ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ ΣΕ ΔΕΔΟΜΕΝΑ ΥΨΗΛΗΣ ΔΙΑΣΤΑΣΗΣ ΓΙΑ ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΟΥ ΚΙΝΔΥΝΟΥ ΤΟΥ COX

ΔΗΜΗΤΡΑΚΟΠΟΥΛΟΣ ΠΑΝΑΓΙΩΤΗΣ

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

Αθήνα, Φεβρουάριος 2014

ΠΕΡΙΛΗΨΗ

Η επιλογή μεταβλητών σε χώρο υψηλής διάστασης έχει αμφισβητήσει πολλά σύγχρονα στατιστικά προβλήματα, τα οποία προέρχονται από πολλούς επιστημονικούς κλάδους. Οι πρόσφατες τεχνολογικές εξελίξεις έχουν καταστήσει δυνατή τη συλλογή ενός τεράστιου ποσού από πληροφορίες συμμεταβλητών, όπως μικροσυστοιχιών, πρωτεομικής και SNP δεδομένων μέσω της τεχνολογίας της βιο-απεικόνισης, καθώς παρατηρούμε τις πληροφορίες επιβίωσης ασθενών σε κλινικές μελέτες. Επίσης, η ίδια πρόκληση εφαρμόζεται στην ανάλυση επιβίωσης προκειμένου να γίνει κατανοητή η σχέση μεταξύ γονιδιωματικών και κλινικών πληροφοριών σχετικά με το χρόνο επιβίωσης. Στην εργασία αυτή, επεκτείνουμε τη διαδικασία του σίγουρου κρησαρίσματος [Fan, J. and Lv, J. (2008)] στο μοντέλο αναλογικού κινδύνου του Cox με μια διαθέσιμη επαναληπτική έκδοση. Μελέτες αριθμητικής προσομοίωσης έχουν δείξει ενθαρρυντικά αποτελέσματα για την προτεινόμενη μέθοδο σε σύγκριση με άλλες τεχνικές, όπως η LASSO. Αυτό αποδεικνύει την χρησιμότητα και την ευελιξία του επαναληπτικού σίγουρου κρησαρίσματος.

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

Στο δεύτερο κεφάλαιο παρουσιάζονται μέθοδοι αντιμετώπισης για τα προβλήματα υψηλής διάστασης. Παρουσιάζεται το «σίγουρο» κρησάρισμα (*sure screening*) και προτείνεται μια μέθοδος πάνω σε αυτό, η οποία βασίζεται στη μάθηση συσχέτισης και καλείται Sure Independence Screening (SIS). Η μέθοδος αυτή χρησιμοποιείται για να μειώσουμε τη διάσταση από υψηλή σε μια μέτριας κλίμακας που είναι μικρότερη από το μέγεθος του δείγματος. Κατόπιν, αναλύεται η επαναληπτική μέθοδος SIS, η οποία καλείται (I)SIS. Τέλος παρουσιάζονται κάποιες προσομοιώσεις, από τις οποίες προκύπτει ότι οι μέθοδοι SIS και (I)SIS λειτουργούν αρκετά ικανοποιητικά στα προβλήματα υψηλής διάστασης.

Το τρίτο κεφάλαιο, πραγματεύεται την ανάλυση επιβίωσης. Συγκεκριμένα, καταγράφονται οι βασικές έννοιες και αναλύεται το μοντέλο αναλογικής διακινδύνευσης του Cox, οι επεκτάσεις του και οι έλεγχοι της υπόθεσης της αναλογικότητας των κινδύνων. Επίσης, παρουσιάζεται η διαδικασία που το μοντέλο αυτό συνδυάζεται με τις μεθόδους ποινικοποιημένης πιθανοφάνειας.

Στο τέταρτο κεφάλαιο, επεκτείνουμε τη διαδικασία του σίγουρου κρησαρίσματος στο μοντέλο αναλογικού κινδύνου του Cox με μια διαθέσιμη επαναληπτική έκδοση. Παρουσιάζονται μελέτες αριθμητικής προσομοίωσης, οι οποίες δείχνουν ενθαρρυντικά αποτελέσματα για την προτεινόμενη μέθοδο σε σύγκριση με άλλες τεχνικές, όπως η LASSO. Αυτό αποδεικνύει τη χρησιμότητα και την ευελιξία του επαναληπτικού σίγουρου κρησαρίσματος (I)SIS.

ABSTRACT

Variable selection in high dimensional space has challenged many contemporary statistical problems from many frontiers of scientific disciplines. Recent technological advances have made it possible to collect a huge amount of covariate information such as microarray, proteomic and SNP data via bioimaging technology while observing survival information on patients in clinical studies. Thus, the same challenge applies in survival analysis in order to understand the association between genomics information and clinical information about the survival time. In this work, it is mentioned the sure screening procedure to Cox's proportional hazards model with an iterative version available. Numerical simulation studies have shown encouraging performance of the proposed method in comparison with other techniques such as LASSO. This demonstrates the utility and versatility of the iterative sure independence screening scheme.

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

In the second chapter, it is proposed a sure screening method, SIS, and it is referred its rationale as well as its connection with other methods of dimensionality reduction. Then, several known techniques for model selection in the reduced feature space are reviewed and two simulations and one real data example are presented, in order to study the performance of SIS-based model selection methods. At the end of the chapter, some extensions of SIS are mentioned and, in particular, an iterative SIS is proposed and illustrated by three simulated examples.

The third chapter deals with survival analysis. Specifically, the basic concepts are cited and the Cox's proportional hazards model is analyzed, in addition with its extensions and tests of proportionality assumption. Furthermore, the procedure of the combination of this model with the penalized likelihood methods is presented.

Finally, in the fourth chapter, it is mentioned the sure screening procedure to Cox's proportional hazards model with an iterative version available. Afterwards, numerical simulation studies are presented, which have shown encouraging performance of the proposed method in comparison with other techniques such as LASSO. This demonstrates the utility and versatility of the iterative sure independence screening scheme.

ΕΥΧΑΡΙΣΤΙΕΣ

Η εκπόνηση και ολοκλήρωση της παρούσας μεταπτυχιακής διπλωματικής εργασίας, πραγματοποιήθηκε υπό την επίβλεψη στελεχών από τον χώρο του Εθνικού Μετσόβιου Πολυτεχνείου. Θα ήθελα συνεπώς σε αυτό το πλαίσιο να ευχαριστήσω θερμά τον Καθηγητή του Ε.Μ.Π., κ. Χρήστο Κουκουβίνο, για την επίβλεψη και καθοδήγηση του, όπως και για την δυνατότητα που μου προσέφερε να ασχοληθώ με ένα θέμα το οποίο ανήκει στα ερευνητικά μου ενδιαφέροντα.

Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Ανδρουλάκη Μάνο, για την πολύτιμη βοήθεια και το συνεχές ενδιαφέρον του καθόλη τη διάρκεια της εκπόνησης της εργασίας αυτής.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου κι ιδιαίτερα τους γονείς μου για τη διαχρονική συμπαράστασή τους.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1ο

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ

1.1 ΕΙΣΑΓΩΓΗ

Πειράματα εκτελούνται σε όλες τις επιστήμες με σκοπό να εξεταστούν και να ανακαλυφθούν νέες διαδικασίες ή καινούρια συστήματα, αλλά και στις βιομηχανίες με στόχο τη βελτιστοποίηση των προϊόντων που παράγουν. Όμως σε κάθε προϊόν επιδρούν διαφορετικοί παράγοντες, ο καθένας σε ξεχωριστό βαθμό. Για το λόγο αυτό κάθε παράγοντας πρέπει να μελετηθεί χωριστά. Είναι εμφανές ότι, θα ήταν ιδιαίτερα χρήσιμο να εντοπιστούν και να εξεταστούν οι παράγοντες που επιδρούν σημαντικά στο προϊόν, με λίγες εκτελέσεις του πειράματος. Στην πραγματικότητα, ο αριθμός των παραγόντων που επηρεάζουν ένα προϊόν είναι μικρός σε σύγκριση με τον αριθμό των παραγόντων που εξετάζονται.

Τα αποτελέσματα και τα συμπεράσματα κάθε πειράματος εξαρτώνται από τον τρόπο με τον οποίο συλλέγονται τα δεδομένα- σχεδιασμός του πειράματος- και από τη μέθοδο που χρησιμοποιείται για την ανάλυση αυτών των δεδομένων. Η διαδικασία σχεδιασμού και εκτέλεσης ενός πειράματος, έτσι ώστε να συλλεχθούν δεδομένα κατάλληλα για στατιστική ανάλυση, τα οποία να μπορούν να δώσουν έγκυρα και αντικειμενικά αποτελέσματα καλείται στατιστικός σχεδιασμός του πειράματος ή πειραματικός σχεδιασμός (*experimental design*).

Για να βελτιωθούν τα αποτελέσματα του πειράματος θα πρέπει να επιλεγεί ένα καλύτερο σύνολο από ανεξάρτητες μεταβλητές. Αυτό σημαίνει ότι απαιτείται η εφαρμογή συστηματικών πειραμάτων, θέτοντας τις ανεξάρτητες μεταβλητές και παρατηρώντας τα αποτελέσματα. Στην προσέγγιση αυτή χρειάζεται ένας πειραματικός σχεδιασμός, που ακολουθείται από στατιστική ανάλυση.

Σύμφωνα με τους Myers και Montgomery (1995), η πρώτη φάση ενός πειράματος είναι η φάση κρησαρίσματος. Το κρησάρισμα είναι η σημαντικότερη

φάση ενός πειραματικού σχεδιασμού, γιατί κάποιο λάθος στο πείραμα θα έχει ως συνέπεια τη λήψη λανθασμένων συμπερασμάτων.

Σε πολλές στατιστικές εφαρμογές συναντάται το πρόβλημα της μελέτης της σχέσης δύο ή περισσότερων τυχαίων μεταβλητών. Η σχέση αυτή μπορεί να προσδιοριστεί με βάση ορισμένες παρατηρήσεις. Σε κάθε σύστημα, στο οποίο οι μεταβλητές ποσότητες αλλάζουν ή μεταβάλλονται, έχει ενδιαφέρον να εξεταστούν οι επιδράσεις που κάποιες μεταβλητές ασκούν (ή φαίνεται ότι ασκούν) σε κάποιες άλλες μεταβλητές. Στις φυσικές διαδικασίες η ύπαρξη μιας τέτοιας απλής σχέσης αποτελεί μάλλον την εξαίρεση παρά τον κανόνα. Συνήθως υπάρχει μια συναρτησιακή σχέση η οποία είναι τόσο πολύπλοκη ώστε δεν μπορεί να γίνει κατανοητή ή να περιγραφεί με απλούς όρους. Σε μια τέτοια περίπτωση, το καλύτερο είναι να γίνει μια προσέγγιση αυτής της πολύπλοκης συναρτησιακής σχέσης με μια απλή μαθηματική συνάρτηση, όπως πολυώνυμο, η οποία να περιλαμβάνει τις κατάλληλες μεταβλητές και να προσεγγίζει την κανονική συνάρτηση για κάποιο περιορισμένο εύρος μεταβλητών που εμπλέκονται στη διαδικασία. Η εξέταση αυτής της απλής μαθηματικής συνάρτησης μπορεί να δώσει περισσότερες πληροφορίες για την υπάρχουσα αληθινή σχέση και να οδηγήσει σε εκτίμηση των ξεχωριστών μεταβλητών αλλά και των επιδράσεων που παράγονται από τις αλλαγές σε συγκεκριμένες σημαντικές μεταβλητές.

Ακόμα και όταν λογικά δεν υπάρχει φυσική σχέση μεταξύ των μεταβλητών, μπορεί να θέλουμε να τις συσχετίσουμε με κάποια μαθηματική εξίσωση. Ενώ η εξίσωση μπορεί να στερείται φυσικής έννοιας, μπορεί να είναι εξαιρετικά πολύτιμη για την πρόβλεψη των τιμών κάποιων μεταβλητών από τις γνώσεις που διαθέτουμε για άλλες μεταβλητές, όταν ισχύουν κάποιες συγκεκριμένες συνθήκες.

Προσδιορίζοντας τη σχέση των μεταβλητών, λέμε ότι έχουμε προσδιορίσει ένα μοντέλο. Η εκτίμηση ενός στατιστικού μοντέλου γίνεται με ανάλυση στατιστικών δεδομένων, δηλαδή παρατηρήσεων της εξαρτημένης μεταβλητής Y σε επιλεγμένα επίπεδά της. Οι στατιστικές τεχνικές που χρησιμοποιούνται για το σκοπό αυτό αναφέρονται ως ανάλυση παλινδρόμησης.

Έστω $\underset{\sim}{Y}$ μια μεταβλητή που μας ενδιαφέρει και $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_p$ ένα σύνολο επεξηγηματικών μεταβλητών ή παραγόντων (*explanatory variables or factors*), τα οποία αποτελούν διανύσματα *n* παρατηρήσεων. Η επιλογή μεταβλητών (*Variable Selection*) χρησιμοποιείται για την πρόβλεψη του «καλύτερου» υποσυνόλου ανεξάρτητων μεταβλητών $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_p$. Σκοπός δηλαδή, είναι να επιλεγούν οι παράγοντες που έχουν σημαντική επίδραση στην απόκριση $\underset{\sim}{Y}$. Με αυτόν τον τρόπο εξηγούμε τα δεδομένα με τον απλούστερο τρόπο, διότι οι περιττές προβλέψεις αφαιρούνται. Η αρχή της μετριοπάθειας (*Law of parsimony*) ορίζει ότι μεταξύ των διαφόρων πιθανών εξηγήσεων για μια περίπτωση, η απλούστερη είναι και η καλύτερη. Η εφαρμογή της στην ανάλυση παλινδρόμησης αποδεικνύει ότι το μικρότερο μοντέλο που ταιριάζει με τα δεδομένα είναι και το καλύτερο. Επιπροσθέτως, οι περιττοί προγνωστικοί παράγοντες θα προσθέσουν περισσότερα εμπόδια στον προσδιορισμό άλλων ποσοτήτων, οι οποίες μας ενδιαφέρουν. Αυτό θα έχει ως αποτέλεσμα να αυξηθούν οι βαθμοί ελευθερίας.

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

1.2. Το γενικό γραμμικό μοντέλο παλινδρόμησης

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

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i,\ i = 1, 2, ..., n \qquad (1.1),$$

όπου

y_i είναι οι τιμές της απόκρισης,

x_{ij} είναι οι τιμές των επεξηγηματικών μεταβλητών (υποθέτουμε, όπως και στο απλό γραμμικό μοντέλο, ότι οι μετρήσεις μας δεν υπόκεινται σε σφάλματα),

β_j είναι οι άγνωστες παράμετροι του μοντέλου οι οποίες και πρέπει να εκτιμηθούν,

ε_i είναι τα σφάλματα ή υπόλοιπα, τα οποία αποτελούν τυχαίες μεταβλητές και υποθέτουμε ότι ικανοποιούν τις παρακάτω σχέσεις:

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

Από τη σχέση (1.1) φαίνεται ότι η απόκριση y_i (*response*) είναι μια γραμμική συνάρτηση των συντελεστών παλινδρόμησης (*regression coefficients*) β_j, με $j = 1, 2, ..., k$.

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

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

1.2.1 Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο ελαχίστων τετραγώνων

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

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

οπότε τελικά προκύπτουν οι εκτιμητές $\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k$. Είναι προτιμότερο να γράψουμε των εξίσωση (1.1) υπό την μορφή πινάκων, δηλαδή

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

όπου

$$\underset{\sim}{Y}=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\qquad \underset{\sim}{X}=\begin{pmatrix}1 & x_{11}\cdots & x_{1k}\\1 & x_{21}\cdots & x_{2k}\\\vdots & \vdots & \vdots\\1 & x_{n1}\cdots & x_{nk}\end{pmatrix}$$

$$\underset{\sim}{\beta}=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{bmatrix},\qquad \underset{\sim}{\varepsilon}=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}.$$

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

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

Έτσι προκύπτει ότι

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

Αναφέρουμε και κάποιες ιδιότητες της εκτιμήτριας ελαχίστων τετραγώνων.

Καταρχήν,

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

καθότι

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

και

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

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

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

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

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

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

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

$$f(x_1, x_2, ..., x_n \mid \underset{\sim}{\theta}) = f(x_1 \mid \underset{\sim}{\theta}) f(x_2 \mid \underset{\sim}{\theta}) \cdots f(x_n \mid \underset{\sim}{\theta}) \qquad (1.2).$$

Στην περίπτωση συγκεκριμένων παρατηρήσεων $x_1, x_2, ..., x_n$ τυχαίου δείγματος, η (1.2) είναι συνάρτηση μόνο της παραμέτρου $\underset{\sim}{\theta}$ και συμβολίζεται ως εξής:

$$L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n) = f(x_1 \mid \underset{\sim}{\theta}) f(x_2 \mid \underset{\sim}{\theta}) \cdots f(x_n \mid \underset{\sim}{\theta}) = \prod_{i=1}^{n} f(x_i \mid \underset{\sim}{\theta}) \qquad (1.3).$$

Η (1.3) καλείται συνάρτηση πιθανοφάνειας (*likelihoodfunction*) του τυχαίου δείγματος $X_1, X_2, ..., X_n$ και εκφράζει το πόσο «πιθανοφανείς», δηλαδή πόσο σύμφωνες με το συγκεκριμένο δείγμα είναι οι διάφορες τιμές της παραμέτρου $\underset{\sim}{\theta}$.

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

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

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

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

Φυσικά, για να είναι η λύση αυτή σημείο μεγίστου, θα πρέπει ο Εσσιανός πίνακας

$$\left[\frac{\partial^2 \log L(\underset{\sim}{\theta})}{\partial \theta_i \partial \theta_j} \right]_{k \times k}$$

να είναι γνήσια αρνητικός για $\underset{\sim}{\theta} = \hat{\underset{\sim}{\theta}}$.

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

1.3. Μέθοδος επιλογής καλύτερου υποσυνόλου

Υπάρχουν περιπτώσεις που καθιστούν αναγκαία την επιλογή ενός μικρού υποσυνόλου από ένα μεγαλύτερο σύνολο μεταβλητών οι οποίες χρησιμοποιούνται για την πρόβλεψη μια εξαρτημένης μεταβλητής. Για παράδειγμα, η διαδικασία πρόβλεψης της εξαρτημένης μεταβλητής Y μπορεί να θεωρείται οικονομικά ασύμφορη και αρκετά χρονοβόρα εφόσον χρησιμοποιηθούν όλες οι τιμές που έχουμε για τις ανεξάρτητες μεταβλητές, για το λόγο αυτό αναζητούμε ένα μικρό υποσύνολο μεταβλητών, το οποίο θα μπορεί με αρκετή ακρίβεια να προβλέψει την τιμή της μεταβλητής Y. Η επιλογή αυτού του συνόλου θα μπορούσε προφανώς στη συνέχεια να χρησιμοποιηθεί για κάποια μελλοντική πρόβλεψη εφόσον τα δεδομένα είναι αντιπροσωπευτικά των συνθηκών υπό τις οποίες θα γίνει η πρόβλεψη. Από την άλλη πλευρά, στην προσπάθειά μας να κατανοήσουμε την επίδραση που έχει η μια μεταβλητή πάνω στην άλλη, ιδιαίτερα όταν τα δεδομένα που έχουμε συλλέξει είναι μέσω παρατήρησης ή μέσω έρευνας και όχι δεδομένα που προήλθαν μέσω πειραμάτων, μπορεί να είναι επιθυμητό να μην χρησιμοποιήσουμε πολλές μεταβλητές οι οποίες έχουν κάποια ισχυρή επίδραση μεταξύ τους.

Κάποιες φορές τα δεδομένα για την πρόβλεψη που θέλουμε να κάνουμε έχουν ήδη συλλεχθεί για προηγούμενες προβλέψεις ή για άλλους σκοπούς με αποτέλεσμα να μην υπάρχει επιπλέον κόστος εάν συμπεριληφθούν στο μοντέλο πρόβλεψης. Σε άλλες περιπτώσεις μπορεί να υπάρχει σημαντικό επιπλέον κόστος εάν χρησιμοποιηθούν όλα τα δεδομένα.

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

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

- Να περιγράψουμε μια σειρά πολυμεταβλητών δεδομένων με μετριότητα (*parsinomy*).
- Να εκτιμήσουμε τους συντελεστές παλινδρόμησης με μικρότερα σφάλματα (ειδικά όταν μερικοί από τους προγνωστικούς παράγοντες σχετίζονται πολύ μεταξύ τους) , (Burak Eksioglu, Riza Demirer, Ismail Capar, 2005).

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

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

Τα κριτήρια πρόβλεψης βασίζονται στα σφάλματα πρόβλεψης. Τα πιο γνωστά είναι το κριτήριο *FPE* (*Final Prediction Error*) (Akaike, 1969) και το κριτήριο $C_p - Mallows$ (Mallows, 1973). Τα κριτήρια που προκύπτουν από τα δεδομένα ή αλλιώς *bootstrap* μέθοδοι, εισήχθησαν από τον Efron και παρέχουν εύκολους και αποτελεσματικούς τρόπους για να εκτιμήσουμε την αναμενόμενη συνολική διαφορά μεταξύ του σωστού και του υποψήφιου μοντέλου. Ο Breiman (1992) πρότεινε ένα τέτοιο κριτήριο και από τις προσομοιώσεις που πραγματοποίησε, έδειξε ότι είναι καλύτερο από το C_p. Όσον αφορά τα κριτήρια καθοδηγούμενα από τα δεδομένα ή μέθοδοι διασταυρωμένης επικύρωσης (*Cross-validation methods*), αυτά βασίζονται στο διαχωρισμό των δεδομένων σε δύο υποσύνολα, όπου το ένα χρησιμοποιείται για την επιλογή του μοντέλου και το άλλο για τον καθορισμό της προβλεπτικής ικανότητας του μοντέλου. Μια τέτοια μέθοδος, είναι η γενικευμένη διασταυρωμένη επικύρωση (*Generalized Cross validation* $-GCV$), η οποία επιλέγει

το μοντέλο με την καλύτερη μέση προβλεπτική ικανότητα, χρησιμοποιώντας διάφορους τρόπους διαχωρισμού του αρχικού συνόλου δεδομένων. Τέλος, οι Μπεϋζιανές μέθοδοι, χρησιμοποιούν μια ιεραρχική εκ των προτέρων κατανομή (*hierarchical prior*), ώστε να αναθέσουν μεγαλύτερες εκ των υστέρων πιθανότητες σε περισσότερο υποσχόμενα μοντέλα.

1.3.1. Κριτήρια βασισμένα στην πιθανοφάνεια ή στην πληροφορία

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

Το κριτήριο *AIC* προτάθηκε από τον Hirotsugu Akaike, με το όνομα «το κριτήριο πληροφοριών» (*information criterion*). Το κριτήριο αυτό στηρίζεται στην ιδέα της εντροπίας των πληροφοριών. Στην πραγματικότητα προσφέρεται ως ένα σχετικό μέτρο για τα στοιχεία που χάνονται σε ένα πραγματικό μοντέλο. Θα μπορούσαμε να πούμε ότι το κριτήριο αυτό περιγράφει τη σχέση μεταξύ της ακρίβειας και της πολυπλοκότητας του μοντέλου. Όμως το κριτήριο αυτό δεν μπορεί να περιγράψει εάν το μοντέλο που μελετάμε ταιριάζει ικανοποιητικά με τα δεδομένα μας.

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

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

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

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

όπου n το μέγεθος του δείγματος και RSS το άθροισμα τετραγώνων των σφαλμάτων. Επίσης, στην περίπτωση των γενικευμένων γραμμικών μοντέλων, παίρνει τη μορφή

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

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

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

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

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

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

Με βάση την εξίσωση του Bayes, ο Schwartz (1978) ανέπτυξε ένα άλλο κριτήριο, το οποίο το ονόμασε κριτήριο πληροφορίας του Bayes επειδή είναι κριτήριο πληροφοριών με βάση τη Bayesian Μέθοδο. Δημιουργήθηκε από τον Gideon E. Schwarz και ουσιαστικά είναι ένα κριτήριο για την επιλογή μοντέλων

μεταξύ μιας ομάδας παραμετρικών μοντέλων με διαφορετικούς αριθμούς παραμέτρων. Διαλέγοντας ένα μοντέλο για τη βελτιστοποίηση του *BIC* είναι μια μορφή κανονικοποίησης.

Το κριτήριο *BIC* δίνεται από τον τύπο

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

Το μοντέλο που επιλέγεται είναι αυτό που δίνει τη μεγαλύτερη τιμή του κριτηρίου. Να σημειώσουμε επίσης, ότι υπάρχει μια κλίμακα (*Raftery's scale*), που ανάλογα με την απόλυτη διαφορά στις τιμές του κριτηρίου *BIC* μεταξύ δύο μοντέλων, κρίνεται κατά πόσο το ένα μοντέλο είναι καλύτερο από το άλλο. Η κλίμακα αυτή αναγράφεται στον παρακάτω Πίνακα 1.Α.

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

Πίνακας 1.Α: Κλίμακα Raftery

1.3.2. Στατιστική συνάρτηση C_p – Mallows

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

$$\text{ΜΤΣ}(\hat{\theta}) = \text{E}(\hat{\theta} - \theta)^2$$

$$= \text{E}[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2$$

$$= \mathrm{E}[\hat{\theta} - E(\hat{\theta})]^2 + 2E[\hat{\theta} - E(\hat{\theta})]\mathrm{E}[E(\hat{\theta}) - \theta] + E[E(\hat{\theta}) - \theta]^2$$

$$= V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

$$= V(\hat{\theta}) + bias^2$$

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

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

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

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

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

όπου $SB(\mathrm{p}) = \sum bias^2$.

