης δεδομένων υψηλής διάστασης

Η ανάλυση δεδομένων υψηλής διάστασης επιφέρει κάποιες προκλήσεις στην ανάλυση τέτοιων δεδομένων οι οποίες είναι οι εξής:

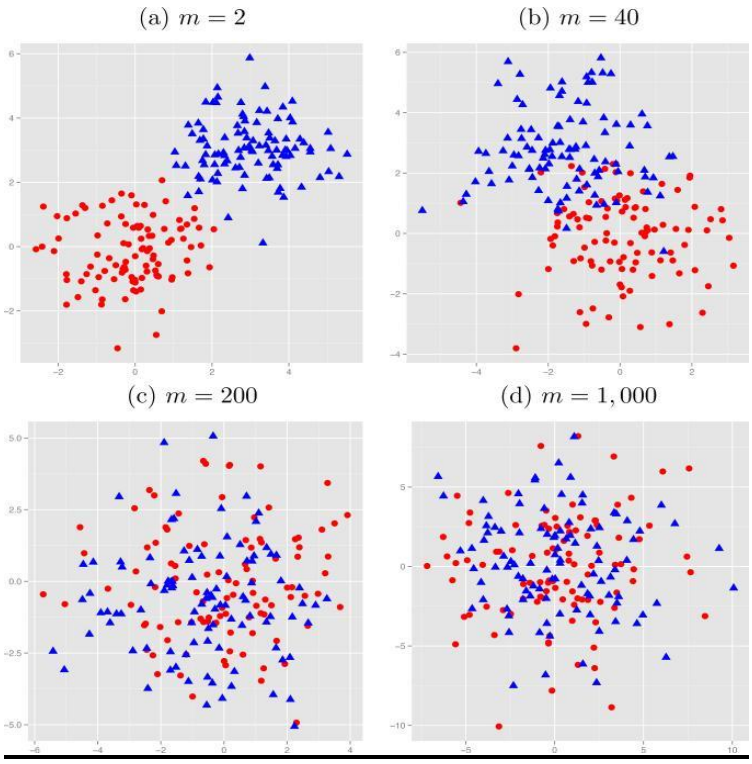
- Συσσώρευση θορύβου (*Noise Accumulation*).
- Ψευδής Συσχέτιση (*Spurious Correlation*).
- Υπολογιστικό Κόστος.

<u>Συσσώρευση θορύβου (*Noise Accumulation*)</u>

Η ανάλυση δεδομένων υψηλής διάστασης απαιτεί πολλές φορές την εκτίμηση και τον έλεγχο πολλών παραμέτρων. Αν ο αριθμός των επεξηγηματικών μεταβλητών είναι τεράστιος, τότε συσσωρεύονται υπολογιστικά σφάλματα τα οποία επηρεάζουν τα επιθυμητά αποτελέσματα. Το αρνητικό αυτό χαρακτηριστικό παρουσιάζεται από τους Han,Liu και Fan (2014) χρησιμοποιώντας ένα παράδειγμα ταξινόμησης. Θεώρησαν ένα δυαδικό πρόβλημα ταξινόμησης χρησιμοποιώντας δύο σύνολα δεδομένων εκπαίδευσης με μέγεθος $n = 100$ και διάσταση $d = 1000$ με:

$$\boldsymbol{X_1}, \boldsymbol{X_2}, \dots \boldsymbol{X}_n \sim N_d(\boldsymbol{\mu_1}, \boldsymbol{I_d}),$$
$$\boldsymbol{Y_1}, \boldsymbol{Y_2}, \dots \boldsymbol{Y}_n \sim N_d(\boldsymbol{\mu_2}, \boldsymbol{I_d}).$$

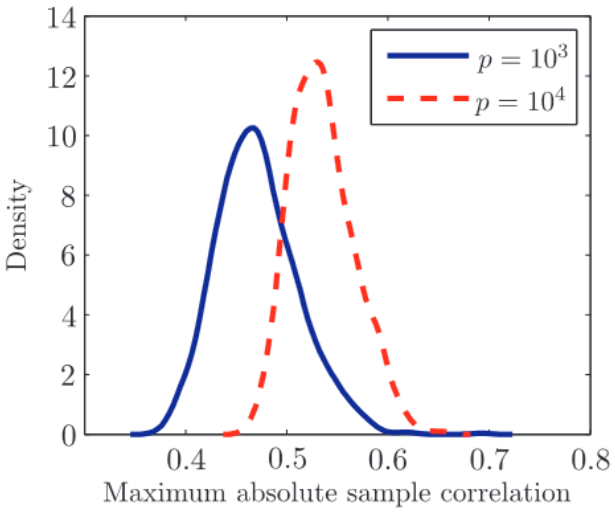
Στην πρώτη κλάση όρισαν $\boldsymbol{\mu_1} = \boldsymbol{0}$ και στην δεύτερη κλάση όρισαν τα πρώτα δέκα χαρακτηριστικά του διανύσματος $\boldsymbol{\mu_2}$ ίσο με 3 και τα υπόλοιπα μηδέν. Ο κανόνας ταξινόμησης καθορίστηκε με κάποιο τρόπο στο δισδιάστατο χώρο όπου οι άξονες αντιστοιχούν στις δύο πρώτες συνιστώσες από τις m πρώτες επεξηγηματικές μεταβλητές όπου $m < d$. Όπως φαίνεται στο σχήμα (2.1.1) για $m = 2$ οι περιπτώσεις των διαφορετικών κλάσεων φαίνονται ξεκάθαρα, όταν όμως το m αυξάνεται, τότε αυξάνεται και το συσσωρευμένο λάθος. Στην περίπτωση που χρησιμοποιήσουμε όλες τις επεξηγηματικές μεταβλητές ($m = p$) τότε οι δύο κλάσεις δεν ξεχωρίζουν καθόλου. Η θεώρηση διακριτών μοντέλων (*sparse models*), δηλαδή η θεώρηση ότι οι συντελεστές μόνο κάποιων συμμεταβλητών δεν είναι μηδέν, αντιμετωπίζει το πιο πάνω πρόβλημα. Επομένως πρέπει να γίνει επιλογή μεταβλητών (*variable selection*) η οποία μπορεί να βελτιώσει την ακρίβεια των εκτιμήσεων επιλέγοντας το σωστό υποσύνολο σημαντικών μεταβλητών. Η επιλογή μεταβλητών αποτελεί πρόκληση για τους στατιστικούς λόγω προβλημάτων όπως η ψευδής συσχέτιση (*spurious correlation*).



Σχήμα 2.1.1: Διάγραμμα διασποράς των προβολών των παρατηρούμενων τιμών (n=100 για κάθε κλάση) πάνω στις δύο κύριες συνιστώσες του m-διάστατου χώρου χαρακτηριστικών. Οι κύκλοι αναφέρονται στην πρώτη κλάση και τα τρίγωνα στην δεύτερη.

<u>**Ψευδής Συσχέτιση (*Spurious Correlation*)**</u>

Όταν υπάρχει έντονη συσχέτιση μεταξύ δύο ή περισσότερων συμμεταβλητών, παρατηρείται το φαινόμενο της πολυσυγραμμικότητας. Όταν έχουμε δεδομένα υψηλής διάστασης, δηλαδή ο αριθμός των συμμεταβλητων είναι πολύ μεγάλος, τότε το φαινόμενο της πολυσυγραμμικότητας γίνεται πολύ έντονο και αυτό οδηγεί στην επιλογή λάθος μοντέλου κατά την επιλογή των επεξηγηματικών μεταβλητών. Οι Fan και Lv (2008) παρουσίασαν το πρόβλημα αυτό δείχνοντας ότι παρόλο που οι συμμεταβλητές μπορεί να είναι ανεξάρτητες, η μέγιστη συσχέτιση δείγματος (*maximum sample correlation),* $\max_{2\le j\le p}\left|corr(X_k, X_j)\right|$ σε σχέση με μια συγκεκριμένη μεταβλητή X_k, μεγαλώνει ανάλογα με τον αριθμό των επεξηγηματικών μεταβλητών (βλέπε σχήμα (2.1.2)). Αυτό έχει ως αποτέλεσμα μια συμμεταβλητή που δεν είναι σημαντική σε ένα μοντέλο μπορεί να έχει υψηλή συσχέτιση με την μεταβλητή απόκρισης εξαιτίας του γεγονότος ότι συσχετίζεται έντονα με άλλες συμμεταβλητές. Λόγω αυτού του προβλήματος, η επιλογή μεταβλητών γίνεται δύσκολη.



Σχήμα 2.1.2: Κατανομή της μέγιστης απόλυτης δειγματικής συσχέτισης συμμεταβλητών $\max_{2 \le j \le p} |corr(X_1, X_j)|$ σε σχέση με την μεταβλητή X_1 για $p = 1000$ και $p = 10000$.
Παρατηρούμαι ότι για μεγαλύτερη διάσταση η συσχέτιση αυξάνεται.

<u>Υπολογιστικό κόστος</u>

Προφανώς όταν υπάρχουν πολλές μεταβλητές σε ένα στατιστικό μοντέλο χρειάζεται μεγαλύτερη υπολογιστική δύναμη. Για παράδειγμα, έστω ότι υπάρχει μια μεταβλητή απόκρισης για την οποία υπάρχει υπόνοια ότι εξαρτάται από k επεξηγηματικές μεταβλητές, όμως ο διαθέσιμος αριθμός μεταβλητών είναι πολύ μεγάλος (p). Ο λογικός τρόπος για να ελεγχθεί ποιο είναι το σωστό υποσύνολο που περιέχει τις k σωστές μεταβλητές, είναι να ελέγξουμε όλα τα πιθανά υποσύνολα και να δούμε ποιο αποδίδει καλύτερα. Επομένως έχουμε να ελέγξουμε $\binom{p}{k}$ υποσύνολα, ο αριθμός των οποίων μεγαλώνει ανάλογα με την τιμή των p και k. Για παράδειγμα αν έχουμε $p = 5000$ με $k = 2$ χρειάζεται να ελεγχθούν 12,5 εκατομμύρια υποσύνολα το οποίο απαιτεί μεγάλη υπολογιστική δύναμη. Επομένως απαιτούνται μέθοδοι που θα μειώσουν την διάσταση έτσι ώστε να μειωθεί το υπολογιστικό κόστος.

Οι Fan και Li (2008) πρότειναν μια μέθοδο σίγουρου κρησαρίσματος και μια παραλλαγή αυτής για την μείωση της διάστασης από p, σε μια σχετικά μεγάλη κλίμακα d, η οποία είναι μικρότερη από το μέγεθος του δείγματος n. Στην συνέχεια για την εύρεση του κατάλληλου μοντέλου εφαρμόζονται μέθοδοι ποινικοποίημενων ελαχίστων τετραγώνων όπως αυτές που είδαμε αναλυτικά στο κεφάλαιο 1.

2.2 Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος (*Sure Independence Screening* (*SIS*))

2.2.1 Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος

Για να αντιμετωπίσουν τα προβλήματα που προκύπτουν από δεδομένα υψηλής διάστασης, οι Fan και Lv (2008) πρότειναν μια μέθοδο που μειώνει την διάσταση του χώρου των επεξηγηματικών μεταβλητών, την μέθοδο σίγουρου ανεξάρτητου κρησαρίσματος (*Sure Independent Screening* (*SIS*)) η οποία βασίζεται στην μάθηση συσχέτισης (*correlation learning*). Η μέθοδος αυτή απομακρύνει τα χαρακτηριστικά που έχουν ασθενή συσχέτιση με την μεταβλητή απόκρισης. Πήρε το όνομα σίγουρου ανεξάρτητου κρησαρίσματος (*Sure Independent Screening* (*SIS*)) επειδή κάθε επεξηγηματική μεταβλητή χρησιμοποιείται ανεξάρτητα για την εύρεση της συσχέτισης με την απόκριση, εφαρμόζοντας παλινδρόμηση για την εκάστοτε μεταβλητή. Με τον όρο σίγουρο κρησάρισμα αναφερόμαστε στην ιδιότητα όπου όλες οι σημαντικές επεξηγηματικές μεταβλητές παραμένουν στο μοντέλο μετά την εφαρμογή της μεθόδου, με πιθανότητα να τείνει στο 1. Μια μέθοδος μείωσης διάστασης είναι καλή αν και μόνο αν έχει την ιδιότητα του σίγουρου κρησαρίσματος.

Έστω το γραμμικό μοντέλο:

$$Y = \sum_{i=1}^{\boldsymbol{p}} \beta_i X_i + \varepsilon, \qquad (2.2.1)$$

με $\boldsymbol{x} = (X_1, \dots, X_p)'$ το διάνυσμα συμμεταβλητών και με πίνακα διασποράς-συνδιασπορας $\boldsymbol{\Sigma} = cov(\boldsymbol{x})$. Θεωρούμε ότι $X_1, \dots, X_p$ είναι τυποποιημένες τυχαίες μεταβλητές με μέση τιμή 0 και διασπορά 1.

Η ιδέα της μεθόδου αυτής είναι πολύ απλή. Αρχικά θεωρούμε ότι το διάνυσμα $\boldsymbol{\beta}$ είναι αραιό (*sparse vector*), δηλαδή μόνο ένας συγκεκριμένος αριθμός από όλες τις μεταβλητές συνεισφέρει στο μοντέλο. Έστω $M = \{1 \leq i \leq p: \beta_\iota \neq 0\}$ είναι το αραιό μοντέλο με τις πραγματικά σημαντικές μεταβλητές με μέγεθος $s = |M|$. Οι υπόλοιπες $p - s$ μεταβλητές που δεν ανήκουν στο M μπορεί να συσχετίζονται με την απόκριση μέσω συσχετίσεων με τις μεταβλητές που ανήκουν στο M.
Για να δούμε τη συσχέτιση μεταξύ μεταβλητών και απόκρισης εκτελούμε κατά συνιστώσες παλινδρόμηση (*componentwise regression*) με την μεταβλητή απόκρισης. Έστω λοιπόν το διάνυσμα $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ που προκύπτει από την κατά συνιστώσες παλινδρόμηση είναι:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}, \qquad (2.2.2)$$

θεωρώντας ότι οι στήλες του πίνακα σχεδιασμού $\boldsymbol{X}$ είναι τυποποιημένες με μέση τιμή 0 και διασπορά 1. Για οποιοδήποτε $\gamma \in (0,1)$, ταξινομούμε τις απόλυτες τιμές των εκτιμήσεων που πήραμε από την κατά συνιστώσες παλινδρόμηση, κατά φθίνουσα σειρά και θεωρούμε ότι συνεισφέρουν στο μοντέλο οι πρώτες $[\gamma \times n]$ μεταβλητές. Δηλαδή ουσιαστικά θεωρούμαι ότι συνεισφέρουν στο μοντέλο οι μεταβλητές που έχουν την μεγαλύτερη συσχέτιση με την μεταβλητή απόκριση. Αυτό προκύπτει αν παρατηρήσουμε

ότι το διάνυσμα $\widehat{\boldsymbol{\beta}}$ μας δίνει την δειγματική συσχέτιση Pearson μεταξύ των συμμεταβλητών και της απόκρισης, πολλαπλασιασμένη με την τυπική απόκλιση της μεταβλητής απόκρισης. Πράγματι:

$$\widehat{cor}(X_i, Y) = \frac{\widehat{cov}(X_i, Y)}{\hat{\sigma}_{X_i}\hat{\sigma}_Y} \approx \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{(n-1)\hat{\sigma}_Y} = \frac{\hat{\beta}_i}{(n-1)\hat{\sigma}_Y}$$

$$=> \hat{\beta}_i \approx \widehat{cor}(X_i, Y)\hat{\sigma}_Y(n-1). \qquad (2.2.3)$$

Αυτός είναι ο λόγος που η μέθοδος σίγουρου ανεξάρτητου κρησαρίσματος SIS ονομάζεται και μάθηση συσχέτισης (*correlation learning)*. Συνεχίζοντας έχουμε:

$$\widehat{M} = \{1 \leq j \leq p\colon \ |\hat{\beta}_j| \ \textit{που είναι στις} \ [\gamma \times n] \ \textit{μεγαλύτερες εκτιμήσεις}\},$$

όπου $[\gamma \times n]$ το ακέραιο μέρος του $\gamma \times n$. Με αυτό τον τρόπο η διάσταση μειώνεται αρκετά και γίνεται ανάλογη των παρατηρούμενων τιμών n. Έτσι αντιμετωπίζονται διάφορα προβλήματα που προκύπτουν από τα υψηλής διάστασης δεδομένα. Το υπολογιστικό κόστος της SIS είναι αρκετά ικανοποιητικό αφού περιορίζεται στον πολλαπλασιασμό ενός πίνακα $p \times n$ με ένα διάνυσμα $n \times 1$ και στην επιλογή των μεγαλύτερων κατά απόλυτη τιμή $[\gamma \times n]$ συνιστωσών ενός διανύσματος $p \times 1$. Επομένως η υπολογιστική πολυπλοκότητα της μεθόδου είναι $O(np)$.

2.2.2 Iteratively Thresholded Ridge Regression Screener (ITRRS)

Η μέθοδος Iteratively Thresholded Ridge Regression Screener (ITRRS) είναι μια επέκταση της μεθόδου SIS η οποία βοηθά στην περαιτέρω κατανόηση της μάθησης συσχέτισης. Γνωρίζουμε ότι όταν ο αριθμός των επεξηγηματικών μεταβλητών είναι μεγαλύτερος από τον αριθμό των παρατηρήσεων τότε ο εκτιμητής ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}} = (\boldsymbol{X}'\boldsymbol{X})^{+}\boldsymbol{X}'\boldsymbol{y}$ **,** όπου $(.)^{+}$ είναι ο Moore-Penrose γενικευμένος αντίστροφος, δεν είναι ακριβής και δίνει εσφαλμένα αποτελέσματα. Όπως είδαμε στο κεφάλαιο 1 ο εκτιμητής παλινδρόμησης κορυφογραμμής είναι:

$$\widehat{\boldsymbol{\beta}}(r) = (\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}, \qquad (2.2.4)$$

όπου $r > 0$ η παράμετρος ποινικοποίησης. Όταν $r \longrightarrow 0$, ισχύει ότι $\widehat{\boldsymbol{\beta}}(r) \longrightarrow \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}$ και επομένως ο εκτιμητής δίνει εσφαλμένα αποτελέσματα. Επομένως είναι λογικό να επιλέξουμε μια παράμετρο r αρκετά μεγάλη έτσι ώστε να μειωθεί η διασπορά του εκτιμητή (βλέπε κεφάλαιο 1). Όταν $r \longrightarrow \infty$, ισχύει ότι $r\widehat{\boldsymbol{\beta}}(r) \longrightarrow \boldsymbol{X}'\boldsymbol{y}$ που αποτελεί τον κατά συνιστώσες εκτιμητή παλινδρόμησης που χρησιμοποιείται από την SIS. Η μέθοδος ITRRS χρησιμοποιεί τον εκτιμητή παλινδρόμησης κορυφογραμμής και επιλέγει υποσύνολα με τις μεγαλύτερες εκτιμήσεις κατά απόλυτη τιμή που προκύπτουν μέχρις ότου το υποσύνολο που θα επιλεχθεί να έχει λιγότερες επεξηγηματικές μεταβλητές από τις παρατηρήσεις.

Πιο συγκεκριμένα για δοθέν $\delta \in (0,1)$, ταξινομούμε τις συνιστώσες του $\widehat{\boldsymbol{\beta}}(r)$ σε φθίνουσα σειρά και ορίζουμε το υποσύνολο:

$$M_{\delta,r}^{1} = \left\{1 \leq i \leq p: \left|\hat{\beta}_i(r)\right| \textit{ που είναι στις } [\delta \times p] \textit{ μεγαλύτερες εκτιμήσεις } \right\}.$$

Στη συνέχεια εφαρμόζουμε την ίδια διαδικασία για το μοντέλο $M_{\delta,r}^{1}$ και παίρνουμε το υποσύνολο $M_{\delta,r}^{2} \subset M_{\delta,r}^{1}$ το οποίο περιέχει $[\delta^2 \times p]$ μεταβλητές. Συνεχίζεται η ίδια διαδικασία μέχρις ότου να επιλεχθεί το υποσύνολο $M_{\delta,,r} \subset M_{\delta,r}^{k}$ με πληθυκότητα $d = [\delta^k \times p] < n$ με $[\delta^{k-1} \times p] \geq n$. Ο αριθμός των επαναλήψεων εξαρτάται από την επιλογή του δ.

Μπορεί να δειχθεί ότι η ITRRS έχει την ιδιότητα σίγουρου κρησαρίσματος κάτω από κάποιες υποθέσεις και με συγκριμένες επιλογές των γ και δ.

2.2.3 Ιδιότητα Σίγουρου Κρησαρίσματος της SIS

Οι Fan και Lv (2008) απέδειξαν ότι υπό κάποιες συγκεκριμένες συνθήκες η μέθοδος SIS ικανοποιεί την ιδιότητα σίγουρου κρησαρίσματος, δηλαδή μετά την εφαρμογή της μεθόδου. το υποσύνολο μεταβλητών που προκύπτει περιέχει όλες τις σημαντικές μεταβλητές με πιθανότητα να τείνει στο ένα.

Πιο συγκεκριμένα, έστω $M = \{1 \leq i \leq p: \beta_\iota \neq 0\}$ είναι το αραιό μοντέλο με τις πραγματικά σημαντικές μεταβλητές, με μέγεθος $s = |M|$ και $\widehat{M} = \left\{1 \leq j \leq p: \left|\beta_j\right| \textit{ που είναι στις } [\gamma \times n] \textit{μεγαλύτερες εκτιμήσεις με } \gamma \in (0,1)\right\}$ το σύνολο που προκύπτει από την κατά συνιστώσες παλινδρόμηση. Οι Fan και Lv (2008) απέδειξαν ότι $P(M \subset \widehat{M}) \xrightarrow{n \to \infty} 1$ υπό κάποιες προϋποθέσεις.

Έστω το γραμμικό μοντέλο:

$$Y = \sum_{i=1}^{p} \beta_i X_i + \varepsilon, \tag{2.2.5}$$

και θεωρούμε τον μετασχηματισμό $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ όπου $\boldsymbol{x} = (X_1, \dots, X_p)'$ το διάνυσμα συμμεταβλητών και $\boldsymbol{\Sigma} = cov(\boldsymbol{x})$. Προφανώς ο πίνακας διασποράς-συνδιασποράς του $\boldsymbol{z}$ είναι ο $\boldsymbol{I}_p$. Επίσης θεωρούμε το μετασχηματισμένο πίνακα σχεδιασμού $\boldsymbol{Z} = \boldsymbol{X\Sigma}^{-1/2}$. Θεωρούμε ότι $X_1, \dots, X_p$ είναι τυποποιημένες τυχαίες μεταβλητές με μέση τιμή 0 και διασπορά 1.