Ο δεύτερος όρος απλοποιείται ως εξής:

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

$$= tr(\mathrm{H}(\sigma^2\,\mathrm{I})\,\mathrm{H}') = \sigma^2 tr(\mathrm{H}^2) = \sigma^2 tr(\mathrm{H}) = \sigma^2 p$$

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

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

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

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

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

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

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

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

αφού

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

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

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

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

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

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

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

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

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

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

Η ποσότητα Γ_p εκτιμάται από την

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

Δηλαδή το $\mathrm{E}(SSE(\mathrm{p}))$ έχει αντικατασταθεί από το παρατηρούμενο άθροισμα τετραγώνων των υπολοίπων του υπό εξέταση μοντέλου και το σ^2 από μια καλή εκτίμηση του $\hat{\sigma}^2$. Συνήθως χρησιμοποιείται το μοντέλο που περιέχει όλες τις υποψήφιες μεταβλητές x_j (και όχι το τρέχον μοντέλο με τους p όρους), δηλαδή $\hat{\sigma}^2 = \frac{SSE(\mathrm{p}')}{n - p'}$, όπου p' = το πλήθος όλων των διαθέσιμων επεξηγηματικών μεταβλητών +1.

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

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

Επομένως, ως βέλτιστο μοντέλο θεωρείται εκείνο για το οποίο ισχύει

$$C_p \simeq p$$

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

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

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

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

Το 1979, οι Hannan και Quinn ασχολήθηκαν με την επιλογή μοντέλων χρονοσειρών, για την περίπτωση όπου ο αριθμός των παραμέτρων στο μοντέλο αυξάνει με το μέγεθος του δείγματος. Οπότε πρότειναν το κριτήριο Hannan και Quinn

$$HQC(j) = n\ln\left(\frac{RSS}{n}\right) + 2p_j c\ln(\ln n),$$

το οποίο και πρέπει να μεγιστοποιηθεί για κάποιο $c > 1$ (στις προσομοιώσεις τους χρησιμοποίησαν την τιμή $c = 1$).

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

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

όπου Y είναι τα δεδομένα, f είναι η πυκνότητα πιθανότητας (της κανονικής), $\underset{\sim}{\hat{\beta}}$ είναι το διάνυσμα των συντελεστών παλινδρόμησης και g είναι το μήκος του μηνύματος που απαιτείται για την κωδικοποίηση των βέλτιστων τιμών των συντελεστών. Συνεπώς, ο τελευταίος όρος θεωρείται μια ποινή στην αύξηση των μεταβλητών. Το κριτήριο Bryant και Cordero-Brana συνίσταται στην ελαχιστοποίηση της ποσότητας

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

όπου

$$G(n) = \frac{n}{2}\left(\frac{n-1}{n}\ln n + \ln\pi\right),$$

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

1.4 Ποινικοποιημένα ελάχιστα τετράγωνα και ποινικοποιημένη πιθανοφάνεια

1.4.1 Εισαγωγή

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

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

- Να είναι ιδιάζουσες (*singular*) στην αρχή ώστε να παράγουν σποραδικές λύσεις (πολλοί εκ των εκτιμηθέντων συντελεστών να έχουν τιμή μηδέν).
- Να ικανοποιούν συγκεκριμένες απαιτήσεις ώστε να παράγουν συνεχή μοντέλα (*continuous models*), οπότε η επιλογή του μοντέλου να χαρακτηρίζεται από σταθερότητα (*stability*).
- Να φράσσονται από μια σταθερά, ώστε να παράγουν σχεδόν αμερόληπτους εκτιμητές για μεγάλους συντελεστές.

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

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

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

1.4.2. Επιλογή μεταβλητών μέσω ποινικοποιημένων ελαχίστων τετραγώνων

Θεωρούμε το γνωστό γραμμικό μοντέλο

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

όπου $\underset{\sim}{Y}$ είναι ένα $n\times 1$ διάνυσμα των παρατηρήσεων, ο $\underset{\sim}{X}$ είναι ένας $n\times d$ πίνακας των επεξηγηματικών μεταβλητών, $\underset{\sim}{\beta}$ είναι ένα $d\times 1$ διάνυσμα των συντελεστών παλινδρόμησης και το $\underset{\sim}{\varepsilon}$ είναι ένα $n\times 1$ διάνυσμα των τυχαίων σφαλμάτων. Όπως και στην περίπτωση του μοντέλου γραμμικής παλινδρόμησης, υποθέτουμε ότι τα y_i είναι υπό συνθήκη ανεξάρτητα, δοθέντων των x_{ij}. Επίσης, υποθέτουμε και ότι οι

στήλες του πίνακα $\underset{\sim}{X}$ είναι ορθοκανονικές (*orthonormal*). Ο υπολογισμός της εκτιμήτριας γίνεται μέσω της ελαχιστοποίησης της ποσότητας

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

η οποία ισοδυναμεί με την ποσότητα

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

όπου

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

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

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

και έστω ότι

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

μια μορφή των ποινικοποιημένων ελαχίστων τετραγώνων είναι η εξής:

$$\frac{1}{2} \| \underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta} \|^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|) = \frac{1}{2} \| \underset{\sim}{Y} - \hat{\underset{\sim}{Y}} \|^2 + \frac{1}{2} \sum_{j=1}^{d} (z_j - \beta_j)^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|) \qquad (1.4).$$

Να σημειωθεί ότι οι συναρτήσεις ποινής p_j στην (1.4) δεν είναι απαραίτητα οι ίδιες για όλα τα j. Για παράδειγμα μπορεί να θέλουμε να κρατήσουμε ορισμένες σημαντικές μεταβλητές σε ένα παραμετρικό μοντέλο και για αυτό το λόγο να μη θέλουμε να ποινικοποιήσουμε τις αντίστοιχες παραμέτρους τους. Για ευκολία όμως, θεωρούμε ότι οι συναρτήσεις ποινής είναι οι ίδιες για όλους τους συντελεστές, και θα συμβολίζονται ως $p(|\cdot|)$. Επίσης, αντί $\lambda p(|\cdot|)$ θα χρησιμοποιούμε το συμβολισμό $p_\lambda(|\cdot|)$, δείχνοντας έτσι ότι το $p(|\cdot|)$ εξαρτάται από το λ.

Το πρόβλημα ελαχιστοποίησης της (1.4) είναι ισοδύναμο με την ελαχιστοποίηση των συνιστωσών. Οπότε θεωρούμε το παρακάτω πρόβλημα ελαχίστων τετραγώνων

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \text{ (1.5).}$$

Εν συνεχεία, χρησιμοποιώντας τη *Hard* συνάρτηση ποινής (βλ. σχήμα 1.1)

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

προκύπτει η *Hard* εκτιμήτρια (βλ. σχήμα 1.2α)

$$\hat{\theta} = zI(|z| > \lambda) \qquad (1.6).$$

Με άλλα λόγια, η λύση της είναι

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

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

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

η οποία και αυτή οδηγεί στη λύση (1.6).

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

- Αμεροληψία: Ο προκύπτων εκτιμητής πρέπει να είναι σχεδόν αμερόληπτος, ιδίως στην περίπτωση όπου η σωστή άγνωστη παράμετρος β_j είναι μεγάλη. Αποφεύγεται έτσι η μεροληψία του μοντέλου.
- Σποραδικότητα: Ο προκύπτων εκτιμητής πρέπει να αποτελεί κανόνα περιορισμού (*thresholdin rule*), ώστε οι εκτιμηθέντες συντελεστές με μικρή τιμή, να μηδενίζονται. Έτσι, μειώνεται η πολυπλοκότητα του μοντέλου.
- Συνέχεια: Ο προκύπτων εκτιμητής πρέπει είναι συνεχής. Αποφεύγεται κατά αυτόν τον τρόπο η αστάθεια στη πρόβλεψη του μοντέλου.

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

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

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

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

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

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

Το παρακάτω γράφημα (Σχήμα 1.3) παρέχει περισσότερες εξηγήσεις σχετικά με αυτό.

Όταν τώρα

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

η παράγωγος της $\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)$ είναι θετική για όλα τα θετικά θ και αρνητική για όλα τα αρνητικά θ. Οπότε σε αυτήν την περίπτωση, ο *PLS* εκτιμητής $\hat{\theta}$ είναι μηδέν. Όταν όμως $|z| > \min_{\theta \neq 0} \{|\theta| + p'_\lambda(|\theta|)\}$, δύο διασταυρώσεις (*crossings*) μπορούν να υπάρξουν, όπως φαίνεται και στο Σχήμα 1.1. Η μεγαλύτερη είναι ο *PLS* εκτιμητής. Αυτό συνεπάγεται ότι ικανή και αναγκαία συνθήκη για την ύπαρξη

συνέχειας είναι το $\min_{\theta}\{|\theta|+p'_{\lambda}(|\theta|)\}$ να πετυχαίνεται στο μηδέν. Από αυτό αντιλαμβανόμαστε πως η συνάρτηση ποινής που ικανοποιεί τις ιδιότητες της σποραδικότητας και της συνέχειας, πρέπει να είναι ιδιάζουσα (*singular*) στην αρχή.

Είναι γνωστό πως η συνάρτηση ποινής L_2

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

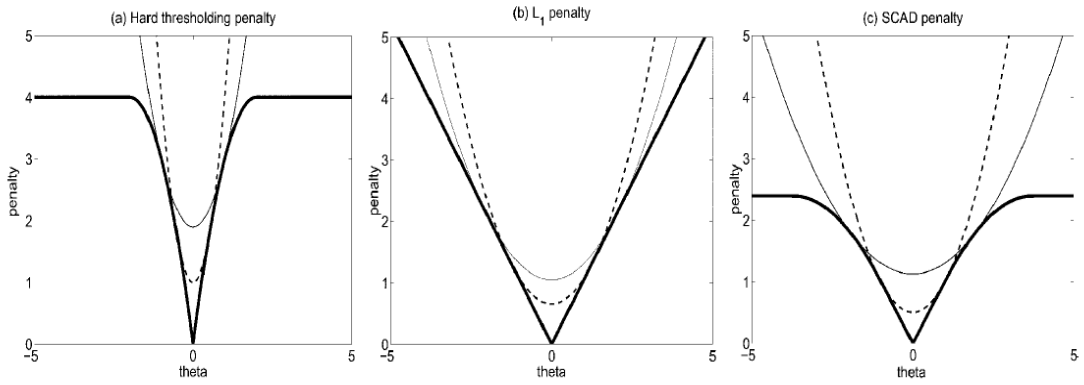
οδηγεί στην παλινδρόμηση κορυφογραμμής. Η συνάρτηση ποινής L_1, οδηγεί στον *soft* οριακό κανόνα

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

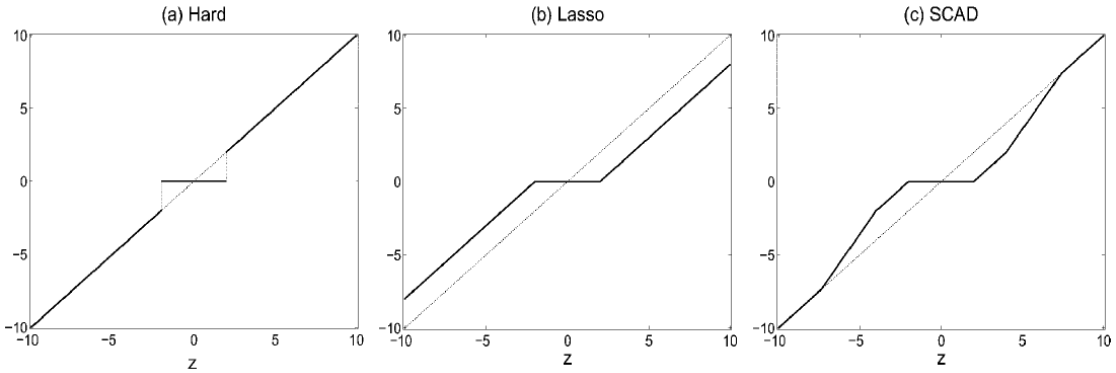
που προτάθηκε από τους Donoho και Johnstone (1994α). Η *LASSO* που προτείνεται από τον Tibshirani (1996, 1997) είναι ο *PLS* εκτιμητής με συνάρτηση ποινής την L_1. Επίσης, η L_q συνάρτηση ποινής

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

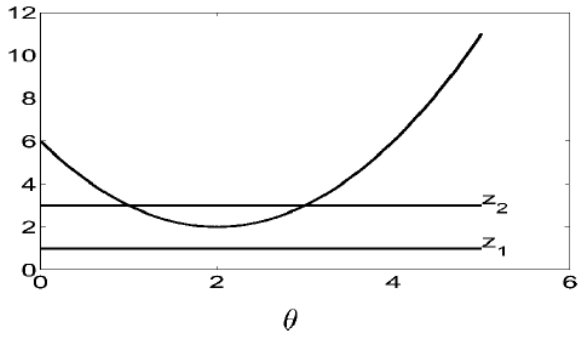
οδηγεί στην παλινδρόμηση *bridge*. Η λύση είναι συνεχής μόνο για $q\geq 1$. Παρόλα αυτά, όταν $q>1$, δεν παράγεται μια σποραδική λύση. Η μόνη συνεχής λύση με κανόνα περιορισμού σε αυτή την οικογένεια συναρτήσεων είναι με τη συνάρτηση ποινής L_1, αυτό όμως προκύπτει μεταβάλλοντας τον εκτιμητή κατά μια σταθερά λ, άρα χάνεται και η αμεροληψία (Σχήμα 1.2(b)). Επίσης για $0\leq q<1$, δεν ικανοποιείται η συνθήκη της συνέχειας.



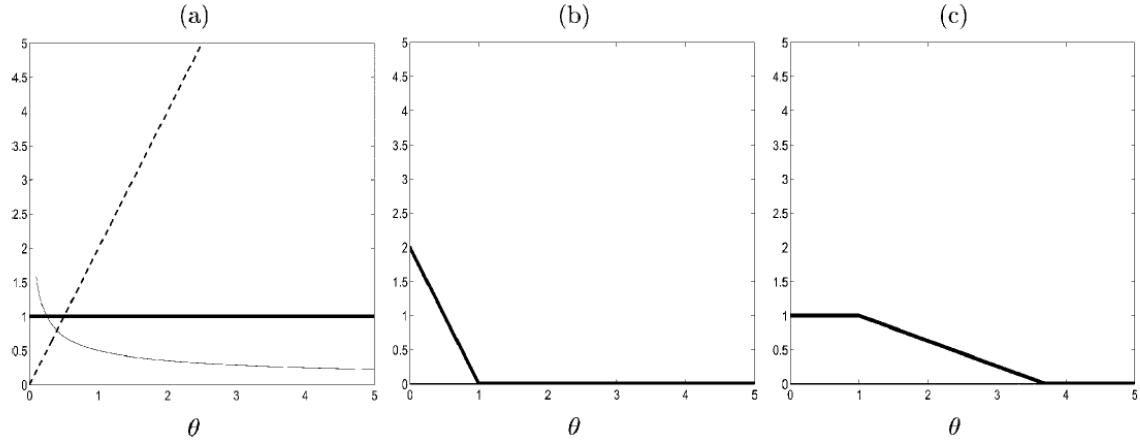
***Σχήμα 1.1**: (a) Οι τρεις συναρτήσεις ποινής και οι τετραγωνικές τους προσεγγίσεις.*



***Σχήμα 1.2**: Οι εκτιμήτριες (thresholding functions) (a) Hard, (b) Soft ή LASSO και (c) SCAD, όπου για την τελευταία λ=2 και a=3.7.*



Σχήμα 1. 3. Η συνάρτηση $\theta + p_\lambda(|\theta|)$ ***ως προς*** θ **.**



Σχήμα 1.4*: Οι συναρτήσεις* $p'_\lambda(|\theta|)$ *ως προς* θ*, για (a) τις συναρτήσεις ποινής* L_q*, (b) τη Hard συνάρτηση ποινής και (c) τη SCAD. Στο (a), η παχιά γραμμή αντιστοιχεί στην* L_1*, η διακεκομμένη στην* $L_{0.5}$ *και η λεπτή γραμμή στην* L_2 *συνάρτηση ποινής.*

1.4.2.1. Η συνάρτηση ποινής SCAD

Οι συναρτήσεις ποινής L_q και Hard δεν ικανοποιούν και τις τρεις απαιτήσεις της αμεροληψίας, της σποραδικότητας και της συνέχειας. Με σκοπό τη βελτίωση της L_1 και της *Hard*, οι Fan και Li (2001) εισήγαγαν μια συνεχής και διαφορίσιμη συνάρτηση ποινής, τη *SCAD* (*Smoothly Clipped Absolute Deviation penalty*) (Σχήμα1.1(c)), η οποία ορίζεται ως

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

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

$$\hat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & |z| \le 2\lambda \\ \{(\alpha-1)z - \text{sgn}(z)\alpha\lambda\}/(\alpha-2), & 2\lambda < |z| \le \alpha\lambda \\ z, & |z| > \alpha\lambda \end{cases} \qquad (1.7)$$

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

Η λύση (1.7) έχει δύο άγνωστες παραμέτρους, α και λ. Στην πράξη θα μπορούσαμε να υπολογίσουμε το βέλτιστο ζεύγος (α, λ) βάσει κάποιων κριτηρίων, όπως της διασταυρωμένης επικύρωσης και της γενικευμένης διασταυρωμένης επικύρωσης. Κάτι που μπορεί να είναι υπολογιστικά χρονοβόρο. Οι Fan και Li (2001), χρησιμοποιώντας εργαλεία Μπεϋζιανής ανάλυσης ρίσκου (*Bayesian risk analysis*), κατέληξαν στην επιλογή του $\alpha = 3.7$.

1.4.2.2. Απόδοση των οριακών κανόνων

Οι Marron et al. (1998), εφαρμόζοντας ανάλυση ρίσκου, προσπάθησαν να κατανοήσουν τη συμπεριφορά των *Hard* και *Soft* οριακών κανόνων, στην περίπτωση

μικρών δειγμάτων. Οι Fan και Li, χρησιμοποίησαν την τιμή $\lambda = 2$ για τον *Hard* οριακό κανόνα. Προσάρμοσαν το λ για τους άλλους δύο οριακούς κανόνες, ώστε να δίνουν τις ίδιες εκτιμήσεις για την περίπτωση όπου $\theta = 3$ και κατέληξαν στο ότι η *SCAD* συμπεριφέρεται εξίσου καλά σε σύγκριση με τους άλλους δύο και διατηρεί τις μαθηματικές ιδιότητές τους. Κάτι που φαίνεται και από το Σχήμα 1.1.

1.4.3. Επιλογή μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας

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

Στο κλασικό μοντέλο παλινδρόμησης οι εκτιμητές ελαχίστων τετραγώνων παράγονται με την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων. Οπότε η (1.4) μπορεί να επεκταθεί για την περίπτωση όπου ο πίνακας σχεδιασμού δεν είναι ορθοκανονικός (*orthonormal*). Μια ισοδύναμη μορφή της (1.4) είναι

$$\frac{1}{2}(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}) + n\sum_{j=1}^{d} p_{\lambda}(|\beta_j|). \qquad (1.8)$$

Ελαχιστοποιώντας την (1.8) ως προς $\underset{\sim}{\beta}$, οδηγούμαστε σε έναν εκτιμητή ποινικοποιημένων ελαχίστων τετραγώνων του $\underset{\sim}{\beta}$.

Είναι γνωστό τώρα ότι ο *OLS* εκτιμητής δεν είναι εύρωστος. Μπορούμε όμως να θεωρήσουμε τη συνάρτηση ψ του Huber (1981), οπότε αντί της ελαχιστοποίησης της (1.8), μπορούμε να ελαχιστοποιήσουμε την

$$\sum_{i=1}^{n}\psi(|\, y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta}\,|) + n\sum_{j=1}^{d} p_\lambda(|\,\beta_j\,|)\,, \qquad (1.9)$$

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

Με τη βοήθεια τώρα του ποινικοποιημένου εκτιμητή μέγιστης πιθανοφάνειας, μπορούμε να επιλέξουμε σημαντικές μεταβλητές. Έχουμε τα εξής: Καταρχήν, έστω ότι τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ έχουν συλλεχθεί ανεξάρτητα. Δεδομένων των $\underset{\sim}{x}_i$, η Y_i έχει συνάρτηση πιθανοφάνειας

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

όπου g είναι μια γνωστή συνάρτηση σύνδεσης. Έστω και ότι

$$l_i = \log f_i$$

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

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

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

$$-\sum_{i=1}^{n} l_i(g(\underset{\sim}{x}_i'\underset{\sim}{\beta}), y_i) + n\sum_{j=1}^{d} p_\lambda(|\,\beta_j\,|) \qquad (1.10)$$

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

1.4.3.1. Δειγματοληπτικές και προβλεπτικές ιδιότητες

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

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

Χωρίς βλάβη της γενικότητας, θεωρούμε ότι

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

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

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

όπου

$$\alpha_n = \max\left\{ p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0 \right\} \qquad (1.11).$$

Αυτό σημαίνει ότι για τις *Hard* και *SCAD* συναρτήσεις ποινής, ο εκτιμητής ποινικοποιημένης πιθανοφάνειας είναι $\sqrt{n}$ -συνεπής (*root-consistent*) αν $\lambda_n \to 0$.

Επιπλέον θα δείξουμε ότι για τον εκτιμητή αυτόν πρέπει να ισχύει ότι

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

και ότι το $\hat{\underset{\sim}{\beta}}_1$ είναι ασυμπτωτικά της κανονικής κατανομής με πίνακα συνδιασποράς I_1^{-1}, αν $n^{1/2}\lambda_n \to \infty$.

Αυτό συνεπάγεται ότι ο εκτιμητής ποινικοποιημένης πιθανοφάνειας συμπεριφέρεται τόσο καλά όσο αν ήταν γνωστό ότι $\underset{\sim}{\beta}_{20} = \underset{\sim}{0}$.

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

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

όπου

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

Ένας υπερ-αποδοτικός εκτιμητής για το μ είναι

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

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

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

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

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

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

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

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

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

και

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

(B) Ο πίνακας πληροφορίας του Fisher

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

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

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

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

Επίσης, υπάρχουν συναρτήσεις M_{jkl} τέτοιες ώστε

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

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

Θεώρημα 1

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

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

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

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

με το α_n να δίνεται από την (1.11). Από το θεώρημα αυτό είναι προφανές ότι με μια σωστή επιλογή του λ_n θα υπάρξει ένας $\sqrt{n}$-συνεπής ποινικοποιημένος εκτιμητής. Θα δείξουμε τώρα ότι ο εκτιμητής αυτός έχει την ιδιότητα της σποραδικότητας $\hat{\underset{\sim}{\beta}}_2=\underset{\sim}{0}$.

Λήμμα 1

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

$$\liminf_{n\to\infty}\ \liminf_{\theta\to 0_+}\ p'_{\lambda_n}(\theta)/\lambda_n>0. \qquad (1.12)$$

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

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

και για κάθε σταθερά C, ισχύει ότι

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

Ορίζουμε τώρα ως

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

και

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

Θεώρημα 2 (Προβλεπτική ιδιότητα)

Θεωρούμε ξανά ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ είναι i.i.d, κάθε ένα με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V},\underset{\sim}{\beta})$ και ότι ικανοποιούν τις υποθέσεις (A)-(C). Έστω επίσης ότι η συνάρτηση ποινής $p_{\lambda_n}(|\theta|)$ ικανοποιεί τη συνθήκη (1.12). Αν $\lambda_n \to 0$ και $\sqrt{n}\lambda_n \to \infty$ όσο το $n \to \infty$, τότε με πιθανότητα που τείνει στο 1, οι $\sqrt{n}$-συνεπείς εκτιμητές $\hat{\underset{\sim}{\beta}} = \begin{pmatrix}\hat{\underset{\sim}{\beta}}_1\\ \hat{\underset{\sim}{\beta}}_2\end{pmatrix}$, του Θεωρήματος 1, πρέπει να ικανοποιούν τα παρακάτω:

- Σποραδικότητα (*sparsity*):

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

- Ασυμπτωτική κανονικότητα (*assymptotic normality*):

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

όπου

$$I_1(\underset{\sim}{\beta}_{10}) = I_1(\underset{\sim}{\beta}_{10}, \underset{\sim}{0})$$

η πληροφορία κατά Fisher, γνωρίζοντας ότι $\underset{\sim}{\beta}_2 = \underset{\sim}{0}$.

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

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

και για τις συναρτήσεις ποινής που αναπτύχθηκαν στην ενότητα 1.4.2, είναι προσεγγιστικά ίσος με

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

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

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

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

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

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

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

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

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

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

τότε ισχύει και η συνθήκη (C). Για γενικότερες συναρτήσεις σύνδεσης, παρόμοιες υποθέσεις πρέπει να ικανοποιούνται ώστε να ισχύουν οι συνθήκες (A)-(C). Τα αποτελέσματα των Θεωρημάτων 1 και 2 μπορούν να προκύψουν και για τις περιπτώσεις των ποινικοποιημένων ελαχίστων τετραγώνων (1.8) και της ποινικοποιημένης εύρωστης γραμμικής παλινδρόμησης (1.9).