<u>Ορισμός:</u> Έστω $\lambda_{max}(.), \lambda_{min}(.)$ η μέγιστη και η ελάχιστη ιδιοτιμή ενός πίνακα αντίστοιχα. Θα λέμε ότι ένα τυχαίος πίνακας $\mathbf{Z}$ έχει την ιδιότητα συγκέντρωσης (concentration property) αν υπάρχουν $c, c_1 > 1$ *και* $C_1 > 0$ έτσι ώστε:

$$P\left\{\lambda_{max}\left(\tilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}'\right) > c_1 \text{ ή } \lambda_{min}\left(\tilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}'\right) \leq 1/c_1\right\} \leq ex\,p(-C_1 n), \tag{2.2.6}$$

για κάθε $n \times \tilde{p}$ πίνακα $\widetilde{\mathbf{Z}}$ του $\mathbf{Z}$ με $cn < \tilde{p} \leq p$.

<u>Θεώρημα:</u> Έστω ότι ισχύουν τα εξής:

1. $p > n$ και $log\, p = O(n^{\xi})$ για κάποιο $\xi \in (0,1-2\kappa)$ όπου το κ δίνεται από το (3.).

2. $\varepsilon \sim N(0, \sigma^2)$ για $\sigma > 0$, το p-διάστατο διάνυσμα συμμεταβλητών $\mathbf{x}$ έχει συμμετρική σφαιρική κατανομή και ο τυχαίος πίνακας $\mathbf{X\Sigma^{-1/2}}$έχει την ιδιότητα συγκέντρωσης (concentration property) που ισχύει για γκαουσιανές κατανομές, όπου $\boldsymbol{\Sigma} = cov(\boldsymbol{x})$.

3. $var(Y) = O(1)$, και για κάποιο $\kappa \geq 0$, $c_2, c_3 > 0$,

$$(i)\ \min_{j\in M}|\beta_j| \geq \frac{c_2}{n^{\kappa}}\ \ \kappa\alpha\iota\ \ (ii)\ \min_{j\in M}|cov(\beta_j^{-1}Y, X_j)| \geq\ c_3.$$

4. $\lambda_{max}(\boldsymbol{\Sigma}) = c_4 n^{\tau}$, $\tau \geq 0, c_4 > 0$.

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

$$P(M \subset \widehat{M}) = 1 - O[exp\{-Cn^{1-2k}/log(n)\}]. \qquad (2.2.7)$$

Για την απόδειξη του θεωρήματος βλέπε Fan και Lv (2008). Το θεώρημα αυτό υποδηλώνει ότι η μέθοδος SIS έχει την ιδιότητα σίγουρου κρησαρίσματος και μπορεί να μειώσει την εκθετικά αυξανόμενη διάσταση p σε μια σχετικά μεγάλη κλίμακα $d = [\gamma \times n] = O\left(n^{1-\theta}\right) < n$, για κάποιο $\theta > 0$. Αν οι συνθήκες (1.)-(4.) ισχύουν τότε μπορούμε να θεωρήσουμε ότι το $d = n - 1$ ή $d = n/\log(n)$.

Η συνθήκη 3(i) καθορίζει το ποσοστό της πιθανότητας λάθους, στο υποσύνολο που θα επιλεγεί να μην υπάρχουν οι πραγματικά σημαντικές μεταβλητές. Η συνθήκη 3(ii) υποθέτει ότι για να ικανοποιείται η ιδιότητα σίγουρου κρησαρίσματος η οριακή συσχέτιση κατά απόλυτη τιμή μεταξύ επεξηγηματικών μεταβλητών και απόκρισης δεν είναι μικρή. Όταν όμως αυτή η συνθήκη δεν ισχύει, δηλαδή όταν μια μεταβλητή είναι οριακά ασυσχέτιστη με την μεταβλητή απόκρισης λόγω των συσχετίσεων με άλλες μεταβλητές, τότε η ιδιότητα σίγουρου κρησαρίσματος δεν ισχύει, κάτι που θα δούμε και αργότερα. Η συνθήκη (4.) προσπαθεί να αποκλείσει το πρόβλημα της πολυσυγραμμικότητας.

2.2.4 Τεχνικές επιλογές μοντέλου βασισμένες στη SIS

Εφαρμόζοντας την μέθοδο SIS επιτυγχάνουμε την μείωση της διάστασης του συνόλου των επεξηγηματικών μεταβλητών με την σημαντική ιδιότητα του σίγουρου κρησαρίσματος. Η μέθοδος αυτή συνδυάζεται με μεθόδους ποινικοποίησης έτσι ώστε να επιταχύνει την επιλογή χαρακτηριστικών (*variable selection*), όταν ο αριθμός των επεξηγηματικών μεταβλητών p είναι πολύ μεγαλύτερος από τον αριθμό των παρατηρήσεων. Παραδείγματα μεθόδων ποινικοποίημενων ελάχιστων τετραγώνων που χρησιμοποιούνται με την SIS είναι η LASSO, η SCAD, και MCP τις οποίες είδαμε εκτενώς στο κεφάλαιο 1.

Σχήμα 2.2.1: Γίνεται ένα μεγάλο κρησάρισμα μεταβλητών για μεγάλης διάστασης επεξηγηματικών μεταβλητών με την μέθοδο SIS, και στη συνέχεια για μέτρια διάσταση μεταβλητών (ανάλογη του n) εφαρμόζονται μέθοδοι μέτριας κλίμακας για περαιτέρω επιλογή σημαντικών μεταβλητών όπως φαίνεται στο σχήμα.

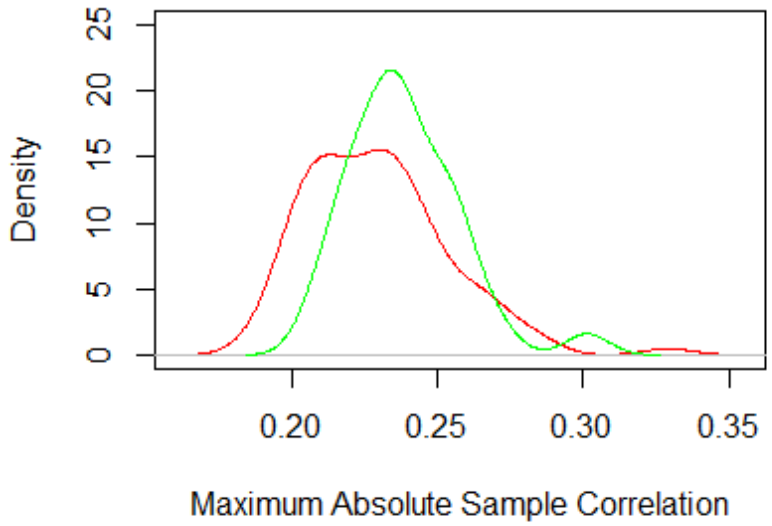
Προσομοιωμένο Παράδειγμα (2.2.1):

Θα προσομοιώσουμε ένα γραμμικό μοντέλο όπου οι επεξηγηματικές μεταβλητές είναι ανεξάρτητες και ακολουθούν την τυποποιημένη κανονική κατανομή με $\varepsilon \sim N(0,1.5)$. Θεωρούμε δύο μοντέλα με $(n,p) = (200,500)$ και $(n,p) = (200,1000)$. Επίσης θεωρούμε ότι ο αριθμός των πραγματικών μεταβλητών στα δύο μοντέλα είναι 3 και 5 αντίστοιχα. Για να δώσουμε τιμές στους συντελεστές των σημαντικών μεταβλητών ακολουθούμε πανομοιότυπη διαδικασία με αυτήν που ακολούθησαν οι Fan και Li (2008). Θεωρούμε $a = \frac{4\log(n)}{\sqrt{n}}$ και $\frac{5\log(n)}{\sqrt{n}}$ αντίστοιχα για τα δύο μοντέλα και επιλέγουμε μη αρνητικούς συντελεστές της μορφής $(-1)^u(a + |z|)$, όπου το u ακολουθεί την κατανομή Bernoulli με παράμετρο 0.3 και το z ακολουθεί την τυποποιημένη κανονική κατανομή. Εκτελέσαμε 100 προσομοιώσεις. Από το σχήμα (2.2.2) παρατηρούμε πως ακόμα και αν οι μεταβλητές μας είναι ανεξάρτητες η συσχέτιση μεταξύ των μεταβλητών (στο σχήμα βλέπουμε $\max_{2\le j\le p}|corr(X_1, X_j)|$) δεν είναι αμελητέα και μεγαλώνει καθώς ο αριθμός των μεταβλητών μεγαλώνει. Η εκτίμηση του διακριτού διανύσματος $\boldsymbol{\beta}$, αφού πρώτα εφαρμόσαμε την SIS και έγινε κρησάρισμα μεταβλητών, έγινε με τρεις διαφορετικές μεθόδους, την LASSO, τη SCAD,και την MCP και τις συμβολίζουμε ως SIS-LASSO, SIS-SCAD και SIS-MCP αντίστοιχα. Τα αποτελέσματα φαίνονται στο πίνακα (2.2.1).

Πίνακας 2.2.1: Αποτελέσματα Προσομοίωσης: ο μέσος αριθμός των επιλεγόμενων μεταβλητών και το μέσο εκτιμώμενο σφάλμα $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ μέσω της L_2 νόρμας.

		$SIS - SCAD$	$SIS - MCP$	$SIS - LASSO$
$p = 500$	Median model size	5.94	5.18	6.73
	$\|\hat{\beta} - \beta\|$	0.20	0.17	0.80
$p = 1000$	Median model size	8.40	7.65	11.45
	$\|\hat{\beta} - \beta\|$	0.26	0.25	0.83

Στο πίνακα παρουσιάζονται ο μέσος αριθμός των επιλεγόμενων μεταβλητών καθώς και το μέσο εκτιμώμενο σφάλμα $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ μέσω της L_2 νόρμας. Παρατηρούμε ότι τόσο στην περίπτωση όπου $p = 500$, όσο και στην περίπτωση όπου $p = 1000$, η $SIS - MCP$ φαίνεται να αποδίδει καλύτερα από τις άλλες δύο μεθόδους, έχοντας το μικρότερο μέσο εκτιμώμενο σφάλμα με μικρή διαφορά από την $SIS - SCAD$. Η μέθοδος $SIS - LASSO$ έχει προφανώς την χειρότερη απόδοση.



Σχήμα 2.2.2: Κατανομή της μέγιστης απόλυτης δειγματικής συσχέτισης συμμεταβλητών $\max_{2 \le j \le p} |corr(X_1, X_j)|$ σε σχέση με την μεταβλητή X_1 για $p = 500$ και $p = 1000$. Παρατηρούμε ότι για μεγαλύτερη διάσταση η συσχέτιση ανεβαίνει.

2.2.3 Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος για ταξινόμηση *(Sure Independent Screening (SIS) for classification)*

Η μέθοδος SIS χρησιμοποιείται κατά κόρον και σε προβλήματα ταξινόμησης τα οποία μπορεί να θεωρηθούν ως ειδικές περιπτώσεις προβλημάτων παλινδρόμησης όπου η απόκριση παίρνει διακριτές τιμές, π.χ. ± 1. Για παράδειγμα όταν θέλουμε να ταξινομήσουμε αν ένας δανειολήπτης είναι αξιόπιστος ή όχι, ανάλογα με διάφορα χαρακτηριστικά (επεξηγηματικές μεταβλητές) όπως ο μηνιαίος μισθός, η περιουσία, η ηλικία κοκ, μπορούμε να θεωρήσουμε ένα ειδικό πρόβλημα παλινδρόμησης με την μεταβλητή απόκρισης να παίρνει την τιμή 1 όταν είναι αξιόπιστος και την τιμή -1 όταν δεν είναι. Όμως ο χώρος των πιθανών χαρακτηριστικών που θα καθορίζουν την αξιοπιστία του δανειολήπτη μπορεί να είναι πολύ μεγάλος, δηλαδή p πολύ μεγάλο, με αποτέλεσμα να αντιμετωπίσουμε προβλήματα συσσώρευσης θορύβου όπως είδαμε στο κεφάλαιο (2.1.1). Οπότε είναι φυσιολογικό να μειώσουμε την διάσταση του προβλήματος μας.

Έστω λοιπόν ότι έχουμε n_1 δανειολήπτες (κλάση 1) που είναι αξιόπιστοι και n_2 που δεν είναι (κλάση 2). Εφαρμόζοντας την κατά συνιστώσες παλινδρόμηση όπως είδαμε στο κεφάλαιο (2.2) τότε θα έχουμε:

$$\boldsymbol{\omega} = \sum_{Y_i=1} \boldsymbol{x_i} - \sum_{Y_i=-1} \boldsymbol{x_i}\,, \tag{2.2.8}$$

το οποίο μπορεί να γραφτεί και ως:

$$\omega_{\boldsymbol{j}} = (n_1\bar{X}_{j,1} - n_2\bar{X}_{j,2})/(\tau\upsilon\pi\iota\kappa\acute{\eta}\ \alpha\pi\acute{o}\kappa\lambda\iota\sigma\eta\ j - o\sigma\tau o\acute{\upsilon}\ \chi\alpha\rho\alpha\kappa\tau\eta\rho\iota\sigma\tau\iota\kappa o\acute{\upsilon}),$$

όπου $\bar{X}_{j,1}$ είναι μέσος του δείγματος του j −χαρακτηριστικού κλάσης 1 και $\bar{X}_{j,1}$ είναι μέσος του δείγματος του j −χαρακτηριστικού κλάσης -1. Αν $n_1 = n_2$ τότε παίρνουμε το *two-sample statistic test* το οποίο χρησιμοποιείται συχνά για την επιλογή χαρακτηριστικών σε προβλήματα μεγάλης διάστασης. Εφαρμόζοντας την μέθοδο SIS όπως την είδαμε στο κεφάλαιο (2.2) μπορούμε να μειώσουμε τον αριθμό των πιθανών επεξηγηματικών μεταβλητών σημαντικά.

2.2.4 Προβλήματα της μεθόδου SIS

Η μέθοδος SIS χρησιμοποιεί μόνο τις οριακές πληροφορίες συσχέτισης των συμμεταβλητών με την απόκριση και έτσι η ιδιότητα σίγουρου κρησαρίσματος μπορεί να αποτύχει όταν κάποιες τεχνικές συνθήκες δεν ικανοποιούνται (βλέπε κεφ. (2.2.1)).

Τρία πιθανά προβλήματα που μπορεί να αντιμετωπίσουμε με την μέθοδο SIS είναι τα εξής:

1. Μια σημαντική επεξηγηματική μεταβλητή η οποία είναι οριακά ασυσχέτιστη με την μεταβλητή απόκρισης λόγω των συσχετίσεων που έχει με τις άλλες σημαντικές μεταβλητές, έχει μικρότερη προτεραιότητα να επιλεχθεί από την SIS. Για παράδειγμα, έστω το διάνυσμα συμμεταβλητών $\boldsymbol{x} = (X_1, \dots, X_p)$' ακολουθεί την τυποποιημένη κανονική κατανομή και η συσχέτιση των μεταβλητών ανά δύο ισούται με ρ. Θεωρούμε το γραμμικό μοντέλο:

   $$Y = \boldsymbol{x}'\boldsymbol{\beta} = X_1 + \cdots + X_j - J\rho X_{j+1}.$$

   Προφανώς η μεταβλητή X_{j+1} είναι ανεξάρτητο του Y όμως η σταθερά παλινδρόμησης $-J\rho$ μπορεί να είναι μεγαλύτερη κατά απόλυτη τιμή από τις άλλες σταθερές. Τέτοια σημαντική μεταβλητή, δεν μπορεί να επιλεγεί από την μέθοδο SIS.

2. Ασήμαντες μεταβλητές οι οποίες έχουν υψηλή συσχέτιση με σημαντικές μεταβλητές έχουν μεγαλύτερη προτεραιότητα να επιλεχθούν από την SIS σε σχέση με μια σημαντική μεταβλητή που σχετίζεται οριακά με την απόκριση. Για παράδειγμα έστω:

   $$Y = \rho X_0 + \cdots + X_j + \varepsilon,$$

όπου η X_0 είναι ανεξάρτητη με τις υπόλοιπες συμμεταβλητές οι οποίες ανά δύο έχουν συσχέτιση ρ. Επομένως $cov(X_j, Y) = J\rho = Jcov(X_0, Y)$ για $j = j + 1, \dots, p$ οπότε η X_0 έχει μικρότερη πιθανότητα να επιλεχθεί από την μέθοδο SIS αφού έχει μικρότερη συσχέτιση με την μεταβλητή απόκρισης σε σχέση με τις ασήμαντες μεταβλητές.

3. Η πολυσυγγραμικότητα μεταξύ των συμμεταβλητών δημιουργεί μεγάλο πρόβλημα στην επιλογή μεταβλητών.

Το πρόβλημα της πολυσυγραμμικότητας μπορεί να αντιμετωπιστεί μετασχηματίζοντας τις μεταβλητές που έχουν υψηλή συσχέτιση. Για παράδειγμα αν ένα μοντέλο περιέχει τρεις επεξηγηματικές μεταβλητές οι οποίες δηλώνουν το βάρος κάποιου ατόμου ανάλογα με την ηλικία του, π.χ για 5, 10, 15 χρονών , προφανώς τα τρία αυτά βάρη θα συσχετίζονται αρκετά μεταξύ τους. Αν αντί τα βάρη σε αυτές τις ηλικίες θεωρήσουμε τις αλλαγές του βάρους του ατόμου σε κάθε ηλικία, οι οποίες εκτιμώνται εύκολα με ένα γραμμικό μετασχηματισμό των βαρών, τότε είναι πιθανόν αυτές οι νέες μεταβλητές να έχουν πιο μικρή συσχέτιση μεταξύ τους. Ένας άλλος τρόπος για να αντιμετωπιστεί το πρόβλημα της πολυσυγραμμικότητας είναι ο χωρισμός των μεταβλητών με υψηλή συσχέτιση σε ομάδες και η εφαρμογή στατιστικής ανάλυσης στις συνιστώσες των ομάδων για να δημιουργηθούν νέες επεξηγηματικές μεταβλητές με χαμηλές συσχετίσεις. Αυτές οι νέες επεξηγηματικές μεταβλητές που προκύπτουν από κάθε ομάδα είναι οι νέες συμμεταβλητές στις οποίες μπορεί να εφαρμοστεί με επιτυχία η μέθοδος σίγουρου ανεξάρτητου κρησαρίσματος.

Για την αντιμετώπιση αυτών των προβλημάτων οι Fan και Li (2008) πρότειναν μια παραλλαγή της μεθόδου SIS, την λεγόμενη επαναληπτική μέθοδο σίγουρου ανεξάρτητου κρησαρίσματος (*Iterative Sure Independence Screening*) την οποία θα δούμε στο αμέσως επόμενο κεφάλαιο.

2.3 Επαναληπτική Μέθοδος Σίγουρου Ανεξάρτητου Κρησαρίσματος (*Iterative Sure Independence Screening* (*ISIS*))

Η ISIS είναι μια επαναληπτική εφαρμογή της μεθόδου SIS για την επιλογή μεταβλητών. Η ιδέα της μεθόδου αυτής είναι να εφαρμοστεί επαναληπτικά ένα κρησάρισμα μεταβλητών μεγάλης κλίμακας, ακολουθούμενο από μιας μέτριας κλίμακας επιλογής μεταβλητών, χρησιμοποιώντας μια από τις μεθόδους που αναφέραμε στο κεφάλαιο 1.

Η ISIS ακολουθεί τα εξής βήματα:

1. Αρχικά επιλέγουμε ένα υποσύνολο κ_1 μεταβλητών, $A_1 = (X_{i_1}, X_{i_2}, \dots, X_{i_{\kappa_1}})$, χρησιμοποιώντας την SIS, επιλέγοντας αρχικά $d_1 = n/\log(n)$ μεταβλητές. Στη συνέχεια με μια μέθοδο μέτριας κλίμακας επιλογής μεταβλητών όπως η SCAD ή η LASSO, παραμένουν στο μοντέλο κ_1 μεταβλητές.

2. Εκτελούμε παλινδρόμηση μεταξύ της απόκρισης Y και των μεταβλητών που επιλέχθηκαν στο βήμα (1.) και θεωρούμε τα υπόλοιπα που προκύπτουν από αυτή την παλινδρόμηση ως τις νέες μας αποκρίσεις. Εφαρμόζουμε τώρα το βήμα (1.) για τις υπόλοιπες $p - \kappa_1$ μεταβλητές και λαμβάνουμε ένα δεύτερο υποσύνολο κ_2 μεταβλητών, $A_2 = (X_{j_1}, X_{j_2}, \dots, X_{j_{\kappa_2}})$.

3. Συνεχίζουμε την ίδια διαδικασία μέχρι να αποκτήσουμε l υποσύνολα, $A_1, \dots, A_l$ των οποίων η ένωση τους, $A = \bigcup_{i=1}^{l} A_i$, έχει μέγεθος d, το οποίο είναι μικρότερο του μεγέθους του δείγματος n.

4. Αφού βρούμε το A μπορούμε πάλι να εφαρμόσουμε μια μέτριας κλίμακας μέθοδο (SCAD,LASSO κ.λπ.) για να γίνει η τελική επιλογή του μοντέλου.

Στο βήμα (2.) η προσαρμογή των υπολοίπων στις υπόλοιπες $p - \kappa_1$ μεταβλητές έχει ως αποτέλεσμα την αντιμετώπιση των προβλημάτων της μεθόδου SIS που αναφέραμε στο κεφάλαιο (2.2.4). Πιο συγκεκριμένα, το βήμα αυτό μπορεί να εξασθενήσει τη προτεραιότητα επιλογής μη σημαντικών μεταβλητών λόγω των συσχετίσεων με τις σημαντικές μεταβλητές, δεδομένου ότι τα υπόλοιπα που χρησιμοποιούνται είναι ασυσχέτιστα με τις μεταβλητές του A_i. Δηλαδή μια μεταβλητή από τις εναπομείναντες $p - \kappa_i$ που έχει υψηλή συσχέτιση με μια από τις επιλεγμένες μεταβλητές του A_i και θα επιλεγόταν λόγω αυτού του το γεγονότος από την SIS, τώρα δεν θα επιλεχθεί γιατί η συσχέτιση αυτή εξαφανίζεται χρησιμοποιώντας τα υπόλοιπα. Επίσης σημαντικές μεταβλητές που είχαν χαθεί σε προηγούμενο βήμα και είναι οριακά ασυσχέτιστες με την απόκριση λόγω των συσχετίσεων τους με άλλες σημαντικές συμμεταβλητές, έχουν προτεραιότητα να επιλεχθούν στο βήμα (2.). Πιο συγκεκριμένα όταν οι μεταβλητές εισαχθούν στο A_i, οι μεταβλητές που είναι οριακά ασθενώς ασυσχέτιστες με την Y εξαιτίας της παρουσίας μεταβλητών στο A_i θα πρέπει να συσχετίζονται με τα υπόλοιπα. Έτσι κατά την εφαρμογή του κρησαρίσματος μεταξύ των υπολοίπων και των εναπομείναντα επεξηγηματικών μεταβλητών, οι μεταβλητές που αρχικά ήταν οριακά ασυσχέτιστες με την απόκριση πλέον έχουν μεγάλη πιθανότητα να επιλεχθούν από την μέθοδο κρησαρίσματος SIS.