1.4.3.2 Ο προτεινόμενος αλγόριθμος

Ο Tibshirani (1996) πρότεινε έναν αλγόριθμο για την επίλυση του προβλήματος ελαχίστων τετραγώνων της *LASSO*, ενώ ο Fu (1998) πρότεινε έναν "shooting" αλγόριθμο για την μέθοδο *LASSO*. Στην ενότητα αυτή θα αναπτύξουμε έναν νέο αλγόριθμο που προτάθηκε από τους Fan και Li (2001), με τη βοήθεια του οποίου επιλύονται τα προβλήματα ελαχιστοποίησης (1.8), (1.9) και (1.10). Αυτό γίνεται μέσω τοπικών τετραγωνικών προσεγγίσεων (*local quadratic approximations*). Ο πρώτος όρος των (1.8), (1.9) και (1.10) μπορεί να θεωρηθεί ως μια συνάρτηση απώλειας (*loss function*) του $\underset{\sim}{\beta}$. Ας την ονομάσουμε $l(\underset{\sim}{\beta})$. Οπότε οι (1.8), (1.9) και (1.10) μπορούν να γραφούν σε μια ενιαία μορφή ως

$$l(\underset{\sim}{\beta}) + n\sum_{j=1}^{d} p_{\lambda}(|\beta_j|). \qquad (1.13)$$

Οι συναρτήσεις ποινής L_1, *SCAD* και *Hard* είναι ιδιάζουσες στην αρχή και δεν έχουν συνεχείς παραγώγους δεύτερης τάξης. Παρόλα αυτά, μπορούν να προσεγγισθούν τοπικά από μια τετραγωνική συνάρτηση ως ακολούθως: Υποθέτουμε ότι έχουμε μια αρχική τιμή $\underset{\sim}{\beta}_0$ η οποία είναι πολύ κοντά στην τιμή που ελαχιστοποιεί την (1.13). Αν το β_{j0} είναι πολύ κοντά στο 0, τότε θέτουμε $\hat{\beta}_j = 0$. Αυτό σημαίνει τη διαγραφή της x_j από το τελικό μοντέλο. Ειδάλλως, χρησιμοποιούμε μια τοπική προσέγγιση της συνάρτησης ποινής $p_\lambda(|\beta_j|)$, βάσει μιας τετραγωνικής συνάρτησης, ήτοι

$$\left[p_\lambda(|\beta_j|)\right]' = p'_\lambda(|\beta_j|)\operatorname{sgn}(\beta_j) \approx \left\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\right\}\beta_j, \text{ όταν } \beta_j \neq 0.$$

Με άλλα λόγια, έχουμε ότι

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\left\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\right\}\left(\beta_j^2 - \beta_{j0}^2\right), \qquad (1.14)$$

για $\beta_j \approx \beta_{j0}$.

Το Σχήμα 1.1 δείχνει τις συναρτήσεις ποινής L_1, *SCAD* και *Hard* καθώς και τις προσεγγίσεις τους βάσει της (1.14), για δύο διαφορετικές τιμές του β_{j0}. Το μόνο μειονέκτημα της προσέγγισης αυτής, είναι ότι από τη στιγμή που κάποιος συντελεστής θα συρρικνωθεί στο 0, θα παραμείνει σε αυτήν την τιμή.

Αν τώρα η $l(\underset{\sim}{\beta})$ είναι η L_1 συνάρτηση απώλειας, όπως στην (1.9), τότε δεν έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $\underset{\sim}{\beta}$. Παρόλα αυτά, η ποσότητα $\psi(|y - \underset{\sim}{x}'\underset{\sim}{\beta}|)$ στην (1.9) μπορεί κατά ανάλογο τρόπο να προσεγγισθεί από την

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

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

καλή. Στην επόμενη ενότητα θα αναλύσουμε κάποιες διαφοροποιήσεις αυτής της προσέγγισης.

Υποθέτουμε στη συνέχεια ότι ο λογάριθμος της πιθανοφάνειας έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $\underset{\sim}{\beta}$. Συνεπώς, είναι εφικτό ο πρώτος όρος της (1.13) να προσεγγισθεί από μια τετραγωνική συνάρτηση. Οπότε, το πρόβλημα ελαχιστοποίησης (1.13) μπορεί να υποβιβασθεί σε ένα τετραγωνικό πρόβλημα ελαχιστοποίησης (*quadratic minimization problem*) και ο αλγόριθμος Newton-Raphson μπορεί να χρησιμοποιηθεί. Πράγματι, η (1.13) προσεγγίζεται (εκτός από έναν σταθερό όρο) από την ποσότητα

$$l(\underset{\sim}{\beta}_0)+\nabla l(\underset{\sim}{\beta}_0)'(\underset{\sim}{\beta}-\underset{\sim}{\beta}_0)+\frac{1}{2}(\underset{\sim}{\beta}-\underset{\sim}{\beta}_0)'\nabla^2 l(\underset{\sim}{\beta}_0)(\underset{\sim}{\beta}-\underset{\sim}{\beta}_0)+\frac{1}{2}n\underset{\sim}{\beta}'\sum\nolimits_{\lambda}(\underset{\sim}{\beta}_0)\underset{\sim}{\beta}, \quad (1.15)$$

όπου

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

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

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

Το τετραγωνικό πρόβλημα ελαχιστοποίησης (1.15), έχει ως λύση την

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

Όταν επέλθει σύγκλιση του αλγορίθμου, ο εκτιμητής ικανοποιεί τη συνθήκη

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

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

τετραγώνων (1.8), η λύση βρίσκεται με επαναληπτικό (*iterative*) υπολογισμό της παλινδρόμησης κορυφογραμμής

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

Ομοίως, η λύση της (1.9) προκύπτει με επαναληπτικό υπολογισμό της

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

όπου

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

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

1.4.3.3. Υπολογισμός του τυπικού σφάλματος

Τα τυπικά σφάλματα των εκτιμηθέντων παραμέτρων μπορούν άμεσα να υπολογισθούν, λόγω του ότι γίνεται ταυτόχρονη εκτίμηση παραμέτρων και επιλογή

μεταβλητών. Ο *sandwich* τύπος μπορεί να χρησιμοποιηθεί για την εκτίμηση της συνδιασποράς του $\hat{\beta}_1$, η μη εξαφανισμένη συνιστώσα του $\hat{\beta}$. Οπότε έχουμε,

$$\mathrm{cov}(\hat{\beta}_1) = \left\{\nabla^2 l(\hat{\beta}_1) + n\sum\nolimits_{\lambda}(\hat{\beta}_1)\right\}^{-1} \mathrm{cov}\left\{\nabla l(\hat{\beta}_1)\right\}\left\{\nabla^2 l(\hat{\beta}_1) + n\sum\nolimits_{\lambda}(\hat{\beta}_1)\right\}^{-1}. \quad (1.16)$$

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

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

$$\{|r_i|^{-1}\}, \text{ με } r_i = y_i - x_i'\beta_0 \text{ και } i = 1,...,n.$$

Οπότε για μια δοθείσα τιμή του β_0, όταν κάποια από τα υπόλοιπα $\{r_i\}$ είναι κοντά στο 0, αυτά τα σημεία αποκτούν πολύ βάρος. Για αυτό το λόγο αντικαθίσταται το βάρος με

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

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

1.4.3.4 Έλεγχος τη σύγκλισης του αλγορίθμου

Οι Fan και Li (2001), απέδειξαν με χρήση του προγράμματος MATLAB ότι όντως ο αλγόριθμος που πρότειναν συγκλίνει στη σωστή λύση. Συγκεκριμένα, χρησιμοποίησαν ένα διάνυσμα β διάστασης 100, αποτελούμενο από 50 μηδενικά και 50 μη μηδενικά στοιχεία που και δημιουργήθηκαν από την κατανομή $N(0,5^2)$. Επίσης χρησιμοποίησαν έναν 100×100 ορθοκανονικό πίνακα σχεδιασμού, για το λόγο ότι τα ποινικοποιημένα ελάχιστα τετράγωνα (PLS) έχουν τότε μαθηματική λύση κλειστής μορφής, οπότε και ήταν εφικτή η σύγκρισή της με αυτήν της

αλγοριθμικής μεθόδου τους. Το διάνυσμα των αποκρίσεων $\underset{\sim}{Y}$ δημιουργήθηκε βάσει του γραμμικού μοντέλου $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$. Τα αποτελέσματα ήταν τα εξής: Το MATLAB χρειάστηκε 0.27, 0.39 και 0.16 sec για να επέλθει σύγκλιση όσον αφορά τα *PLS* με τη *SCAD*, L_1 και *Hard* συνάρτηση ποινής αντίστοιχα. Επίσης, ο αριθμός των επαναλήψεων ήταν 30, 30 και 5 αντίστοιχα. Να σημειωθεί, ότι στη δέκατη επανάληψη, ο *PLS* εκτιμητής ήταν ήδη αρκετά κοντά στη σωστή τιμή.

1.4.4. Αριθμητικές συγκρίσεις

Στην ενότητα αυτή, θα συγκρίνουμε την απόδοση των προτεινόμενων μεθόδων με τις ήδη υπάρχουσες και θα ελέγξουμε την ακρίβεια της μεθόδου εύρεσης του τυπικού σφάλματος. Επίσης θα αναφέρουμε και κάποιες μελέτες προσομοίωσης (*simulation studies*) που έκαναν οι Fan και Li (2001) χρησιμοποιώντας τις ποινικοποιημένες μεθόδους.

1.4.4.1 Σφάλμα πρόβλεψης και σφάλμα μοντέλου

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

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

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

Ο παραπάνω τύπος μπορεί να αναλυθεί ως

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

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

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

1.4.4.2. Επιλογή των οριακών παραμέτρων

Οι Fan και Li (2001), προκειμένου να εκτιμήσουν τη ρυθμιστική (*tuning*) παράμετρο $\underset{\sim}{\theta}$, όπου $\underset{\sim}{\theta} = (\lambda, \alpha)$ για τη *SCAD* συνάρτηση ποινής και τη $\underset{\sim}{\theta} = \lambda$ για τη *LASSO* και *Hard*, χρησιμοποίησαν δύο μεθόδους. Την πενταπλή (*fivefold*) διασταυρωμένη επικύρωση και τη γενικευμένη διασταυρωμένη επικύρωση. Θα αναπτύξουμε τις δύο αυτές διαδικασίες για την περίπτωση των γραμμικών μοντέλων παλινδρόμησης. Η επέκταση των διαδικασιών αυτών σε εύρωστα γραμμικά μοντέλα παλινδρόμησης καθώς και σε γραμμικά μοντέλα βασισμένα στην πιθανοφάνεια, δεν εμπεριέχει ιδιαίτερες δυσκολίες.

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

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

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

Στη μέθοδο της γενικευμένης διασταυρωμένης επικύρωσης, μετατρέπουμε τη λύση ως

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

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

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

και μπορούμε να θεωρήσουμε ως πίνακα προβολής τον

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

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

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

το κριτήριο της γενικευμένης διασταυρωμένης επικύρωσης είναι

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

και

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

1.4.4.3. Προσομοιώσεις

Οι Fan και Li στα ακόλουθα παραδείγματα προσομοιώσεων, σύγκριναν τις προτεινόμενες μεθόδους επιλογής μεταβλητών με τις ακόλουθες μεθόδους:

Α) Ελάχιστα τετράγωνα.

Β) Παλινδρόμηση κορυφογραμμής.

Γ) Επιλογή καλύτερου υποσυνόλου.

Δ) *Garrote*

Οι προσομοιώσεις έγιναν με χρήση του MATLAB. Χρησιμοποιήθηκε επίσης η γενικευμένη διασταυρωμένη επικύρωση για την εκτίμηση των οριακών παραμέτρων.

Προσομοίωση 1-Γραμμική παλινδρόμηση: Δημιουργήθηκαν 100 σύνολα δεδομένων, αποτελούμενα από n παρατηρήσεις, βάσει του μοντέλου

$$\underset{\sim}{Y} = \underset{\sim}{x}'\underset{\sim}{\beta} + \sigma\underset{\sim}{\varepsilon},$$

όπου τα $\underset{\sim}{x}$ και $\underset{\sim}{\varepsilon}$ είναι της τυποποιημένης κανονικής κατανομής και $\underset{\sim}{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. Η συσχέτιση μεταξύ των x_i και x_j είναι $\rho^{|i-j|}$ με $\rho = 0.5$. Αρχικά, έγινε η επιλογή του $n = 40$ και του $\sigma = 3$. Έπειτα, μειώθηκε το σ σε 1 και το n αυξήθηκε στις 60 παρατηρήσεις. Το σφάλμα του μοντέλου συγκρίθηκε με αυτό του εκτιμητή ελαχίστων τετραγώνων. Η διάμεσος των σχετικών σφαλμάτων του μοντέλου (*Median of Relative Model Errors – MRME*) από 100 προσομοιωμένα σύνολα δεδομένων, υπάρχει στον πίνακα Α. Επίσης, στον ίδιο πίνακα φαίνεται και ο μέσος αριθμός των μηδενικών συντελεστών, με τη στήλη «correct» να αντιστοιχεί στο μέσο αριθμό των σωστά εκτιμηθέντων ως μηδενικοί συντελεστών, ενώ η στήλη «incorrect» αντιστοιχεί σε αυτούς που λανθασμένα εκτιμήθηκαν ως μηδενικοί.

Method	MRME (%)	Avg. No. of 0 Coefficients	
		Correct	Incorrect
$n=40, \sigma=3$			
SCAD[1]	72.90	4.20	.21
SCAD[2]	69.03	4.31	.27
LASSO	63.19	3.53	.07
Hard	73.82	4.09	.19
Ridge	83.28	0	0
Best subset	68.26	4.50	.35
Garrote	76.90	2.80	.09
Oracle	33.31	5	0
$n=40, \sigma=1$			
SCAD[1]	54.81	4.29	0
SCAD[2]	47.25	4.34	0
LASSO	63.19	3.51	0
Hard	69.72	3.93	0
Ridge	95.21	0	0
Best subset	53.60	4.54	0
Garrote	56.55	3.35	0
Oracle	33.31	5	0
$n=60, \sigma=1$			
SCAD[1]	47.54	4.37	0
SCAD[2]	43.79	4.42	0
LASSO	65.22	3.56	0
Hard	71.11	4.02	0
Ridge	97.36	0	0
Best subset	46.11	4.73	0
Garrote	55.90	3.38	0
Oracle	29.82	5	0

***Πίνακας 1.4**: Αποτελέσματα προσομοιώσεων για το γραμμικό μοντέλο παλινδρόμησης. Για τη SCAD[1] το α επιλέχθηκε βάσει της GCV και για τηSCAD[2] έχει την τιμή 3.7.*

Από τον παραπάνω πίνακα, παρατηρούμε ότι όταν ο θόρυβος είναι υψηλός και το μέγεθος του δείγματος μικρό, η *LASSO* έχει την καλύτερη απόδοση. Επίσης μειώνει σημαντικά τόσο το σφάλμα του μοντέλου όσο και την πολυπλοκότητά του. Αυτό ισχύει και για τις υπόλοιπες μεθόδους επιλογής μεταβλητών, ενώ αντιθέτως, η παλινδρόμηση κορυφογραμμής μειώνει μόνο το σφάλμα του μοντέλου. Όταν όμως μειώθηκε ο θόρυβος, η *SCAD* είναι αποδοτικότερη από τη *LASSO* και τη *Hard* . Η παλινδρόμηση κορυφογραμμής έχει κακή απόδοση ενώ η μέθοδος επιλογής καλύτερου υποσυνόλου έχει παρόμοια απόδοση με τη *SCAD*. Επίσης, η *garrote* έχει γενικά καλή απόδοση. Να σημειώσουμε και ότι η *SCAD* είχε πολύ καλά αποτελέσματα με επιλογή του $\alpha = 3.7$, η οποία τιμή χρησιμοποιήθηκε και στις επόμενες προσομοιώσεις. Τελειώνοντας, συμπεραίνουμε ότι αναμένεται η *SCAD* να έχει τόσο καλά αποτελέσματα όσο αυτά του *oracle* εκτιμητή (ο οποίος επίσης

χρησιμοποιήθηκε ώστε να συγκριθεί με τις προτεινόμενες μεθόδους), καθώς το μέγεθος του δείγματος αυξάνει.

Όσον αφορά τώρα την ακρίβεια της μεθόδου υπολογισμού του τυπικού σφάλματος (1.16), έχουμε τα εξής: Η διάμεσος των απολύτων τιμών της απόκλισης των 100 εκτιμηθέντων συντελεστών των 100 συνόλων δεδομένων, διαιρεμένη με 0.6745, συμβολιζόμενη ως *SD* και μπορεί να θεωρηθεί ως το πραγματικό τυπικό σφάλμα. Η διάμεσος των 100 αυτών εκτιμηθέντων *SDs*, συμβολίζεται με SD_m και η διάμεσος των απολύτων τιμών του σφάλματος της απόκλισης των 100 εκτιμημένων τυπικών σφαλμάτων διαιρεμένη με 0.6745, συμβολίζεται με SD_{mad} αποτελούν μια αποτίμηση της συνολικής απόδοσης της (1.16). Ο πίνακας (1.5) περιέχει τα αποτελέσματα για τους μη μηδενικούς συντελεστές, στην περίπτωση όπου $n = 60$. Στην περίπτωση όπου $n = 40$, είχαμε παρόμοια αποτελέσματα. Βάσει του πίνακα αυτού, συμπεραίνουμε ότι ο *sandwich* τύπος είναι αρκετά αποτελεσματικός.

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
SCAD[1]	.166	.161 (.021)	.170	.160 (.024)	.148	.145 (.022)
SCAD[2]	.161	.161 (.021)	.164	.161 (.024)	.151	.143 (.023)
LASSO	.164	.154 (.019)	.173	.150 (.022)	.153	.142 (.021)
Hard	.169	.161 (.022)	.174	.162 (.025)	.178	.148 (.021)
Best subset	.163	.155 (.020)	.152	.154 (.026)	.152	.139 (.020)
Oracle	.155	.154 (.020)	.147	.153 (.024)	.146	.137 (.019)

***Πίνακας 1.5**: Τυπικές αποκλίσεις των εκτιμητών στο γραμμικόμοντέλο παλινδρόμησης (n=60).*

Προσομοίωση 2-Εύρωστη γραμμική παλινδρόμηση: Δημιουργήθηκαν 100 σύνολα δεδομένων αποτελούμενα από 60 παρατηρήσεις, βάσει του μοντέλου

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

με τα ίδια $\underset{\sim}{\beta}$ και $\underset{\sim}{x}$ όπως και στην προηγούμενη προσομοίωση. Το ε είναι της τυποποιημένης κανονικής κατανομής με ένα ποσοστό 10% άτυπων σημείων (*outliers*) της κατανομής *Cauchy*. Τα αποτελέσματα βρίσκονται στον πίνακα (1.6). Βλέπουμε ότι την καλύτερη απόδοση την έχει η *SCAD*. Επίσης, οι αληθείς και οι

εκτιμώμενες βάσει της (1.6) τυπικές αποκλίσεις των εκτιμητών βρίσκονται στον πίνακα 1.7, όπου και καταδεικνύεται η πολύ καλή απόδοση της μεθόδου.

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
SCAD ($a = 3.7$)	35.52	4.71	0
LASSO	52.80	4.29	0
Hard	47.22	4.70	0
Best subset	41.53	4.85	.18
Oracle	23.33	5	0

***Πίνακας 1.6**: Αποτελέσματα προσομοίωσης για το εύρωστο γραμμικό μοντέλο παλινδρόμησης.*

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
SCAD	.167	.171 (.018)	.185	.176 (.022)	.165	.155 (.020)
LASSO	.158	.165 (.022)	.159	.167 (.020	.182	.154 (.019)
Hard	.179	.168 (.018)	.176	.176 (.025)	.157	.154 (.020)
Best subset	.198	.172 (.023)	.185	.175 (.024)	.199	.152 (.023)
Oracle	.163	.199 (.040)	.156	.202 (.043)	.166	.177 (.037)

***Πίνακας 1.7**: Τυπικές αποκλίσεις των εκτιμητών για το εύρωστο γραμμικό μοντέλο παλινδρόμησης.*

Προσομοίωση 3-Λογιστική παλινδρόμηση: Δημιουργήθηκαν 100 σύνολα δεδομένων αποτελούμενα από 200 παρατηρήσεις, βάσει του μοντέλου

$$Y \sim Bernoulli\{p(\underset{\sim}{x}'\underset{\sim}{\beta})\},$$

όπου

$$p(u) = \frac{\exp(u)}{1+\exp(u)},$$

με τις πρώτες 6 συνιστώσες των $\underset{\sim}{\beta}$ και $\underset{\sim}{x}$ να είναι οι ίδιες με αυτές της πρώτης προσομοίωσης. Οι δύο τελευταίες συνιστώσες του $\underset{\sim}{x}$ ήταν i.i.d. από την *Bernoulli* κατανομή με πιθανότητα επιτυχίας 0.5. Επίσης, όλες οι μεταβλητές ήταν κανονικοποιημένες. Τα σφάλματα του μοντέλου υπολογίσθηκαν μέσω 1000 *Monte Carlo* προσομοιώσεων. Τα αποτελέσματα βρίσκονται στους πίνακες (1.8) και (1.9). Η εκτιμήτρια ποινικοποιημένης πιθανοφάνειας με χρήση της *SCAD* είχε καλύτερη απόδοση από αυτήν της *LASSO* και της *Hard*. Επιπλέον, είχε παρόμοια απόδοση συγκριτικά με τον *oracle* εκτιμητή όσον αφορά το *MRME* και την ακρίβεια των εκτιμώμενων τυπικών σφαλμάτων.

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
SCAD ($a = 3.7$)	26.48	4.98	.04
LASSO	53.14	3.76	0
Hard	59.06	4.27	0
Best subset	31.63	4.84	.01
Oracle	25.71	5	0

***Πίνακας 1.8**: Αποτελέσματα προσομοίωσης για τη λογιστική παλινδρόμηση.*

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
SCAD ($a = 3.7$)	.571	.538 (.107)	.383	.372 (.061)	.432	.398 (.065)
LASSO	.310	.379 (.037)	.285	.284 (.019)	.244	.287 (.019)
Hard	.675	.561 (.126)	.428	.400 (.062)	.467	.421 (.079)
Best subset	.624	.547 (.121)	.398	.383 (.067)	.468	.412 (.077)
Oracle	.553	.538 (.103)	.374	.373 (.060)	.432	.398 (.064)

***Πίνακας 1.9**: Τυπικές αποκλίσεις των εκτιμητών για τη λογιστική παλινδρόμηση.*

Παρατηρούμε ότι οι εκτιμώμενες τυπικές αποκλίσεις για τον L_1 εκτιμητή ποινικοποιημένης πιθανοφάνειας (*LASSO*) είναι μικρότερες από αυτές της *SCAD*, αλλά με το συνολικό *MRME* μεγαλύτερο. Αυτό σημαίνει ότι η μεροληψία των εκτιμητών της *LASSO* είναι μεγάλη. Κάτι που ισχύει και για όλες τις προαναφερθείσες προσομοιώσεις.

1.4.5. Συμπεράσματα

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

1.5. Μια επεξήγηση της διακλιμάκωσης

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

μια διακλιμάκωση (*scaling*) στις τιμές των συντελεστών, κάτι που θα οδηγήσει σε καλύτερες εκτιμήσεις αυτών.

1.6. Η διαδικασία PRESS

Η διαδικασία επιλογής μεταβλητών του προβλεπόμενου αθροίσματος τετραγώνων των υπολοίπων PRESS (*prediction sum of squares*) προτάθηκε από τον Allen (1971). Η διαδικασία εφαρμογής του κριτηρίου ξεκινά με τη διαγραφή του πρώτου συνόλου παρατηρήσεων για την εξαρτημένη και τις ανεξάρτητες μεταβλητές και προσαρμόζοντας το μοντέλο στις επόμενες παρατηρήσεις. Εν συνεχεία, χρησιμοποιούμε το κάθε γραμμικό μοντέλο για να υπολογίσουμε το $\hat{Y}_1$ και κατόπιν εκτιμούμε το εκτιμώμενο σφάλμα $Y_1 - \hat{Y}_1$. Επαναλαμβάνουμε την ίδια διαδικασία διαγράφοντας το δεύτερο σύνολο παρατηρήσεων για την εξαρτημένη και τις ανεξάρτητες μεταβλητές, έτσι ώστε να προβλέψουμε το Y_2 και να πάρουμε τις τιμές του εκτιμώμενου σφάλματος $Y_2 - \hat{Y}_2$ και συνεχίζουμε με τον ίδιο τρόπο μέχρι να πάρουμε τόσες διαγραφές όσες και οι παρατηρήσεις μας. Για κάθε ένα από τα προκύπτοντα μοντέλα παλινδρόμησης, υπολογίζουμε το παρακάτω άθροισμα τετραγώνων των σφαλμάτων πρόβλεψης:

$$PRESS = \sum_{i=1}^{n} (\mathrm{Y}_i - \hat{Y}_{i(\mathrm{i})})^2$$

Αφού ολοκληρωθεί η διαδικασία του κριτηρίου *PRESS* επιλέγουμε ως βέλτιστο μοντέλο παλινδρόμησης εκείνο που δίνει τη μικρότερη τιμή για το άθροισμα τετραγώνων των εκτιμώμενων σφαλμάτων και ταυτόχρονα δεν περιλαμβάνει πολλές ανεξάρτητες μεταβλητές.

1.7. Η μέθοδος *nonnegative garrote*

Η μέθοδος *nonnegative garrote* δημιουργήθηκε υιοθετώντας τον *NNLS* (*nonnegative least squares*) κώδικα των Lawson και Hanson (1974) και είναι μια κλιμακωτή έκδοση της αρχής ελαχίστων τετραγώνων. Σύμφωνα με τη μέθοδο αυτή ελαχιστοποιείται η ποσότητα $\sum_{i=1}^{N}(\mathrm{Y}_i-\alpha-\sum_j c_j\hat{\beta}_j X_{ij})^2$, όπου $c_j \geq 0\ \forall j$ και $\sum_j |\, c_j\, | \leq t$ για κάποια t .

Επίσης, $\beta_j, \mathrm{j}=1,....,\mathrm{k}$ είναι οι συντελεστές παλινδρόμησης της μεθόδου ελαχίστων τετραγώνων για το πλήρες μοντέλο. Αξίζει να σημειωθεί ότι διαφορετικός συντελεστής συρρίκνωσης c_j επιβάλλεται σε κάθε συντελεστή. Αν k είναι το πλήθος των προβλεπουσών μεταβλητών, τότε μας ενδιαφέρουν οι τιμές για τις οποίες ισχύει $t < k$.