Στο κεφάλαιο 5 γίνεται εφαρμογή της επαναληπτικής μεθόδου SIS για διάφορα στατιστικά μοντέλα.

ΚΕΦΑΛΑΙΟ 3

ΠΡΟΒΟΛΗ ΚΑΝΟΝΙΚΩΝ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ ΣΕ ΥΨΗΛΗΣ ΔΙΑΣΤΑΣΗΣ ΔΕΔΟΜΕΝΑ ΓΙΑ ΚΡΗΣΑΡΙΣΜΑ ΜΕΤΑΒΛΗΤΩΝ (*HIGH DIMENSIONAL ORDINARY LEAST SQUARES PROJECTION FOR SCREENING VARIABLES*)

3.1 Εισαγωγή

Η επιλογή μεταβλητών (*variable selection*) σε εφαρμογές της στατιστικής γίνεται αρκετά δύσκολη όταν ο αριθμός των πιθανών επεξηγηματικών μεταβλητών ξεπερνά κατά πολύ τον αριθμό των παρατηρήσεων. Στο κεφάλαιο 2 αναφερθήκαμε στη μέθοδο Σίγουρου Ανεξάρτητου Κρησαρίσματος (*SIS*) η οποία μειώνει δραστικά την διάσταση του προβλήματος και διατηρεί τις πραγματικά σημαντικές μεταβλητές με μεγάλη πιθανότητα. Η ιδέα όμως της SIS βασίζεται στην υπόθεση ότι η συσχέτιση μεταξύ απόκρισης και μεταβλητών είναι μεγάλη, κάτι που δεν ισχύει απαραίτητα. Σε δεδομένα υψηλής διάστασης, ασήμαντες μεταβλητές που έχουν υψηλή συσχέτιση με την απόκριση, λόγω της συσχέτισης τους με σημαντικές μεταβλητές, έχουν μεγαλύτερη προτεραιότητα να επιλεχθούν σε σχέση με κάποιες άλλες σημαντικές μεταβλητές. Επίσης σημαντικές μεταβλητές που έχουν ασθενή συσχέτιση με την απόκριση λόγω των συσχετίσεων τους με τις άλλες σημαντικές μεταβλητές αποτυγχάνουν να επιλεχθούν από την μέθοδο SIS. Οι Whang και Leng (2015) πρότειναν μια νέα μέθοδο κρησαρίσματος την Προβολή Κανονικών Ελάχιστων Τετραγώνων σε Υψηλής Διάστασης Δεδομένα *(High Dimensional Ordinary Least Square Screening (HOLP)*) η οποία δεν υποθέτει την υψηλή συσχέτιση μεταξύ απόκρισης και μεταβλητών, όπως η μέθοδος SIS. Η μέθοδος αυτή βασίζεται στην ιδιότητα αυστηρής διαγώνιας υπεροχής που χαρακτηρίζει τους πίνακας. Οι Whang και Leng (2015) απέδειξαν ότι κάτω από παρόμοιες υποθέσεις με την SIS, η μέθοδος HOLP έχει την ιδιότητα του σίγουρου κρησαρίσματος.

3.2 Προβολή Κανονικών Ελάχιστων σε Υψηλής Διάστασης Δεδομένα (*High Dimensional Ordinary Least Square Screening* (*HOLP*))

3.2.1 Εισαγωγικά στοιχεία

Ορισμός (Moore-Penrose Inverse): Για $A \in R^{m\times n}$, ένας Moore-Penrose ψευδό-αντίστροφος πίνακας ορίζεται να είναι ένας πίνακας $A^{+} \in R^{n\times m}$ έτσι ώστε:

$$AA^{+}A = A,\ A^{+}AA^{+} = A^{+}, (AA^{+})^{*} = AA^{+}, (A^{+}A)^{*} = A^{+}A,$$

όπου A^{*} είναι ο συζευγμένος του A.

Ορισμός (Stiefel Manifold): Ο χώρος που αποτελείτε από $p \times n$ πίνακες έτσι ώστε $\boldsymbol{X'X} = \boldsymbol{I_n}$ ονομάζεται χώρος Stiefel και συμβολίζεται ως εξής:

$$V_{n,p} = \{\boldsymbol{X} \in R^{p\times n}: \boldsymbol{X'X} = \boldsymbol{I_n}\}.$$

Έστω το γραμμικό μοντέλο:

$$Y = \sum_{i=1}^{\boldsymbol{p}} \beta_i X_i + \varepsilon, \tag{3.2.1}$$

όπου $\boldsymbol{x} = (X_1, \dots, X_p)'$ το διάνυσμα συμμεταβλητών και $\boldsymbol{\Sigma} = cov(\boldsymbol{x})$ ο πίνακας διασποράς-συνδιασποράς. Θεωρούμε ότι ο πίνακας $\boldsymbol{X'X}$ είναι αντιστρέψιμος για $n > p$ και ο πίνακας $\boldsymbol{XX'}$ είναι αντιστρέψιμος για $n < p$. Έστω $M = \{X_1, \dots, X_p\}$ είναι το μοντέλο που περιέχει όλες τις πιθανές μεταβλητές και M_s είναι το αραιό μοντέλο με τις πραγματικά σημαντικές μεταβλητές με δείκτες $S = \{\{1 \le i \le p: \beta_\iota \neq 0\}$ και πληθυκότητα $s = |S|$. Μια γενική μορφή εκτιμητών του $\boldsymbol{\beta}$ είναι:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{Ay}, \tag{3.2.2}$$

όπου $\boldsymbol{A} \in \boldsymbol{R}^{p\times n}$. Παρατηρούμε ότι για $\boldsymbol{A} = \boldsymbol{X'}$ παίρνουμε τον εκτιμητή της μεθόδου SIS. Επειδή σκοπός μιας μεθόδου κρησαρίσματος είναι να διατηρήσει της σημαντικές μεταβλητές στο μοντέλο, πρέπει ο εκτιμητής $\widehat{\boldsymbol{\beta}}$ του $\boldsymbol{\beta}$ όχι να είναι ακριβής αλλά να διατηρεί την διάταξη των $\boldsymbol{\beta}$ κατά απόλυτη τιμή έτσι ώστε οι μη μηδενικές τιμές του $\boldsymbol{\beta}$ να έχουν σχετικά μεγάλη τιμή στο $\widehat{\boldsymbol{\beta}}$. Δηλαδή για παράδειγμα αν οι συνιστώσες του πραγματικού $\boldsymbol{\beta}$ κατά απόλυτη τιμή σε φθίνουσα σειρά είναι $(|\beta_3|, |\beta_5|, |\beta_1|, \dots, |\beta_p|)$, θα θέλαμε οι εκτιμήσεις να προκύψουν με την ίδια σειρά δηλαδή $(|\hat{\beta}_3|, |\hat{\beta}_5|, |\hat{\beta}_1|, \dots, |\hat{\beta}_P|)$.

Παρατηρούμαι ότι :

$$\boldsymbol{Ay} = \boldsymbol{A}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = (\boldsymbol{AX})\boldsymbol{\beta} + \boldsymbol{A\varepsilon}, \tag{3.2.3}$$

όπου το $\boldsymbol{A\varepsilon}$ αποτελείτε από ένα γραμμικό συνδυασμό των ε_i με μέση τιμή μηδέν. Το $\boldsymbol{AX}$ αποτελεί το συστηματικό μέρος, εννοώντας ότι παρέχει τις συστηματικές πληροφορίες που χρειαζόμαστε.

Θα επιθυμούσαμε το συστηματικό μέρος να είναι μεγαλύτερο από το κομμάτι του σφάλματος για να πάρουμε όσο μεγαλύτερη πληροφορία μπορούμε. Μια κατάλληλη επιλογή του $\boldsymbol{A}$ για να το πετύχουμε αυτό, είναι να η επιλογή του να είναι τέτοια έτσι ώστε $\boldsymbol{AX} = \boldsymbol{I}$. Επομένως επιλέγοντας ως $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, ισχύει $\boldsymbol{AX} = \boldsymbol{I}$ και προφανώς ο εκτιμητής του $\boldsymbol{\beta}$ προκύπτει να είναι ο γνωστός μας εκτιμητής ελαχίστων τετραγώνων. Όμως ο $\boldsymbol{X}'\boldsymbol{X}$ είναι αντιστρέψιμος μόνο όταν ο αριθμός των παρατηρήσεων είναι μικρότερος από την διάσταση του χώρου των συμμεταβλητών ($p < n$), οπότε σε δεδομένα υψηλής διάστασης είναι ακατάλληλος.

Για δεδομένα όπου $p > n$ οι Whang και Leng (2015) πρότειναν την χρησιμοποίηση ενός γενικευμένου αντίστροφου του πίνακα σχεδιασμού $\boldsymbol{X}$, τον αντίστροφο πίνακα Moore-Penrose. Αποδεικνύεται ότι ο αντίστροφος Moore-Penrose του $\boldsymbol{X}$ ισούται με $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ όταν $p < n$ και με $\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}$ όταν $p > n$. Αν λοιπόν επιλέξουμε $\boldsymbol{A} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}$ προφανώς ο $\boldsymbol{AX}$ δεν μας δίνει το μοναδιαίο πίνακα, παρόλα αυτά αποδεικνύεται ότι ο $\boldsymbol{AX}$ χαρακτηρίζεται από την ιδιότητα της αυστηρά διαγώνιας υπεροχής. Το γεγονός αυτό έχει ως αποτέλεσμα τα $\hat{\beta}_i \in S$ να εκμεταλλεύονται τους μεγάλους όρους στην διαγώνιο για να υπερνικούν τα $\hat{\beta}_i$ που δεν ανήκουν στο S τα οποία είναι απλά γραμμικοί συνδυασμοί των μη διαγώνιων όρων του πίνακα $\boldsymbol{AX}$.

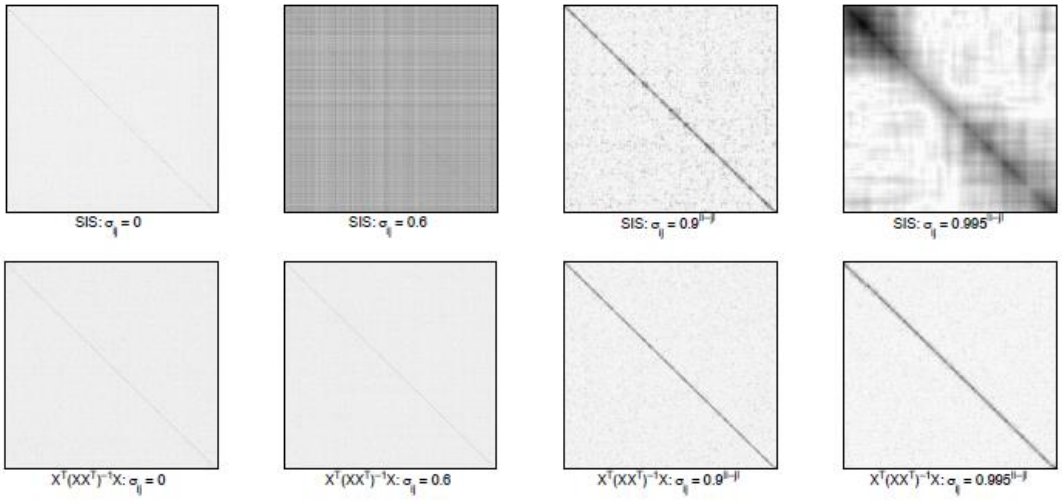
Για να γίνει κατανοητό όμως γιατί ο πίνακας $\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}$ έχει αυστηρή διαγώνια υπεροχή θεωρούμε ένα αποτέλεσμα της γραμμικής άλγεβρας, την αποσύνθεση μοναδική τιμής (*Singular Values Decomposition*) που λέει ότι ο πίνακας $\boldsymbol{X}$ μπορεί να διασπαστεί σε γινόμενο τριών πινάκων, δηλαδή $\boldsymbol{X} = \boldsymbol{VDU}'$, όπου $\boldsymbol{V}$ είναι ένας $n \times n$ ορθογωναδιαίος πίνακάς με $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$, $\boldsymbol{D}$ είναι ένας $n \times n$ διαγώνιος πίνακα και $\boldsymbol{U}$ είναι ένας $p \times n$ με $\boldsymbol{U} \in V_{n,p}$. Τότε αποδεικνύεται εύκολα ότι:

$$\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}'. \tag{3.2.4}$$

Ο πίνακας $\boldsymbol{U}\boldsymbol{U}'$ αποδεικνύεται ότι έχει αυστηρά διαγώνια υπεροχή με μεγάλη πιθανότητα. Στο σχήμα (3.2.1) εμφανίζονται οι θερμικοί χάρτες του πίνακα $\boldsymbol{X}'\boldsymbol{X}$, ο οποίος χρησιμοποιείται από τη SIS, και του πίνακα $\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}$. Δηλαδή εμφανίζουμε οπτικά που βρίσκονται τα μεγαλύτερα στοιχεία των πινάκων. Θεωρούμε ότι έχουμε n= 50 παρατηρήσεις και $p = 1000$ επεξηγηματικές μεταβλητές με $\boldsymbol{x} = (X_1, \ldots, X_{1000}) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Έστω ότι ο πίνακας διασπορά-συνδιασποράς για $i, j = 1,2,3, \ldots, p$, ικανοποιεί τα εξής:

$$(i)\ \boldsymbol{\Sigma} = \boldsymbol{I_p}, \quad (ii)\sigma_{ii} = 1, \sigma_{ij} = 0.6, i \neq j, \quad (iii)\ \sigma_{ij} = 0.9^{|i-j|}, \quad (iv)\ \sigma_{ij} = 0.995^{|i-j|}.$$

Όπως μπορούμε να δούμε από το σχήμα (3.2.1) ο πίνακας $\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}$ έχει αυστηρή διαγώνια υπεροχή σε όλες τις περιπτώσεις κάτι που δεν ισχύει για τον πίνακα $\boldsymbol{X}'\boldsymbol{X}$ οπότε ο πίνακας $\boldsymbol{A} = \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}$ μπορεί να χρησιμοποιηθεί στον εκτιμητή κρησαρίσματος τον οποίο θα ορίσουμε πιο κάτω.



Σχήμα 3.2.1: Θερμικοί χάρτες του πίνακα $\boldsymbol{X'X}$, και του πίνακα $\boldsymbol{X'(XX')^{-1}X}$ για διαφορετικά στοιχεία του πίνακα διασποράς και συνδιασποράς. Το σκούρο χρώμα υποδηλώνει τα μεγαλύτερα σε μέγεθος στοιχεία του πίνακα.

3.2.2 Ορισμός Μεθόδου HOLP

Έστω $M = \{X_1, \dots, X_p\}$ είναι το μοντέλο με όλες τις πιθανές μεταβλητές και M_s είναι το αραιό μοντέλο που περιέχει τις πραγματικά σημαντικές μεταβλητές με δείκτες $S = \{\{1 \leq i \leq p: \beta_\iota \neq 0\}$ και πληθυκότητα $s = |S|$. Θεωρούμε τον ακόλουθο εκτιμητή για το διάνυσμα $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{X'(XX')^{-1}y}. \tag{3.2.5}$$

Ο εκτιμητής αυτός ονομάζεται εκτιμητής προβολής κανονικών ελάχιστων τετραγώνων σε υψηλής διάστασης δεδομένα (*High Dimensional Ordinary Least Square Screening (HOLP)*). Ο εκτιμητής αυτός έχει σκοπό να μειώσει την διάσταση και όχι να υπολογίσει με ακρίβεια το $\boldsymbol{\beta}$. Ακολουθούμε την ίδια διαδικασία με την μέθοδο SIS, δηλαδή επιλέγουμε τις $d < n$ μεγαλύτερες εκτιμήσεις κατά απόλυτη τιμή από το $\widehat{\boldsymbol{\beta}}$ για να παραμείνουν στο μοντέλο. Επομένως επιλέγουμε το υποσύνολο:

$$\widehat{M}_d = \{1 \leq j \leq p\colon \left|\hat{\beta}_j\right| \textit{ που είναι στις } d \textit{ μεγαλύτερες εκτιμήσεις}\}, \tag{3.2.6}$$

ή ισοδύναμα:

$$\widehat{M}_d = \{1 \leq j \leq p\colon \left|\hat{\beta}_j\right| > \gamma \textit{ για κάπιο } \gamma > 0\}. \tag{3.2.7}$$

Παρατηρούμε ότι:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{X'(XX')^{-1}X\beta} + \boldsymbol{X'(XX')^{-1}\varepsilon}, \tag{3.2.8}$$

και βλέποντας τον πρώτο όρο βλέπουμε γιατί ο εκτιμητής $\widehat{\boldsymbol{\beta}}$ θεωρείται προβολή του $\boldsymbol{\beta}$.

Την μέθοδο HOLP μπορούμε να την δούμε και από μια άλλη οπτική όπως ακριβώς και την μέθοδο SIS. Θεωρούμε τον εκτιμητή παλινδρόμησης κορυφογραμμής με παράμετρο r ο οποίος ισούται με:

$$\widehat{\boldsymbol{\beta}}(r) = (\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{3.2.9}$$

Όταν $r \longrightarrow \infty$, ισχύει ότι $r\widehat{\boldsymbol{\beta}}(r) \longrightarrow \boldsymbol{X}'\boldsymbol{y}$ το οποίο όπως είδαμε στο κεφάλαιο 2 αποτελεί τον κατά συνιστώσες εκτιμητή παλινδρόμησης που χρησιμοποιείται από την μέθοδο SIS.

<u>Πρόταση:</u> Ισχύει η πιο κάτω ανίσωση:

$$\widehat{\boldsymbol{\beta}}(r) = (\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{I_p})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{I_n})^{-1}\boldsymbol{y}. \tag{3.2.10}$$

<u>Απόδειξη:</u> Εφαρμόζουμε την ισότητα των Sherman-Morrison-Woodbury η οποία είναι η εξής:

$$(\boldsymbol{A} + \boldsymbol{UDV})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{D}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}. \tag{3.2.11}$$

Επομένως εφαρμόζοντας την για $\boldsymbol{A} = \frac{\boldsymbol{I_p}}{\boldsymbol{r}}, \boldsymbol{U} = \boldsymbol{X}', \boldsymbol{V} = \boldsymbol{X}$ έχουμε:

$$r(r\boldsymbol{I}_p + \boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{I_p} - \boldsymbol{X}'\left(\boldsymbol{I_n} + \frac{1}{r}\boldsymbol{X}\boldsymbol{X}'\right)^{-1}\boldsymbol{X}\frac{1}{r} = \boldsymbol{I_p} - \boldsymbol{X}'(r\boldsymbol{I_n} + \boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X} \tag{3.2.12}$$

Πολλαπλασιάζοντας και τα δύο μέρη με $\boldsymbol{X}'\boldsymbol{y}$ έχουμε:

$$r(r\boldsymbol{I}_p + \boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'(r\boldsymbol{I_n} + \boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{y}. \tag{3.2.13}$$

Το δεξιό μέλος ισούται με:

$$\begin{aligned} &\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'(r\boldsymbol{I_n} + \boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{y} = \\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'(r\boldsymbol{I_n} + \boldsymbol{X}\boldsymbol{X}')^{-1}((r\boldsymbol{I_n} + \boldsymbol{X}\boldsymbol{X}') - r\boldsymbol{I_n})\boldsymbol{y} \\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} + r(r\boldsymbol{I_n} + \boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{y} \\ &= r\boldsymbol{X}'(r\boldsymbol{I_n} + \boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{y}. \end{aligned} \tag{3.2.14}$$

Και άρα:

$$(\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{I_p})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X} + r\boldsymbol{I_n})^{-1}\boldsymbol{y}.$$

■

Για $r \longrightarrow 0$, ισχύει:

$$\widehat{\boldsymbol{\beta}}(r) \longrightarrow \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{y}, \tag{3.2.15}$$

το οποίο αποτελεί τον εκτιμητή HOLP. Άρα ο εκτιμητής αυτός μπορεί να θεωρηθεί ως μια ακραία μορφή του εκτιμητή παλινδρόμησης κορυφογραμμής στέλνοντας το r στο μηδέν αντί στο άπειρο όπως είδαμε στην μέθοδο SIS.

Σχετικά με το υπολογιστικό κόστος της μεθόδου, η υπολογιστική πολυπλοκότητα της μεθόδου είναι $O(n^2p)$ ενώ της SIS ήταν $O(np)$ επομένως σε δεδομένα υψηλής διάστασης η HOLP έχει λίγο μεγαλύτερο υπολογιστικό κόστος από ότι η SIS.

3.2.3 Ιδιότητα Σίγουρου Κρησαρίσματος της HOLP

Έστω το γραμμικό μοντέλο:

$$Y = \sum_{i=1}^{p} \beta_i X_i + \varepsilon, \tag{3.2.16}$$

και θεωρούμε τον μετασχηματισμό $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$, όπου $\boldsymbol{x} = (X_1, \dots, X_p)'$ το διάνυσμα συμμεταβλητών και $\boldsymbol{\Sigma} = cov(\boldsymbol{x})$. Προφανώς ο πίνακας διασποράς-συνδιασποράς του $\boldsymbol{z}$ είναι ο $\boldsymbol{I}_P$. Επίσης θεωρούμε το μετασχηματισμένο πίνακα σχεδιασμού $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$. Θεωρούμε ότι $X_1, \dots, X_p$ είναι τυποποιημένες τυχαίες μεταβλητές με μέση τιμή 0 και διασπορά 1.

<u>Ορισμός:</u> Μια κατανομή F με μηδενική μέση τιμή έχει μια $q - εκθετική\ ουρά$ αν οποιεσδήποτε $n \geq 1$ ανεξάρτητες τυχαίες μεταβλητές $\varepsilon_i \sim F$ ικανοποιούν για οποιοδήποτε $\alpha \in R^n$ με $\|\alpha\|_2 = 1$ την ανισότητα:

$$P\left(\left|\sum_{i=1}^{n} a_i \varepsilon_i\right| > t\right) \leq \exp\big(1 - q(t)\big), \tag{3.2.17}$$

για οποιοδήποτε $t > 0$ και για κάποια συνάρτηση $q(.)$.Σημειώνουμε ότι αν $\varepsilon_i \sim N(0,1) \Rightarrow \sum_{i=1}^{n} a_i \varepsilon_i \sim N(0,1)$ και μπορεί να δειχθεί ότι για $q(t) = t^2/2$ η κανονική κατανομή έχει τετραγωνική εκθετική ουρά.

Θεωρούμε τώρα τις πιο κάτω υποθέσεις:

1. Το μετασχηματισμένο διάνυσμα μεταβλητών $\mathbf{z}$ έχει σφαιρική συμμετρική κατανομή και υπάρχουν $c_1 > 1$ και $C_1 > 0$ έτσι ώστε:

   $$P\{\lambda_{max}(p^{-1}\boldsymbol{Z}\boldsymbol{Z}') > c_1 \text{ ή } \lambda_{min}(p^{-1}\boldsymbol{Z}\boldsymbol{Z}') \leq 1/c_1\} \leq ex\,p(-C_1 n),$$

   όπου $\lambda_{\max}(.), \lambda_{\min}(.)$ η μέγιστη και η ελάχιστη ιδιοτιμή ενός πίνακα αντίστοιχα και $p > c_0 n$ για καποιοι $c_0 > 1$.

2. Η τυχαία μεταβλητή ε έχει μηδενική μέση τιμή και τυπική απόκλιση ίση με $\sigma > 0$, ανεξάρτητη από το p-διάστατο διάνυσμα συμμεταβλητών $\boldsymbol{x}$. Το τυποποιημένο ε/σ έχει $q - εκθετική$ ουρά για κάποια συνάρτηση $q(.)$.

3. $var(Y) = O(1)$, και για κάποιο $\kappa \geq 0, , \tau \geq 0, \nu \geq 0$ και $c_2, c_3, c_4 > 0$,

   $$(i) \min_{j \in S} |\beta_j| \geq \frac{c_2}{n^{\kappa}}\ , (ii)\ s \leq c_3 n^{\nu}\ και\ (iii) cond(\boldsymbol{\Sigma}) \leq c_4 n^{\tau},$$

   όπου $cond(\boldsymbol{\Sigma}) = \lambda_{max}(\Sigma)/\lambda_{min}(\Sigma)$.

Παρατηρούμαι ότι οι υποθέσεις είναι παρόμοιες με αυτές της μεθόδου σίγουρου ανεξάρτητου κρησαρίσματος SIS, με μία ουσιαστική διαφορά. Η υπόθεση:

$$\min_{j\in S}\left|cov(\beta_j^{-1}Y, X_j)\right| \geq c_3\,,\ c_3 > 0, \tag{3.2.18}$$

η οποία υποθέτει την υψηλή συσχέτιση μεταξύ επεξηγηματικών μεταβλητών και απόκρισης δεν χρειάζεται πλέον, επομένως η μέθοδος HOLP αποδεικνύεται ότι έχει την ιδιότητα σίγουρου κρησαρίσματος με λιγότερες υποθέσεις από ότι η SIS.

<u>Θεώρημα 1:</u> Υπό τις υποθέσεις (1.)-(3.), αν διαλέξουμε γ_n έτσι ώστε:

$$\frac{p\gamma_n}{n^{1-\tau-\kappa}} \xrightarrow{n\to\infty} 0 \ \kappa\alpha\iota\ \frac{p\gamma_n\sqrt{\log(n)}}{n^{1-\tau-\kappa}} \xrightarrow{n\to\infty} \infty,$$

για το $C_1 > 0$ που ορίζεται από την υπόθεση (1.) ισχύει:

$$\begin{aligned} &\mathrm{P}\left(M_S \subset \widehat{M}_{\gamma_n}\right) = \\ &= 1 - O\left\{\exp\left(-C_1\frac{n^{1-2\kappa-5\tau-\nu}}{2\log(n)}\right)\right\} - s\cdot exp\left\{1 - q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\} \\ &+3s\exp(-C_1 n). \end{aligned} \tag{3.2.19}$$

Παρατηρούμε ότι δεν γίνεται καμία υπόθεση για το p αρκεί το $p > c_0 n$.

<u>Θεώρημα 2:</u> Αν ισχύουν οι υποθέσεις του θεωρήματος 1 και για το p ισχύει:

$$logp = o(min\left\{\frac{n^{1-2\kappa-5\tau}}{2\log n}, q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\},$$

τότε για το ίδιο γ_n που ορίζεται στο θεώρημα 1 και για το ίδιο C_1 που καθορίζεται από την υπόθεση (1.) έχουμε:

$$\begin{aligned} &(i)\quad P\left(\min_{j\in S}\left|\hat{\beta}_j\right| > \gamma_n > \max_{j\notin S}\left|\hat{\beta}_j\right|\right) = \\ &= 1 - O\left\{\exp\left\{-C_1\frac{n^{1-2\kappa-5\tau-\nu}}{2\log(n)}\right\} - exp\left\{1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\}\right\}. \end{aligned} \tag{3.2.20}$$

Ισοδύναμα, μπορούμε να επιλέξουμε ένα υποσύνολο M_d με $d = n^{\iota}$ με $\iota \in (\nu, 1]$ έτσι ώστε:

$$\begin{aligned} &(ii)\quad P(M_S \subset M_d) = \\ &= 1 - O\left\{\exp\left\{-C_1\frac{n^{1-2\kappa-5\tau-\nu}}{2\log(n)}\right\} - exp\left\{1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\}\right\}. \end{aligned} \tag{3.2.21}$$

Σημειώνουμε ότι το (i) αναφέρει ότι αν ο αριθμός των επεξηγηματικών μεταβλητών ικανοποιεί την συνθήκη (1.), τότε οι σημαντικές μεταβλητές διαχωρίζονται από τις μη σημαντικές από μια σταθερά γ_n, με μεγάλη πιθανότητα. Ισοδύναμα το (ii) αναφέρει ότι με μεγάλη πιθανότητα αν επιλέξουμε ένα υποσύνολο από το σύνολο όλων των πιθανών επεξηγηματικών μεταβλητών με διάσταση μικρότερη του αριθμού των παρατηρήσεων τότε αυτό θα περιέχει όλες τις σημαντικές μεταβλητές. Αν $d = s$ τότε η HOLP επιλέγει τις πραγματικά σημαντικά μεταβλητές πάλι με πολύ μεγάλη πιθανότητα. Για τις αποδείξεις των πιο πάνω θεωρημάτων βλέπε παράρτημα 1.

3.3 Τεχνικές επιλογές μοντέλου βασισμένες στη HOLP

Όπως και στη μέθοδο SIS έτσι και στη μέθοδο HOLP μετά το κρησάρισμα των μεταβλητών μπορούμε να εφαρμόσουμε επιλογή μεταβλητών με διάφορες ποινικοποιημένες μεθόδους που αναλύσαμε στο κεφάλαιο 1, όπως η Lasso, η SCAD, και η MCP.

<u>Προσομοιωμένο Παράδειγμα (3.3.1):</u>

Χρησιμοποιούμε τα ίδια δεδομένα με το προσομοιωμένο παράδειγμα (2.2.1) στο κεφάλαιο (2.2.4). Υπενθυμίζουμε ότι θεωρούμε ένα γραμμικό μοντέλο όπου οι επεξηγηματικές μεταβλητές είναι ανεξάρτητες και ακολουθούν την τυποποιημένη κανονική κατανομή. Επίσης θεωρούμε ότι $\varepsilon \sim N(0,1.5)$. Προσομοιώνουμε δύο μοντέλα, $(n,p) = (200,500)$ και $(n,p) = (200,1000)$ όπου ο αριθμός των πραγματικών μεταβλητών είναι 3 και 5 αντίστοιχα και οι συντελεστές αυτών ορίζονται να είναι της μορφής $(-1)^u(a + |z|)$. Το $a = \frac{4\log(n)}{\sqrt{n}}$ και $\frac{5\log(n)}{\sqrt{n}}$ αντίστοιχα για τα δύο μοντέλα, το u ακολουθεί την κατανομή Bernoulli με παράμετρο 0.4 και το z ακολουθεί την τυποποιημένη κανονική κατανομή. Εφαρμόζουμε τις μεθόδους HOLP-SCAD, HOLP-MCP, και HOLP-LASSO και επαναλαμβάνουμε το πείραμα 100 φορές. Τα αποτελέσματα φαίνονται στο πίνακα (3.3.1).

Πίνακας 3.3.1: Αποτελέσματα Προσομοίωσης: ο μέσος αριθμός των επιλεγόμενων μεταβλητών και το μέσο εκτιμώμενο σφάλμα $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ μέσω της $l2$ νόρμας.

		$\boldsymbol{HOLP-SCAD}$	$\boldsymbol{HOLP-MCP}$	$\boldsymbol{HOLP-LASSO}$
$\boldsymbol{p = 500}$	Median model size	7.00	6.54	8.19
	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.19	0.19	0.71
$\boldsymbol{p = 1000}$	Median model size	7.29	7.59	11.5
	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$	0.26	0.24	0.82

Παρατηρούμε πανομοιότυπα αποτελέσματα με αυτά της μεθόδου SIS όπως είδαμε στο κεφάλαιο (2.2.4), στα οποία είναι ξεκάθαρο ότι οι μέθοδοι HOLP-SCAD και HOLP-MCP έχουν πολύ καλύτερη απόδοση απ' ότι η HOLP-LASSO.

3.4 Προβλήματα της μεθόδου HOLP

Όταν ο αριθμός των παρατηρήσεων ισούται με τον αριθμό των πιθανών επεξηγηματικών μεταβλητών ($n \approx p$) τότε η HOLP ίσως αντιμετωπίσει προβλήματα, αφού ο πίνακας $\boldsymbol{XX'}$ είναι κοντά στο να γίνει μη αντιστρέψιμος τετραγωνικός πίνακας. Μπορεί ο πίνακας $\boldsymbol{X'(XX')^{-1}X} = \boldsymbol{UU'}$ να διατηρεί την αυστηρή διαγώνια υπεροχή του αλλά ο όρος των σφαλμάτων $\boldsymbol{X'(XX')^{-1}\varepsilon}$ μπορεί να μεγαλώσει πολύ με αποτέλεσμα να επισκιάσει την επίδραση του πίνακα $\boldsymbol{X'(XX')^{-1}X}$ και άρα η απόδοση της HOLP να επηρεάζεται αρνητικά.

Υπάρχουν δύο τρόποι να αντιμετωπιστεί το πρόβλημα αυτό:

1. **Ridge-HOLP:** Χρησιμοποιώντας την παλινδρόμηση κορυφογραμμής όπως είδαμε στο κεφ. (3.2.2), δηλαδή προσθέτοντας την παράμετρο ποινικοποίησης r μπορούμε να ελέγξουμε την αύξηση του μεγέθους το όρου των σφαλμάτων.

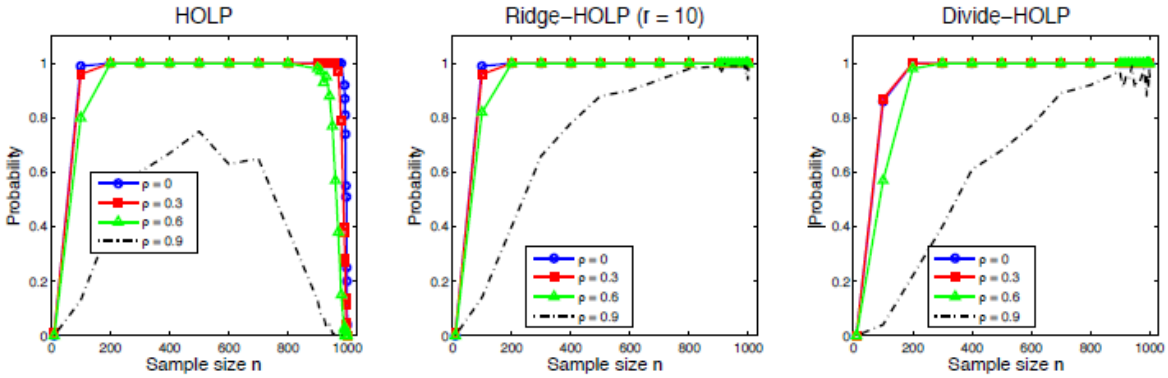
2. **Divide-HOLP:** Ο δεύτερος τρόπος είναι να χωρίσουμε τις παρατηρήσεις μας σε m τυχαία υποσύνολα και να εφαρμόσουμε την μέθοδο HOLP παίρνοντας m μειωμένα μοντέλα. Ενώνοντας τα μοντέλα αυτά παίρνουμε το τελικό μοντέλο με μειωμένη διάσταση. Με αυτό τον τρόπο είναι σίγουρο ότι η υπόθεση (1.) που είδαμε στο κεφάλαιο (3.2.3) ικανοποιείται για κάθε υποσύνολο και επίσης το υπολογιστικό κόστος μειώνεται από $O(n^2p)$ σε $O(\frac{n^2p}{m})$.

Για να δείξουν ότι όντως αυτές οι δύο μέθοδοι λειτουργούν καλύτερα από την HOLP στην περίπτωση όπου $n \approx p$ οι Whang και Leng (2015) προσομοίωσαν το γραμμικό μοντέλο:

$$Y = 5X_1 + 5X_2 + 5X_3 + 5X_4 + 5X_5 + \varepsilon,$$

όπου οι επεξηγηματικές μεταβλητές ακολουθούν την τυποποιημένη κανονική κατανομή, και έχουν μεταξύ τους όλες την ίδια συσχέτιση ρ. Επίσης αριθμός των επεξηγηματικών μεταβλητών είναι ίσος με $p = 1000$. Εφάρμοσαν και τις τρεις μεθόδους για διαφορετικό αριθμό παρατηρήσεων n και για διαφορετικές τιμές συσχέτισης ρ. Κάθε μέθοδος επιλέγει ένα υποσύνολο με πληθυκότητα $\min\{n, 100\}$. Η επίδοση της κάθε μεθόδου είναι ανάλογη με το ποσοστό εύρεση όλων των πραγματικά σημαντικών μεταβλητών. Για την μέθοδο Ridge-Holp χρησιμοποιήθηκε η ποινικοποιημένη παράμετρος $r = 10$ ενώ για την Divide-Holp η διαμέριση των παρατηρήσεων έγινε σε δύο υποσύνολα ($m = 2$). Τα αποτελέσματα φαίνονται στο σχήμα (3.3.1).

Παρατηρούμε ότι η μέθοδος HOLP όταν $n \approx p$ και όταν το n είναι μικρό έχει πολύ κακή απόδοση. Οι άλλες δύο μέθοδοι αποδίδουν εξίσου καλά όταν $n \approx p$ με την Ridge-HOLP να αποδίδει καλύτερα από την Divide-HOLP όταν το n είναι μικρό. Τέλος παρατηρούμε ότι όταν το $\rho = 0.9$, η απόδοση των τριών μεθόδων μειώνεται, παρόλα αυτά η Ridge-Holp και η Divide-Holp αποδίδουν στο μέγιστο όταν $n \approx p$.



Σχήμα 3.3.1: Απόδοση των μεθόδων Holp, Ridge-Holp, Divide-Holp, για $p = 1000$ και για διαφορετικό αριθμό παρατηρήσεων.

3.5 Προσομοιώσεις

Στο κεφάλαιο αυτό θα εφαρμόσουμε και θα συγκρίνουμε τις μεθόδους HOLP, Divide HOLP και SIS. Σημειώνουμε ότι οι μέθοδοι HOLP και Divide HOLP όπως και η μέθοδος SIS δημιουργήθηκαν με την βοήθεια του στατιστικού πακέτου R και οι κώδικες είναι διαθέσιμοι στο Παράρτημα 2.

Έστω το γραμμικό μοντέλο:

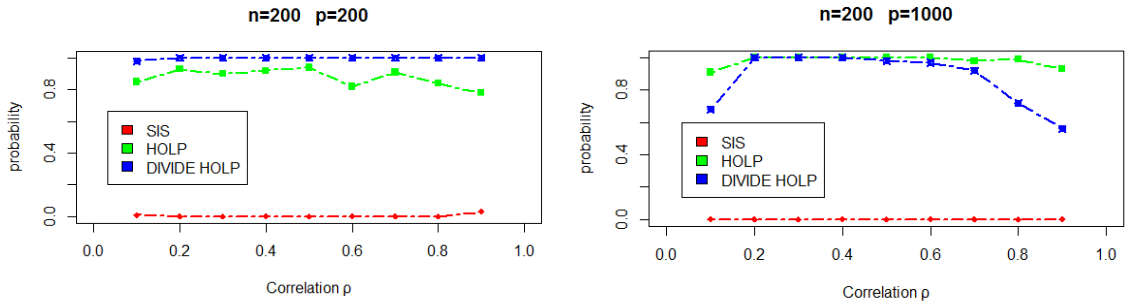
$$Y = 5X_1 + 5X_2 + 5X_3 + 5X_4 - 20\rho X_5 + \varepsilon, \qquad (3.5.1)$$

Θεωρούμε ότι το διάνυσμα των επεξηγηματικών μεταβλητών ακολουθεί την κανονική κατανομή $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ με τον πίνακα διασποράς-συνδιασποράς $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ να έχει τα εξής στοιχεία: $\sigma_{ii} = 1$, και $\sigma_{ij} = \rho\ i \neq j, i = 1, \dots p$, όπου ρ η συσχέτιση Pearson. Επίσης έστω ότι $\varepsilon \sim N(0,1)$. Προφανώς η μεταβλητή X_5 είναι ασυσχέτιστη με την μεταβλητή απόκρισης λόγω των συσχετίσεων με τις άλλες επεξηγηματικές μεταβλητές. Θεωρούμε τα μοντέλα $(n, p, \rho) = (200,200,0.2)$, $(n, p, \rho) = (200,1000,0.2)$, $(n, p, \rho) = (200,1000,0.5)$ και $(n, p, \rho) = (200,1000,0.8)$. Τα αποτελέσματα παρουσιάζονται στο πίνακα (3.5.1).

Παρατηρούμε ότι οι μέθοδοι HOLP και Divide HOLP καταφέρνουν να διατηρήσουν στο μοντέλο την μεταβλητή X_5 με μεγάλη πιθανότητα αντίθετα με την μέθοδο SIS που την διατηρεί μόνο τυχαία. Αυτό οφείλεται στο γεγονός ότι η μέθοδος HOLP δεν χρησιμοποιεί την συσχέτιση μεταβλητών για το κρησάρισμα των μεταβλητών με αποτέλεσμα όταν μια σημαντική μεταβλητή είναι οριακά ασυσχέτιστη με την μεταβλητή απόκρισης, αυτό να μην επηρεάζει το κρησάρισμα. Όπως μπορούμε να δούμε στο σχήμα (3.5.1) η μέθοδος HOLP εντοπίζει όλες τις σημαντικές μεταβλητές με μεγάλη πιθανότητα όταν το ρ είναι σχετικά μικρό και όσο αυτό αυξάνεται τόσο η πιθανότητα αυτή μειώνεται. Επίσης αξίζει να σημειώσουμε ότι όταν $n \approx p$ η επίδοση της μεθόδου HOLP μειώνεται, όμως η μέθοδος Divide HOLP αποδίδει τα μέγιστα. Αυτό έρχεται σε συμφωνία με τα όσα προαναφέραμε.

Πίνακας 3.5.1: Αποτελέσματα από την προσομοίωση του γραμμικού μοντέλου (3.5.1). Επίδοση των μεθόδων SIS, HOLP, Divide HOLP καθορίζεται από την πιθανότητα να περιλαμβάνουν το πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4, X_5\}$ σε 100 προσομοιώσεις.

		Model (3.5.1)					
		P_s					P_α
(ρ, p)	Methods	X_1	X_2	X_3	X_4	X_5	ALL
(0.8,200)	SIS	1.00	1.00	1.00	1.00	0.01	0.01
	HOLP	0.95	0.92	0.90	0.96	1.00	0.77
	Divide HOLP	1.00	1.00	1.00	1.00	1.00	1.00
(0.2,1000)	SIS	1.00	1.00	1.00	1.00	1.00	1.00
	HOLP	1.00	1.00	1.00	1.00	1.00	1.00
	Divide HOLP	1.00	1.00	1.00	1.00	0.99	0.99
(0.5,1000)	SIS	1.00	1.00	1.00	1.00	1.00	1.00
	HOLP	1.00	1.00	1.00	1.00	1.00	1.00
	Divide HOLP	1.00	1.00	1.00	0.99	1.00	0.99
(0.8,1000)	SIS	0.97	1.00	0.97	0.98	0.00	0.00
	HOLP	0.98	1.00	0.98	0.96	1.00	0.92
	Divide HOLP	0.90	0.97	0.88	0.93	1.00	0.74



Σχήμα 3.5.1: Στο x-άξονα έχουμε τις τιμές συσχέτισης Pearson και στον y-άξονα την πιθανότητα όλες οι σημαντικές μεταβλητές να συμπεριληφθούν στο μοντέλο. Γίνεται σύγκριση μεταξύ των μεθόδων HOLP,Divide HOLP και SIS . Εμφανώς οι HOLP και Divide HOLP αποδίδουν καλύτερα από την μέθοδο SIS. Για $n \approx p$ βλέπουμε ότι πέφτει η απόδοση της HOLP με την Divide HOLP να αποδίδει καλύτερα.

ΚΕΦΑΛΑΙΟ 4

ΚΡΗΣΑΡΙΣΜΑ ΜΕΤΑΒΛΗΤΩΝ ΜΕΣΩ ΤΗΣ ΣΥΣΧΈΤΙΣΗ ΤΗΣ ΑΠΟΣΤΑΣΗΣ (*FEATURE SCREENING VIA DISTANCE CORRELATION*)

4.1 Εισαγωγή

Η επιλογή σημαντικών χαρακτηριστικών που επιδρούν σε ένα στατιστικό μοντέλο αποτελεί μεγάλη πρόκληση ειδικά στην περίπτωση όπου ο αριθμός των πιθανών χαρακτηριστικών είναι πάρα πολύ μεγάλος. Την τελευταία δεκαετία πολλές μέθοδοι προτάθηκαν για να αντιμετωπιστεί αυτή η πρόκληση. Σε προηγούμενα κεφάλαια αναφερθήκαμε στις μεθόδους SIS και HOLP, που πρότειναν οι Fan και Li (2008) και Whang και Leng (2015) αντίστοιχα. Η πρώτη βασίζεται στην μάθηση συσχέτισης, ούτως ώστε να γίνει ένα κρησάρισμα των πιθανών μεταβλητών, μειώνοντας την διάσταση του χώρου των επεξηγηματικών μεταβλητών. Η δεύτερη βασίζεται στην ιδιότητα της αυστηρά διαγώνιας υπεροχής ενός πίνακα. Οι μέθοδοι όμως αυτές λειτουργούν κάτω από συγκεκριμένες υποθέσεις και όταν δεν ισχύουν αυτές δεν είναι εφαρμόσιμες. Για να λειτουργήσει οι μέθοδοι SIS και HOLP μια κύρια υπόθεση είναι η γραμμικότητα του στατιστικού μοντέλου κάτι που πολλές φορές στην πραγματικότητα δεν ισχύει. Οπότε χρειάζονται μέθοδοι με λιγότερα απαιτούμενα για να εφαρμόζονται σε μεγαλύτερης ευρύτητας προβλήματα.