Το 1993 ο Breiman βελτίωσε την παραπάνω μέθοδο χρησιμοποιώντας μια *barrier* μέθοδο για να εισάγει έναν επιπλέον περιορισμό στο άθροισμα των συντελεστών. Η *nonnegative garotte* ελαχιστοποιεί το άθροισμα των τετραγώνων των υπολοίπων, υπό τον περιορισμό ότι το άθροισμα των απολύτων τιμών των συντελεστών να είναι μικρότερο από μια σταθερά. Η *garotte* ξεκινά με τους συντελεστές ελαχίστων τετραγώνων και τους συρρικνώνει με μη αρνητικούς παράγοντες, των οποίων το άθροισμα περιορίζεται. Σε εκτεταμένες μελέτες προσομοίωσης ο Breiman έδειξε ότι η *garotte* έχει με συνέπεια μικρότερο σφάλμα πρόβλεψης από την επιλογή υποσυνόλου (*subset selection*) και είναι εξίσου ανταγωνιστική με τη *ridge regression* (παλινδρόμηση διασέλου), εκτός από την περίπτωση κατά την οποία το μοντέλο έχει πολλούς μικρούς μη μηδενικούς συντελεστές. Ένα μειονέκτημα της *garotte* είναι ότι οι λύσεις της εξαρτώνται τόσο από το πρόσημο όσο και από το μέγεθος των OLS εκτιμητών. Συνεπώς, σε περιπτώσεις κατά τις οποίες οι OLS εκτιμητές (εκτιμητές ελαχίστων τετραγώνων) δεν αποδίδουν καλά, η *garotte* επηρεάζεται.

1.8. Μέθοδος LASSO

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

1.8.1. Περιγραφή της μεθόδου LASSO

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

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

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

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

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

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

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

Η παράμετρος t ελέγχει το μέγεθος της συρρίκνωσης, στο οποίο υπόκεινται οι συντελεστές. Έστω $\hat{\beta}_j^0$ είναι οι εκτιμητές ελαχίστων τετραγώνων και $t_0 = \sum_j |\hat{\beta}_j^0|$, τότε οι τιμές των t, οι οποίες είναι μικρότερες από t_0 θα προκαλέσει συρρίκνωση των λύσεων προς το μηδέν και ορισμένοι από τους συντελεστές μπορεί να γίνουν ακριβώς μηδέν. Για παράδειγμα, αν $t = \frac{t_0}{2}$, τότε το αποτέλεσμα θα είναι σχεδόν ίδιο με το να βρούμε το καλύτερο υποσύνολο μεγέθους $\frac{p}{2}$. Επίσης, αξίζει να σημειωθεί ότι ο πίνακας σχεδιασμού δεν είναι απαραίτητο να είναι πλήρους βαθμού.

1.8.2. Η περίπτωση του ορθοκανονικού πίνακα σχεδιασμού.

Καλύτερη εικόνα σχετικά με τη φύση της συρρίκνωσης μπορεί να μας δώσει ο ορθοκανονικός πίνακας σχεδιασμού. Έστω $\underset{\sim}{X}$ ένας $n \times p$ πίνακας σχεδιασμού, ο οποίος είναι ορθοκανονικός, δηλαδή ισχύει ότι $\underset{\sim}{X}^T \underset{\sim}{X} = I$. Οι λύσεις της εξίσωσης $(\hat{\alpha}, \hat{\underset{\sim}{\beta}}) = \text{argmin}\left\{\sum_{i=1}^{N}(\text{y}_i - \alpha - \sum_j \beta_j x_{ij})^2\right\}$ είναι $\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+$, όπου το γ καθορίζεται από τη σχέση $\sum|\beta_j| = t$.

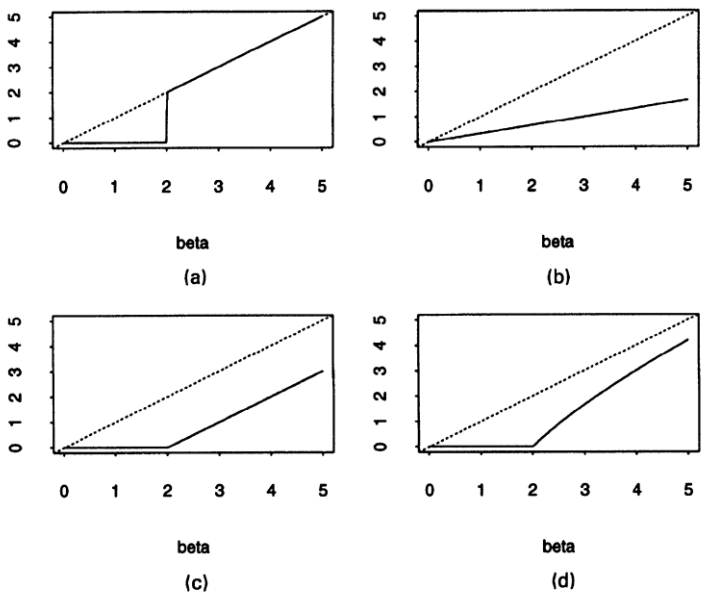
Στην περίπτωση ορθοκανονικού πίνακα σχεδιασμού, η μέθοδος επιλογής καλύτερου υποσυνόλου μεγέθους k, επιλέγει τους k μεγαλύτερους συντελεστές κατά απόλυτη τιμή και ταυτόχρονα θέτει τους υπόλοιπους ίσους με μηδέν. Για

κάποια επιλογή του λ, αυτό είναι ισοδύναμο με το να θέσουμε $\hat{\beta}_j = \hat{\beta}_j^0$ όταν $|\hat{\beta}_j^0| > \lambda$ και $\hat{\beta}_j = 0$ όταν $|\hat{\beta}_j^0| \leq \lambda$.

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

Οι λύσεις (*ridge solutions*) που προκύπτουν είναι $\frac{1}{1+\gamma}\hat{\beta}_j^0$, όπου το γ εξαρτάται από το λ ή το t. Τέλος, οι εκτιμητές *garotte* είναι $(1 - \frac{\gamma}{\hat{\beta}_j^{02}})^+ \hat{\beta}_j^0$.

Στα παρακάτω σχήματα, βλέπουμε τη μορφή των λύσεων αυτών.



<u>Σχήμα 1.10</u>: (a) Μέθοδος επιλογής καλύτερου υποσυνόλου, (b) παλινδρόμηση διασέλου, (c) μέθοδος LASSO και (d) μέθοδος garrote : ── μορφή της συρρίκνωσης των συντελεστών στην περίπτωση του ορθοκανονικού πίνακα σχεδιασμού.

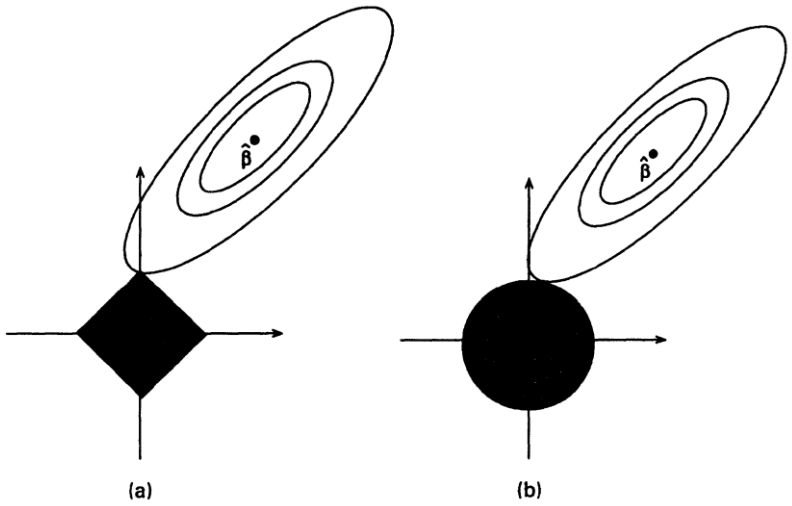
Παρατηρούμε ότι η μέθοδος παλινδρόμησης διασέλου κλιμακώνει τους συντελεστές κατά ένα σταθερό παράγοντα, ενώ η *LASSO* τους μεταθέτει κατά ένα σταθερό παράγοντα, περικόπτοντας κάποιους μηδενικούς. Η *garrote* είναι παρόμοια

με τη *LASSO* αλλά επιβάλλει λιγότερη συρρίκνωση στους μεγαλύτερους συντελεστές. Όπως έχουν δείξει οι προσομοιώσεις, οι διαφορές μεταξύ της *garrote* και της *LASSO* γίνονται μεγάλες στην περίπτωση κατά την οποία ο πίνακας σχεδιασμού δεν είναι ορθογώνιος ($X'X = nI_n$).

1.8.3. Η γεωμετρία της LASSO

Είναι φανερό από το σχήμα 1.10 για ποιο λόγο η μέθοδος LASSO δίνει συχνά συντελεστές με τιμή ακριβώς μηδέν. Τα ερωτήματα, όμως, που τίθενται είναι γιατί αυτό συμβαίνει και στη γενική περίπτωση κατά την οποία ο πίνακας σχεδιασμού δεν είναι ορθογώνιος, καθώς επίσης για ποιο λόγο η *ridge regression* (παλινδρόμηση διασέλου), η οποία χρησιμοποιεί τον περιορισμό $\sum \beta_j^2 \le t$, έχει την ίδια ιδιότητα.

Η εξήγηση προκύπτει κατά κάποιο τρόπο από το Σχήμα 1.11, στην περίπτωση κατά την οποία p=2.



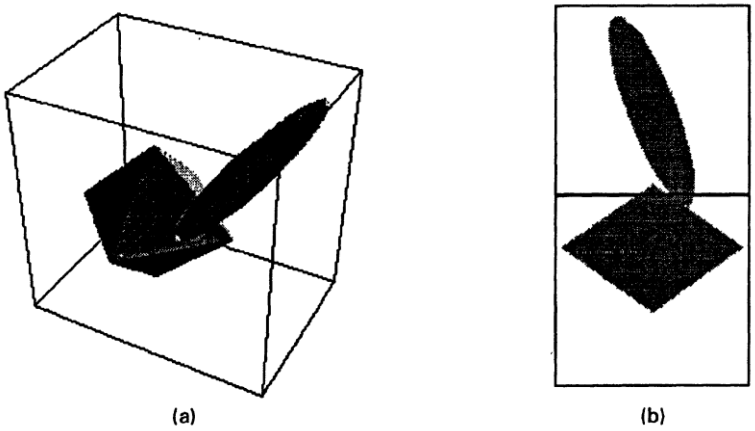
<u>Σχήμα 1.11</u>: (a) Γεωμετρία της LASSO για p=2, (b) γεωμετρία της παλινδρόμησης διασέλου για p=2.

Το κριτήριο $\sum_{i=1}^{N}(\mathrm{y}_i-\sum_j \beta_j x_{ij})^2$ ισοδυναμεί με την τετραγωνική συνάρτηση

$(\underset{\sim}{\beta}-\underset{\sim}{\hat{\beta}}^0)^{\mathrm{T}}(\underset{\sim}{X}^T\underset{\sim}{X})(\underset{\sim}{\beta}-\underset{\sim}{\hat{\beta}}^0)$ (συν μια σταθερά).

Είναι κεντραρισμένα στους *OLS* εκτιμητές. Επίσης, η περιοριστική περιοχή είναι ο χρωματισμένος ρόμβος. Η λύση που δίνει η *LASSO*, είναι το πρώτο σημείο όπου τα περιγράμματα ακουμπούν το ρόμβο. Αυτό κάποιες φορές συμβαίνει σε κάποια γωνία, οπότε έχουμε και την περίπτωση μηδενικού συντελεστή. Στο Σχήμα 1.11(b) φαίνεται η εναλλακτική περίπτωση της παλινδρόμησης διασέλου, όπου δεν υπάρχουν γωνίες για να συναντήσουν οι ελλείψεις, οπότε και είναι αρκετά δύσκολο να προκύψουν μηδενικοί συντελεστές.

Συνεχίζουμε, θέτοντας το εξής ερώτημα: Υπάρχει περίπτωση τα πρόσημα των εκτιμηθέντων συντελεστών της μεθόδου *LASSO* να διαφέρουν από αυτά των *OLS* εκτιμητών $\hat{\beta}_j^0$; Από τη στιγμή που είναι κανονικοποιημένες οι μεταβλητές, για την περίπτωση όπου $p=2$, οι κύριοι άξονες των ελλείψεων είναι σε κλίση $\pm 45^o$ ως προς τους άξονες συντεταγμένων και μπορεί να δειχθεί ότι οι ελλείψεις πρέπει να συναντήσουν το ρόμβο στο τεταρτημόριο που περιέχει το $\underset{\sim}{\hat{\beta}}^0$. Αν παρόλα αυτά το $p>2$ και υπάρχει έστω και μια μέτρια συσχέτιση στα δεδομένα, αυτό δεν είναι σίγουρο ότι ισχύει. Στο επόμενο Σχήμα 1.12 φαίνεται ένα παράδειγμα στις τρεις διαστάσεις. Με τη βοήθεια της κάτοψης (b) βλέπουμε ότι η έλλειψη ακουμπά την περιοριστική περιοχή σε οκτημόριο διαφορετικό από αυτό όπου βρίσκεται το κέντρο της.Οπότε, παρόλο που η μέθοδος garotte διατηρεί το πρόσημο κάθε $\hat{\beta}_j^0$, αυτό είναι δυνατόν να μην ισχύει για τη *LASSO*.



***Σχήμα 1.12*: (a) Περίπτωση όπου η LASSO εκτιμήτρια βρίσκεται σε οκτημόριο διαφορετικό από εκείνο της εκτιμήτριας ελαχίστων τετραγώνων, (b) κάτοψη.**

1.8.4. Τυπικά σφάλματα και εκτίμηση της παραμέτρου t.

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

Μια προσεγγιστική κλειστής μορφής εκτίμηση μπορεί να αποκτηθεί γράφοντας τη συνάρτηση ποινής $\sum_j |\beta_j|$ ως $\sum_j \beta_j^2 / |\beta_j|$. Συνεπώς, βάσει της *LASSO* εκτιμήτριας $\underset{\sim}{\widehat{\beta}}$, μπορούμε να προσεγγίσουμε τη λύση με χρήση παλινδρόμησης διασέλου (*ridge regression*) της μορφής

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

όπου $\underset{\sim}{W}$ είναι ο διαγώνιος πίνακας με στοιχεία $|\widehat{\beta}_j|$, ο $\underset{\sim}{W}^-$ είναι ο γενικευμένος αντίστροφος του $\underset{\sim}{W}$ και το λ έχει επιλεγεί κατά τρόπον ώστε $\sum_j |\beta_j|^* = t$. Ο

πίνακας συνδιασποράς των εκτιμητών μπορεί τότε να προσεγγισθεί από την ποσότητα

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

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

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

1.8.5. Adaptive LASSO

Η LASSO στο Tibshirani (1996) έχει χρησιμοποιηθεί ευρέως εξαιτίας της κυρτότητάς της. Όμως, δημιουργεί μεροληψία εκτίμησης. Το πρόβλημα αυτό επισημάνθηκε από τους Fan & Li (2001) και επισήμως φαίνεται από το Zou (2006) ακόμη και σε μια πεπερασμένη ρύθμιση παραμέτρων. Για ξεπεραστεί αυτό το πρόβλημα μεροληψίας, ο Zou (2006) πρότεινε μια προσαρμοσμένη LASSO και ο Meinshausen (2007) πρότεινε μια «χαλαρή» (relaxed) LASSO.

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

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

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

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

1.9. Dantzig selector

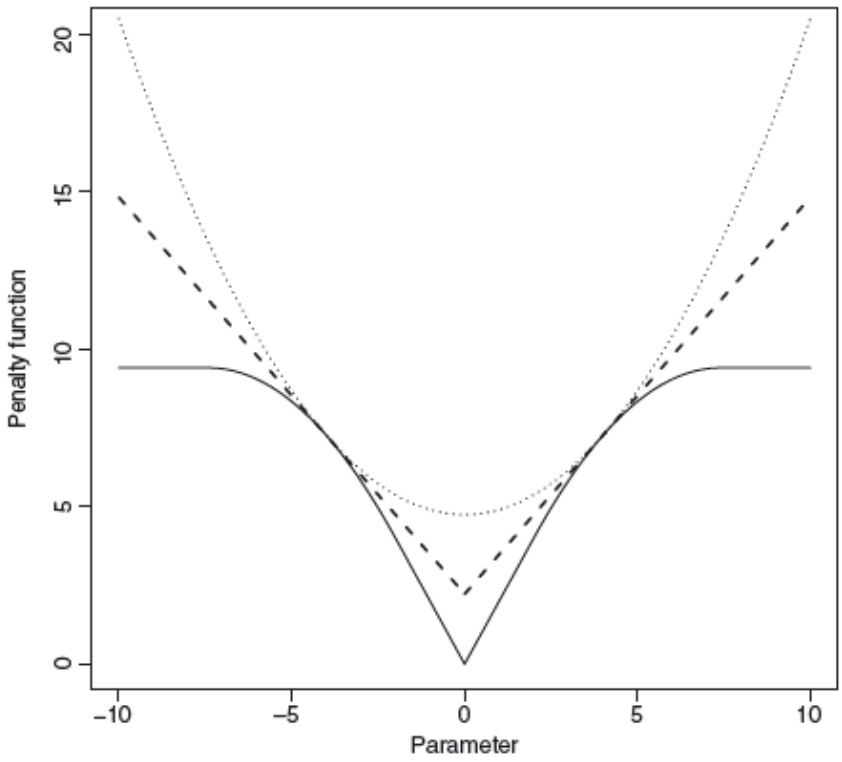
Σε πολλές σημαντικές στατιστικές εφαρμογές, ο αριθμός των μεταβλητών ή παραμέτρων *p* είναι πολύ μεγαλύτερος από τον αριθμό των παρατηρήσεων *n*. Για παράδειγμα, στη ραδιολογία και στη βιοϊατρική απεικόνιση υπάρχουν πολλές περιπτώσεις στις οποίες $p \gg n$. Υποθέτουμε ότι έχουμε παρατηρήσεις για τις οποίες ισχύει $Y = X\beta + z$, όπου $\beta \in \mathbb{R}$ είναι ένα διάνυσμα παραμέτρων, *X* είναι ένας πίνακας με πολύ λιγότερες ενδεχομένως γραμμές από στήλες ($n \ll p$) και $z \sim N(0, \sigma^2 \mathrm{I}_n)$ είναι ένα διάνυσμα ανεξάρτητων κανονικών τυχαίων μεταβλητών. Για να υπολογίσουμε το β, εισάγουμε ένα νέο εκτιμητή, που προτάθηκε από τους Candes & Tao (2007) για να ανακτήσει ένα αραιό υψηλής διάστασης διάνυσμα παραμέτρων στο γραμμικό μοντέλο, ο οποίος λέγεται επιλογέας *Dantzig* (*Dantzig selector*) και είναι μια λύση του προβλήματος της l_1 κανονικοποίησης $\min_{\tilde{\beta} \in \mathbb{R}^p} \|\tilde{\beta}\|_{l_1}$ δεδομένου ότι $\|X_r^*\|_{\infty} \leq (1 + \mathrm{t}^{-1})\sqrt{2 \log p} \cdot \sigma$ όπου r είναι το διάνυσμα υπολοίπων $Y - X\beta$.

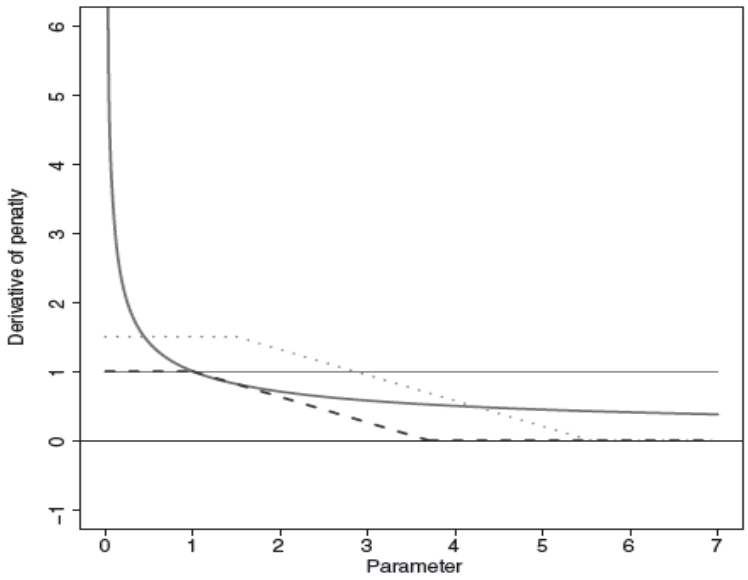
Έστω ότι το X υπακούει σε μια ενιαία αρχή αβεβαιότητας και η πραγματική παράμετρος β είναι αρκετά αραιή, τότε με πολύ μεγάλη πιθανότητα ισχύει ότι:

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

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

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





<u>Σχήμα 1.13</u> (a) SCAD ποινή (—) και οι τοπικές γραμμικές (- - - -) και τετραγωνικές $(\cdots\cdot)$ προσεγγίσεις και (b) $p'_\lambda(\cdot)$ για την ποινικοποιημένη L_1 (—), SCAD με λ=1 (- - - -) και λ=1.5 $(\cdots\cdot)$ και προσαρμοσμένη LASSO (—) με γ=0.5.

ΚΕΦΑΛΑΙΟ 2ο

ΜΕΘΟΔΟΣ ΣΙΓΟΥΡΟΥ ΚΡΗΣΑΡΙΣΜΑΤΟΣ (SURE INDEPENDENCE SCREENING (SIS))

2.1. Εισαγωγή

Η επιλογή μεταβλητών παίζει ένα σημαντικό ρόλο στα στατιστικά μοντέλα μεγάλης διάστασης, τα οποία εμφανίζονται σε πολλούς τομείς και είναι το κλειδί για πολλές επιστημονικές ανακαλύψεις. Σε προβλήματα με μεγάλη διάσταση p, η ακρίβεια των εκτιμήσεων και το υπολογιστικό κόστος είναι δύο σημαντικοί παράγοντες.

Για τα προβλήματα μεγάλης διάστασης, παρουσιάστηκε το «σίγουρο» κρησάρισμα (*sure screening*) και προτάθηκε μια μέθοδος πάνω σε αυτό, που βασίζεται στη μάθηση συσχέτισης και καλείται Sure Independence Screening. Η μέθοδος αυτή χρησιμοποιείται για να μειώσουμε τη διάσταση από υψηλή σε μια μέτριας κλίμακας που είναι μικρότερη από το μέγεθος του δείγματος.

Σε ένα αρκετά γενικό πλαίσιο, η μάθηση συσχέτισης φαίνεται ότι έχει την ιδιότητα του σίγουρου κρησαρίσματος ακόμα και για διάσταση που αυξάνεται εκθετικά.

Θεωρούμε το πρόβλημα της εκτίμησης ενός p-διανύσματος των παραμέτρων β από το γραμμικό μοντέλο

$$y = X\beta + \varepsilon \qquad (2.1)$$

όπου $y = (\mathrm{Y}_1, ..., \mathrm{Y}_n)^T$ είναι ένα *n*-διάστατο διάνυσμα αποκρίσεων, $X = (\mathrm{x}_1, ..., \mathrm{x}_n)^T$ ένας $n \times p$ τυχαίος πίνακας με ανεξάρτητα και ταυτόσημα κατανεμημένα (i.i.d.) $\mathrm{x}_1, ..., \mathrm{x}_n$, $\beta = (\beta_1, ..., \beta_p)^T$ ένα διάνυσμα παραμέτρων *p*-διάστασης και $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ ένα διάνυσμα από ταυτόσημα κατανεμημένα τυχαία σφάλματα *n*-διάστασης.

Όταν η διάσταση *p* είναι μεγάλη, τότε μπορούμε να θεωρήσουμε ότι μόνο ένας μικρός αριθμός επεξηγηματικών μεταβλητών μεταξύ των $X_1, ..., X_p$ μπορεί να οδηγήσει στην απόκριση, κάτι το οποίο μπορεί να μας οδηγήσει στο συμπέρασμα ότι το παραμετρικό διάνυσμα β είναι αραιό. Με την αραιότητα, η επιλογή μεταβλητών μπορεί να βελτιώσει την ακρίβεια της εκτίμησης προσδιορίζοντας αποτελεσματικά το υποσύνολο των σημαντικών επεξηγηματικών μεταβλητών και ταυτόχρονα να ενισχύσει επεξηγηματικά το μοντέλο με φειδωλή εκπροσώπηση.

Η αραιότητα συχνά συνδέεται με δεδομένα μεγάλης διάστασης, τα οποία συναντάμε σε πολλούς τομείς της σύγχρονης στατιστικής. Τα προβλήματα αυτά συχνά εμφανίζονται στη γονιδιωματική ιατρική, όπως στην έκφραση γονιδίων, στη βιοϊατρική απεικόνιση, στη λειτουργία της μαγνητικής τομογραφίας, στην επεξεργασία σήματος, στην ανάλυση εικόνας και στα οικονομικά, όπου ο αριθμός των μεταβλητών ή των παραμέτρων *p* μπορεί να είναι πολύ μεγαλύτερος από το μέγεθος του δείγματος *n*. Για παράδειγμα, κάποιος θα μπορούσε να επιθυμεί να ταξινομήσει τους όγκους χρησιμοποιώντας εκφράσεις γονιδίων με συστοιχίες ή με πρωτεϊνικά δεδομένα ή θα μπορούσε να συνδέσει τις συγκεντρώσεις πρωτεϊνών με την έκφραση των γονιδίων ή για την πρόβλεψη ορισμένων κλινικών προγνώσεων (π.χ. χρόνο επιβίωσης) χρησιμοποιώντας δεδομένα γονιδιακής έκφρασης. Για αυτού του είδους τα προβλήματα, η διάσταση μπορεί να είναι μεγαλύτερη από το μέγεθος του δείγματος, κάτι που μας οδηγεί σε νέες ή εκτεταμένες στατιστικές μεθόδους και θεωρίες.

Στο μοντέλο που θεωρήσαμε αρχικά (2.1), μια πρόκληση είναι να βρούμε δέκα σημαντικές μεταβλητές από τις χιλιάδες επεξηγηματικές μεταβλητές, με αριθμό παρατηρήσεων συνήθως σε δεκάδες ή εκατοντάδες.