Οι Li, Zhong και Zhu (2012) πρότειναν μια νέα μέθοδο η οποία βασίζεται στην συσχέτιση απόστασης (*distance correlation)* για την κατάταξη της σημαντικότητας του εκάστοτε χαρακτηριστικού σε οποιοδήποτε είδος στατιστικού μοντέλου, δηλαδή όχι μόνο γραμμικού. Η *distance correlation* προτάθηκε από τους Szekely, Rizzo και Bakirov (2007) για την μέτρηση της εξάρτησης μεταξύ δύο τυχαίων διανυσμάτων. Οι Li, Zhong και Zhu (2012) ονόμασαν την μεθοδολογία αυτή ως Συσχέτιση της Απόστασης βασισμένη στο Σίγουρο Ανεξάρτητο Κρησάρισμα (*Distance Correlation-based Sure Independence Screening* (*DC-SIS*)). Δύο τυχαία διανύσματα θεωρούνται ανεξάρτητα αν και μόνο αν η συσχέτιση της απόσταση τους (*distance correlation*) είναι ίση με μηδέν. Λόγω αυτού η DCSIS δεν απαιτεί συγκεκριμένη σχέση μεταξύ της απόκριση και των συμμμεταβλητών. Αυτό το χαρακτηριστικό είναι πολύ ελκυστικό γιατί είναι πολύ δύσκολο να καθορίσεις ένα παραμετρικό μοντέλο για δεδομένα υψηλής διάστασης. Επίσης η *distance correlation* μεταξύ δύο τυχαίων μεταβλητών κανονικής κατανομής είναι μια αυστηρώς αύξουσα

συνάρτηση της απόλυτης τιμής της συσχέτισης Pearson, επομένως η μέθοδος DC-SIS αποδίδει το ίδιο καλά με την μέθοδο SIS για γκαουσιανα γραμμικά μοντέλα και εφαρμόζεται το ίδιο εύκολα. Επίσης λόγω του ότι η *distance correlation* ορίζεται για δύο τυχαία διανύσματα με διαφορετικές διαστάσεις, η DC-SIS μπορεί να εφαρμοστεί σε πολυμεταβλητές αποκρίσεις. Τέλος η DC-SIS μπορεί να εφαρμοστεί σε συνεχείς, κατηγορικές και διακριτές αποκρίσεις. Όλα αυτά συνηγορούν ότι η μέθοδος αυτή μπορεί να χρησιμοποιηθεί σε μεγαλύτερο εύρος προβλημάτων.

Παρόλα αυτά η DC-SIS παρουσιάζει παρόμοια προβλήματα με την SIS όπως αυτά που αναφέραμε στο κεφ. (2.2.4), οπότε οι Zhong και Zhu (2015) πρότειναν μια επαναληπτική DC-SIS την *Iterative* DC-SIS, για να επιλυθούν τα προβλήματα αυτά. Στο υπόλοιπο του κεφαλαίου θα αναφέρουμε την *distance correlation* με τα αρχικά DC.

4.2 Θεωρητικό Υπόβαθρο-Συσχέτιση της Απόστασης (*Distance Correlation*)

Έστω δύο τυχαία διανύσματα $\boldsymbol{u}$ και $\boldsymbol{v}$ με χαρακτηριστικές συναρτήσεις $\varphi_{\boldsymbol{u}}(\boldsymbol{t})$ και $\varphi_{\boldsymbol{v}}(\boldsymbol{s})$ αντίστοιχα. Η από κοινού χαρακτηριστική τους συνάρτηση ορίζεται να είναι η $\varphi_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{t},\boldsymbol{s})$. Οι Szekely, Rizzo και Bakirov (2007) όρισαν ως συσχέτιση συνδιακύμανσης (*distance covariance*) μεταξύ δύο διανυσμάτων $\boldsymbol{u}$ και $\boldsymbol{v}$, τον μη αρνητικό αριθμό $dcov(\boldsymbol{u},\boldsymbol{v})$ ο οποίος δίνεται από την εξίσωση:

$$\begin{aligned} dcov^2(\boldsymbol{u},\boldsymbol{v}) &= \left\|\varphi_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{t},\boldsymbol{s}) - \varphi_{\boldsymbol{u}}(\boldsymbol{t})\varphi_{\boldsymbol{v}}(\boldsymbol{s})\right\|^2 \\ &= \int_{R^{d_u d_v}} \left|\varphi_{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{t},\boldsymbol{s}) - \varphi_{\boldsymbol{u}}(\boldsymbol{t})\varphi_{\boldsymbol{v}}(\boldsymbol{s})\right|^2 w(\boldsymbol{t},\boldsymbol{s})d\boldsymbol{t}d\boldsymbol{s}, \end{aligned} \tag{4.2.1}$$

όπου d_u, d_v είναι οι διαστάσεις των $\boldsymbol{u}$ και $\boldsymbol{v}$ και:

$$w(\boldsymbol{t},\boldsymbol{s}) = \left\{c_{d_u} c_{d_v} \|\boldsymbol{t}\|_{d_u}^{\boldsymbol{1}+d_u} \|\boldsymbol{s}\|_{d_v}^{1+d_v}\right\}^{-\boldsymbol{1}}, \tag{4.2.2}$$

με $c_d = \pi^{\frac{\boldsymbol{1}+\boldsymbol{d}}{\boldsymbol{2}}} / \Gamma\{\frac{1+d}{2}\}$.

<u>Λήμμα:</u> Αν $0 < \alpha < 2$ τότε για όλα τα $x \in R^d$ ισχύει:

$$\int_{R^d} \frac{1 - \cos\langle t, x\rangle}{|t|_d^{d+a}} = C(d,a)|x|^a, \tag{4.2.3}$$

όπου $C(d,a) = \frac{2\pi^{\frac{d}{2}}\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^{\alpha}\Gamma\left(\frac{d+a}{2}\right)}$.

Παρατηρούμε ότι $c_d = C(d,1) = \pi^{\frac{\boldsymbol{1}+\boldsymbol{d}}{\boldsymbol{2}}} / \Gamma\{\frac{1+d}{2}\}$, έτσι το λήμμα αυτό ήταν αφορμή για να επιλεχθεί η συγκεκριμένη συνάρτηση $w(\boldsymbol{t},\boldsymbol{s})$.

Η ιδέα της συσχέτισης της απόστασης (*distance correlation*) προήλθε από μια ιδιότητα που χαρακτηρίζει τις χαρακτηριστικές συναρτήσεις η οποία λέει ότι αν δύο μεταβλητές είναι ανεξάρτητες, η από κοινού χαρακτηριστική τους συνάρτηση ισούται με το γινόμενο των χαρακτηριστικών τους συναρτήσεων.

Η συσχέτιση της απόστασης (*distance correlation*) δύο τυχαίων διανυσμάτων ορίζεται είναι:

$$dcor(\boldsymbol{u},\boldsymbol{v}) = \frac{dcov(\boldsymbol{u},\boldsymbol{v})}{\sqrt{dcov(\boldsymbol{u},\boldsymbol{u})dcov(\boldsymbol{v},\boldsymbol{v})}} \ . \tag{4.2.4}$$

Η DC έχει δύο σημαντικές ιδιότητες που έδωσαν το κίνητρο στους Li, Zhong και Zhu (2012) να την χρησιμοποιήσουν ως μέθοδο κρησαρίσματος χαρακτηριστικών:

1. Για δύο μονοπαραμετρικές τυχαίες μεταβλητές U και V με σταθερά συσχέτισης Pearson ρ ισχυεί:

$$dcorr(U,V) = \\ = \left\{ \frac{\rho \arcsin(\rho) + \sqrt{1-\rho^2} - \rho \arcsin\left(\frac{\rho}{2}\right) - \sqrt{4-\rho^2} + 1}{1 + \frac{\pi}{3} - \sqrt{3}} \right\}^{1/2} . \tag{4.2.5}$$

   Η συνάρτηση αυτή, είναι μια αυστηρώς αύξουσα συνάρτηση της απόλυτης τιμής της συσχέτισης Pearson $|\rho|$, και αυτό υπονοεί ότι η μέθοδος κρησαρίσματος βασισμένη στην DC είναι ισοδύναμη με την μάθηση συσχέτισης Pearson για γραμμικά μοντέλα η οποία χρησιμοποιήθηκε από τους Fan και Li (2008) .

2. Δύο τυχαία διανύσματα $\boldsymbol{u},\boldsymbol{v}$ είναι ανεξάρτητα αν και μόνο αν $dcor(\boldsymbol{u},\boldsymbol{v}) = 0$. Είναι γνωστό ότι δύο τυχαίες μεταβλητές U και V είναι ανεξάρτητες αν και μόνο αν U και $T(V)$, μια αυστηρώς μονότονη συνάρτηση του V, είναι ανεξάρτητες. Αυτό υπονοεί ότι όταν U και V έχουν μη γραμμική σχέση, η μέθοδος κρησαρίσματος βασισμένη στην DC είναι πιο αποτελεσματική από την μέθοδο κρησαρίσματος βασισμένη στην μάθηση Pearson συσχέτισης.

Η εξίσωση (3.2.1) μπορεί να γραφεί και ως:

$$dcov^2(\boldsymbol{u},\boldsymbol{v}) = S_1 + S_2 - 2S_3, \tag{4.2.6}$$

όπου S_1, S_2, S_3 ορίζονται να είναι:

$$S_1 = E\{\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_{d_u} \|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|_{d_v}\}, \tag{4.2.7}$$

$$S_2 = E\left(\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_{d_u}\right) E\left(\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|_{d_v}\right), \tag{4.2.8}$$

$$S_3 = E\{E\left(\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|_{d_u} | \boldsymbol{u}\right) E\left(\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|_{d_v} | \boldsymbol{v}\right)\}, \tag{4.2.9}$$

όπου $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ είναι ένα ανεξάρτητο αντίγραφο του $(\boldsymbol{u},\boldsymbol{v})$**,** δηλαδή $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ είναι ανεξάρτητες των $(\boldsymbol{u},\boldsymbol{v})$**,** με την ίδια κατανομή.

Αν $\{(\boldsymbol{u_i}, \boldsymbol{v_i}), i = 1, \dots, n\}$ είναι ένα δείγμα από τον πληθυσμό $(\boldsymbol{u}, \boldsymbol{v})$ τότε οι εκτιμήσεις των S_1, S_2, S_3 προκύπτουν να είναι οι εξής:

$$\hat{S}_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\boldsymbol{u_i} - \boldsymbol{u_j}\|_{d_u} \|\boldsymbol{v_i} - \boldsymbol{v_j}\|_{d_v}, \tag{4.2.10}$$

$$\hat{S}_2 = \frac{1}{n^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\boldsymbol{u_i} - \boldsymbol{u_j}\|_{d_u} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\boldsymbol{v_i} - \boldsymbol{v_j}\|_{d_v}, \tag{4.2.11}$$

$$\hat{S}_3 = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \|\boldsymbol{u_i} - \boldsymbol{u_l}\|_{d_u} \|\boldsymbol{v_i} - \boldsymbol{v_l}\|_{d_v}, \tag{4.2.12}$$

Επομένως ο εκτιμητής του $dcov^2(\boldsymbol{u}, \boldsymbol{v})$ είναι:

$$\widehat{dcov}^2(\boldsymbol{u}, \boldsymbol{v}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3, \tag{4.2.13}$$

και επομένως ο εκτιμητής του $dcor(\boldsymbol{u}, \boldsymbol{v})$ είναι:

$$\widehat{dcor}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\widehat{dcov}(\boldsymbol{u}, \boldsymbol{v})}{\sqrt{\widehat{dcov}(\boldsymbol{u}, \boldsymbol{u}) \widehat{dcov}(\boldsymbol{v}, \boldsymbol{v})}}, \tag{4.2.14}$$

Για περισσότερες θεωρητικές πληροφορίες παραπέμπουμε στο άρθρο των Szekely, Rizzo και Bakirov (2007).

4.3 Μέθοδος Συσχέτιση της Απόστασης βασισμένη στο Σίγουρο Ανεξάρτητο Κρησάρισμα (*Distance Correlation-based Sure Independence Screening* (*DC-SIS*))

4.3.1 Ορισμός Μεθόδου DC-SIS

Έστω $\boldsymbol{y} = (Y_1, \dots, Y_q)'$ το διάνυσμα απόκρισης και $\boldsymbol{x} = (X_1, \dots, X_p)'$ οι επεξηγηματικές μεταβλητές . Θεωρούμε ότι το q είναι ένας σταθερός αριθμός και το p πιθανόν είναι πολύ μεγαλύτερο από το μέγεθος του δείγματος n. Έστω $\boldsymbol{x}_A = \{X_k : k \in A\}$ είναι το υποσύνολο του $\boldsymbol{x}$ με τις πραγματικά σημαντικές συμμεταβλητές, αν η δεσμευμένη πυκνότητα πιθανότητας του $\boldsymbol{y}$ εξαρτάται από το $\boldsymbol{x}$ μέσω του $\boldsymbol{x}_A$. Επομένως $A = \{k: o\ \delta\acute{\epsilon}\iota\kappa\tau\eta\varsigma\ \pi\rho\alpha\gamma\mu\alpha\tau\iota\kappa\acute{\alpha}\ \sigma\eta\mu\alpha\nu\tau\iota\kappa\acute{\omega}\nu\ \mu\epsilon\tau\alpha\beta\lambda\eta\tau\acute{\omega}\nu\}$. Αντίστοιχα ορίζουμε ως $\boldsymbol{x}_{A^c} = \{X_j : j \in A^c\}$ το υποσύνολο του $\boldsymbol{x}$ με τις μη σημαντικές μεταβλητές. Όπως και στις μεθόδους SIS και HOLP, ο στόχος είναι να μειώσουμε την διάσταση του χώρου των επεξηγηματικών μεταβλητών αλλά και να κρατήσουμε όλες τις πραγματικά σημαντικές μεταβλητές στο μοντέλο. Προφανώς αφού το $\boldsymbol{x}_A = \{X_j : j \in A\}$ είναι το υποσύνολο του $\boldsymbol{x}$ με τις πραγματικά σημαντικές συμμεταβλητές είναι λογικό να υποθέσουμε ότι το $\boldsymbol{x}_A$ έχει μεγαλύτερη συσχέτιση με το $\boldsymbol{y}$ σε σχέση με το $\boldsymbol{x}_{A^c}$.

Έστω $\{(\boldsymbol{x_i}, \boldsymbol{y_i}), i = 1, \ldots, n\}$ είναι ένα δείγμα από τον πληθυσμό $(\boldsymbol{x}, \boldsymbol{y})$. Η DC-SIS αυτό που ουσιαστικά κάνει είναι να ταξινομεί τις συμμεταβλητές ανάλογα με τη συσχέτιση της απόστασης (*distance correlation*) που έχουν με την μεταβλητή απόκρισης ($dcor(X_k, \boldsymbol{y})$). Η μέθοδος διατηρεί στο μοντέλο τις μεταβλητές που έχουν την υψηλότερη DC με την μεταβλητή απόκρισης. Επομένως έχουμε το σύνολο:

$$\hat{A} = \{k: \widehat{dcor}^2(X_k, \boldsymbol{y}) \geq cn^{-\kappa}\ 1 \leq k \leq p\}, \tag{4.3.1}$$

όπου :

$$\widehat{dcor}(X_k, \boldsymbol{y}) = \frac{\widehat{dcov}(X_k, \boldsymbol{y})}{\sqrt{\widehat{dcov}(X_k, X_k)\widehat{dcov}(\boldsymbol{y}, \boldsymbol{y})}}, \tag{4.3.2}$$

$$\widehat{dcov}^2(\boldsymbol{u}, \boldsymbol{v}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3, \tag{4.3.3}$$

$$\hat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_{\boldsymbol{ki}} - X_{\boldsymbol{kj}}\|_1 \|\boldsymbol{y_i} - \boldsymbol{y_j}\|_q, \tag{4.3.4}$$

$$\hat{S}_2 = \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_{\boldsymbol{ki}} - X_{\boldsymbol{kj}}\|_1 \sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{y_i} - \boldsymbol{y_j}\|_q, \tag{4.3.5}$$

$$\hat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|X_{\boldsymbol{ki}} - X_{\boldsymbol{kl}}\|_1 \|\boldsymbol{y_i} - \boldsymbol{y_l}\|_q, \tag{4.3.6}$$

Σημειώνουμε ότι $\|\alpha\|_d$ είναι η Ευκλείδεια νόρμα ενός $\alpha \in R^d$. Τα c και κ θα οριστούν πιο μετά από την συνθήκη (2.) του επόμενου κεφαλαίου. Αντί να αναζητούμε το σύνολο (3.3.1) είναι ισοδύναμο να αναζητήσουμε το σύνολο:

$\hat{A}_* = \{k: \eta\ \widehat{dcor}(X_k, \boldsymbol{y})\ \textit{να είναι στις πρώτες } d \textit{ τιμες απο όλες τις } p \textit{ μεταβλητες}\}$,

όπου η τιμή d προκαθορίζεται όπως ακριβώς και στη SIS.

4.3.2 Ιδιότητα Σίγουρου Κρησαρίσματος της DC-SIS

Θα δούμε κάτω υπό ποιες συνθήκες ισχύει η ιδιότητα σίγουρου κρησαρίσματος, δηλαδή αν η DC-SIS έχει την ιδιότητα μετά το κρησάρισμα των μεταβλητών να διατηρεί στο μοντέλο όλες τις σημαντικές μεταβλητές, με πιθανότητα να τείνει στο 1. Στη συνέχεια της ανάλυσης για ευκολία γραφής ορίζουμε τα $\omega_k = dcor^2(X_k, \boldsymbol{y})$ και $\hat{\omega}_k = \widehat{dcor}^2(X_k, \boldsymbol{y})$. Ορίζουμε τις πιο κάτω συνθήκες οι οποίες μπορεί να μην είναι πιο ασθενείς συνθήκες για να ισχύει η ιδιότητα σίγουρου κρησαρίσματος, αλλά είναι αυτές χρησιμοποιήθηκαν από τους Li, Zhong και Zhu (2012) έτσι ώστε να απλοποιηθεί η απόδειξη του παρακάτω θεωρήματος.

1. Υπάρχει μια θετική σταθερά s_0 έτσι ώστε για όλα τα $0 < s \leq s_0$ να ισχύουν:

$$\sup_p \max_{1\leq k\leq p} E\{\exp(s\|X_k\|_1^2\} < \infty,$$

$$E\{\exp(s\|\boldsymbol{y}\|_q^2\} < \infty.$$

2. Η ελάχιστη συσχέτιση της απόστασης (DC) των σημαντικών μεταβλητών πρέπει να ικανοποιεί:

$$\min_{k\in A} \omega_k \geq cn^{-\kappa},$$

για κάποια σταθερά $c > 0$ και $0 < \kappa < 1/2$.

Η συνθήκη (1.) ικανοποιείται αυτόματα όταν οι $\boldsymbol{y}$ και $\boldsymbol{x}$ ακολουθούν την πολυμεταβλητή κανονική κατανομή, η οποία είναι μια υπόθεση που χρησιμοποιείται συχνά σε δεδομένα υψηλής διάστασης. Όσον αφορά την συνθήκη (2.), αυτό που ουσιαστικά μας αναφέρει είναι ότι η οριακή DC των πραγματικά σημαντικών μεταβλητών δεν μπορεί να είναι πολύ μικρή.

<u>Θεώρημα:</u> Υπό την συνθήκη (1.), για οποιοδήποτε $0 < \gamma < \frac{1}{2} - \kappa$ υπάρχουν θετικές σταθερές $c_1 > 0$ και $c_2 > 0$ έτσι ώστε:

$$P\left(\max_{1\leq k\leq p} |\widehat{\omega}_k - \omega_k| \geq cn^{-\kappa}\right) \leq$$

$$\leq O\left(p\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + nexp(-c_2 n^{\gamma})\right]\right). \qquad (4.3.7)$$

Υπό τις συνθήκες (1.) και (2.) ισχύει ότι:

$$\mathrm{P}(A \subset \hat{A}) \geq 1 - O\left(s_n\left[\exp\left\{-c_1 n^{1-2(\kappa+\gamma)}\right\} + nexp(-c_2 n^{\gamma})\right]\right), \qquad (4.3.8)$$

όπου s_n η πληθυκότητα του A.

Παρατηρούμε ότι η ιδιότητα σίγουρου κρησαρίσματος για DC-SIS ισχύει υπό ασθενέστερες συνθήκες από αυτές τις SIS. Σημειώνουμε ότι η DC μπορεί να χρησιμοποιηθεί σε αυθαίρετα μοντέλα παλινδρόμησης, γραμμικά ή μη γραμμικά, επιτρέπει πολυμεταβλητές αποκρίσεις, ανεξάρτητα του αν είναι συνεχείς, διακριτές ή κατηγορικές και επιτρέπει ομαδοποιημένες συμμεταβλητές.

4.3.3 Προβλήματα της μεθόδου DC-SIS

Όπως ακριβώς και η SIS, η DC-SIS αντιμετωπίζει κάποια προβλήματα στην εύρεση των πραγματικών μεταβλητών. Κατά το κρησάρισμα είναι πιθανόν πραγματικά σημαντικές μεταβλητές να μην παραμείνουν στο μοντέλο λόγω των συσχετίσεων τους με τις υπόλοιπες σημαντικές μεταβλητές. Το πιο κάτω παράδειγμα είναι χαρακτηριστικό.

Σημειώνουμε ότι η ανάλυση του παραδείγματος θα γίνει με κώδικα στην R που εμείς δημιουργήσαμε για την μέθοδο DC-SIS (βλέπε Παράρτημα 2).

Παράδειγμα (4.3.1): Έστω το γραμμικό μοντέλο:

$$Y = 2X_1 + 3X_2 - 5\sqrt{\rho}X_3,$$

και έστω ότι έχουμε 1000 πιθανές επεξηγηματικές μεταβλητές, δηλαδή $\boldsymbol{x} = (X_1, \dots, X_{1000})'$. Κάθε $X_k \sim N(0,1)$ και όλες οι συμμεταβλητές έχουν μεταξύ τους την ίδια συσχέτιση ρ εκτός από την X_3 η οποία έχει συσχέτιση με τις άλλες $p-1$ ίση με $\sqrt{\rho}$. Θεωρούμε ότι $\rho = 0.2$ και $\varepsilon \sim N(0,1.5)$. Επίσης θεωρούμε ότι το μέγεθος του δείγματος ισούται με $n = 200$ και ο αριθμός των μεταβλητών που θα παραμείνουν στο μοντέλο μετά το κρησάρισμα είναι $d = 2\left[\frac{n}{\log(n)}\right] = 43$. Δημιουργήσαμε 100 προσομοιώσεις εξαιτίας της χαμηλής υπολογιστικής δύναμης που διαθέτουμε. Παρατηρούμε ότι οι σημαντικές μεταβλητές X_1 και X_2 επιλέγονται και τις 100 φορές από την μέθοδο αντίθετα η X_3 καμία. Αυτό συμβαίνει γιατί λόγω των συσχετίσεων της X_3 με τις άλλες σημαντικές μεταβλητές, ισχύει ότι $dcor(X_k, Y) = 0$, δηλαδή η μεταβλητή X_3 είναι ανεξάρτητη της απόκρισης. Επομένως έχει μικρότερη προτεραιότητα να επιλεχθεί από το μοντέλο.