2.1.1. Μείωση διάστασης

Η μείωση της διάστασης ή η επιλογή χαρακτηριστικών είναι μια αποτελεσματική στρατηγική για την αντιμετώπιση των προβλημάτων μεγάλης διάστασης. Μειώνοντας τη διάσταση, το «βάρος» των υπολογισμών μπορεί να μειωθεί αισθητά. Εν τω μεταξύ, μπορούμε να λάβουμε ακριβείς εκτιμήσεις χρησιμοποιώντας μερικές μεθόδους χαμηλής διάστασης που έχουν αναπτυχθεί. Επομένως, τίθεται το εξής ζήτημα:

«Η μείωση της διάστασης p από μια μεγάλη κλίμακα (έστω $\exp\{O(n^{\xi})\}$ **για κάποιο** $\xi > 0$ ***) σε μια άλλη σχετικά μεγάλη κλίμακα d (***$o(n)$ ***) με μία γρήγορη και αποτελεσματική μέθοδο».***

Για να το πετύχουμε αυτό, μπορούμε να εισάγουμε την ιδέα του σίγουρου κρησαρίσματος και να προτείνουμε μια μέθοδο σίγουρου κρησαρίσματος, η οποία βασίζεται στη μάθηση συσχέτισης και η οποία απομακρύνει τα χαρακτηριστικά που έχουν ασθενή συσχέτιση με την απόκριση. Ένα τέτοιο κρησάρισμα συσχέτισης καλείται Sure Independence Screening (*SIS*). Από δω και στο εξής, με τον όρο σίγουρο κρησάρισμα θα εννοούμε την ιδιότητα όπου όλοι οι σημαντικοί παράγοντες επιβιώνουν μετά από το κρησάρισμα των μεταβλητών με πιθανότητα που τείνει στο 1. Αυτό περιορίζει σημαντικά τον έλεγχο για σημαντικές επεξηγηματικές μεταβλητές.

2.2. Μέθοδος Σίγουρου Κρησαρίσματος: μάθηση συσχέτισης

Με τον όρο σίγουρο κρησάρισμα εννοούμε την ιδιότητα όπου όλες οι σημαντικές μεταβλητές επιβιώνουν μετά την εφαρμογή της διαδικασίας κρησαρίσματος μεταβλητών με πιθανότητα που τείνει στο 1. Η μέθοδος για τη μείωση της διάστασης είναι επιθυμητή αν έχει την ιδιότητα του σίγουρου

κρησαρίσματος. Θα παρουσιάσουμε μια απλή μέθοδο σίγουρου κρησαρίσματος που χρησιμοποιεί τη μάθηση συσχέτισης. Εστιάζουμε σε κάθε μεταβλητή εισόδου έτσι ώστε η παρατηρούμενη μέση τιμή να είναι 0 και περιορίζουμε κάθε επεξηγηματική μεταβλητή, έτσι ώστε η τυπική απόκλιση του δείγματος να είναι 1. Έστω ότι $M_* = \{1 \le i \le p : \beta_i \ne 0\}$ είναι το πραγματικό αραιό μοντέλο με μη-αραιό μέγεθος $s = |M_*|$. Οι υπόλοιπες $\mathrm{p}-\mathrm{s}$ μεταβλητές μπορούν να συσχετιστούν με τη μεταβλητή απόκρισης μέσω της σύνδεσης με τις επεξηγηματικές μεταβλητές που περιέχονται στο μοντέλο.

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

$$\omega = X^T y \tag{2.2}$$

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

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

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

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

Μια τέτοια μάθηση συσχέτισης κατατάσσει τη σημαντικότητα των χαρακτηριστικών σύμφωνα με την οριακή συσχέτιση με τη μεταβλητή απόκρισης και απορρίπτει αυτά που έχουν αδύναμες οριακές συσχετίσεις με τη μεταβλητή απόκρισης. Αυτή τη μέθοδο κρησαρίσματος συσχέτισης την ονομάζουμε SIS, αφού κάθε χαρακτηριστικό χρησιμοποιείται ανεξάρτητα ως μια επεξηγηματική μεταβλητή, ώστε να αποφασίσει πόσο χρήσιμο είναι να προβλεφθεί η μεταβλητή απόκρισης.

Αυτή η έννοια είναι ευρύτερη από τη συσχέτιση κρησαρίσματος και εφαρμόζεται σε γενικευμένα γραμμικά μοντέλα, καθώς και σε προβλήματα ταξινόμησης με διάφορες λειτουργίες απώλειας ή ζημιάς .

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

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

Η SIS είναι μια δύσκολη μέθοδος. Για ορθογώνιους σχεδιασμούς πινάκων είναι κατανοητή. Όμως για γενικές μεθόδους σχεδιασμού πινάκων, δεν υπάρχει θεωρητικό υπόβαθρο. Γι' αυτό, παρόλο που συχνά χρησιμοποιείται στις εφαρμογές, είναι σημαντικό να ορίσουμε τις συνθήκες κάτω από τις οποίες οι ιδιότητες του σίγουρου κρησαρίσματος ισχύουν. Με άλλα λόγια, επιλέγουμε $[\gamma_n]$ μεταβλητές, που έχουν το μεγαλύτερο μέγεθος συσχέτισης με τη μεταβλητή απόκρισης, χωρίς να υπολογίσουμε τη συμβολή των άλλων συμμεταβλητών στην παλινδρόμηση. Αποδεικνύεται ότι μια τέτοια απλή διαδικασία έχει την ιδιότητα του σίγουρου κρησαρίσματος με την εξής έννοια:

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

για δοθέν γ. Πιο συγκεκριμένα, έχουμε:

Θεώρημα 1. Υποθέτουμε ότι:

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

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

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

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

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

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

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

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

Από το θεώρημα συνεπάγεται ότι ο μη αραιός αριθμός $s < [\gamma_n]$. Αποδεικνύει έτσι ότι η SIS μπορεί να μειώσει τη διάσταση p από την εκθετικά αυξανόμενη σε μια σχετικά μεγάλη κλίμακα $d = [\gamma_n] = O(n^{1-\theta}) < n$ για κάποιο $\theta > 0$ και το μειωμένο μοντέλο M_γ εξακολουθεί να περιέχει όλες τις μεταβλητές που υπάρχουν στο πραγματικό μοντέλο με συντριπτική πιθανότητα. Πιο συγκεκριμένα, μπορούμε να επιλέξουμε το μέγεθος του υπομοντέλου $d = n-1$ ή $n / \log n$, να είναι συντηρητικό. Παρά το γεγονός ότι η SIS προτείνεται για τη μείωση του αριθμού των μεταβλητών του μοντέλου από p σε d, όπου το d είναι μικρότερο από το μέγεθος του δείγματος n,

τίποτα δεν μας σταματά από το να το εφαρμόσουμε με τελικό μέγεθος μοντέλου $d \geq n$. Είναι προφανές ότι μεγαλύτερο d σημαίνει και μεγαλύτερη πιθανότητα να περιλαμβάνει το πραγματικό μοντέλο M_* στο τελικό μοντέλο M_γ.

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

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

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

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

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

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

2.3.Σύνδεση με άλλες μεθόδους μείωσης διάστασης

Η SIS χρησιμοποιεί τις οριακές πληροφορίες της συσχέτισης για να μειώσει τη διάσταση.

Θα μελετήσουμε τη SIS στα πλαίσια της ταξινόμησης, στην οποία η ιδέα του ανεξάρτητου κρησαρίσματος εμφανίζεται φυσικά και έχει χρησιμοποιηθεί ευρέως.

Το πρόβλημα της ταξινόμησης μπορούμε να το δούμε ως μια ειδική περίπτωση παλινδρόμησης με μεταβλητές απόκρισης με διακριτές τιμές, όπως ± 1. Για τα προβλήματα μεγάλης διάστασης, όπως η ταξινόμηση του όγκου με χρήση γονιδιακής έκφρασης ή πρωτεϊνικών δεδομένων είναι προτιμότερο να μην ταξινομήσουμε τα δεδομένα χρησιμοποιώντας το πλήρες διάστημα χαρακτηριστικών

γνωρισμάτων, λόγω της συσσώρευσης του θορύβου και της επεξηγηματικότητας. Επίσης, πολλά από τα χαρακτηριστικά έρχονται στο προσκήνιο μέσω της σύνδεσης με τις σημαντικές επεξηγηματικές μεταβλητές. Παρόλα αυτά, η επιλογή χαρακτηριστικών είναι σημαντική για την ταξινόμηση υψηλής διάστασης.

Η SIS μπορεί να χρησιμοποιηθεί για να μειώσει το χώρο των χαρακτηριστικών. Υποθέτουμε ότι έχουμε n_1 δείγματα από την κλάση 1 και n_2 δείγματα από την κλάση -1. Τότε ο κατά συνιστώσες εκτιμητής παλινδρόμησης (2.2) γίνεται:

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

όπου κάθε μεταβλητή έχει ομαλοποιηθεί οριακά.

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

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

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

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

2.4.Τεχνικές επιλογής μοντέλου βασισμένες στη SIS

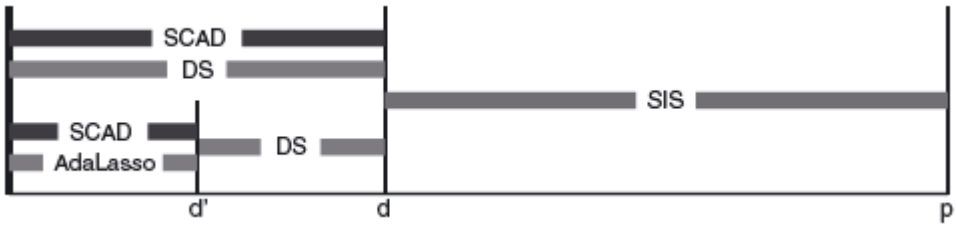
Με τη μάθηση συσχέτισης, μπορούμε να συρρικνώσουμε ολόκληρο το μοντέλο $\{1,....,\mathrm{p}\}$ σε ένα υποσύνολο του μοντέλου $\mathrm{M} = \mathrm{M}_\gamma$ με μέγεθος $d = [\gamma_n]$. Το αρχικό πρόβλημα της εκτίμησης του αραιού *p*-διανύσματος β στο μοντέλο (2.1) μειώνεται στο να εκτιμηθεί το αραιό *d*-διάνυσμα $\beta = (\beta_1,..., \beta_d)^T$ που βασίζεται τώρα στο πολύ μικρότερο υπομοντέλο

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

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

2.5. Μέθοδοι επιλογής μοντέλου βασισμένες στη SIS

Για τα προβλήματα επιλογής μεταβλητών μεγάλης διάστασης, αρχικά προτείνουμε να χρησιμοποιηθεί μια μέθοδος σίγουρου κρησαρίσματος όπως η SIS για τη μείωση της διάστασης από p σε σχετικά μεγάλη κλίμακα d κάτω από το μέγεθος του δείγματος n. Στη συνέχεια χρησιμοποιούμε μια μέθοδο επιλογής μοντέλου χαμηλότερης διάστασης όπως η SCAD, ο Dantzig επιλογέας, η LASSO ή η adaptive LASSO. Καλούμε τη SIS που ακολουθείται από τη SCAD και τον Dantzig επιλογέα ως SIS-SCAD και SIS-DS αντίστοιχα, για χάρη συντομίας. Σε ορισμένες περιπτώσεις, μπορεί να θέλουμε να μειώσουμε περισσότερο το μέγεθος του μοντέλου σε $\mathrm{d}' < \mathrm{d}$ χρησιμοποιώντας μια μέθοδο, όπως ο Dantzig επιλογέας ή η LASSO με κατάλληλο συντονισμό, και τελικά να επιλέξουμε ένα μοντέλο με μια πιο εκλεπτυσμένη μέθοδο όπως η SCAD ή η adaptive LASSO, που θα καλούνται SIS-DS-SCAD και SIS-DS-AdaLASSO αντίστοιχα για ευκολία. Στο παρακάτω σχήμα 2.1 έχουμε τη σχηματική απεικόνιση αυτών των προσεγγίσεων.



<u>Σχήμα 2.1</u>. Μέθοδοι της επιλογής μοντέλου με υψηλή διάσταση.

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

2.6. Αριθμητικές μελέτες

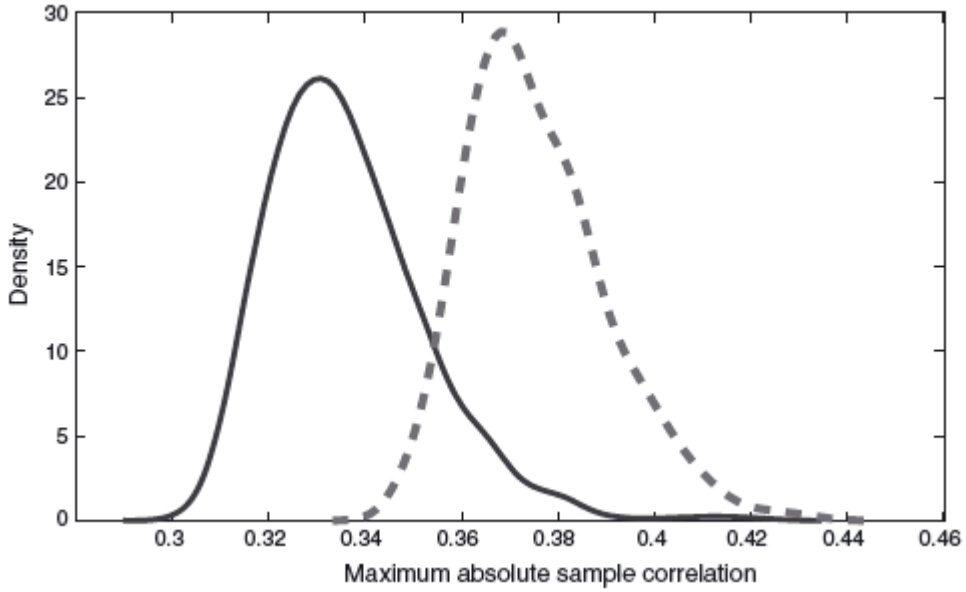
Για τη μελέτη της απόδοσης των μεθόδων επιλογής μοντέλου που βασίζονται στη SIS και είδαμε παραπάνω, θα παρουσιάσουμε τις παρακάτω προσοποιώσεις.

2.6.1. Προσομοίωση Ι: «ανεξάρτητα χαρακτηριστικά»

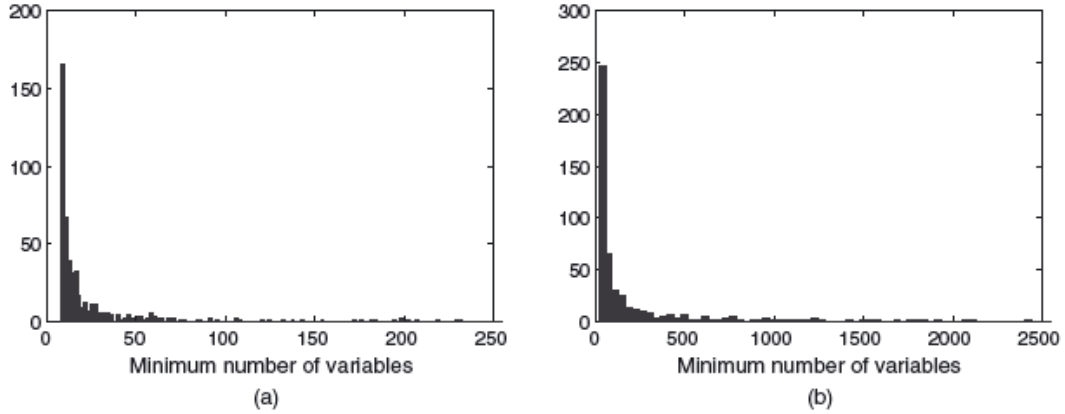
Για την πρώτη προσομοίωση, χρησιμοποιήσαμε το γραμμικό μοντέλο (2.1) με i.i.d. κανονικές επεξηγηματικές μεταβλητές και κανονικό θόρυβο με τυπική απόκλιση $\sigma = 1.5$. Θεωρήσαμε δύο τέτοια μοντέλα με $(\mathrm{n}, \mathrm{p}) = (200, 1000)$ και $(\mathrm{n}, \mathrm{p}) = (800, 20000)$. Τα μεγέθη s των πραγματικών μοντέλων, δηλαδή οι αριθμοί των μη μηδενικών συντελεστών, επιλέχθηκαν να είναι 8 και 18, καθώς και οι μη μηδενικές συνιστώσες των p-διανυσμάτων β επιλέχθηκαν τυχαία ως εξής: Ορίζουμε $\alpha = 4\log(\mathrm{n})/\mathrm{n}^{1/2}$ και $5\log(\mathrm{n})/\mathrm{n}^{1/2}$ αντίστοιχα και επιλέγουμε μη μηδενικούς συντελεστές της μορφής $(-1)^u\left(\alpha + |z|\right)$ για κάθε μοντέλο, όπου το u προέρχεται από μια κατανομή Bernoulli με παράμετρο 0.4 και το z προέρχεται από την τυπική κανονική κατανομή. Ειδικότερα, οι l_2- νόρμες $\|\beta\|$ των δύο προσομοιωμένων μοντέλων είναι 6.795 και 8.908. Για κάθε μοντέλο προσομοιώσαμε 200 σύνολα δεδομένων. Ακόμη και με τις i.i.d. κανονικές επεξηγηματικές μεταβλητές, αυτές οι ρυθμίσεις είναι μη τετριμμένες αφού υπάρχει μη αμελητέα συσχέτιση δείγματος μεταξύ των επεξηγηματικών μεταβλητών, η οποία αντικατοπτρίζει τη δυσκολία της επιλογής μεταβλητών υψηλής διάστασης. Ως απόδειξη, αναφέρουμε στο Σχ. 2.2 τις κατανομές της μέγιστης απόλυτης συσχέτισης δείγματος όταν $n = 200$ και $p = 1000$ και $p = 5000$. Αυτό αποκαλύπτει σημαντική συσχέτιση δείγματος μεταξύ

των επεξηγηματικών μεταβλητών. Η πολλαπλή κανονική συσχέτιση μεταξύ δύο ομάδων επεξηγηματικών μεταβλητών μπορεί να είναι μεγαλύτερη.

Για να εκτιμηθούν τα αραιά p-διανύσματα β, εφαρμόζουμε έξι μεθόδους: τον επιλογέα Dantzig χρησιμοποιώντας έναν αρχέγονο διπλό αλγόριθμο, τη LASSO χρησιμοποιώντας τον LARS αλγόριθμο και τις μεθόδους SIS-SCAD, SIS-DS, SIS-DS-SCAD και SIS-DS-AdaLASSO (βλ. Σχ. 2.1). Για τις SIS-SCAD και SIS-DS μεθόδους , επιλέξαμε $d = \lfloor n/\log(n) \rfloor$ και για τις δύο τελευταίες μεθόδους επιλέξαμε $d = n-1$ και $d' = \lfloor n/\log(n) \rfloor$, ενώ στο μεσαίο βήμα χρησιμοποιήθηκε ο επιλογέας Dantzig για να μειώσει το μέγεθος του μοντέλου περισσότερο από το d στο d', επιλέγοντας μεταβλητές με τα d' μεγαλύτερα κατά συνιστώσες μεγέθη του εκτιμώμενου d-διανύσματος (βλ. Σχ. 2.1).



Σχήμα 2.2. Κατανομές της μέγιστης απόλυτης συσχέτισης δείγματος όταν $n = 200$ και $p = 1000$ (—) και $n = 200$ και $p = 5000$ (-----).



Σχήμα 2.3. Κατανομή του μικρότερου αριθμού επιλεγμένων μεταβλητών που απαιτούνται να συμπεριληφθούν στο πραγματικό μοντέλο χρησιμοποιώντας τη SIS όταν (a) $n = 200$ και $p = 1000$ **και (b)** $n = 800$ και $p = 20000$ **στην προσομοίωση I.**

p	*Results for the following methods:*					
	Dantzig selector	*Lasso*	*SIS–SCAD*	*SIS–DS*	*SIS–DS–SCAD*	*SIS–DS–AdaLasso*
1000	10^3	62.5	15	37	27	34
	(1.381)	(0.895)	(0.374)	(0.795)	(0.614)	(1.269)
20000	—	—	37	119	60.5	99
	—	—	(0.288)	(0.732)	(0.372)	(1.014)

Πίνακας 2.1. Αποτελέσματα της προσομοίωσης I: οι μέσοι των μεγεθών των επιλεγμένων μοντέλων και σφαλμάτων εκτίμησης (στην παρένθεση).

Τα αποτελέσματα της προσομοίωσης συνοψίζονται στο Σχήμα 2.3 και στον Πίνακα 2.1., ο οποίος παρήχθη πάνω στη βάση των 500 προσομοιώσεων, εικονίζει την κατανομή του ελάχιστου αριθμού των επιλεγμένων μεταβλητών, δηλαδή το επιλεγμένο μέγεθος μοντέλου, που απαιτείται για το σίγουρο κρησάρισμα με τη χρήση της SIS. Δείχνει σαφώς ότι και στις δύο ρυθμίσεις είναι ασφαλές να συρρικνωθεί το πλήρες μοντέλο σε ένα υπομοντέλο του μεγέθους $\left[n/\log(n)\right]$ με τη SIS, το οποίο είναι συνεπές με την ιδιότητα του σίγουρου κρησαρίσματος της SIS που φαίνεται στο Θεώρημα 1. Για παράδειγμα, στην περίπτωση όπου $n = 200$ και p=1000, μειώνοντας το μέγεθος του μοντέλου σε 50 περιλαμβάνει τις μεταβλητές στο πραγματικό μοντέλο με υψηλή πιθανότητα και για την περίπτωση $n = 800$ και p=20000, είναι ασφαλές να μειώσουμε τη διάσταση σε περίπου 500. Για καθεμία από τις παραπάνω έξι μεθόδους, έχουμε εκθέσει στον Πίνακα 2.1 τη διάμεση τιμή των επιλεγμένων μεγεθών μοντέλου και τη διάμεσο των σφαλμάτων εκτίμησης $\left\|\hat{\beta} - \beta\right\|$ στη l_2-νόρμα. Τέσσερις καταχωρήσεις στον Πίνακα 2.1 λείπουν λόγω της περιορισμένης υπολογιστικής ισχύος και του λογισμικού που χρησιμοποιήθηκε. Σε σύγκριση, η SIS μειώνει σημαντικά την υπολογιστική επιβάρυνση.

Από τον Πίνακα 2.1 φαίνεται ότι ο Dantzig επιλογέας δίνει μη-αραιές λύσεις και η LASSO χρησιμοποιώντας τη διασταυρωμένη επικύρωση για την επιλογή της παραμέτρου συντονισμού παράγει μεγάλα μοντέλα. Αυτό μπορεί να οφείλεται στο ότι οι μεροληψίες στη LASSO απαιτούν ένα μικρό εύρος ζώνης στη διασταυρωμένη επικύρωση. Όμως ένα μικρό εύρος ζώνης οδηγεί σε έλλειψη της «*sparsistency*» σύμφωνα με την ορολογία του Ravikumar et al. (2007). Αυτό έχει επίσης παρατηρηθεί και αποδειχθεί στην εργασία των Lam & Fan (2007) στο πλαίσιο της εκτίμησης αραιής συνδιακύμανσης ή πινάκων ακριβείας. Θα πρέπει να επισημάνουμε εδώ ότι μια παραλλαγή του επιλογέα Dantzig, ο Gauss-Dantzig επιλογέας στους Candes & Tao (2007), θα πρέπει να δώσει πολύ μικρότερα μοντέλα. Από όλες τις μεθόδους, η SIS-SCAD εκτελείται καλύτερα και παράγει πολύ μικρότερα και πιο ακριβή μοντέλα. Παρατηρώντας τα σφάλματα εκτίμησης, γίνεται σαφές ότι η SCAD δίνει πιο ακριβείς εκτιμήσεις από την προσαρμοσμένη LASSO (adaptive LASSO). Επίσης, η SIS ακολουθούμενη από τον Dantzig επιλογέα, και μόνο από αυτόν,

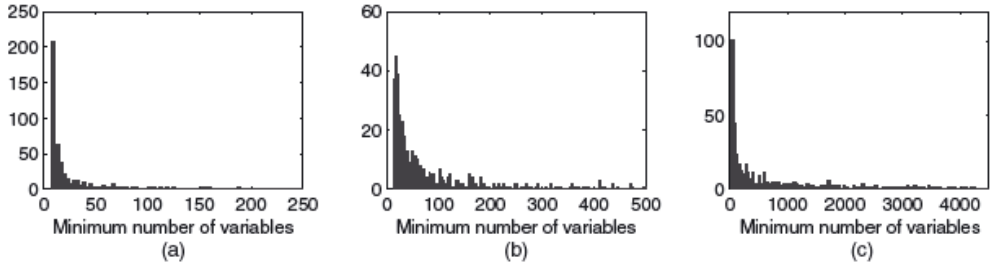
βελτιώνει την ακρίβεια της εκτίμησης, γεγονός το οποίο είναι σύμφωνο με τα θεωρητικά μας αποτελέσματα.

2.6.2. Προσομοίωση ΙΙ: «εξαρτημένα» χαρακτηριστικά

Για τη δεύτερη προσομοίωση, χρησιμοποιήσαμε παρόμοια μοντέλα με εκείνα στην προσομοίωση Ι εκτός από το ότι οι επεξηγηματικές μεταβλητές τώρα συσχετίζονται μεταξύ τους. Θεωρήσαμε τρία μοντέλα με $(n, p, s) = (200, 1000, 5)$, $(200, 1000, 8)$ και $(800, 20000, 14)$, όπου το s υποδηλώνει το μέγεθος του πραγματικού μοντέλου, δηλαδή τον αριθμό των μη μηδενικών συντελεστών. Τα τρία p-διανύσματα β παρήχθησαν κατά τον ίδιο τρόπο όπως και στην προσομοίωση Ι. Θέτουμε $(\sigma, \alpha) = \left(1, 2\log(\mathrm{n})/n^{1/2}\right), \left(1.5, 4\log(\mathrm{n})/n^{1/2}\right), \left(2, 4\log(\mathrm{n})/n^{1/2}\right)$. Ειδικότερα, οι l_2-νόρμες $\|\beta\|$ των τριών προσομοιωμένων μοντέλων είναι 3.304, 6.795 και 7.257. Για να εισάγουμε τη συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών, αρχικά χρησιμοποιήσαμε τη MATLAB συνάρτηση *sprandsym* για να δημιουργήσουμε τυχαία ένα $s \times s$ συμμετρικό θετικά ορισμένο πίνακα A με $n^{1/2}/\log(\mathrm{n})$ και δημιουργήσαμε δείγματα από s επεξηγηματικές μεταβλητές $X_1, \ldots, X_s$ από $N(0, \mathrm{A})$. Στη συνέχεια, πήραμε $\mathrm{Z}_{s+1}, \ldots, Z_p \sim N(0, \mathrm{I}_{p-s})$ και ορίσαμε τις εναπομείνασες επεξηγηματικές μεταβλητές ως $\mathrm{X}_i = Z_i + rX_{i-s}$, $i = s+1, \ldots, 2s$, και $\mathrm{X}_i = Z_i + (1-r)X_1$, $i = 2s+1, \ldots, \mathrm{p}$ με $\mathrm{r} = 1 - 4\log(\mathrm{n})/p$, $1 - 5\log(\mathrm{n})/p$ και $1 - 5\log(\mathrm{n})/p$. Για κάθε μοντέλο προσομοιώσαμε 200 σύνολα δεδομένων.

Εφαρμόσαμε τις ίδιες έξι μεθόδους, όπως αυτές στην προσομοίωση Ι, για να εκτιμήσουμε τα αραιά p-διανύσματα β. Για τις SIS-SCAD και SIS-DS μεθόδους, επιλέξαμε $d = \left[\frac{3}{2}n \Big/ \log(\mathrm{n})\right], \left[\frac{3}{2}n \Big/ \log(\mathrm{n})\right]$ και $\left[n/\log(\mathrm{n})\right]$, και για τις δύο τελευταίες μεθόδους, επιλέξαμε $d = n-1$ και $d' = \left[\frac{3}{2}n \Big/ \log(\mathrm{n})\right], \left[\frac{3}{2}n \Big/ \log(\mathrm{n})\right]$ και $\left[n/\log(\mathrm{n})\right]$. Τα αποτελέσματα προσομοίωσης συνοψίζονται στο Σχ. 2.4 (το οποίο βασίζεται σε 500 προσομοιώσεις) και στον Πίνακα 2.2. Παρόμοια συμπεράσματα όπως αυτά από την

προσομοίωση Ι μπορούν να εξαχθούν. Επίσης, όπως και στην προσομοίωση Ι, για λόγους απλότητας δεν συμπεριλάβαμε τον Gauss-Dantzig επιλογέα. Είναι ενδιαφέρον να παρατηρήσουμε ότι στην πρώτη ρύθμιση, η LASSO δίνει μεγάλα μοντέλα και τα σφάλματα εκτίμησής τους είναι αισθητά σε σύγκριση με τη νόρμα των πραγματικών συντελεστών β.



Σχήμα 2.4. Κατανομή του ελάχιστου αριθμού επιλεγμένων μεταβλητών που απαιτούνται να συμπεριληφθούν στο πραγματικό μοντέλο με τη χρήση της SIS όταν (a) $n = 200$ και p=1000 και s=5 **(b)** $n = 200$ και p=1000 και s=8 **(c)** $n = 800$ **και** p=20000 **και** s=8 **στην προσομοίωση ΙΙ.**

p	*Results for the following methods:*					
	Dantzig selector	*Lasso*	*SIS–SCAD*	*SIS–DS*	*SIS–DS–SCAD*	*SIS–DS–AdaLasso*
1000	10^3	91	21	56	27	52
($s=5$)	(1.256)	(1.257)	(0.331)	(0.727)	(0.476)	(1.204)
($s=8$)	10^3	74	18	56	31.5	51
	(1.465)	(1.257)	(0.458)	(1.014)	(0.787)	(1.824)
20000	—	—	36	119	54	86
	—	—	(0.367)	(0.986)	(0.743)	(1.762)

Πίνακας 2.2. Αποτελέσματα προσομοίωσης ΙΙ: οι διάμεσοι των επιλεγμένων μεγεθών μοντέλου και εκτιμητές σφάλματος (μέσα στην παρένθεση).

2.7. Μερικές επεκτάσεις της μάθησης συσχέτισης

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

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

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

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

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

2.8. Επαναληπτική SIS – Επαναληπτική μάθηση συσχέτισης

Θα δείξουμε ότι όταν οι υποθέσεις των μοντέλων ικανοποιούνται, οι οποίες αποκλείουν τα τρία προαναφερθέντα προβλήματα, η SIS μπορεί με ακρίβεια να μειώσει τη διάσταση από υψηλή σε μια μέτριας κλίμακας, κάτω από το μέγεθος του δείγματος. Αλλά όταν οι υποθέσεις αποτυγχάνουν, θα μπορούσε η SIS να χάσει κάποιες από τις σημαντικές επεξηγηματικές μεταβλητές. Για να ξεπεράσουμε αυτό

το πρόβλημα προτείνουμε την ISIS, ούτως ώστε να ενισχύσουμε την ισχύ της μεθοδολογίας. Η ISIS είναι μια επαναληπτική εφαρμογή της προσέγγισης SIS στην επιλογή μεταβλητών. Η ουσία είναι να εφαρμοστεί επαναληπτικά ένα κρησάρισμα μεταβλητών μεγάλης κλίμακας, ακολουθούμενο από μιας μέτριας κλίμακας προσεκτική επιλογή μεταβλητών.

Η ISIS λειτουργεί ως εξής: Αρχικά, επιλέγουμε ένα υποσύνολο k_1 μεταβλητών $A_1 = \left\{ X_{i_1}, ..., X_{i_{k_1}} \right\}$ χρησιμοποιώντας μια μέθοδο επιλογής μοντέλου βασισμένο στη SIS, όπως είναι οι μέθοδοι SIS-SCAD ή SIS-LASSO. Αυτές οι μεταβλητές επιλέχθηκαν, χρησιμοποιώντας τη SCAD ή τη LASSO, επί τη βάσει των κοινών πληροφοριών των $\left[n/\log(n) \right]$ μεταβλητών, οι οποίες επιβίωσαν μετά το κρησάρισμα συσχέτισης. Κατόπιν, έχουμε ένα n-διάνυσμα υπολοίπων από την παλινδρόμηση της απόκρισης Y πάνω στα $X_{i_1}, ..., X_{i_{k_1}}$. Στο επόμενο βήμα, μεταχειριζόμαστε αυτά τα υπόλοιπα ως τις νέες αποκρίσεις και εφαρμόζουμε την ίδια μέθοδο όπως στο προηγούμενο βήμα στις εναπομείνασες $p - k_1$ μεταβλητές, το οποίο καταλήγει σε ένα υποσύνολο k_2 μεταβλητών $A_2 = \left\{ X_{j_1}, ..., X_{j_{k_2}} \right\}$. Σημειώνουμε ότι η προσαρμογή των υπολοίπων από το προηγούμενο βήμα στο $\left\{ X_1, ..., X_p \right\} / A_1$ μπορεί να εξασθενίσει σημαντικά την προτεραιότητα αυτών των μη σημαντικών μεταβλητών, που είναι αρκετά συσχετισμένες με την απόκριση μέσω των σχέσεων τους με τα $X_{i_1}, ..., X_{i_{k_1}}$, δεδομένου ότι τα υπόλοιπα είναι ασυσχέτιστα με αυτές τις επιλεγμένες μεταβλητές στο A_1. Αυτό βοηθάει στο να λυθεί το πρώτο ζήτημα. Επίσης, καθιστά δυνατό αυτές οι σημαντικές επεξηγηματικές μεταβλητές που έχουν χαθεί στο προηγούμενο βήμα να επιβιώσουν. Αυτό διευθύνει το δεύτερο ζήτημα. Στην πραγματικότητα, μόλις οι μεταβλητές στο A_1 εισαχθούν στο μοντέλο, εκείνες που είναι οριακά ασθενώς συσχετισμένες με την Y, λόγω της παρουσίας των μεταβλητών στο A_1 θα πρέπει να συσχετιστούν με τα υπόλοιπα. Μπορούμε να συνεχίσουμε να το κάνουμε αυτό μέχρι να αποκτήσουμε l υποσύνολα $A_1, ..., A_l$ των οποίων η ένωση $A = \bigcup_{i=1}^{l} A_i$ έχει μέγεθος d, το οποίο είναι μικρότερο του n. Με πρακτική εφαρμογή, μπορούμε να επιλέξουμε, για παράδειγμα, το μεγαλύτερο l

υποσύνολο, για το οποίο ισχύει $|A| < n$. Από τα επιλεγμένα χαρακτηριστικά στο A, μπορούμε να επιλέξουμε τα χαρακτηριστικά χρησιμοποιώντας μιας μέτριας κλίμακας μέθοδο, όπως είναι η SCAD, η LASSO ή ο επιλογέας Dantzig.

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

2.9. Ομαδοποίηση και μετασχηματισμός των μεταβλητών εισόδου.

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

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

είναι ο *subject-related* μετασχηματισμός και ο άλλος είναι ο *στατιστικός* μετασχηματισμός.

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

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

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

μοντέλων. Ο Ravikumar et al. (2007) εισήγαγε αραιά προσθετικά μοντέλα για την αντιμετώπιση της επιλογής μη γραμμικών χαρακτηριστικών.

2.10. Αριθμητικά στοιχεία

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

2.10.1 Προσομοιωμένο παράδειγμα

Για το προσομοιωμένο παράδειγμα, χρησιμοποιήσαμε ένα γραμμικό μοντέλο

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

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

Για κάθε μοντέλο, εφαρμόσαμε τη SIS και την ISIS για να επιλέξουμε n μεταβλητές και δοκιμάσαμε την ακρίβειά τους στο να περιλαμβάνουν το πραγματικό

μοντέλο $\{X_1, X_2, X_3\}$. Για την ISIS, η SIS-SCAD μέθοδος με $d = [n/\log(n)]$ χρησιμοποιήθηκε σε κάθε βήμα και συνεχίσαμε να συλλέγουμε μεταβλητές σε αυτές τις ασυνεχείς A_j μέχρι να έχουμε n μεταβλητές (αν υπήρχαν περισσότερες μεταβλητές από αυτές που χρειάζονται στο τελικό στάδιο, θα περιλαμβάναμε μόνο εκείνες με τους μεγαλύτερους απόλυτους συντελεστές). Στον Πίνακα 2.3, αναφέρουμε τα ποσοστά της SIS, LASSO και ISIS που περιλαμβάνει το πραγματικό μοντέλο. Όλες αυτές οι τρεις μέθοδοι επιλέγουν $n-1$ μεταβλητές, για να κάνουν δίκαιες συγκρίσεις. Είναι σαφές ότι η συγγραμμικότητα (μεγάλη τιμή του ρ) και η υψηλή διάσταση επιδεινώνει την απόδοση της SIS και της LASSO, και η LASSO υπερτερεί της SIS κατά κάποιο τρόπο. Ωστόσο, όταν το μέγεθος του δείγματος είναι 50 ή μεγαλύτερο, η διαφορά στην απόδοση είναι πολύ μικρή, αλλά η SIS έχει πολύ λιγότερο υπολογιστικό κόστος. Σε αντίθεση, η ISIS βελτιώνει δραματικά την απόδοση αυτής της απλής SIS και LASSO. Πράγματι, σε αυτή την προσομοίωση, η ISIS επιλέγει πάντα όλες τις πραγματικές μεταβλητές. Μπορεί ακόμη και να έχει λιγότερο υπολογιστικό κόστος από ότι η LASSO, όταν η LASSO χρησιμοποιείται κατά την εφαρμογή της ISIS.

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

Πίνακας 2.3. Αποτελέσματα του παραδείγματος προσομοίωσης: ακρίβεια της SIS, της LASSO και της ISIS που περιλαμβάνονται στο πραγματικό μοντέλο $\{X_1, X_2, X_3\}$.

2.10.2. Προσομοιωμένο Παράδειγμα ΙΙ

Για το δεύτερο προσομοιωμένο παράδειγμα, χρησιμοποιήσαμε την ίδια διάταξη όπως στο παράδειγμα Ι εκτός από το ότι το ρ καθορίζεται να είναι 0,5 για λόγους ευκολίας και απλότητας. Επιπλέον, έχουμε προσθέσει μια τέταρτη μεταβλητή X_4 στο μοντέλο και το γραμμικό μοντέλο είναι πλέον

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + \varepsilon$$

όπου η $X_4 \sim N(0,1)$ και έχει συσχέτιση $\rho^{1/2}$ με όλες τις άλλες p-1 μεταβλητές. Ο τρόπος, με τον οποίο η X_4 έχει εισαχθεί, είναι για να το καταστήσει ασυσχέτιστο με την απόκριση Y. Ως εκ τούτου, η SIS δεν μπορεί να πάρει το πραγματικό μοντέλο εκτός από τύχη.

Και πάλι προσομοιώσαμε 200 σύνολα δεδομένων για κάθε μοντέλο. Στον Πίνακα 2.4, αναφέρουμε τα ποσοστά της SIS, της LASSO και της ISIS που περιλαμβάνουν το πραγματικό μοντέλο των τεσσάρων μεταβλητών. Σε αυτό το παράδειγμα προσομοίωσης, η SIS συμπεριφέρεται κατά κάποιο τρόπο καλύτερα από τη LASSO στο κρησάρισμα μεταβλητών, και η ISIS υπερτερεί σημαντικά της απλής SIS και LASSO. Σε αυτή την προσομοίωση πάντα παίρνει όλες τις πραγματικές μεταβλητές. Αυτό αποδεικνύει ότι η ISIS μπορεί να χειριστεί αποτελεσματικά το δεύτερο πρόβλημα που αναφέρθηκε παραπάνω.

p	*Method*	*Results for the following values of n:*		
		$n=20$	$n=50$	$n=70$
100	SIS	0.025	0.490	0.740
	Lasso	0.000	0.360	0.915
	ISIS	1	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	1	1	1

†$\rho = 0.5$.

Πίνακας 2.4. Αποτελέσματα του παραδείγματος προσομοίωσης ΙΙ: ακρίβεια της SIS, της LASSO και της ISIS που περιλαμβάνονται στο πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4\}$.

2.10.3. Παράδειγμα Προσομοίωσης ΙΙΙ

Για το τρίτο παράδειγμα προσομοίωσης, χρησιμοποιήσαμε την ίδια διάταξη όπως και στο παράδειγμα ΙΙ εκτός από το ότι προσθέσαμε στο μοντέλο μια πέμπτη μεταβλητή X_5 και το γραμμικό μοντέλο τώρα γίνεται

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + X_5 + \varepsilon$$

όπου $X_5 \sim N(0,1)$ και δεν σχετίζεται με όλες τις άλλες p-1 μεταβλητές. Και πάλι η X_4 είναι ασυσχέτιστη με την απόκριση Y. Ο τρόπος με τον οποίο η X_5 έχει εισαχθεί, ήταν για να καταστήσει να έχει μια πολύ μικρή συσχέτιση με την απόκριση και στην πραγματικότητα η μεταβλητή X_5 έχει την ίδια αναλογία συνεισφοράς στην απόκριση όπως και ο θόρυβος ε. Για αυτό το συγκεκριμένο παράδειγμα, η X_5 έχει ασθενή οριακή συσχέτιση με την Y από ότι οι $X_6,...,X_p$ και ως εκ τούτου έχει λιγότερη προτεραιότητα στο να επιλεχθεί από τη SIS.

Για κάθε μοντέλο προσομοιώσαμε 200 σύνολα δεδομένων. Στον Πίνακα 2.5, αναφέρουμε την ακρίβεια σε ποσοστό της SIS, της LASSO και της ISIS στο να περιλαμβάνουν το πραγματικό μοντέλο. Είναι σαφές ότι η ISIS μπορεί να βελτιώσει σημαντικά σε σχέση με την απλή SIS και τη LASSO και πάντα παίρνει όλες τις πραγματικές μεταβλητές. Αυτό δείχνει και πάλι ότι η ISIS μπορεί να επιλέξει δύο διαφορετικές μεταβλητές X_4 και X_5, η οποία αντιμετωπίζει ταυτόχρονα το δεύτερο και τρίτο πρόβλημα που αναφέραμε παραπάνω.

p	*Method*	*Results for the following values of n:*		
		n = 20	*n* = 50	*n* = 70
100	SIS	0.000	0.285	0.645
	Lasso	0.000	0.310	0.890
	ISIS	1	1	1
1000	SIS	0.000	0.000	0.000
	Lasso	0.000	0.000	0.000
	ISIS	1	1	1

†$\rho = 0.5$.

<u>Πίνακας 2.5</u>. Αποτελέσματα παραδείγματος προσομοίωσης ΙΙΙ: ακρίβεια της SIS, της LASSO και της ISIS που περιλαμβάνονται στο πραγματικό μοντέλο $\{X_1, X_2, X_3, X_4, X_5\}$.

p	*Results for simulation I*		*Results for simulation II*
1000	13 (0.329)	(*s* = 5)	11 (0.223)
		(*s* = 8)	13.5 (0.366)
20000	31 (0.246)		27 (0.315)

<u>Πίνακας 2.6</u>. Προσομοιώσεις Ι και ΙΙ: οι διάμεσοι των μεγεθών του μοντέλου που επιλέχθηκαν και τα σφάλματα εκτίμησης (μέσα στην παρένθεση) για τη ISIS-SCAD μέθοδο.

2.10.4. Προσομοιώσεις I και II της προηγούμενης ενότητας

Για καθένα από τα δύο προηγούμενα παραδείγματα, χρησιμοποιήσαμε την τεχνική της ISIS με SCAD και $d = \lfloor n/\log(n) \rfloor$ για να επιλέξουμε $q = \lfloor n/\log(n) \rfloor$ μεταβλητές. Μετά από αυτό, εκτιμήσαμε το q-διάνυσμα β με τη χρήση της SCAD. Αυτή η μέθοδος αναφέρεται ως ISIS-SCAD. Αναφέρουμε στον Πίνακα 2.6 τη διάμεσο των επιλεγμένων μεγεθών μοντέλου και τη διάμεσο των σφαλμάτων εκτίμησης $\left\| \hat{\beta} - \beta \right\|$ στη l_2-νόρμα. Μπορούμε εύκολα να διαπιστώσουμε ότι η ISIS βελτιώνεται σε σχέση με την απλή SIS. Οι βελτιώσεις είναι πιο δραστικές για την προσομοίωση II στην οποία οι συμμεταβλητές είναι πιο συσχετισμένες και οι επιλογές μεταβλητών δημιουργούν περισσότερες προκλήσεις.

2.10.5. Συμπερασματικά σχόλια.

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

ΚΕΦΑΛΑΙΟ 3^{ο}

ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΗΣ ΔΙΑΚΙΝΔΥΝΕΥΣΗΣ ΤΟΥ COX

3.1 Εισαγωγικά στοιχεία

Η ανάλυση δεδομένων διάρκειας ζωής είναι ο κλάδος της Στατιστικής που ασχολείται με δεδομένα, τα οποία αντιπροσωπεύουν χρόνο μέχρι την εμφάνιση ενός γεγονότος. Όταν ασχολούμαστε με τεχνικά συστήματα είναι γνωστή ως θεωρία αξιοπιστίας (*Reliability Theory*), ενώ όσον αφορά τις βιοϊατρικές εφαρμογές ως ανάλυση επιβίωσης (*Survival Analysis*).

Το χαρακτηριστικό γνώρισμα των χρόνων επιβίωσης (*Survival Analysis*) είναι ότι δεν ακολουθούν σχεδόν ποτέ την κανονική κατανομή. Αυτή είναι μια από τις αιτίες που χρησιμοποιούνται διαφορετικές μέθοδοι στατιστικής ανάλυσης από τις συνηθισμένες. Γενικά, η ανάλυση επιβίωσης εστιάζεται στην εκτίμηση της πιθανότητας επιβίωσης ενός ατόμου για ένα δεδομένο χρονικό διάστημα.

Ένας επιπλέον λόγος, εξαιτίας του οποίου, τα δεδομένα επιβίωσης δεν αναλύονται με την χρήση συνηθισμένων στατιστικών τεχνικών, είναι ότι οι χρόνοι επιβίωσης ορισμένων παρατηρήσεων είναι αποκομμένοι.

3.1.1. Συνάρτηση κατανομής

Η συνάρτηση κατανομής (*distribution function*), *F*(t), ορίζεται ως η πιθανότητα να έχουμε αποτυχία πριν το χρόνο *t*, δηλαδή

$$F(t) = P(T \le t) = \int_0^t f(u)du .$$

Εξ' ορισμού, η F(t) είναι αύξουσα, με $\lim_{t\to 0} F(t) = 0$ και $\lim_{t\to\infty} F(t) = 1$.

3.1.2. Συνάρτηση πυκνότητας πιθανότητας (*probability density function ή density function)*

Η συνάρτηση πυκνότητας πιθανότητας (*probability density function*) $f(t)$ του χρόνου επιβίωσης T, εκτιμάται ως το όριο της πιθανότητας ένα άτομο να αποτύχει σε ένα μικρό χρονικό διάστημα (t, t+Δt) ανά μονάδα πλάτους *Δt*, δηλαδή:

$$f(\mathrm{t}) = \lim_{dt \to 0} \frac{P(\mathrm{t} < \mathrm{T} < \mathrm{t} + \mathrm{dt})}{dt}$$

Η καμπύλη της *f(t)* ονομάζεται *καμπύλη πυκνότητας (density curve).* Για τη συνάρτηση πυκνότητας πιθανότητας του χρόνου *T* ισχύουν οι εξής ιδιότητες:

$f(t) \geq 0$ για $t \geq 0$ και $f(t) = 0$ για $t < 0$

Το εμβαδόν μεταξύ της καμπύλης πιθανότητας και του άξονα του t ισούται με 1.

Κατόπιν υπολογισμών προκύπτει ότι:

$$f(u) = \frac{d}{dt} F(t) = -\frac{d}{dt} S(t),$$

όπου S(t) είναι η συνάρτηση επιβίωσης (*survival function*).

3.1.3. Συνάρτηση επιβίωσης

Η συνάρτηση επιβίωσης (*survival function*), *S*(t), ορίζεται ως η πιθανότητα ένα άτομο να επιβιώσει για χρόνο μεγαλύτερο από t.

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du.$$

Η S(t) είναι μια φθίνουσα συνάρτηση του t με τις εξής ιδιότητες:

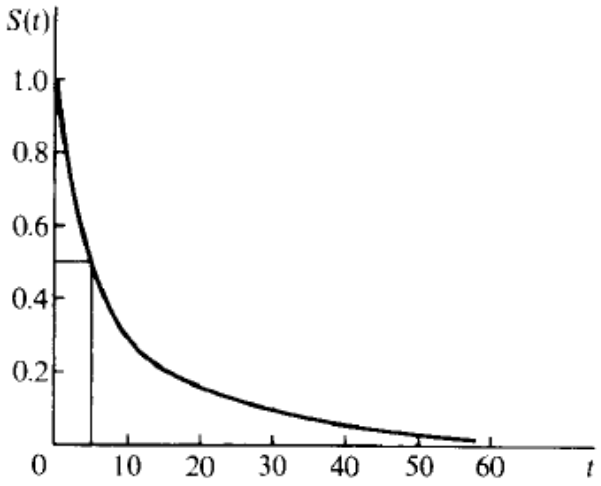
S(t)=1 για t=0

S(t)=0 για $t = \infty$

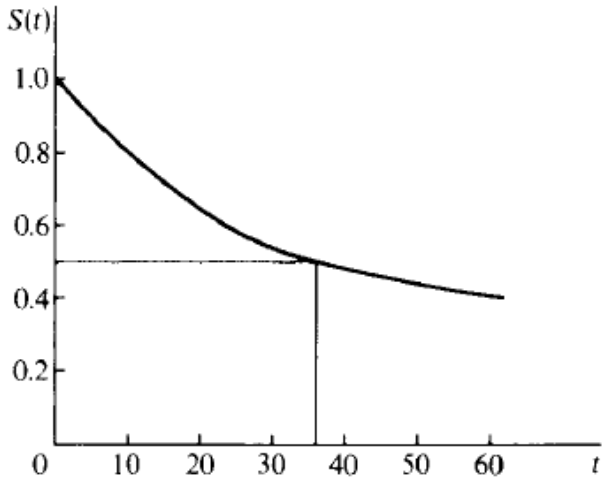
Δηλαδή, η πιθανότητα το άτομο να επιβιώσει τουλάχιστον σε χρόνο 0 είναι 1 και η πιθανότητα επιβίωσης σε άπειρο χρόνο είναι 0.

Η γραφική παράσταση της S(t) συναρτήσει του *t* ονομάζεται *καμπύλη επιβίωσης (survival curve).*

Μια απότομα φθίνουσα καμπύλη, όπως αυτή του παρακάτω γραφήματος, υποδεικνύει χαμηλό ποσοστό επιβίωσης ή μικρή διάρκεια επιβίωσης.



Αντιθέτως, μια βαθμιαία φθίνουσα καμπύλη, όπως της παρακάτω γραφικής παράστασης, υποδεικνύει υψηλό ποσοστό ή μεγαλύτερη διάρκεια επιβίωσης.



Η συνάρτηση επιβίωσης ή η καμπύλη επιβίωσης χρησιμοποιείται για την εύρεση της διαμέσου του χρόνου επιβίωσης (*median survival time*). Είναι ο χρόνος στον οποίο το 50% των υπό μελέτη ασθενών επιβιώνει.