Για την επίλυση του προβλήματος αυτού προτάθηκε μια επαναληπτική μέθοδος της DC-SIS από τους Zhong και Zhu (2015), η επαναληπτική μέθοδος συσχέτισης της απόστασης βασισμένη στο σίγουρο ανεξάρτητο κρησάρισμα (*Iterative Distance Correlation-Based Sure Independence Screening*) την οποία θα συμβολίζουμε ως DC-ISIS.

4.4 Επαναληπτική Μέθοδος Συσχέτιση της Απόστασης βασισμένη στο Σίγουρο Ανεξάρτητο Κρησάρισμα (*Iterative Distance Correlation-based Sure Independence Screening* (*DC-ISIS*)).

Οι Zhong και Zhu (2015) ήθελαν να εφαρμόσουν κάτι παρόμοιο με την ISIS αλλά η γενίκευση της μεθόδου αυτής δεν μπορούσε να γινεί λόγω του ότι η DC-SIS εφαρμόζεται σε αυθαίρετου είδους μοντέλα παλινδρόμησης. Η ιδέα της επαναληπτικής μεθόδου συσχέτισης της Απόστασης βασισμένη στο Σίγουρο Ανεξάρτητο Κρησάρισμα (*Iterative Distance Correlation-based Sure Independence Screening* (*DC-ISIS*)) είναι αρκετά απλή. Θα την εξηγήσουμε θεωρώντας το παράδειγμα (4.3.1). Παρατηρούμε ότι, αφού εξαιτίας της συσχέτισης της X_3 με τις X_1 και X_2, προκύπτει η οριακή ανεξαρτησία μεταξύ της X_3 με το Y, καλό θα ήταν αυτή η συσχέτιση να εξαλειφθεί. Ένας απλός τρόπος για να εξαλειφθούν αυτές οι συσχετίσεις είναι να εκτελέσουμε παλινδρομήσεις έχοντας ως αποκρίσεις τις X_k, $k = 3, \dots, p$, και ως επεξηγηματικές μεταβλητές τις X_1 και X_2. Τα υπόλοιπα που θα προκύψουν από αυτή την παλινδρόμηση θα είναι ασυσχέτιστα με τις X_1 και X_2. Επομένως χρησιμοποιώντας τα υπόλοιπα αντί τις X_k, $k = 3, \dots, p$, επιλύουμε το πρόβλημα μας και ξανατρέχοντας την μέθοδο DC-SIS, είναι πολύ πιθανόν η μέθοδος να εντοπίσει την μεταβλητή X_3 που αδυνατούσε να βρει προηγουμένως.

Γενικεύοντας την πιο πάνω διαδικασία η μέθοδος DC-ISIS αποτελείται από τρία βήματα το οποία είναι:

1. Αρχικά εφαρμόζουμε την μέθοδο DC-SIS για τα $\boldsymbol{y}$ και $\boldsymbol{x}$, επιλέγοντας p_1 μεταβλητές, έστω τις $\boldsymbol{x}_{A_1} = \left\{X_1^{(1)}, \dots, X_{p_1}^{(1)}\right\}$, με $p_1 < d$, όπου το d καθορίζεται από τον χρήστη. Θεωρούμε ότι $d = 2\left[\frac{n}{\log(n)}\right]$.

2. Έστω $\boldsymbol{X}_1$ ο πίνακας σχεδιασμού με στήλες μόνο αυτές που επιλέχθηκαν στο προηγούμενο βήμα δηλαδή τις $\boldsymbol{x}_{A_1}$ και έστω $\boldsymbol{X}_1^c$ ο πίνακας σχεδιασμού για τις υπόλοιπες $n - p_1$ μεταβλητές. Θεωρούμε ως νέο πίνακα σχεδιασμού τον πίνακα $\boldsymbol{X}_{new} = \{I_n - \boldsymbol{X}_1(\boldsymbol{X}_1{}'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\}\boldsymbol{X}_1^c$. Εφαρμόζουμε ξανά την DC-SIS για τα $\boldsymbol{y}$ και $\boldsymbol{x}_{new}$, επιλέγοντας p_2 μεταβλητές, έστω τις $\boldsymbol{x}_{A_2} = \left\{X_1^{(2)}, \dots, X_{p_2}^{(2)}\right\}$.

3. Θεωρούμε ότι $A_1 = A_1 \cup A_2$ και επαναλαμβάνουμε το βήμα (2.) μέχρι να επιλεχθούν d μεταβλητές.

Με το βήμα (2.) η μέθοδος προσπαθεί να εξαλείψει την οριακή ανεξαρτησία που προκύπτει μεταξύ της απόκρισης και μιας επεξηγηματικής μεταβλητής που προκαλείτε μέσω των συσχετίσεων με άλλες μεταβλητές. Επίσης με το βήμα αυτό μη σημαντικές μεταβλητές που είχαν προτεραιότητα να επιλεχθούν από την DC-SIS, λόγω των υψηλών τους συσχετίσεων με πραγματικά σημαντικές μεταβλητές, πλέον χάνουν αυτή την προτεραιότητα αφού στο πρώτο βήμα πιθανόν να επιλεχθούν οι μεταβλητές με τις οποίες συσχετίζονται και έτσι στο επόμενο βήμα πραγματικά σημαντικά μεταβλητές που δεν επιλέχθηκαν λόγω του γεγονότος αυτού, θα έχουν προτεραιότητα να επιλεχθούν από την μέθοδο. Όπως και για την μέθοδο DC-SIS, δημιουργήσαμε κώδικα με την βοήθεια του στατιστικού πακέτου R για την μέθοδο DC-ISIS (βλέπε Παράρτημα 2).

Επιστρέφοντας στο παράδειγμα (3.3.2) εφαρμόζουμε την μέθοδο DC-ISIS. Με τα ίδια δεδομένα όπως και στο παράδειγμα (3.3.2) εφαρμόσαμε 100 προσομοιώσεις. Παρατηρούμαι ότι η μέθοδος μας επίλεξε την μεταβλητή X_3 και τις 100 φορές, άρα φαίνεται ότι η μέθοδος επιλύει το πρόβλημα της οριακής ανεξαρτησίας μεταξύ απόκρισης και σημαντικών επεξηγηματικών μεταβλητών.

4.4.1 Εύρεση σημαντικών αλληλεπιδράσεων με την βοήθεια της DC-ISIS

Σε πολλά στατιστικά προβλήματα οι αλληλεπιδράσεις μεταξύ μεταβλητών επηρεάζουν σημαντικά την μεταβλητή απόκρισης, για αυτό είναι σημαντικό να μπορούμε να εντοπίσουμε αυτές τις αλληλεπιδράσεις. Για το λόγο αυτό προτάθηκε ένας επαναληπτικός αλγόριθμος της DC-ISIS που αποτελείται από τέσσερα βήματα. Σημειώνουμε ότι οι αλληλεπιδράσεις που προσπαθεί να εντοπίσει ο αλγόριθμος περιορίζονται σε δύο μεταβλητές και δεν μπορεί να εντοπίσει σημαντικές αλληλεπιδράσεις μεταξύ περισσοτέρων μεταβλητών.

Ο αλγόριθμος αποτελείται από τα εξής βήματα:

1. Εφαρμόζουμε την DC-ISIS για όλες τις πιθανές επεξηγηματικές μεταβλητές και έστω ότι παραμένουν $\boldsymbol{x}_A$ μεταβλητές μετά το κρησάρισμα.

2. Σχηματίζουμε τις αλληλεπιδράσεις ανά δύο μεταβλητών μεταξύ αυτών που επέζησαν από το βήμα (1.). Επομένως παίρνουμε το σύνολο:
   $\boldsymbol{X}_{int} = \{X_i X_j, i, j \in A \; \mu\varepsilon \; i < j\}$.

3. Θεωρούμε το $\boldsymbol{X}_{int}^* = \{\boldsymbol{I_n} - \boldsymbol{X}_A(\boldsymbol{X}_A'\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A')\}\, \boldsymbol{X}_{int}$ το οποίο βασίζεται στην ίδια ιδέα με το βήμα (2.) της DC-ISIS.

4. Εφαρμόζουμε ξανά την DC-ISIS για $\boldsymbol{y}$ και $\boldsymbol{x}_{int}^*$,γίνεται ένα κρησάρισμα για τις αλληλεπιδράσεις και κρατάμε τις σημαντικές, όπως αυτές τις βρήκε η μέθοδος.

Σημειώνουμε ότι ο αλγόριθμος υποθέτει ότι οι σημαντικές αλληλεπιδράσεις αποτελούνται από σημαντικές επεξηγηματικές μεταβλητές αφού δεν ελέγχει όλες τις αλληλεπιδράσεις παρά μόνο αυτές που προκύπτουν από τις μεταβλητές που επιλέχθηκαν στο βήμα (1.). Επίσης οι αλληλεπιδράσεις που προκύπτουν στο βήμα (2.) είναι $\binom{d}{2} = \frac{d(d-1)}{2}$ και επομένως ο αριθμός αυτός αυξάνεται ραγδαία καθώς το d αυξάνει, όπου d ο αριθμός των μεταβλητών που επιλέχθηκαν στο πρώτο κρησάρισμα. Για το λόγο αυτό ίσως χρειαστεί να γίνει πρώτα ένα κρησάρισμα για τις αλληλεπιδράσεις.

ΚΕΦΑΛΑΙΟ 5

ΕΦΑΡΜΟΓΕΣ ΜΕΘΟΔΩΝ ΚΑΙ ΣΥΓΚΡΙΣΕΙΣ

5.1 Εισαγωγή

Σκοπός του κεφαλαίου αυτού είναι να γίνει μια σύγκριση μεταξύ των μεθόδων που παρουσιάστηκαν στα προηγούμενα κεφάλαια. Πιο συγκεκριμένα θα εφαρμόσουμε τις μεθόδους LASSO, SIS, ISIS, HOLP, DC-SIS και DC-ISIS σε διαφορετικά μοντέλα για να δούμε ποιες μέθοδοι υπερτερούν και ποιες υστερούν και σε ποιες περιπτώσεις. Όλες οι πιο κάτω προσομοιώσεις έγιναν με την βοήθεια του στατιστικού πακέτου R. Οι αλγόριθμοι για τις μεθόδους SIS, HOLP, DC-SIS και DC-ISIS δημιουργήθηκαν στην R και παρατίθενται στο παράρτημα 2. Σημειώνουμε ότι για την δημιουργία των μεθόδων αυτών χρησιμοποιήθηκαν τα έτοιμα πακέτα MASS, energy, ncvreg. Για τις μεθόδους LASSO, ISIS χρησιμοποιήσαμε τα έτοιμα πακέτα που παρέχονται από την R, glmnet και SIS αντίστοιχα.

Για όλες τις μεθόδους ορίζουμε τον αριθμό των μεταβλητών που θα παραμείνουν στο μοντέλο μετά το κρησάρισμα να είναι $d = (\frac{2n}{\log(n)})$, ενώ για την μέθοδο DC-ISIS θεωρούμε $p_1 = 20$. Κάθε παράδειγμα θα επαναλαμβάνεται 100 φορές και η επίδοση των μεθόδων θα ορίζεται από τα εξής κριτήρια:

1. Το ποσοστό όπου κάθε σημαντική επεξηγηματική μεταβλητή ξεχωριστά επιλέγεται από την κάθε μέθοδο σε 100 επαναλήψεις και θα συμβολίζεται με P_s.

2. Το ποσοστό όπου όλες οι σημαντικές μεταβλητές επιλέγονται από τις μεθόδους σε 100 επαναλήψεις και θα συμβολίζεται με P_α.

Όσο μεγαλύτερα είναι αυτά τα ποσοστά τόσο μεγαλύτερη είναι και η επίδοση της εκάστοτε μεθόδου. Προφανώς η ιδανική περίπτωση είναι τα δύο ποσοστά να είναι ίσα με 1.

5.2 Γραμμικά Μοντέλα

5.2.1 Ανεξάρτητες Επεξηγηματικές Μεταβλητές

Θεωρούμε τα εξής δύο γραμμικά μοντέλα:

$$(i) \quad Y = c_1X_1 + c_2X_2 + c_3X_3 + c_4X_4 + \varepsilon_1,$$

$$(ii) \quad Y = c_1X_1 + c_2X_2 + c_3X_3 + c_4X_4 + \varepsilon_2,$$

όπου τα $c_i = (-1)^u(a + |z|),\ i = 1,2,3,4$ με το u να ακολουθεί την κατανομή Bernoulli με παράμετρο 0.4, το z να ακολουθεί την τυποποιημένη κανονική κατανομή και $a = \frac{4\log(n)}{\sqrt{n}}$. Επίσης θεωρούμε ότι $\varepsilon_1 \sim N(0,1)$ και $\varepsilon_2 \sim N(0,6.5)$ έτσι ώστε για το δεύτερο μοντέλο να υπάρχει ασθενέστερο *signal to noise*. Θα προσομοιώσουμε αυτά τα δύο μοντέλα με $(n,p) = (200,1000)$. Έστω ότι οι επεξηγηματικές μεταβλητές είναι ανεξάρτητες και ακολουθούν την τυποποιημένη κανονική κατανομή. Εφαρμόζουμε τις έξι προαναφερθέντες μεθόδους και τα αποτελέσματα φαίνονται στον πίνακα (5.2.1).

Πίνακας 5.2.1: Αποτελέσματα από την προσομοίωση των γραμμικών μοντέλων (i) και (ii). Η επίδοση των μεθόδων LASSO, SIS, ISIS, DC-SIS, DC-ISIS καθορίζεται από την πιθανότητα να περιλαμβάνουν το πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4\}$ σε 100 προσομοιώσεις.

		Model (i)					Model (ii)				
		P_s				P_α	P_s				P_α
(ρ, p)	Methods	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
(0.2,1000)	LASSO	1.00	1.00	1.00	1.00	1.00	0.59	0.88	0.29	0.33	0.14
	SIS	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.80	0.87	0.67
	ISIS	1.00	1.00	1.00	1.00	1.00	0.94	1.00	0.78	0.74	0.56
	DC-SIS	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.75	0.79	0.57
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	0.94	1.00	0.69	0.72	0.56
	HOLP	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.79	0.81	0.64

Παρατηρούμε ότι όταν το *signal to noise* είναι μεγάλο, δηλαδή όταν τα σφάλματα δεν επηρεάζουν το μοντέλο μας και παίρνουμε σχεδόν όλη την πληροφορία από το συστηματικό κομμάτι του μοντέλου, τότε όλες οι μέθοδοι αποδίδουν εξίσου καλά. Αντίθετα όταν έχουμε ασθενές *signal to noise* τότε οι αποδόσεις μειώνονται. Οι μέθοδοι SIS και HOLP φαίνεται να υπερτερούν σε σχέση με τις άλλες μεθόδους ενώ η μέθοδος LASSO έχει αρκετά κακή επίδοση. Οι DC-SIS και DC-ISIS αποδίδουν λίγο χειρότερα από τη SIS και την HOLP αλλά έχουν την ίδια επίδοση με την ISIS. Οι μέθοδοι SIS, ISIS και HOLP χρησιμοποιούν την πληροφορία του γραμμικού μοντέλου, δηλαδή στην ανάλυση τους θεώρησαν ότι υπάρχει γραμμική σχέση μεταξύ της απόκρισης και των επεξηγηματικών μεταβλητών αντίθετα με τις DC-SIS και DC-ISIS. Παρόλα αυτά η απόδοση των DC-SIS και DC-ISIS είναι κοντά στην απόδοση των SIS, ISIS και HOLP.

5.2.2 Ψευδώς Αρνητικά (*False Negative*)

Έστω τα γραμμικά μοντέλα :

$$(iii)\ \ Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon,$$

$$(iv)\ Y = 2.5X_1 + 2.5X_2 + 2.5X_3 - 7.5\sqrt{\rho}X_4 + \varepsilon,$$

Έστω ότι $\varepsilon \sim N(0,1)$. Επίσης θεωρούμε ότι το διάνυσμα των επεξηγηματικών μεταβλητών ακολουθεί την κανονική κατανομή $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ με τον πίνακα διασποράς-συνδιασποράς $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ να αποτελείται από τα εξής στοιχεία για $i, j = 1, 2, 3, \ldots, p$:

$$\sigma_{ii} = 1, \qquad \sigma_{4j} = \sqrt{\rho}, j \neq 4, \qquad \sigma_{i4} = \sqrt{\rho}, i \neq 4, \qquad \sigma_{ij} = \rho,\ i \neq j \neq 4,$$

όπου ρ η συσχέτιση Pearson. Δηλαδή υποθέτουμε ότι όλες οι μεταβλητές έχουν την ίδια συσχέτιση ρ μεταξύ τους εκτός από την X_4 που έχει συσχέτιση με τις υπόλοιπες $p-1$ ίση με $\sqrt{\rho}$. Θεωρούμαι τρεις διαφορετικές τιμές για το ρ η οποίες είναι η εξής: $(i)\ \rho = 0.2,\ (ii)\ \rho = 0.5, (iii)\ \rho = 0.8$. Εύκολα μπορούμε να δούμε και στα δύο μοντέλα ότι λόγω της συσχέτισης της σημαντικής μεταβλητής X_4 με τις υπόλοιπες μεταβλητές, η X_4 είναι ασυσχέτιστη με την απόκριση. Εφαρμόζουμε τις μεθόδους κρησαρίσματος για τα μοντέλα $(n, p, \rho) = (200,1000,0.2)$, $(n, p, \rho) = (200,1000,0.5)$ και $(n, p, \rho) = (200,1000,0.8)$. Τα αποτελέσματα παρουσιάζονται στον πίνακα (5.2.2).

Πίνακας 5.2.2: Αποτελέσματα από την προσομοίωση των γραμμικών μοντέλων (iii), (iv). Η επίδοση των μεθόδων LASSO, SIS, ISIS, DC-SIS, DC-ISIS καθορίζεται από την πιθανότητα να περιλαμβάνουν το πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4\}$ σε 100 προσομοιώσεις .

		Model (iii)					Model (iv)				
		P_s				P_α	P_s				P_α
(ρ, p)	Methods	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
(0.2,1000)	LASSO	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.84	0.84
	SIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.01	0.01
	ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.96
	DC-SIS	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.03	0.03
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	HOLP	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
(0.5,1000)	LASSO	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
	SIS	1.00	1.00	1.00	0.01	0.01	1.00	1.00	0.99	0.00	0.00
	ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
	DC-SIS	1.00	1.00	1.00	0.02	0.02	1.00	1.00	1.00	0.01	0.01
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	HOLP	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
(0.8,1000)	LASSO	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
	SIS	0.97	0.94	0.96	0.00	0.00	0.92	0.93	0.93	0.00	0.00
	ISIS	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.74	0.74
	DC-SIS	0.93	0.93	0.92	0.00	0.00	0.91	0.95	0.91	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.93
	HOLP	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00

Για το μοντέλο (iii) παρατηρούμε ότι η μέθοδος LASSO καταφέρνει να εντοπίσει την μεταβλητή X_4 μόνο όταν $\rho = 0.2$. Οι μέθοδοι SIS και DC-SIS, όπως αναμενόταν, καταφέρνουν να εντοπίσουν με επιτυχία όλες τις μεταβλητές πλην της X_4 με μεγάλη πιθανότητα. Η απόδοση τους μειώνεται ελάχιστα όταν η συσχέτιση των μεταβλητών γίνεται αρκετά μεγάλη ($\rho = 0.8$). Η μέθοδος HOLP επίσης εντοπίζει τις μεταβλητές X_1, X_2, X_3 αλλά αποτυγχάνει να εντοπίσει την X_4 κάτι που δεν αναμενόταν αφού σε παρόμοιο πρόβλημα όπου υπήρχε ασυσχέτιστη σημαντική μεταβλητή η μέθοδος απόδιδε εξαιρετικά και αυτό είναι κάτι που αξίζει διερεύνησης. Οι επαναληπτικές μέθοδοι DC-ISIS και ISIS επιτυγχάνουν να εντοπίσουν όλες τις μεταβλητές με πολύ μεγάλη πιθανότητα. Όσο αφορά το μοντέλο (iv), όπου το *signal to noise* είναι ασθενέστερο, η απόδοση των μεθόδων μειώνεται. Ιδιαίτερα η απόδοση της ISIS μειώνεται όταν το ρ είναι μεγάλο. Αντίθετα η μέθοδος DC-ISIS παρόλο που ως μέθοδος δεν χρησιμοποιεί οποιαδήποτε πληροφορία παλινδρόμησης, δηλαδή δεν χρησιμοποιεί την υπόθεση της γραμμικότητας όπως οι άλλες μέθοδοι, εξακολουθεί να έχει υψηλή απόδοση με ελάχιστη πτώση.

5.2.3 Ψευδώς Θετικά (*False Positive*)

Θεωρούμε το γραμμικό μοντέλο:

$$(v) \quad Y = 5X_1 + 5X_2 + 5X_3 - 15\rho X_4 + X_5 + \varepsilon.$$

Έστω ότι $\varepsilon \sim N(0,1)$ και έστω ότι το διάνυσμα των επεξηγηματικών μεταβλητών ακολουθεί την κανονική κατανομή $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ με τον πίνακα διασποράς-συνδιασποράς $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ να αποτελείται από τα εξής στοιχεία για $i, j = 1, 2, 3, \ldots, p$:

$$\sigma_{ii} = 1, \qquad \sigma_{5j} = 0, j \neq 5, \qquad \sigma_{i5} = 0, i \neq 5, \qquad \sigma_{ij} = \rho, i \neq j \neq 5,$$

όπου ρ η συσχέτιση Pearson. Είναι προφανές ότι η συσχέτιση της X_5 με τη μεταβλητή απόκρισης είναι μικρότερη από όλες τις άλλες συμμεταβλητές ακόμα και από αυτές που δεν είναι πραγματικά σημαντικές. Επίσης πάλι η X_4 είναι ασυσχέτιστη με την απόκριση λόγω της συσχέτισης της με τις άλλες μεταβλητές. Εφαρμόζουμε τις μεθόδους κρησαρίσματος για τα μοντέλα $(n, p, \rho) = (200,1000,0.5)$, $(n, p, \rho) = (200,1000,0.8)$. Τα αποτελέσματα φαίνονται στον πίνακα (5.2.3).

Οι μέθοδοι SIS και DC-SIS αδυνατούν να εντοπίσουν την μεταβλητή X_5 λόγω της μικρής της συσχέτισης με την μεταβλητή απόκρισης αλλά και την μεταβλητή X_4 για τον ίδιο λόγο με το παράδειγμα (5.2.2). Αντίθετα η DC-ISIS δυσκολεύτε να εντοπίσει μόνο τη μεταβλητή X_5. Η HOLP έχει αρκετά καλή απόδοση αλλά και αυτή δυσκολεύεται να εντοπίσει την X_5 λόγω της μικρής τιμής του συντελεστή της. Παρατηρούμε ότι, αντίθετα με το προηγούμενο παράδειγμα, εδώ η HOLP καταφέρνει να εντοπίσει την ασυσχέτιστη με την απόκριση μεταβλητή X_4. Τέλος οι ISIS και LASSO έχουν την καλύτερη απόδοση αφού εντόπισαν όλες τις μεταβλητές και στις 100 επαναλήψεις.

Πίνακας 5.2.3: Αποτελέσματα από την προσομοίωση του γραμμικού μοντέλου (v). Η επίδοση των μεθόδων LASSO, SIS, ISIS, DC-SIS, DC-ISIS καθορίζεται από την πιθανότητα να περιλαμβάνουν το πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4, X_5\}$ σε 100 προσομοιώσεις.

		Model (v)					
		P_s					P_α
(ρ, p)	Methods	X_1	X_2	X_3	X_4	X_5	ALL
(0.8,1000)	LASSO	1.00	1.00	1.00	1.00	1.00	0.00
	SIS	1.00	1.00	1.00	0.00	0.00	0.00
	ISIS	1.00	1.00	1.00	1.00	1.00	1.00
	DC-SIS	1.00	1.00	1.00	0.00	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	0.36	0.36
	HOLP	1.00	1.00	1.00	1.00	0.70	0.70
(0.8,1000)	LASSO	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	1.00	1.00	1.00	0.01	0.00	0.00
	ISIS	1.00	1.00	1.00	1.00	1.00	1.00
	DC-SIS	0.99	1.00	1.00	0.01	0.01	0.01
	DC-ISIS	1.00	1.00	1.00	1.00	0.56	0.56
	HOLP	1.00	1.00	1.00	1.00	0.76	0.76

5.2.4 Έντονη Συσχέτιση

Στόχος είναι να ελέγξουμε πως αποδίδουν οι μέθοδοι κρησαρίσματος όταν υπάρχει έντονη συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών. Έστω ότι έχουμε ένα δείγμα με $n = 200$ παρατηρήσεις και $p = 1000$ πιθανές επεξηγηματικές μεταβλητές. Έστω οι σημαντικές μεταβλητές είναι οι μεταβλητές με δείκτες $S = \{1,2,3,4,5\}$ και έστω $Z_i \sim N(0,1)$ και $W_i \sim N(0,1)$. Θεωρούμε ότι οι σημαντικές μεταβλητές δίνονται από την εξίσωση:

$$X_i = \frac{Z_i + W_i}{\sqrt{2}},\ i = 1, \ldots, 5,$$

και οι μη σημαντικές δίνονται από τις εξισώσεις:

$$X_{i+s}, X_{i+2s} = X_i + N(0,0.01), i = 1, \ldots, 5,$$

$$X_i = \frac{Z_i + \sum_{j=1}^{5} W_j}{\sqrt{2}}, i = 16, \ldots, p.$$

Θεωρούμε ότι $\boldsymbol{\beta} = (5,5,5,5,5)$. Επομένως έχουμε το γραμμικό μοντέλο:

$$(vi) \quad Y = 5X_1 + 5X_2 + 5X_3 + 5X_4 + 5X_5 + \varepsilon,$$

Μπορεί να δειχθεί ότι οι συσχετίσεις σημαντικών μεταβλητών με την απόκριση είναι μικρότερες από τις συσχετίσεις της απόκρισης με μη σημαντικές μεταβλητές. Επαναλάβαμε το πείραμα 100 φορές και πήραμε τα αποτελέσματα που παρουσιάζονται στον πίνακα (5.2.4).

Πίνακας 5.2.4: Αποτελέσματα από την προσομοίωση του γραμμικού μοντέλου (vi). Η επίδοση των μεθόδων LASSO, SIS, ISIS, DC-SIS, DC-ISIS καθορίζεται από την πιθανότητα να περιλαμβάνουν το πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4, X_5\}$ σε 100 προσομοιώσεις.

		Model (vi)					
		P_s					P_α
(ρ, p)	Methods	X_1	X_2	X_3	X_4	X_5	ALL
(200,1000)	LASSO	1.00	1.00	1.00	1.00	1.00	1.00
	SIS	0.00	0.00	0.00	0.00	0.00	0.00
	ISIS	0.00	0.00	0.00	0.00	0.00	0.00
	DC-SIS	0.00	0.00	0.00	0.00	0.00	0.00
	DC-ISIS	0.99	0.98	1.00	0.98	0.99	0.97
	HOLP	1.00	1.00	1.00	1.00	1.00	1.00

Παρατηρούμε ότι οι μέθοδοι SIS, ISIS και DC-ISIS αποτυγχάνουν πλήρως να εντοπίσουν οποιαδήποτε σημαντική μεταβλητή και αυτό οφείλεται στην έντονη συσχέτιση μεταξύ των μεταβλητών. Αντίθετα οι μέθοδοι LASSO, DC-ISIS και HOLP αποδίδουν αρκετά καλά. Πιο συγκεκριμένα η χρησιμοποίηση της DC-ISIS έτσι όπως την περιγράψαμε, εξουδετερώνει την υψηλή συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών. Επίσης το γεγονός ότι η HOLP δεν χρησιμοποιεί την συσχέτιση των συμμμεταβλητών με την απόκριση για το κρησάρισμα των μεταβλητών έχει ως αποτέλεσμα την καλή επίδοση της μεθόδου στο συγκεκριμένο μοντέλο.

5.3 Κατηγορική Μεταβλητή Απόκρισης

Στο κεφάλαιο αυτό θα ελέγξουμε την απόδοση των μεθόδων μας σε περίπτωση που έχουμε κατηγορική μεταβλητή απόκρισης. Έστω ότι:

$$Y^* = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon,$$

Και θεωρούμε το μοντέλο:

$$(vii)\ Y = \begin{cases} 1, \alpha\nu\ Y^* \in (-\infty, -3) \\ 2, \alpha\nu\ Y^* \in [-3,0) \\ 3, \alpha\nu\ Y^* \in [0,3) \\ 4, \alpha\nu\ Y^* \in [3, +\infty) \end{cases}$$

Θεωρούμε ότι το διάνυσμα των επεξηγηματικών μεταβλητών ακολουθεί την κανονική κατανομή $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ με τον πίνακα διασπορά-συνδιασποράς $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ να αποτελείται από τα εξής στοιχεία για $i, j = 1, 2, 3, \ldots, p$:

$$\sigma_{ii} = 1, \qquad \sigma_{4j} = \sqrt{\rho}, j \neq 4, \qquad \sigma_{i4} = \sqrt{\rho}, i \neq 4, \qquad \sigma_{ij} = \rho,\ i \neq j \neq 4,$$

Εφαρμόζουμε τις μεθόδους κρησαρίσματος για τα μοντέλα $(n, p, \rho) = (200,1000,0.5)$, $(n, p, \rho) = (200,1000,0.8)$. Τα αποτελέσματα φαίνονται στο πίνακα (5.3.1).

Πίνακας 5.3.1: Αποτελέσματα από την προσομοίωση του μοντέλου (vii): Επίδοση των μεθόδων LASSO, SIS, ISIS, DC-SIS, DC-ISIS καθορίζεται από την πιθανότητα να περιλαμβάνουν το πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4\}$ σε 100 προσομοιώσεις. Στο παράδειγμα αυτό μόνο οι μέθοδοι DC-SIS και DC-ISIS μπορούν να χρησιμοποιηθούν για αυτό το λόγο μόνο αυτές αξιολογούνται.

		Model (vii)				
		P_s				P_α
(ρ, p)	Methods	X_1	X_2	X_3	X_4	ALL
(0.5,1000)	LASSO	-	-	-	-	-
	SIS	-	-	-	-	-
	ISIS	-	-	-	-	-
	DC-SIS	1.00	1.00	1.00	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00
	HOLP	-	-	-	-	-
(0.8,5000)	LASSO	-	-	-	-	-
	SIS	-	-	-	-	-
	ISIS	-	-	-	-	-
	DC-SIS	0.94	0.93	0.92	0.00	0.00
	DC-ISIS	1.00	1.00	1.00	1.00	1.00
	HOLP	-	-	-	-	-

Οι μόνες μέθοδοι που μπορούν να εφαρμοστούν σε κατηγορική μεταβλητή απόκρισης είναι οι DC-SIS και DC-ISIS γιατί η συσχέτιση της απόστασης (*distance correlation*) που παρουσιάσαμε στο κεφάλαιο 4 μπορεί να χρησιμοποιηθεί σε οποιοδήποτε είδος μεταβλητών. Προφανώς η DC-ISIS αποδίδει καλύτερα αφού εντοπίζει όλες τις σημαντικές μεταβλητές με πιθανότητα 1, αντίθετα με την μέθοδο DC-SIS η οποία όχι μόνο αποτυγχάνει να επιλέξει την X_4 αλλά μειώνεται και η απόδοση της καθώς η συσχέτιση των μεταβλητών αυξάνεται.

5.4 Μη γραμμικά Μοντέλα

Αφού μελετήσαμε τα γραμμικά μοντέλα θα θέλαμε να δούμε την επίδοση των μεθόδων μας και στα μη γραμμικά. Θεωρούμε το μη γραμμικό μοντέλο:

$$(viii)\ \ Y = log\left(1 + \left|4 + 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon\right|\right),$$

Έστω και πάλι ότι το διάνυσμα των επεξηγηματικών μεταβλητών ακολουθεί την κανονική κατανομή $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ με τον πίνακα διασπορά συνδιασποράς $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ να αποτελείται απ' τα εξής στοιχεία για $i, j = 1, 2, 3, \ldots, p$:

$$\sigma_{ii} = 1, \qquad \sigma_{4j} = \sqrt{\rho}, j \neq 4, \qquad \sigma_{i4} = \sqrt{\rho}, i \neq 4, \qquad \sigma_{ij} = \rho,\ i \neq j \neq 4,$$

όπου ρ η συσχέτιση Pearson. Θεωρούμε ότι $\rho = 0.5$. Επομένως θα μελετήσουμε το μοντέλο: $(n, p, \rho) = (200,1000,0.5)$. Τα αποτελέσματα φαίνονται στον πίνακα (5.4.1).

Πίνακας 5.4.1: Αποτελέσματα από την προσομοίωση του μη γραμμικού μοντέλου (vii). Η επίδοση των μεθόδων LASSO, SIS, ISIS, DC-SIS, DC-ISIS καθορίζεται από την πιθανότητα να περιλαμβάνουν το πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4\}$ σε 100 προσομοιώσεις.

		Model $(viii)$				
		P_s				P_α
(ρ, p)	Methods	X_1	X_2	X_3	X_4	ALL
(0.5,1000)	LASSO	0.86	0.85	0.88	0.00	0.00
	SIS	0.95	0.98	0.98	0.00	0.00
	ISIS	0.74	0.72	0.69	0.84	0.35
	DC-SIS	0.96	0.96	0.98	0.01	0.01
	DC-ISIS	1.00	1.00	0.99	0.90	0.90
	HOLP	0.49	0.42	0.47	0.00	0.00

Είναι εμφανές ότι οι μέθοδοι DC-SIS και DC-ISIS αποδίδουν καλύτερα από τις άλλες μεθόδους επειδή δεν θεωρούν ότι υπάρχει γραμμική συσχέτιση μεταξύ απόκρισης και επεξηγηματικών μεταβλητών. Φυσικά η DC-ISIS υπερτερεί της DC-SIS λόγω του ότι επιτυγχάνει να εντοπίσει την μεταβλητή X_4 όπως είδαμε και στα προηγούμενα παραδείγματα.

5.5 Επίδραση Αλληλεπιδράσεων

Τέλος μελετάμε την δυνατότητα επιλογής αλληλεπιδράσεων από τις μεθόδους κατά το κρησάρισμα μεταβλητών.

Θεωρούμε τα μοντέλα:

$$(ix)\ \ Y = 3X_1 + X_{101} + 0.8X_{201} + 0.6X_{202} - 1.5X_1X_{101} - 2X_{201}X_{202} + \varepsilon,$$

$$(x)\ Y = 3X_1 + X_{101} + 0.8X_{201} + 0.6X_{202} - 1.5(X_1X_{101})^2 - 2X_{201}X_{202} + \varepsilon.$$

Έστω ότι $\varepsilon \sim N(0,1)$ και έστω ότι το διάνυσμα των επεξηγηματικών μεταβλητών ακολουθεί την κανονική κατανομή $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ με τον πίνακα διασπορά-συνδιασποράς $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ να αποτελείται απ' τα εξής στοιχεία: $\sigma_{ij} = \rho^{|i-j|},\ i,j = 1,2,3,\ldots,p$, όπου ρ η συσχέτιση Pearson. Επίσης αντικαθιστούμε κάθε X_k με την $X_k = 0.8X_1 + \zeta_k$, όπου $\zeta_k \sim N(0,1)\ \gamma\iota\alpha\ \kappa = 2,3,\ldots,100$, έτσι ώστε οι πρώτες 99 μεταβλητές να έχουν υψηλή συσχέτιση με την X_1. Λόγω της συσχέτισης αυτής, οι μέθοδοι κρησαρίσματος που χρησιμοποιούν την συσχέτιση ως χαρακτηριστικό για το κρησάρισμα μεταβλητών δύσκολα εντοπίζουν οποιαδήποτε σημαντική μεταβλητή εκτός της X_1. Στον πίνακα (5.5.1) παρουσιάζονται τα αποτελέσματα από τις μεθόδους SIS, ISIS, DC-SIS, DC-ISIS. Σημειώνουμε ότι στο παράδειγμα αυτό ο συμβολισμός DC-ISIS αναφέρεται στην επαναληπτική DC-ISIS που αναφέραμε στο κεφάλαιο (4.4.1) και σχετίζεται με τον

εντοπισμό των σημαντικών αλληλεπιδράσεων. Επίσης για την μέθοδο ISIS χρησιμοποιείται πανομοιότυπη μέθοδος με την επαναληπτική DC-ISIS.

Πίνακας 5.5.1: Αποτελέσματα από την προσομοίωση των μοντέλων (ix), (x). Σημειώνουμε ότι στην επαναληπτική DC-ISIS θέσαμε $d_1 = \frac{d}{2} = [\frac{n}{logn}]$ και $d_2 = \frac{d}{2} = [\frac{n}{logn}]$.

		Model (ix)						
		P_s						P_α
(ρ, p)	Methods	X_1	X_{101}	X_{201}	X_{202}	X_1X_{101}	$X_{201}X_{202}$	ALL
(0.5,1000)	SIS	1.00	0.00	0.03	0.01	0.00	0.01	0.00
	ISIS	0.68	0.91	0.88	0.77	0.39	0.69	0.28
	DC-SIS	1.00	0.02	0.03	0.02	0.02	0.01	0.00
	DC-ISIS	1.00	1.00	1.00	0.99	0.99	0.99	0.98
(0.8,1000)	SIS	1.00	0.01	0.07	0.07	0.01	0.06	0.00
	ISIS	0.65	0.83	0.75	0.67	0.30	0.48	0.17
	DC-SIS	1.00	0.03	0.20	0.17	0.02	0.13	0.01
	DC-ISIS	1.00	0.99	0.99	0.98	0.90	0.97	0.87
		Model (x)						
(0.5,1000)	SIS	0.99	0.08	0.07	0.03	0.01	0.02	0.00
	ISIS	0.93	0.61	0.64	0.52	0.06	0.36	0.02
	DC-SIS	1.00	0.05	0.09	0.03	0.05	0.02	0.00
	DC-ISIS	1.00	1.00	0.98	0.98	0.99	0.94	0.93
(0.8,1000)	SIS	0.98	0.11	0.14	0.13	0.02	0.09	0.01
	ISIS	0.92	0.55	0.57	0.47	0.06	0.27	0.02
	DC-SIS	1.00	0.05	0.27	0.24	0.05	0.20	0.03
	DC-ISIS	1.00	0.98	0.98	0.97	0.96	0.94	0.90

Σύμφωνα με τον πίνακα (5.5.1) η επαναληπτική μέθοδος DC-ISIS συμπεριφέρεται εξαιρετικά αφού εντοπίζει τόσο τις κύριες σημαντικές μεταβλητές καθώς και τις σημαντικές αλληλεπιδράσεις αντίθετα με την παρόμοια μέθοδο ISIS. Ιδιαίτερα στο μοντέλο (x) όπου υπάρχει ο τετραγωνικός όρος $(X_1X_{101})^2$ η επαναληπτική DC-ISIS διατηρεί στο μοντέλο την αλληλεπίδραση X_1X_{101} με μεγάλη πιθανότητα αντίθετα με την ISIS που μόνο τυχαία μπορεί να την διατηρήσει.

5.6 Συμπεράσματα

Το γεγονός ότι η μέθοδος DC-ISIS παρουσιάζει καλύτερη απόδοση σε μη γραμμικά μοντέλα από ότι οι υπόλοιπες μέθοδοι, καθώς και το γεγονός ότι μπορεί να χρησιμοποιηθεί σε μεταβλητές οποιασδήποτε μορφής, κάνει την μέθοδο πολύ επιθυμητή και κατά τη γνώμη μας αποτελεί την καλύτερη μέθοδο κρησαρίσματος σε σχέση με τις μεθόδους που παρουσιάστηκαν στην παρούσα διπλωματική εργασία. Θεωρούμε την μέθοδο DC-ISIS ως την καλύτερη μέθοδο κρησαρίσματος και όχι την DC-SIS, γιατί παρόλο που έχουν τις ίδιες ιδιότητες λόγω της χρησιμοποίησης της συσχέτισης της απόστασης, η πρώτη επιλύει προβλήματα που παρουσιάζει η δεύτερη που έχουν να κάνουν με την συσχέτιση των μεταβλητών. Επίσης με μια μικρή παραλλαγή της DC-ISIS, καθιστά δυνατό το επιτυχές

κρησάρισμα των αλληλεπιδράσεων μεταξύ των σημαντικών μεταβλητών. Σχετικά με τις υπόλοιπες μεθόδους, η SIS παρουσιάζει προβλήματα που προκύπτουν από την συσχέτιση των επεξηγηματικών μεταβλητών όπως είδαμε εκτενώς στην παρούσα διπλωματική εργασία, προβλήματα που επιλύει η επαναληπτική μέθοδος ISIS. Παρόλα αυτά η ISIS στα μη γραμμικά μοντέλα καθώς και σε μοντέλα όπου το signal to noise είναι ασθενές αποδίδει χειρότερα από την SIS. Τέλος η HOLP αποδίδει αρκετά καλά σε γραμμικά μοντέλα αφού επιλύει προβλήματα που εμφανίζονται στις μεθόδους συσχέτισης, SIS και DC-SIS και έχουν να κάνουν με την συσχέτιση των μεταβλητών. Η απόδοση της μειώνεται δραστικά όταν η σχέση της απόκρισης με τις σημαντικές επεξηγηματικές μεταβλητές παύει να είναι γραμμική.

ΒΙΒΛΙΟΓΡΑΦΙΑ (*REFERENCES*)

1. P. Oikonomou and Ch. Karwni (2010) Στατιστικά Μοντέλα Παλινδρόμησης.

2. G.Kokolakis and D. Fouskakis (2009) Στατιστική Θεωρία και Εφαρμογές.

3. Jianqing Fan and Runze Li (2001) Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

4. Jianqing Fan and Jinchi Lv (2010) A Selective Overview of Variables in High Dimensional Features, *Statistica Sinica*, **20**, 101-148.

5. Hui Zou (2006) The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, **101**, 1418-1429.

6. Robert Tibshiriani (1996) Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**, 267-288.

7. Knigh, K. and Fu, W. (2000) Asymptotic for Lasso type estimators, *Ann. Statistics*, **28**, 1356-1378.

8. Efron, Hastie, Johnstone & Tibshirani (2004) Least angle regression, *The Annals of Statistics*, **32**, 407-499.

9. Jianqing Fan and Fang Han & Han Liu (2014) Challenges of Big Data Analysis, *National Science Review*.

10. Fan, J. and Lv, J. (2008) Sure Independence Screening for ultra-high dimensional feature space, *J. Roy. Statistic. Soc. Ser. B*, **70**, 849-911.

11. Gabor, J. Szekely, Maria, L.Rizzo and Nail, K. Bakirov (2007) Measuring And Testing Dependence by correlation of Distances, *The Annals of Statistics*, **35**, 2769-2794

12. Wei Zhong and Liping Zhu (2015) An iterative approach to distance correlation-based sure independence screening, *Journal of Statistical Computation and Simulation*, **85**, 2331-2345

13. Li, R., Zhong, W., Zhu, L. (2012) Feature screening via distance correlation learning', *Journal of American Statistical Association*, **107**, 1129-1139

14. Xiangyu Wang and Chenlei Leng (2015) High-dimensional Ordinary Least-squares Projection for Screening Variables.

15. Meinhausen, N. and Buhlmann, P. (2008). High Dimensional Graphs and variable selection with the Lasso, *The Annals of Statistics*, **34**, 1436-1462

16. Donoho, D. L (2000) High Dimensional data Analysis: the curses and blessings of dimensionality, *American Mathematical Society Conf. Math Challenges of 21th Century.*

17. Donoho, D. L. and Elad, M. (2003) Maximal Sparsity representation via l_1 minimization, *Proc. Natn. Acad. Sci USA*, **100**, 2197-2202

18. Fan, J. and Li, R. (2006) Statistical Challenges with high dimensionality: feature selection in knowledge of discovery, *Proc Int. Congr. Mathematicians*, vol **III**, 595-622

19. Greenshtein E (2006) Best Subset selection, persistence in high dimensional statistical learning and optimization under l_1 constraint, *Ann. Statistics*, **34**, 2367-2386

20. Zhang, C.-H. (2009). Penalize linear unbiased selection. *Ann. Statist.*

21. Bickel, P. J. (2008) Discussion of "Sure Independence Screening for Ultra high dimensional feature space", *J. Roy. Statist. Soc. Ser. B*, **70**, 883-884

22. Bickel, P. J. and Li, B. (2006) Regularizations in statistics (with discussion). Test **15**, 271-344

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΡΤΗΜΑ 1: ΑΠΟΔΕΙΞΕΙΣ ΘΕΩΡΗΜΑΤΩΝ ΤΗΣ ΜΕΘΟΔΟΥ HOLP

Έστω το γραμμικό μοντέλο:

$$Y = \sum_{i=1}^{p} \beta_i X_i + \varepsilon, \tag{6.1.1}$$

και θεωρούμε τον μετασχηματισμό $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ όπου $\boldsymbol{x} = (X_1, \dots, X_p)'$ το διάνυσμα συμμεταβλητών και $\boldsymbol{\Sigma} = cov(\boldsymbol{x})$. Προφανώς ο πίνακας διασποράς-συνδιασποράς του $\boldsymbol{z}$ είναι ο $\boldsymbol{I_p}$. Επίσης θεωρούμε το μετασχηματισμένο πίνακα σχεδιασμού $\boldsymbol{Z} = \boldsymbol{X\Sigma}^{-1/2} \Longrightarrow \boldsymbol{X} = \boldsymbol{Z\Sigma}^{1/2}$. Θεωρούμε ότι οι $X_1, \dots, X_p$ είναι τυποποιημένες τυχαίες μεταβλητές με μέσο 0 και διασπορά 1.