3.1.4. Συνάρτηση διακινδύνευσης

Η συνάρτηση διακινδύνευσης ή συνάρτηση βαθμού κινδύνου (*hazard function*), $h(t)$, ορίζεται ως:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{[S(t) - S(t+\delta t)] / S(t)}{\delta t} \right] = \frac{f(t)}{S(t)}.$$

Εκφράζει την πιθανότητα ένα άτομο ηλικίας *t* να βιώσει το γεγονός στην αμέσως επόμενη χρονική στιγμή.

Η συνάρτηση διακινδύνευσης είναι δυνατό να αυξάνει, να φθίνει, να μένει σταθερή ή να δηλώνει μια πιο σύνθετη διαδικασία.

3.1.5. Σωρευτική συνάρτηση διακινδύνευσης

Η σωρευτική συνάρτηση διακινδύνευσης (*cumulative hazard function*), συμβολίζεται με H(t) κι ορίζεται ως

$$H(t) = \int_0^t h(u)du.$$

Από τους παραπάνω ορισμούς προκύπτει ότι:

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{f(u)}{S(u)} du = \int_0^t \frac{-S'(u)}{S(u)} du = \left[-\ln(S(u)) \right]_0^t = -\ln S(t),$$

άρα

$$S(t) = \exp\{-H(t)\}.$$

Από τις παραπάνω σχέσεις φαίνεται ότι οι συναρτήσεις $S(t)$, $h(t)$, $f(t)$, $F(t)$ και $H(t)$ είναι μαθηματικά ισοδύναμες, διότι γνωρίζοντας μία από αυτές, είναι δυνατό να υπολογίσουμε και τις υπόλοιπες.

3.2. Το μοντέλο αναλογικής διακινδύνευσης του Cox

Το μοντέλο αναλογικής διακινδύνευσης του Cox, είναι ένα μοντέλο παλινδρόμησης που μοντελοποιεί τη συνάρτηση διακινδύνευσης $h(t)$ και επιτρέπει τη σύγκριση των επεξηγηματικών μεταβλητών, ώστε να γίνει δυνατή η επιλογή των στατιστικά σημαντικότερων εξ' αυτών. Το μοντέλο αυτό έχει πολλές εφαρμογές κυρίως στη βιοϊατρική επιστήμη, καθώς και σε προβλήματα πιστωτικού κινδύνου τραπεζών.

3.2.1. Περιγραφή του μοντέλου του Cox

Έστω $\underset{\sim}{x} = (\underset{\sim}{x}_1, \underset{\sim}{x}_2 ..., \underset{\sim}{x}_p)$ είναι το διάνυσμα των μεταβλητών μιας μελέτης, η οποία αποτελείται από n άτομα και $\underset{\sim}{x}_i = (x_{i1}, x_{i2} ..., x_{ip})$, $i = 1, 2, ..., n$, είναι το διάνυσμα με τις τιμές των συμμεταβλητών που αντιστοιχεί στο i άτομο. Αν οι τιμές των συμμεταβλητών παραμένουν σταθερές με την πάροδο του χρόνου (δηλαδή είναι ανεξάρτητες από τον χρόνο), τότε το μοντέλο αναλογικής διακινδύνευσης του Cox δίνεται από τη σχέση:

$$h(t, \underset{\sim}{x}) = h_0(t) e^{\underset{\sim}{x}' \underset{\sim}{\beta}} \tag{3.1}$$

όπου $h(t, \underset{\sim}{x})$ η συνάρτηση διακινδύνευσης στο χρόνο t, η $h_0(t)$ είναι γνωστή ως βασική συνάρτηση διακινδύνευσης (*baseline hazard function*) στο χρόνο t και εκφράζει τον κίνδυνο θανάτου ή αποτυχίας, όταν όλες οι επεξηγηματικές μεταβλητές x_j *είναι ίσες με μηδέν*, για *j=1,2,….,k* και το $\underset{\sim}{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ είναι το διάνυσμα των συντελεστών παλινδρόμησης, οι οποίοι εκφράζουν ποσοτικά την επίδραση κάθε μεταβλητής και πρέπει να εκτιμηθούν.

Θεωρούμε τώρα το λόγο:

$$\frac{h(t, \underset{\sim}{x}_i)}{h(t, \underset{\sim}{x}_{i+1})} = \frac{h_0(t) e^{\underset{\sim}{x}_i' \underset{\sim}{\beta}}}{h_0(t) e^{\underset{\sim}{x}_{i+1}' \underset{\sim}{\beta}}} = e^{(\underset{\sim}{x}_i - \underset{\sim}{x}_{i+1})' \underset{\sim}{\beta}}$$

που είναι γνωστός ως αναλογία κινδύνου *(hazard ratio).*

Το διάνυσμα x_i αντιστοιχεί στο διάνυσμα των *k* επεξηγηματικών μεταβλητών για το *i*-οστό άτομο που συμμετέχει στο πείραμα.

Από την παραπάνω σχέση παρατηρούμε ότι η τιμή των επεξηγηματικών μεταβλητών δεν εξαρτάται από τον χρόνο *t* κι ότι η ποσότητα $e^{(\underset{\sim}{x}_i' - \underset{\sim}{x}_{i+1})\underset{\sim}{\beta}}$ είναι σταθερή στο χρόνο. Εξαιτίας αυτού του γεγονότος το μοντέλο του Cox είναι γνωστό ως *μοντέλο αναλογικής διακινδύνευσης*.

Ισοδύναμη έκφραση του μοντέλου είναι η εξής: $\ln \frac{h(\mathrm{t}, \underset{\sim}{\mathrm{x}})}{h_0(\mathrm{t})} = \beta_1 x_1 + ... + \beta_k x_k$.

Η τελευταία μορφή απλοποιεί σημαντικά την ερμηνεία των συντελεστών.

Συγκεκριμένα, το β_i ισοδυναμεί με το λογάριθμο του σχετικού κινδύνου όταν έχουμε αύξηση μιας μονάδας στη μεταβλητή x_i, ενώ οι άλλες παραμένουν σταθερές. Με άλλα λόγια, η σχέση e^{β_i} εκφράζει το σχετικό ρίσκο που λαμβάνεται όταν η επεξηγηματική μεταβλητή x_i αυξάνεται κατά μια μονάδα, ενώ οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Γενικά, η συνάρτηση διακινδύνευσης $h(\mathrm{t}, \underset{\sim}{\mathrm{x}})$ εξαρτάται από δύο παράγοντες: τη συνάρτηση $h_0(\mathrm{t})$ και το διάνυσμα $\underset{\sim}{\beta} = (\beta_1, ..., \beta_k)$. Η συνάρτηση $h_0(\mathrm{t})$, η οποία εξαρτάται μόνο από το χρόνο, στο μοντέλο του Cox αφήνεται αυθαίρετη και θεωρείται ίδια για το σύνολο των *n* ατόμων της μελέτης. Για το λόγο αυτό, το μοντέλο αναλογικού κινδύνου του Cox θεωρείται ημι-παραμετρικό (*semi-parametric*), αφού δεν καθορίζει τη μορφή της $h_0(\mathrm{t})$, αλλά υποθέτει ότι οι επιδράσεις των μεταβλητών παραμένουν σταθερές στο χρόνο και είναι προσθετικές σε μια συγκεκριμένη κλίμακα.

Έτσι γνωρίζουμε ότι $H(t, \underset{\sim}{x}) = \int_0^t h_0(u) e^{\underset{\sim}{x}'\underset{\sim}{\beta}} du = H_0(t) e^{\underset{\sim}{x}'\underset{\sim}{\beta}}$.

Με τη βοήθεια της σχέσης: $S(t) = \exp\{-H(t)\}$ προκύπτει ότι $S(t) = \exp\left\{-H_0(t)e^{\underset{\sim}{x}'\underset{\sim}{\beta}}\right\} = \left[S_0(t)\right]^{e^{\underset{\sim}{x}'\underset{\sim}{\beta}}}$.

Συνεπώς προκύπτει ότι

$$S(t) = \left[S_0(t)\right]^{e^{\underset{\sim}{x}'\underset{\sim}{\beta}}}, \tag{3.2}$$

όπου $S_0(t) = \exp\left[-H_0(t)\right]$ είναι η βασική συνάρτηση επιβίωσης (*baseline survival function*).

Με τη βοήθεια της σχέσης (3.2), μπορεί να εκτιμηθεί η συνάρτηση επιβίωσης οποιουδήποτε ατόμου που συμμετέχει στη μελέτη.

3.2.2. Συνάρτηση μερικής πιθανοφάνειας

Ας θεωρήσουμε ότι μετά τη λήξη ενός πειράματος έχουμε *n* αριθμό παρατηρήσεων, εκ των οποίων οι *k* παρατηρήσεις είναι μη αποκομμένες, δηλαδή είναι πλήρεις. Άρα οι *n-k* είναι αποκομμένες. Θεωρούμε τους αντίστοιχους χρόνους αποτυχίας $t_{(1)}, t_{(2)}, ..., t_{(k)}$. Τότε αν R_j είναι το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή t_j, η συνάρτηση πιθανοφάνειας (*partial likelihood*) ορίζεται ως εξής:

$$L(\underset{\sim}{\beta}) = \prod_{j=1}^{k} \frac{e^{\underset{\sim}{x}_j'\underset{\sim}{\beta}}}{\sum_{i \in R_j} e^{\underset{\sim}{x}_j'\underset{\sim}{\beta}}}$$

Το γεγονός ότι στην παραπάνω διαδικασία δεν προσδιορίζεται η $h_0(t)$ εξηγεί τον όρο «μερική πιθανοφάνεια». Αξίζει να σημειωθεί ότι η μοναδική συμβολή των πειραματικών δεδομένων στην πιθανοφάνεια, είναι οι χρονικές στιγμές στις οποίες παρατηρούνται τα μη αποκομμένα δεδομένα.

Όπως απέδειξε ο Cox, η μερική πιθανοφάνεια μπορεί να χρησιμοποιηθεί ως μια συνηθισμένη συνάρτηση πιθανοφάνειας για τον υπολογισμό της εκτίμησης $\hat{\beta}$

των $\underset{\sim}{\beta}_i$. Όμως είναι αποτελεσματική στην περίπτωση που κάθε χρόνος επιβίωσης εμφανίζεται μόνο μια φορά στα πειραματικά μας δεδομένα.

Σε αντίθετη περίπτωση οι απαραίτητοι υπολογισμοί για την εκτίμηση των συντελεστών είναι αρκετά πολύπλοκοι. Τα στατιστικά πακέτα που χρησιμοποιούνται για την εκτίμηση αυτών, χρησιμοποιούν συχνά μια προσέγγιση που δόθηκε από μαθηματικό R. Peto το 1990, η οποία αποδεικνύεται υπό την προϋπόθεση ο αριθμός των ταυτόχρονων γεγονότων να είναι μικρός σε σύγκριση με τα άτομα που βρίσκονται σε κίνδυνο για κάθε χρονική στιγμή. Στην περίπτωση πολλαπλών ταυτόχρονων θανάτων αποδεικνύονται αποτελεσματικές οι προσεγγίσεις της μερικής πιθανοφάνειας των Breslow και Efron.

Συγκεκριμένα, έστω ότι έχουμε d_j αποτυχίες (π.χ. θανάτους) στο χρόνο $t_{(j)}$ και ότι D_j είναι το σύνολο των ατόμων που αποτυγχάνουν στο χρόνο $t_{(j)}$. Η προσέγγιση του Breslow είναι:

$$L(\underset{\sim}{\beta}) = \prod_{j=1}^{k} \left\{ \frac{\Psi_j}{[\sum_{i \in R_j} e^{\underset{\sim}{x}_i' \underset{\sim}{\beta}}]^{d_j}} \right\},$$

όπου $\Psi_j = \exp[(\sum_{i \in D_j} \underset{\sim}{x}_i)' \underset{\sim}{\beta}]$

Επιπλέον, η προσέγγιση του Efron είναι:

$$L(\underset{\sim}{\beta}) = \prod_{j=1}^{k} \left\{ \frac{\Psi_j}{\prod_{l=1}^{d_j} \left[\sum_{i \in R_j} e^{\underset{\sim}{x}_{(i)}' \underset{\sim}{\beta}} - \frac{l-1}{d_j} \right] \sum_{i \in D_j} e^{\underset{\sim}{x}_{(i)}' \underset{\sim}{\beta}}} \right\}$$

Μεγιστοποιώντας λοιπόν την $L(\underset{\sim}{\beta})$, θα προκύψει η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\beta}$ ως εξής: Αρχικά λογαριθμούμε τη σχέση $L(\underset{\sim}{\beta}) = \prod_{j=1}^{k} \frac{e^{x_j' \underset{\sim}{\beta}}}{\sum_{i \in R_j} e^{x_j' \underset{\sim}{\beta}}}$ και έχουμε: $l(\underset{\sim}{\beta}) = \sum_{j=1}^{k} \underset{\sim}{\beta}' \underset{\sim}{x}_{(j)} - \sum_{j=1}^{k} \ln \left\{ \sum_{i \in R_j} e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{(i)}} \right\}$.

Οι πρώτες μερικές παράγωγοι είναι

$$\frac{\partial l}{\partial \beta_r} = \sum_{j=1}^{k} x_{(j)r} - \sum_{j=1}^{k} \left[\sum_{i \in R_j} x_{(i)r} e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{(i)}} / \sum_{i \in R_j} e^{\underset{\sim}{\beta}' \underset{\sim}{x}_{(i)}} \right] , \qquad r = 1,2,\ldots,k$$

Η παραπάνω σχέση αποτελεί σύστημα *p*-εξισώσεων, το οποίο λύνεται με επαναληπτικές μεθόδους, όπως είναι η μέθοδος Newton-Raphson.

3.3. Μοντέλο διακινδύνευσης του Cox στην περίπτωση χρονοεξαρτημένων επεξηγηματικών μεταβλητών.

Στο μοντέλο αναλογικής διακινδύνευσης του Cox οι μεταβλητές θεωρούνται σταθερές ως προς τον χρόνο. Υπάρχει, όμως, περίπτωση αυτές να μεταβάλλονται με την πάροδο του χρόνου, με αποτέλεσμα να έχουμε δύο κατηγορίες χρονοεξαρτημένων μεταβλητών (*time dependent*). Αυτές είναι οι εσωτερικές (*internal*) και οι εξωτερικές (*external*).

Ας θεωρήσουμε ένα πρόβλημα επιβίωσης που αφορά την εξέλιξη των ασθενών με όγκο στο ήπαρ. Στην κατηγορία των εσωτερικών μεταβλητών ανήκουν εκείνες, των οποίων οι τιμές μπορούν να ληφθούν μόνο όσο ο ασθενής βρίσκεται εν ζωή, ενώ στην κατηγορία των εξωτερικών ανήκουν οι μεταβλητές, οι οποίες μπορούν να πάρουν τιμή ανεξάρτητα αν ο ασθενής είναι ζωντανός ή όχι. Για παράδειγμα, το μέγεθος του όγκου του ασθενούς κι ο αριθμός των ερυθρών του αιμοσφαιρίων αποτελούν εσωτερικές χρονοεξαρτώμενες μεταβλητές, ενώ η θερμοκρασία του αέρα και η συγκέντρωση διοξειδίου του άνθρακα (CO_2) είναι εξωτερικές χρονοεξαρτώμενες μεταβλητές.

3.3.1. Συνάρτηση διακινδύνευσης για το μοντέλο του Cox με χρονοεξαρτώμενες μεταβλητές

Έστω ότι υπάρχουν *n* άτομα σε μια μελέτη κι ότι το σύνολο των συμμεταβλητών (*covariates*), δηλαδή των επεξηγηματικών μεταβλητών που εισάγονται στο πρόβλημα, είναι *p*.

Τότε η συνάρτηση διακινδύνευσης για το *i*-οστό άτομο που βρίσκεται υπό μελέτη δίνεται από τη σχέση:

$$h_i(\mathrm{t}) = \mathrm{h}_o(\mathrm{t}) \cdot \mathrm{e}^{\sum_{j=1}^{p} \beta_j x_{ji}},$$

όπου x_{ji} είναι η τιμή της j-οστής επεξηγηματικής μεταβλητής για το *i*-οστό άτομο.

Γενικεύοντας το παραπάνω μοντέλο για την περίπτωση παρουσίας χρονοεξαρτώμενων μεταβλητών έχουμε τον ακόλουθο μετασχηματισμό:

$$h_i(\mathrm{t}) = \mathrm{h}_o(\mathrm{t}) \cdot \mathrm{e}^{\sum_{j=1}^{p} \beta_j x_{ji}(\mathrm{t})},$$

όπου $i = 1, 2, \ldots, n$ και $j = 1, 2, \ldots, \mathrm{p}$ και το $x_{ji}(\mathrm{t})$ είναι χρονοεξαρτώμενο.

3.3.2. Εκτιμήτρια μερικής πιθανοφάνειας

Ο λογάριθμος μερικής πιθανοφάνειας στην περίπτωση χρονοεξαρτώμενων μεταβλητών δίνεται από τη σχέση :

$$\ln \mathrm{L}(\underset{\sim}{\beta}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} x_{ji}(\mathrm{t}_i) \beta_j - \log\left(\sum_{k \in R(\mathrm{t}_i)} e^{\beta_j x_{jk}(\mathrm{t}_i)} \right) \right),$$

όπου t_i είναι η χρονική στιγμή θανάτου του i-οστού ατόμου, $R(t_i)$ είναι το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή *t*, δ_i είναι η χαρακτηριστική συνάρτηση αποκοπής για την οποία αν *C* είναι το σύνολο των

αποκομμένων παρατηρήσεων και *U* είναι το σύνολο των μη αποκομμένων παρατηρήσεων έχουμε:

$$\delta_i = \begin{cases} 0, i \in C \\ 1, \mathrm{i} \in U \end{cases}$$

Είναι προφανές ότι στην περίπτωση των χρονοεξαρτημένων μεταβλητών και η σχετική διακινδύνευση $\frac{h_i(\mathrm{t})}{h_0(\mathrm{t})}$ εξαρτάται από τον χρόνο. Συνεπώς, στην περίπτωση αυτή δεν ισχύει η υπόθεση της αναλογικής διακινδύνευσης.

3.4. Το στρωματοποιημένο μοντέλο του Cox

Στην περίπτωση που σε μια στατιστική μελέτη έχουμε μια ή περισσότερες κατηγορικές μεταβλητές, υπάρχει περίπτωση να μην ισχύει η υπόθεση της αναλογικότητας και γι' αυτό το λόγο πρέπει να ελεγχθεί.

Έστω ότι σε ένα πείραμα μία από τις ανεξάρτητες μεταβλητές είναι το φύλο του ασθενούς. Η συνάρτηση διακινδύνευσης μπορεί να διαμορφώνεται διαφορετικά για τις δύο τιμές της μεταβλητής, έτσι ώστε να μην ισχύει η υπόθεση της αναλογικότητας. Ο τρόπος αντιμετώπισης αυτού του φαινομένου είναι η λεγόμενη στρωματοποίηση (*stratification*) ως προς τη συγκεκριμένη κατηγορική μεταβλητή. Στην περίπτωση αυτή, θεωρούμε ότι τα άτομα που ανήκουν σε διαφορετικό επίπεδο (στρώμα) της μεταβλητής έχουν διαφορετική συνάρτηση διακινδύνευσης κι ότι όλες οι υπόλοιπες επεξηγηματικές μεταβλητές ικανοποιούν την υπόθεση της αναλογικότητας των κινδύνων σε κάθε στρώμα (*stratum*). Έστω ότι το πλήθος των στρωμάτων είναι *s*, τότε για κάθε στρώμα λαμβάνουμε διαφορετική βασική συνάρτηση διακινδύνευσης $h_{0k}(\mathrm{t})$, όπου $\mathrm{k} = \{1, 2, ..., \mathrm{s}\}$.

Η συνάρτηση διακινδύνευσης του i-οστού ατόμου είναι: $h_{ik}(\mathrm{t}) = \mathrm{h}_{0k}(\mathrm{t}) \cdot \mathrm{e}^{\underset{\sim}{\beta} \chi_{ik}}$, όπου k είναι το στρώμα στο οποίο ανήκει το άτομο και x_{ik} είναι το διάνυσμα τιμών των επεξηγηματικών μεταβλητών για το άτομο αυτό.

Από την παραπάνω σχέση παρατηρούμε ότι οι συντελεστές $\underset{\sim}{\beta}$ είναι οι ίδιοι σε κάθε στρώμα. Η προσαρμογή του μοντέλου γίνεται με χρήση στατιστικών πακέτων.

3.5. Έλεγχοι καταλληλότητας του μοντέλου

Το μοντέλο του Cox αποτελεί μεν το γνωστότερο μοντέλο ανάλυσης δεδομένων επιβίωσης με συμμεταβλητές, αλλά πάντα υποθέτουμε ότι ισχύει η υπόθεση της αναλογικής διακινδύνευσης, διαφορετικά υπάρχει περίπτωση να προκύψουν λανθασμένα συμπεράσματα. Αυτό σημαίνει ότι είναι απαραίτητο να γίνεται έλεγχος καταλληλότητας του μοντέλου για κάθε συντελεστή. Ο έλεγχος πραγματοποιείται με βάση δύο τρόπους, οι οποίοι είναι οι εξής:

- Με χρήση της γραφικής μεθόδου
- Με χρήση ελέγχου υπολοίπων

3.5.1.Γραφικός έλεγχος

Ο γραφικός έλεγχος καταλληλότητας πραγματοποιείται ως εξής:

Λογαριθμούμε τη συνάρτηση επιβίωσης:

$$S(t,\underset{\sim}{x}) = \exp\{-H(t,\underset{\sim}{x})\} = \exp\{-H_0(t)\,e^{\underset{\sim}{x}\underset{\sim}{\beta}}\}$$

κι έχουμε :

$$\ln\{-\ln S(t,\underset{\sim}{x})\} - \ln H_0(t) = \underset{\sim}{x}\underset{\sim}{\beta}$$

Αυτό γραφικά σημαίνει ότι οι καμπύλες $\ln\{-\ln S(t,\underset{\sim}{x})\}$ και $\ln H_0(t)$ πρέπει να είναι παράλληλες για κάθε $\underset{\sim}{x}$ και $t>0$. Συνεπώς, όλες οι καμπύλες $\ln\{-\ln S(t,\underset{\sim}{x}_i)\}$ για

διάφορα $\underset{\sim}{x}_i$ θα πρέπει να είναι παράλληλες για να ισχύει η αναλογικότητα του κινδύνου.

Με βάση την παραπάνω διαπίστωση μπορούμε να αποφανθούμε γραφικά, σχετικά με την καταλληλόλητα του μοντέλου, πραγματοποιώντας έλεγχο για κάθε μεταβλητή που εμπεριέχεται στο αρχικό πρόβλημα. Για να είναι η μέθοδος αξιόπιστη πρέπει ο αριθμός των συμμεταβλητών (*covariates*) να είναι περιορισμένος. Όμως, το σημαντικότερο πλεονέκτημα της μεθόδου είναι ότι μπορεί να χρησιμοποιηθεί για οποιοδήποτε μοντέλο αναλογικής διακινδύνευσης κι όχι μόνο για το μοντέλο του Cox.

Στην περίπτωση που η μεταβλητή είναι ποσοτική, πρέπει πρώτα να γίνει κατηγοριοποίηση. Για να γίνει όμως η γραφική παράσταση της συνάρτησης $\ln\{-\ln S(t,\underset{\sim}{x})\}$ με το χρόνο, πρέπει πρώτα να προσδιορισθούν οι εκτιμήτριες της συνάρτησης επιβίωσης για επιλεγμένες τιμές των $\underset{\sim}{x}$. Ισχύει όμως ότι

$$\hat{S}(t,\underset{\sim}{x}) = \{\hat{S}_0(t)\}^{\exp(\underset{\sim}{x}'\underset{\sim}{\beta})} \text{ και } \hat{S}_0(t) = e^{-\hat{H}_0(t)}.$$

Άρα, απαιτείται η εκτίμηση της βασικής σωρρευτικής συνάρτησης διακινδύνευσης για κάθε διαφορετικό στρώμα-επίπεδο της μεταβλητής. Αυτό γίνεται κάνοντας χρήση της μη παραμετρικής εκτιμήτριας του Breslow

$$\hat{H}_0(t) = \sum_{t(j)\le t}\left(\frac{d_j}{\sum_{i\in R(t_j)} e^{\underset{\sim}{\beta}'\underset{\sim}{x}_i}}\right)$$

με $R(t_j)$ το σύνολο των ατόμων σε κίνδυνο το χρόνο t_j.

3.5.2. Έλεγχος καταλληλότητας μοντέλου μέσω υπολοίπων

Ένας άλλος τρόπος ελέγχου της καταλληλότητας του μοντέλου είναι μέσω υπολοίπων, τα οποία αποτελούν ένα μέτρο συμφωνίας των προβλέψεων της στατιστικής ανάλυσης με πραγματικές τιμές.

Εν συνεχεία, υποθέτουμε το μοντέλο του Cox με *k* επεξηγηματικές μεταβλητές, τέτοιες ώστε να ισχύει :

$$\hat{\underset{\sim}{\beta}}'\underset{\sim}{x}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$$

και θεωρούμε ένα δείγμα *n* ατόμων, όπου οι *n-r* χρόνοι επιβίωσης είναι δεξιά αποκομμένοι. Η συνάρτηση διακινδύνευσης για το *i* άτομο, με *i*=1,2,…,*n* είναι η εξής:

$$\hat{h}_i(\mathrm{t}) = \exp(\hat{\underset{\sim}{\beta}}'\underset{\sim}{x}_i)\hat{h}_0(\mathrm{t}) \ ,$$

με $\hat{h}_0(\mathrm{t})$ να είναι η εκτίμηση της βασικής συνάρτησης διακινδύνευσης. Κατόπιν, ορίζουμε κάποιες κατηγορίες υπολοίπων.