Όπως είδαμε στο κεφάλαιο 3, ο εκτιμητής κρησαρίσματος HOLP είναι ο εξής:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'(\boldsymbol{XX}')^{-1}\boldsymbol{y} = \boldsymbol{X}'(\boldsymbol{XX}')^{-1}\boldsymbol{X\beta} + \boldsymbol{X}'(\boldsymbol{XX}')^{-1}\boldsymbol{\varepsilon} := \xi + \eta. \tag{6.1.2}$$

Θεωρούμε την αποσύνθεση μοναδική τιμής (*singular value decomposition)* του $\boldsymbol{Z}$ δηλαδή $\boldsymbol{Z} = \boldsymbol{VDU}'$ όπου $\boldsymbol{V}$ είναι ένας $n \times n$ ορθογωναδιαίος πίνακάς, $\boldsymbol{U} \in V_{\boldsymbol{n,p}}$ και $\boldsymbol{D}$ είναι ένα $n \times n$ διαγώνιος πίνακας. Επομένως ισχύει:

$$\boldsymbol{X} = \boldsymbol{Z\Sigma}^{1/2} = \boldsymbol{VDU}'\boldsymbol{\Sigma}^{1/2}. \tag{6.1.3}$$

Ο πίνακας προβολής $\boldsymbol{X}'(\boldsymbol{XX}')^{-1}\boldsymbol{X}$ γράφεται ως εξής:

$$\begin{aligned} \boldsymbol{X}'(\boldsymbol{XX}')^{-1}\boldsymbol{X} &= \boldsymbol{\Sigma}^{1/2}\boldsymbol{UV}'(\boldsymbol{VDU}'\boldsymbol{\Sigma UDV}')^{-1}\boldsymbol{VDU}'\boldsymbol{\Sigma}^{1/2} \\ &= \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{\Sigma U})^{-1}\boldsymbol{U}'\boldsymbol{\Sigma}^{1/2}\boldsymbol{HH}', \end{aligned} \tag{6.1.4}$$

όπου $\boldsymbol{H} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{\Sigma U})^{-1/2}$, και ικανοποιεί την ισότητα $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$.

Επομένως μπορούμε να γράψουμε το ξ ως:

$$\xi = \boldsymbol{HH}'\boldsymbol{\beta}. \tag{6.1.5}$$

Στη συνέχεια της ανάλυσης μας θεωρούμε ότι $\boldsymbol{e_i} = (0, \dots, 1, 0, \dots, 0)'$ αποτελεί την $i - o\sigma\tau$ή φυσική βάση του p −διαστατου χώρου. Επίσης θα αναφέρουμε τρία λήμματα τα οποία θα χρησιμοποιηθούν χωρίς απόδειξη, για την απόδειξη των θεωρημάτων της

μεθόδου HOLP. Για την απόδειξη των λημμάτων αυτών βλέπε το άρθρο των Wang και Len (2015).

<u>Λήμμα 1:</u> Αν οι υποθέσεις (1.), (3.) που είδαμε στο κεφάλαιο (3.2.3) ισχύουν τότε για $C > 0$ και για οποιοδήποτε σταθερό διάνυσμα $\boldsymbol{v}$ με $\|\boldsymbol{v}\| = 1$, υπάρχουν σταθερές c_1' , c_2', με $0 < c_1' < 1 < c_2'$ έτσι ώστε:

$$P\left(\boldsymbol{v}'\boldsymbol{H}\boldsymbol{H}'\boldsymbol{v} < c_1'\frac{n^{1-\tau}}{p} \text{ ή } \boldsymbol{v}'\boldsymbol{H}\boldsymbol{H}'\boldsymbol{v} > c_2'\frac{n^{1+\tau}}{p}\right) < 4\exp(-Cn). \tag{6.1.6}$$

Για $\boldsymbol{v} = \boldsymbol{\beta}$ το οποίο δεν έχει νόρμα ίση με 1 ισχύει:

$$P\left(\boldsymbol{\beta}'\boldsymbol{H}\boldsymbol{H}'\boldsymbol{\beta} > c_2'\frac{n^{1+\tau}}{p}\right) < 2\exp(-Cn). \tag{6.1.7}$$

<u>Λήμμα 2:</u> Αν οι υποθέσεις (1.), (3.) που είδαμε στο κεφάλαιο (3.2.3) ισχύουν τότε για οποιοδήποτε $C > 0$ υπάρχουν $c, \tilde{c} > 0$, έτσι ώστε για κάθε $i \in S$ να ισχύει:

$$P\left(|\boldsymbol{e_i}\boldsymbol{H}\boldsymbol{H}'\boldsymbol{\beta}| <. c\frac{n^{1-\tau-\kappa}}{p}\right) \leq O\left\{exp\left(\frac{-Cn^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\}. \tag{6.1.8}$$

και για κάθε $i \notin S$ να ισχύει:

$$P\left(|\boldsymbol{e_i}\boldsymbol{H}\boldsymbol{H}'\boldsymbol{\beta}| > \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) \leq O\left\{exp\left(\frac{-Cn^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\}, \tag{6.1.9}$$

με τ, κ, ν να είναι οι παράμετροι που καθορίζονται από την υπόθεση (3.).

<u>Λήμμα 3:</u> Αν οι υποθέσεις (1.)-(3.) που είδαμε στο κεφάλαιο (3.2.3) ισχύουν τότε για οποιοδήποτε $i \in \{1,2,\ldots,n)$ ισχύει:

$$P\left(|\eta_i| > \frac{\sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) <$$
$$< exp\left(1 - q\left(\frac{\sqrt{C_1}n^{1-5\tau-2\kappa-\nu}}{\sqrt{\log(n)}}\right)\right) + 3\exp(-C_1 n), \tag{6.1.10}$$

όπου C_1, c_1, c_4 καθορίζονται από τις υποθέσεις και το c_2' από το λήμμα 1 και σ η τυπική απόκλιση του σφάλματος.

<u>Θεώρημα 1:</u> Υπό τις υποθέσεις (1.)-(3.) αν διαλέξουμε γ_n έτσι ώστε:

$$\frac{p\gamma_n}{n^{1-\tau-\kappa}} \xrightarrow{n\to\infty} 0 \; \kappa\alpha\iota \; \frac{p\gamma_n\sqrt{\log(n)}}{n^{1-\tau-\kappa}} \xrightarrow{n\to\infty} \infty,$$

για το $C_1 > 0$ που ορίζεται από την υπόθεση (1.) ισχύει:

$$\mathrm{P}\left(M_S \subset \widehat{M}_{\gamma_n}\right) =$$

$$= 1 - O\left\{\exp\left(-C_1 \frac{n^{1-2\kappa-5\tau-\nu}}{2\log(n)}\right)\right\} - s \cdot exp\left\{1 - q\left(\frac{\sqrt{C_1} n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\}$$
$$+3s\exp(-C_1 n). \tag{6.1.11}$$

<u>Απόδειξη:</u> Εφαρμόζοντας τα λήμματα 2 και 3 για όλα τα $i \in S$ έχουμε:

$$P\left(\min_{i\in S}|\xi_i| < c\frac{n^{1-\tau-\kappa}}{p}\right) = O\left\{sexp\left(\frac{-Cn^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\}, \tag{6.1.12}$$

και

$$P\left(\max_{i\in S}|\eta_i| > \frac{\sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) =$$

$$= s \cdot exp\left(1 - q\left(\frac{\sqrt{C_1} n^{1-5\tau-2\kappa-\nu}}{\sqrt{\log(n)}}\right)\right) + 3s \cdot \exp(-C_1 n). \tag{6.1.13}$$

Επειδή $s = c_3 n^{\nu}$ $\mu\varepsilon$ $\nu < 1$, θεωρώντας ότι $C = C_1$ η (1) μετατρέπεται ως εξής:

$$P\left(\min_{i\in S}|\xi_i| < c\frac{n^{1-\tau-\kappa}}{p}\right) = O\left\{exp\left(\frac{-C_1 n^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\}. \tag{6.1.14}$$

Αν διαλέξουμε γ_n έτσι ώστε:

$$\frac{\tilde{c} + \sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p} < \gamma_n < \frac{c}{2}\frac{n^{1-\tau-\kappa}}{p},$$

ή αν ικανοποιείται :

$$\frac{p\gamma_n}{n^{1-\tau-\kappa}} \xrightarrow{n\to\infty} 0 \ \kappa\alpha\iota \ \frac{p\gamma_n\sqrt{\log(n)}}{n^{1-\tau-\kappa}} \xrightarrow{n\to\infty} \infty,$$

Τότε ισχύει:

$$P\left(\min_{i\in S}|\beta_i| < \gamma_n\right) = P\left(\min_{i\in S}|\xi_i + \eta_i| < \gamma_n\right) \leq$$

$$\leq P\left(\min_{i\in S}|\xi_i| < c\frac{n^{1-\tau-\kappa}}{p}\right) + P\left(\max_{i\in S}|\eta_i| > \frac{\sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right)$$

$$= O\left\{exp\left(\frac{-C_1 n^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\} + s \cdot exp\left(1 - q\left(\frac{\sqrt{C_1} n^{1-5\tau-2\kappa-\nu}}{\sqrt{\log(n)}}\right)\right) \tag{6.1.15}$$
$$+ 3s\exp(-C_1 n).$$

Επομένως έχουμε:

$$\mathrm{P}\left(M_S \subset \widehat{M}_{\gamma_n}\right) = P\left(\min_{i\in S}|\beta_i| \geq \gamma_n\right) = 1 - P\left(\min_{i\in S}|\beta_i| < \gamma_n\right) \geq$$

$$\geq 1 - O\left\{exp\left(\frac{-C_1 n^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\} + s \cdot exp\left(1 - q\left(\frac{\sqrt{C_1}n^{1-5\tau-2\kappa-\nu}}{\sqrt{\log(n)}}\right)\right)$$

$$+3s\exp(-C_1 n).$$

■

<u>Θεώρημα 2:</u> Αν ισχύουν οι υποθέσεις του θεωρήματος 1 και για το p ισχύει:

$$logp = o(min\left\{\frac{n^{1-2\kappa-5\tau}}{2\log n}, q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\},$$

τότε για το ίδιο γ_n που ορίζεται στο θεώρημα 1 και για το ίδιο C_1 που καθορίζεται από την υπόθεση (1.) έχουμε:

$$(i)\quad P\left(\min_{j\in S}|\hat{\beta}_j| > \gamma_n > \max_{j\notin S}|\hat{\beta}_j|\right) =$$

$$= 1 - O\left\{\exp\left\{-C_1\frac{n^{1-2\kappa-5\tau-\nu}}{2\log(n)}\right\} - exp\left\{1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\}\right\}. \qquad (6.1.16)$$

Ισοδύναμα, μπορούμε να επιλέξουμε ένα υποσύνολο M_d με $d = n^{\iota}$ με $\iota \in (\nu, 1]$ έτσι ώστε:

$$(ii)\quad P(M_S \subset M_d) =$$

$$= 1 - O\left\{\exp\left\{-C_1\frac{n^{1-2\kappa-5\tau-\nu}}{2\log(n)}\right\} - exp\left\{1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\}\right\}. \qquad (6.1.17)$$

<u>Απόδειξη:</u> Από το λήμμα 2 για κάθε $i \notin S$ και κάθε $C > 0$ υπάρχει ένα $\tilde{c} > 0$ έτσι ώστε:

$$P\left(|\boldsymbol{e_i HH'\beta}| > \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) \leq O\left\{exp\left(\frac{-Cn^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\}, \qquad (6.1.18)$$

Χρησιμοποιώντας την ανισότητα Bonferroni ισχύει:

$$P\left(\max_{i\notin S}|\xi_i| < \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) = P\left(|\boldsymbol{e_i HH'\beta}| > \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right)$$

$$< O\left\{pexp\left(\frac{-Cn^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\}. \qquad (6.1.19)$$

Όμοια από το λήμμα 6 παίρνουμε:

$$P\left(\max_{i\in S}|\eta_i| > \frac{\sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) <$$
$$< p\cdot exp\left(1 - q\left(\frac{\sqrt{C_1}n^{1-5\tau-2\kappa-\nu}}{\sqrt{\log(n)}}\right)\right) + 3p\cdot\exp(-C_1 n). \qquad (6.1.20)$$

Αν ισχύει ότι:

$$logp = o(min\left\{\frac{n^{1-2\kappa-5\tau}}{2\log n}, q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\},$$

τότε για το C_1 που καθορίστηκε στην υπόθεση (2.) ισχύει:

$$P\left(\max_{i\notin S}|\xi_i| < \frac{\tilde{c}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) < O\left\{exp\left(\frac{-Cn^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right)\right\}, \qquad (6.1.21)$$

$$P\left(\max_{i\in S}|\eta_i| > \frac{\sigma\sqrt{C_1 c_1 c_2' c_4}}{\sqrt{\log(n)}}\frac{n^{1-\tau-\kappa}}{p}\right) <$$
$$< O\left\{exp\left(1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{1-5\tau-2\kappa-\nu}}{\sqrt{\log(n)}}\right)\right) + \exp\left(-\frac{C_1}{2}n\right)\right\}. \qquad (6.1.22)$$

Αν διαλέξουμε γ_n όπως το θεώρημα 1 τότε:

$$P\left(\min_{i\notin S}|\beta_i| > \gamma_n\right) <$$
$$< O\left\{exp\left(\frac{-Cn^{1-5\tau-2\kappa-\nu}}{2\log(n)}\right) + exp\left(1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{1-5\tau-2\kappa-\nu}}{\sqrt{\log(n)}}\right)\right)\right\}. \qquad (6.1.23)$$

Ενώνοντας το προηγούμενο αποτέλεσμα με το θεώρημα 1 και παρατηρώντας ότι $s < p$ τότε:

$$P\left(\min_{j\in S}|\hat{\beta}_j| > \gamma_n > \max_{j\notin S}|\hat{\beta}_j|\right) =$$
$$= 1 - O\left\{\exp\left\{-C_1 n^{1-2\kappa-5\tau-\frac{\nu}{2\log(n)}}\right\} - s\,exp\left\{1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\}\right\}, \qquad (6.1.24)$$

Επομένως επιλέγοντας υποσύνολο με μέγεθος d ≥ s τότε:

$$P(M_S \subset M_d) =$$
$$= 1 - O\left\{\exp\left\{-C_1\frac{n^{1-2\kappa-5\tau-\nu}}{2\log(n)}\right\} - exp\left\{1 - \frac{1}{2}q\left(\frac{\sqrt{C_1}n^{\frac{1}{2}-2\tau-\kappa}}{\sqrt{logn}}\right)\right\}\right\}.$$

■

ΠΑΡΑΡΤΗΜΑ 2: ΚΩΔΙΚΕΣ R

Μέθοδος SIS

```
SIS1<-function (x, y, d, penalty= "")
{ x1<-scale(x); yy<-y/(sd(y)*(n-1))
  b<-t(x1)%*%yy;  b_abs=abs(b)
  imp_pred<-sort(b_abs, decreasing=TRUE, index.return=TRUE)
  h<-sort(imp_pred$ix[1:d])
  z1=0; b1=0; z2=0; b2=0; z2=0; b2=0, z3=0; b3=0
 cv=cv.ncvreg(x[,h],y)
  #penalty choose
if (penalty=="SCAD")
  {
    fitscad = ncvreg(x[,h],y,family="gaussian",penalty="SCAD",lambda=cv$lambda.min)
    w<-rep(0,d)
    for (i in 2:(d+1))
      {if (fitscad$beta[i]!=0)
        {w[i-1]<-i-1}}

    w<-w[w!=0]
    k<-1
    z1<-rep(0,length(h))
    for (i in w)
    { z1[k]=h[i]
      k<-k+1}
    z1<-z1[z1!=0]
    b1<-fitscad$beta[w+1]}

  if (penalty=="MCP")
  {
    fitmcp = ncvreg(x[,h],y,family="gaussian",penalty="MCP",lambda=cv$lambda.min)
    w<-rep(0,d)
    for (i in 2:(d+1))
     {if (fitmcp$beta[i]!=0)
      {w[i-1]<-i-1}}
    w<-w[w!=0]
    k<-1
    z2<-rep(0,length(h))
    for (i in w)
    { z2[k]=h[i]
      k<-k+1}
    z2<-z2[z2!=0]
    b2<-fitmcp$beta[w+1]}
  if (penalty=="LASSO")

{fitlasso=ncvreg(x[,h],y,type.gaussian="covariance",penalty="lasso",lambda=cv$lambda.min)
    w<-rep(0,d)
    for (i in 2:(d+1))
```

```
    {if (fitlasso$beta[i]!=0)
   w<-w[w!=0]
   k<-1
   z3<-rep(0,length(h))
   for (i in w)
   { z3[k]=h[i]
     k<-k+1}
  #results
     list(imp_pred=sort(imp_pred$ix[1:d]),sort_imp=imp_pred$ix[1:d],
      SIS_SCAD=z1,SIS_MCP=z2,SIS_LASSO=z3,
      SIS_SCAD_b=b1,SIS_MCP_b=b2,SIS_LASSO_b=b3)
}
```

Μέθοδος HOLP

```
HOLP<-function (x, y, d, penalty=" ")
{
  x1<-scale(x)
  b<-t(x1)%*%ginv(x1%*%t(x1))%*%y; b_abs=abs(b)
  imp_pred<-sort(b_abs, decreasing=TRUE, index.return=TRUE)
  h<-sort(imp_pred$ix[1:d])
  z1=0; b1=0; z2=0;b2=0; z3=0; b3=0
 cv=cv.ncvreg(x[,h],y)
  #penalty choose
if (penalty=="SCAD")
  {
   fitscad = ncvreg(x[,h],y,family="gaussian",penalty="SCAD",lambda=cv$lambda.min)
   w<-rep(0,d)
   for (i in 2:(d+1))
     {if (fitscad$beta[i]!=0)
       {w[i-1]<-i-1}}

   w<-w[w!=0]
   k<-1
   z1<-rep(0,length(h))
   for (i in w)
   { z1[k]=h[i]
     k<-k+1}
   z1<-z1[z1!=0]
   b1<-fitscad$beta[w+1]}
  if (penalty=="MCP")
  {fitmcp = ncvreg(x[,h],y,family="gaussian",penalty="MCP",lambda=cv$lambda.min)
   w<-rep(0,d)
   for (i in 2:(d+1))
    {if (fitmcp$beta[i]!=0)
     {w[i-1]<-i-1}}
   w<-w[w!=0]
   k<-1
   z2<-rep(0,length(h))
   for (i in w)
```

```
   { z2[k]=h[i]
     k<-k+1}
   z2<-z2[z2!=0]
   b2<-fitmcp$beta[w+1]}
 if (penalty=="LASSO")
fitlasso=ncvreg(x[,h],y,type.gaussian="covariance",penalty="lasso",lambda=cv$lambda.min)
   w<-rep(0,d)
   for (i in 2:(d+1))
     {if (fitlasso$beta[i]!=0)
     {w[i-1]<-i-1}}
   w<-w[w!=0]
   k<-1
   z3<-rep(0,length(h))
   for (i in w)
   { z3[k]=h[i]
     k<-k+1}
   z3<-z3[z3!=0]
 #results
 list(imp_pred=sort(imp_pred$ix[1:d]),sorted_imp_pred=imp_pred$ix[1:d],
      HOLP_SCAD=z1,HOLP_MCP=z2,HOLP_LASSO=z3,
      HOLP_SCAD_b=b1,HOLP_MCP_b=b2,HOLP_LASSO_b=b3)
}
```

Μέθοδος Divide HOLP

```
DIVIDE_HOLP<-function (x, y, d, m)
 {g<-c(1:n)
  h<-c(1:n)
  important_pred1<-rep(0,ncol(x))
  j<-1
  m<-2
  for (i in 1:m)
{   l<-h[1:(n/m)]
   y1<-y[l]
   x1<-x[l,]
   b1<-t(x1)%*%ginv(x1%*%t(x1))%*%y1
   b1_abs=abs(b1)
   imp_pred1<-sort(b1_abs, decreasing=TRUE, index.return=TRUE)
   v<-min(n/m,d/m)
   index1<-sort(imp_pred1$ix[1:v])
   important_pred1=union(important_pred1,index1)
   g1<-g[-l]
#results
  list(imp_pred=sort(important_pred1[important_pred1!=0]))
}
```

Μέθοδος DC-SIS

```
DCSIS<-function (x, y, d)
{distcors<-rep(0,ncol(x))
for(i in 1:ncol(x))
{distcors[i]=dcor(y,x[,i])}
sorted_distcors=sort(distcors,decreasing=TRUE, index.return=TRUE)
sorted_imp_pred_by_distcor=sorted_distcors$ix[c(1:d)]
important_predictors=sort(sorted_imp_pred_by_distcor)
#results
list(imp_pred=important_predictors,
    sorted_pred_by_cor=sorted_imp_pred_by_distcor)
}
```

Μέθοδος DC-ISIS

```
DCISIS<-function (x, y, d, p1, iter)
{
   predictors1=DCSIS(x,y,p1)
  I=diag(nrow(x))
  z=predictors1$imp_pred
  z1<-0
  i=1
  while (i<iter)
  {
   candind=setdiff(1:p,z)
   x1= x[,z]
   x1c=x[,-z]
   xnew=(I-x1%*%(ginv(t(x1)%*%x1))%*%t(x1))%*%x1c
   predictors2=DCSIS(xnew,y,(d/iter))
   newix=candind[predictors2$imp_pred]
   sorted_newix=candind[predictors2$sorted_pred_by_cor]
   z1=union(z1,sorted_newix)
   z=union(z,newix)
   i=i+1
  }
#results
  list(imp_pred=z,
      sorted_pred_by_cor=z1)
}
```