Υπόλοιπα Cox-Snell

Το Cox-Snell υπόλοιπο για το *i* άτομο δίνεται από τον τύπο:

$$\mathrm{r}_{CSi} = \exp(\hat{\underset{\sim}{\beta}}'\underset{\sim}{x}_i)\hat{H}_0(\mathrm{t}_i) \, ,$$

όπου t_i είναι ο παρατηρούμενος χρόνος επιβίωσης του *i*-οστού ατόμου.

Τα υπόλοιπα αυτά χρησιμοποιούνται κυρίως στα παραμετρικά μοντέλα κι όχι τόσο στο ημι-παραμετρικό μοντέλο του Cox.

Υπόλοιπα Martingale

Τα υπόλοιπα Martingale ορίζονται βάσει των υπολοίπων Cox-Snell ως εξής:

$$\mathrm{r}_{Mi} = \delta_i - \mathrm{r}_{CSi} \, ,$$

όπου δ_i=0 για αποκομμένη παρατήρηση και δ_i=1 για μη αποκομμένη παρατήρηση.

Τα υπόλοιπα Martingale παίρνουν τιμές μεταξύ του $-\infty$ και της μονάδας και χρησιμεύουν για την εύρεση της συναρτησιακής μορφής μιας μεταβλητής που πρόκειται να εισαχθεί στο μοντέλο του Cox.

Υπόλοιπα Deviance

Τα υπόλοιπα Deviance ορίζονται ως εξής:

$$r_{Di} = \text{sgn}(\text{r}_{Mi})[-2\{\text{r}_{Mi} + \delta_i \log(\delta_i - \text{r}_{Mi})\}]^{1/2},$$

με $\text{sgn}(\text{x}) = \begin{cases} 1, x > 0 \\ -1, x < 0 \end{cases}$ και χρησιμοποιούνται αντί των *Martingale*, επειδή είναι πιο εύχρηστα στη γραφική τους ερμηνεία.

Υπόλοιπα Schoenfeld

Τα υπόλοιπα Schoenfeld δεν ορίζονται για όλα τα άτομα του δείγματος, αλλά μόνο για τις μη αποκομμένες παρατηρήσεις, ενώ δε δίνουν ακριβή τιμή για κάθε μονάδα αλλά ένα σύνολο τιμών. Για τον υπολογισμό τους δεν χρειάζεται εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης.

Επίσης, η πιθανότητα αποτυχίας της j-οστής υπό μελέτη μονάδας είναι η εξής:

$$p_j = \frac{e^{\underset{\sim}{\beta}' \underset{\sim}{x}_k}}{\sum_{i \in R_j} e^{\underset{\sim}{\beta}' \underset{\sim}{x}_i}},$$

όπου R_j είναι το σύνολο των ατόμων που βρίσκονται σε κίνδυνο την χρονική στιγμή t_j.

Το διάνυσμα των επεξηγηματικών μεταβλητών έχει αναμενόμενη τιμή:

$$E(\underset{\sim}{\text{x}} \mid \text{R}_j) = \sum_{k \in R_j} \underset{\sim}{x}_k p_k = \frac{\sum_{k \in R_j} \underset{\sim}{x}_k e^{\underset{\sim}{\beta}' \underset{\sim}{x}_k}}{\sum_{i \in R_j} e^{\underset{\sim}{\beta}' \underset{\sim}{x}_i}}.$$

Τελικά τα υπόλοιπα *Schoenfeld* ορίζονται ως εξής:

$$\hat{\underset{\sim}{r}}_j = \underset{\sim}{x}_j - E(\underset{\sim}{\text{x}} \mid \text{R}_j),$$

όπου η εκτιμήτρια της αναμενόμενης μέσης τιμής προκύπτει με αντικατάσταση των εκτιμητριών $\hat{\underset{\sim}{\beta}}$.

ΚΕΦΑΛΑΙΟ 4ο

ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΥΨΗΛΗΣ ΔΙΑΣΤΑΣΗΣ ΓΙΑ ΤΟ ΜΟΝΤΕΛΟ ΑΝΑΛΟΓΙΚΩΝ ΚΙΝΔΥΝΩΝ ΤΟΥ COX

4.1 Εισαγωγή

Η επιλογή μεταβλητών σε χώρο υψηλής διάστασης έχει αμφισβητήσει πολλά σύγχρονα στατιστικά προβλήματα από πολλούς επιστημονικούς κλάδους. Οι πρόσφατες τεχνολογικές εξελίξεις έχουν καταστήσει δυνατή τη συλλογή μιας τεράστιας ποσότητας από μεταβλητές πληροφορίας, όπως μικροσυστοιχιών, πρωτεομική και SNP δεδομένων μέσω της τεχνολογίας της βιο-απεικόνισης, παρατηρώντας ταυτόχρονα τις πληροφορίες για την επιβίωση των ασθενών σε κλινικές μελέτες. Έτσι, η ίδια πρόκληση εφαρμόζεται στην ανάλυση επιβίωσης προκειμένου να γίνει κατανοητή η σχέση μεταξύ γονιδιωματικών και κλινικών πληροφοριών σχετικά με το χρόνο επιβίωσης. Στο κεφάλαιο αυτό, επεκτείνουμε τη διαδικασία του σίγουρου κρησαρίσματος [Fan, J. and Lv, J. (2008)] στο μοντέλο αναλογικού κινδύνου του Cox με μια διαθέσιμη επαναληπτική έκδοση. Μελέτες αριθμητικής προσομοίωσης έχουν δείξει ενθαρρυντικά αποτελέσματα για την προτεινόμενη μέθοδο σε σύγκριση με άλλες τεχνικές, όπως η LASSO. Αυτό αποδεικνύει τη χρησιμότητα και την ευελιξία του επαναληπτικού σίγουρου κρησαρίσματος.

Η ανάλυση επιβίωσης είναι μια κοινώς χρησιμοποιούμενη μέθοδος για την ανάλυση του χρόνου αποτυχίας όπως του βιολογικού θανάτου ή της μηχανικής βλάβης. Στο πλαίσιο αυτό, ο θάνατος ή η αποτυχία/ανεπάρκεια αναφέρεται επίσης και ως «γεγονός». Η ανάλυση επιβίωσης προσπαθεί να διαμορφώσει/μοντελοποιήσει την επιβίωση χρόνου δεδομένων, η οποία συνήθως υπόκεινται σε λογοκρισία λόγω του τερματισμού της μελέτης. Ο βασικός σκοπός είναι να μελετήσουμε την εξάρτηση του χρόνου T στις μεταβλητές συνδιασποράς $X = (\mathrm{X}_1, ..., \mathrm{X}_p)^T$, όπου το p δηλώνει τη

διάσταση του χώρου συνδιασποράς. Ένας κοινός τρόπος για την επίτευξη αυτού του στόχου είναι ο κίνδυνος παλινδρόμησης, η οποία μελετά πως η υποθετική συνάρτηση κινδύνου T εξαρτάται από τη συμμεταβλητή $X = x$, η οποία ορίζεται ως εξής:

$$h(\mathrm{t}|\mathrm{x}) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P\left\{t \le T < t + \Delta t \middle| T \ge t, X = x\right\}$$

Σύμφωνα με τον ορισμό, η υποθετική συνάρτηση κινδύνου δεν είναι τίποτα άλλο από το στιγμιαίο ρυθμό αποτυχίας σε χρόνο t, που δίνει μια συγκεκριμένη τιμή του x για τη συμμεταβλητή X. Το μοντέλο αναλογικού κινδύνου είναι πολύ δημοφιλές, εν μέρει λόγω της απλότητας και της ευκολίας του στην αντιμετώπιση της λογοκρισίας.

Το μοντέλο υποθέτει ότι

$$h(\mathrm{t}|\mathrm{x}) = h_0(\mathrm{t})\Psi(\mathrm{x})$$

στην οποία το $h_0(\mathrm{t})$ είναι η συνάρτηση κινδύνου και $\Psi(\mathrm{x})$ είναι το αποτέλεσμα της συμμεταβλητότητας. Σημειώνουμε ότι το μοντέλο αυτό δεν είναι μοναδικά καθορισμένο στο $ch_0(\mathrm{t})$ και η $\Psi(\mathrm{x})/\mathrm{c}$ δίνει το ίδιο μοντέλο για οποιοδήποτε $\mathrm{c} > 0$. Έτσι, μια συνθήκη ταυτοποίησης πρέπει να προσδιορίζεται. Όταν η συνθήκη ταυτοποίησης $\Psi(0) = 1$ έχει επιβληθεί, η συνάρτηση $h_0(\mathrm{t})$, δηλαδή η υποθετική συνάρτηση κινδύνου του T, δίνει $X = 0$ και καλείται βασική συνάρτηση κινδύνου (baseline). Αν ξαναπαραμετρικοποιήσουμε $\Psi(\mathrm{x}) = \mathrm{e}^{\psi(\mathrm{x})}$, ο Cox [(1972), (1975)] εισήγαγε το αναλογικό μοντέλο κινδύνου

$$h(\mathrm{t}|\mathrm{x}) = h_0(\mathrm{t})\,\mathrm{e}^{\psi(\mathrm{x})}$$

Εδώ, η βασική συνάρτηση κινδύνου $h_0(\mathrm{t})$ είναι συνήθως εντελώς απροσδιόριστη και χρειάζεται να εκτιμηθεί μη παραμετρικά. Μπορεί να γίνει μια παραδοχή του γραμμικού μοντέλου $\psi(\mathrm{x}) = \mathrm{x}^T\beta$, όπου $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ είναι το διάνυσμα παραμετρικής παλινδρόμησης. Κατά τη διεξαγωγή της ανάλυσης επιβίωσης, δεν χρειάζεται μόνο να εκτιμήσουμε το β αλλά πρέπει ακόμα να εκτιμήσουμε και τη βασική συνάρτηση κινδύνου $h_0(\mathrm{t})$ μη παραμετρικά.

Οι πρόσφατες τεχνολογικές εξελίξεις έχουν καταστήσει δυνατή τη συλλογή ενός τεράστιου ποσού από πληροφορίες συμμεταβλητών, όπως μικροσυστοιχιών, πρωτεομικής και SNP δεδομένων μέσω της τεχνολογίας της βιο-απεικόνισης, καθώς παρατηρούμε τις πληροφορίες επιβίωσης ασθενών σε κλινικές μελέτες. Ωστόσο, είναι πολύ πιθανό ότι δεν θα σχετίζονται όλες οι διαθέσιμες συμμεταβλητές με την κλινική έκβαση, όπως ο χρόνος επιβίωσης. Στην πραγματικότητα, συνήθως ένα μικρό κλάσμα συμμεταβλητών σχετίζεται με τον κλινικό χρόνο. Αυτή είναι η έννοια της αραιότητας και συνεπώς καλεί για την ταυτοποίηση των σημαντικών παραγόντων κινδύνου και την ίδια στιγμή υπολογίζει τη συμβολή του κινδύνου όταν αναλύουμε τα δεδομένα επιβίωσης χρόνου με πολλές επεξηγηματικές μεταβλητές. Από μαθηματικής άποψης, αυτό σημαίνει ότι θα πρέπει να προσδιορίσουμε ποια β_j είναι μη μηδενικά και επίσης να εκτιμήσουμε αυτά τα μη μηδενικά β_j.

Οι πιο κλασσικές τεχνικές επιλογής μοντέλου έχουν επεκταθεί από τη γραμμική παλινδρόμηση στην ανάλυση επιβίωσης, οι οποίες αναφέρονται σε προηγούμενο κεφάλαιο.

Πρόσφατα έχουμε δει μια απότομη αύξηση του ενδιαφέροντος στην επιλογή μεταβλητής με εξαιρετικά υψηλή διάσταση. Με εξαιρετικά υψηλή διάσταση [Fan, J. and Lv, J. (2008)] εννοούμε ότι η διάσταση αυξάνεται εκθετικά στο μέγεθος του δείγματος, δηλαδή $\log(p) = O(n^a)$, για κάποιο $\alpha \in (0,1/2)$. Για τη γραμμική παλινδρόμηση υψηλής διάστασης, έχει προταθεί η SIS που βασίζεται στην ταξινόμηση οριακής συσχέτισης, όπως είδαμε και σε προηγούμενη ενότητα. Η ασυμπτωτική θεωρία έχει καταφέρει να αποδείξει ότι, με μεγάλη πιθανότητα, η SIS διατηρεί όλες τις επεξηγηματικές μεταβλητές με ποσοστό εσφαλμένης επιλογής που εξαφανίζεται. Μια σημαντική επέκταση της SIS, η ISIS, προτάθηκε επίσης για να αντιμετωπίσει τις δύσκολες επεξηγηματικές περιπτώσεις, όπως όταν κάποιες επεξηγηματικές μεταβλητές είναι οριακά ασυσχέτιστες με την απόκριση. Προκειμένου να αντιμετωπιστεί το πρόβλημα με τα πιο σύνθετα πραγματικά δεδομένα, οι SIS και ISIS επεκτάθηκαν σε πιο γενικά μοντέλα απώλειας [Fan, J., Samworth, R. and Wu, Y. (2009)], όπως τα γενικευμένα γραμμικά μοντέλα,η ισχυρή υποχώρηση και ταξινόμηση, και βελτιώθηκαν κάποια σημαντικά βήματα της αρχικής ISIS. Ειδικότερα, προτάθηκε η ιδέα της υπό όρους οριακής παλινδρόμησης κι έτσι

μια νέα παραλλαγή της μεθόδου, η οποία βασίζεται σε δείγματα διαχωρισμού, δημιουργήθηκε. Ένα μη ασυμπτωτικό θεωρητικό αποτέλεσμα δείχνει ότι η νέα παραλλαγή διαχωρισμού, μπορεί να μειώσει το ποσοστό εσφαλμένης ανακάλυψης. Παρά το γεγονός ότι η επέκταση καλύπτει ένα ευρύ φάσμα των στατιστικών μοντέλων, δεν κατέστη δυνατό να διερευνηθεί κατά πόσο η ISIS μέθοδος μπορεί να επεκταθεί στην παλινδρόμηση κινδύνου με αποκομμένο (*censoring*) χρόνο. Θα επικεντρωθούμε λοιπόν στο αναλογικό μοντέλο κινδύνου του Cox και θα επεκτείνουμε τη SIS και ISIS αναλόγως.

Το υπόλοιπο κεφάλαιο είναι οργανωμένο ως εξής: Στη δεύτερη ενότητα δίνονται οι λεπτομέρειες της μεθόδου του Cox, στην τρίτη ενότητα γίνεται επισκόπηση της επιλογής μεταβλητών μέσω ποινικοποιημένης προσέγγισης. Στην τέταρτη ενότητα, γίνεται επέκταση της SIS και ISIS, στην πέμπτη ενότητα παίρνουμε τα αποτελέσματα της προσομοίωσης ενώ στην έκτη και τελευταία ενότητα αποδεικνύουμε την αποτελεσματικότητα της SIS.

4.2. Το μοντέλο αναλογικών κινδύνων του Cox

Έστω *T, C, X* να δηλώνουν το χρόνο επιβίωσης, το χρόνο αποκοπής και τις συσχετιζόμενες συμμεταβλητές, αντίστοιχα. Επίσης, αντίστοιχα συμβολίζουμε με $Y = \min\{\mathrm{T}, \mathrm{C}\}$ τον παρατηρούμενο χρόνο και με $\delta = \mathrm{I}(\mathrm{T} \le \mathrm{C})$ το δείκτη αποκοπής ή λογοκρισίας. Για λόγους απλότητας, υποθέτουμε ότι *T* και *C* είναι δυνητικά ανεξάρτητες δοθέντος *X* και δεδομένου ότι ο μηχανισμός αποκοπής είναι non-informative.

Το παρατηρούμενο σύνολο δεδομένων μας $\{(\mathrm{x}_i, \mathrm{y}_i, \delta_i) : \mathrm{x}_i \in \mathbb{R}^p, \mathrm{y}_i \in \mathbb{R}^+,$ $\delta_i \in \{0,1\}, \mathrm{i} = 1, 2, ..., \mathrm{n}\}$ είναι ανεξάρτητο και ιδανικά κατανεμημένο τυχαίο δείγμα από ένα ορισμένο πληθυσμό $(\mathrm{X}, \mathrm{Y}, \delta)$. Ορίζουμε $C = \{\mathrm{i} : \delta_i = 0\}$ και $U = \{\mathrm{i} : \delta_i = 1\}$ να είναι το σύνολο των δεικτών που έχουν αποκοπεί και που δεν έχουν αποκοπεί αντίστοιχα. Τότε η πλήρης πιθανότητα του παρατηρούμενου συνόλου δεδομένων δίνεται από τη σχέση:

$$L = \prod_{i \in U} f(y_i / x_i) \prod_{i \in C} \overline{F}(y_i / x_i) = \prod_{i \in U} h(y_i / x_i) \prod_{i=1}^{n} \overline{F}(y_i / x_i)$$

όπου $f(t/x)$ είναι η δεσμευμένη συνάρτηση πυκνότητας, $\overline{F}(t/x) = \int_t^{\infty} f(s/x)\,ds$ η δεσμευμένη συνάρτηση επιβίωσης και $h(t/x) = f(t/x) / \overline{F}(t/x)$ η δεσμευμένη συνάρτηση διακινδύνευσης του X δοσμένου $X=x$.

Έστω $t_1^0 < t_2^0 < ... < t_N^0$ είναι τα διακεκριμένα διατεταγμένα παρατηρούμενα χρονικά διαστήματα αποτυχίας. Έστω *(j)* ο δείκτης που συνδέεται με τις συμμεταβλητές $x_{(j)}$ και $R(t)$ είναι ο κίνδυνος πριν από το χρόνο *t*: $R(t) = \{i : y_i \geq t\}$. Θεωρούμε το μοντέλο αναλογικού κινδύνου,

$$h(t/x) = h_0(t) \exp(x^T \beta), \qquad (4.1)$$

όπου $h_0(t)$ είναι η βασική συνάρτηση διακινδύνευσης. Σε αυτό το μοντέλο, τόσο η $h_0(t)$ όσο και η β είναι άγνωστες και πρέπει να εκτιμηθούν. Σύμφωνα με το μοντέλο (4.1), η πιθανότητα γίνεται:

$$L = \prod_{j=1}^{N} h_0(y_{(j)}) \exp(x_{(j)}^T \beta) \prod_{i=1}^{N} \exp\{-H_0(y_i) \exp(x_i^T \beta)\}$$

Όπου $H_0(t) = \int_0^t h_0(s)\,ds$ είναι η αντίστοιχη αθροιστική βασική συνάρτηση διακινδύνευσης. Ακολουθώντας την ιδέα του Breslow, θεωρούμε τη λιγότερο *informative* μη παραμετρική μοντελοποίηση για $H_0(t)$, στο οποίο $H_0(t)$ έχει ένα άλμα h_j στον παρατηρηθέντα χρόνο αποτυχίας t_j^0, δηλαδή $H_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. Τότε

$$H_0(y_i) = \sum_{j=1}^{N} h_j I(i \in R(t_j^0)) \qquad (4.2)$$

Συνεπώς, η λογαριθμική πιθανοφάνεια γίνεται

$$\sum_{j=1}^{N} \{\log(h_j) + x_{(j)}^T \beta\} - \sum_{i=1}^{n} \left\{ \sum_{j=1}^{N} h_j I(i \in R(t_j^0)) \exp(x_i^T \beta) \right\}.$$

Η μεγιστοποίηση της h_j δίνεται από τη σχέση:

$$\hat{h}_j(\beta) = \{ \sum_{i \in R(t_j^0)} \exp(x_i^T \beta) \}^{-1} \quad (4.3)$$

Αντικαθιστώντας τη μεγιστοποιημένη h_j στη λογαριθμική πιθανοφάνεια έχουμε:

$$\sum_{j=1}^{N} [x_{(j)}^T \beta - \log\{ \sum_{i \in R(t_j^0)} \exp(x_i^T \beta) \}]$$

που είναι ισοδύναμη με

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i x_i^T \beta - \sum_{i=1}^{n} \delta_i \log\left\{ \sum_{j \in R(y_i)} \exp(x_j^T \beta) \right\} \quad (4.4)$$

[Fan, J. and Li, R. (2001)].

Μεγιστοποιώντας την $\ell(\beta)$ στη (4.4) σε σχέση με τη β, μπορούμε να πάρουμε μια εκτίμηση $\hat{\beta}$ της παραμέτρου της παλινδρόμησης. Μόλις το $\hat{\beta}$ είναι διαθέσιμο, μπορούμε να το αντικαταστήσουμε στην (4.3) για να πάρουμε το $\hat{h}_j(\hat{\beta})$. Αυτό το πρόσφατο $\hat{h}_j(\hat{\beta})$ μπορεί να αντικατασταθεί στην (4.2) για να πάρουμε την μη-παραμετρική εκτίμηση της αθροιστικής συνάρτησης κινδύνου.

4.3. Επιλογή μεταβλητών για το μοντέλο αναλογικού κινδύνου του Cox μέσω ποινικοποίησης

Στο σύστημα εκτίμησης που παρουσιάστηκε, κανένας από τους συντελεστές παλινδρόμησης που εκτιμήθηκαν δεν είναι ακριβώς μηδέν, αφήνοντας όλες τις συμμεταβλητές στο τελικό μοντέλο. Συνεπώς, δεν είναι κατάλληλο για την επιλογή σημαντικών μεταβλητών και το χειρισμό της υπόθεσης με $p > n$. Για να επιτύχουμε επιλογή μεταβλητών, κλασσικές τεχνικές όπως η επιλογή του καλύτερου υποσυνόλου, η σταδιακή επιλογή και οι διαδικασίες εκκίνησης, επεκτάθηκαν ανάλογα έτσι ώστε να επιτευχθεί καλύτερος χειρισμός του μοντέλου αναλογικών κινδύνων του Cox.

Σ’ αυτή την ενότητα, θα επικεντρωθούμε σε κάποιες προηγμένες τεχνικές για την επιλογή μεταβλητών μέσω ποινικοποίησης. Η επιλογή μεταβλητών μέσω ποινικοποίησης έχει λάβει ιδιαίτερη προσοχή τελευταία. Βασικά, χρησιμοποιεί κάποια επιλογή μεταβλητών με δυνατότητα συνάρτησης ποινής προκειμένου να κανονικοποιήσει την αντικειμενική συνάρτηση κατά την εκτέλεση της βελτιστοποίησης. Έχουν προταθεί πολλές μέθοδοι επιλογής μεταβλητών με συναρτήσεις ποινής. Ένα πολύ γνωστό παράδειγμα είναι η L_1 ποινή $\lambda\sum_{j=1}^{p}|\beta_j|$ που είναι επίσης γνωστή ως ποινή LASSO [Tibshirani, R. J. (1996)].

Θεωρούμε μια γενική συνάρτηση ποινής με $p_\lambda(\cdot)$, όπου λ>0 είναι μια παράμετρος κανονικοποίησης. Από τις απαγωγές στην τελευταία ενότητα, η ποινικοποιημένη πιθανοφάνεια είναι ισοδύναμη με την ποινικοποιημένη μερική πιθανοφάνεια: Καθώς μεγιστοποιούμε το $\ell(\beta)$ στη (4.4), κάποιος μπορεί να το κανονικοποιήσει χρησιμοποιώντας το $\sum_{j=1}^{p} p_\lambda(\beta_j)$. Ισοδύναμα, επιλύουμε $\min -\ell(\beta)+\sum_{j=1}^{p} p_\lambda(\beta_j)$ συμπεριλαμβάνοντας το αρνητικό πρόσημο (-) μπροστά από το $\ell(\beta)$ [Tibshirani, R. (1997)].

Κατόπιν, θα χρησιμοποιήσουμε τη SCAD ποινή για τις επεκτάσεις της SIS και ISIS, όποτε είναι αναγκαίο. Η συνάρτηση SCAD είναι μια τετραγωνική συνάρτηση spline και συμμετρική γύρω από την αρχική. Μπορεί να οριστεί σε όρους της πρώτης τάξης παραγώγου

$$p'_\lambda(|\beta|) = \lambda\left\{\mathbb{1}_{\{|\beta|\le\lambda\}} + \frac{(a\lambda-|\beta|)_+}{(a-1)\lambda}\mathbb{1}_{\{|\beta|>\lambda\}}\right\},$$

για κάποια $\alpha>2$ και $\beta\neq 0$. Εδώ, το α είναι μια παράμετρος και οι Fan και Li πρότειναν να χρησιμοποιηθεί το $\alpha=3.7$ που βασίζεται στο Μπεϋζιανό επιχείρημα [Fan, J. and Li, R. (2001)]. Για τη SCAD ποινικοποιημένη βελτιστοποίηση, οι Fan και Li πρότειναν την τοπική τετραγωνική προσέγγιση, οι Zou και Li (2001) πρότειναν την τοπική γραμμική προσέγγιση, οι Wu και Liu (2009) παρουσίασαν το διαφορετικό

κυρτό αλγόριθμο. Εδώ, όταν είναι απαραίτητο, χρησιμοποιούμε τον τοπικό γραμμικό αλγόριθμο προσέγγισης για την επίλυση της SCAD ποινικοποιημένης βελτιστοποίησης.

4.4. SIS και ISIS για το μοντέλο αναλογικού κινδύνου του Cox

Η ποινή με τις κυμαινόμενες τεχνικές επιλογής μεταβλητών λειτουργεί τέλεια με ένα μέτριο πλήθος συμμεταβλητών. Ωστόσο, η χρησιμότητά τους είναι περιορισμένη καθώς έρχεται αντιμέτωπη με υψηλής διάστασης [Fan, J. and Lv, J. (2008)]. Στην περίπτωση γραμμικής παλινδρόμησης, προτάθηκε η ταξινόμηση των συμμεταβλητών ανάλογα με την απόλυτη τιμή της οριακής τους συσχέτισης με τη μεταβλητή απόκρισης, επιλέγοντας τις συμμεταβλητές που είναι σε υψηλή κατάταξη. Για να είναι σίγουρο ότι αυτή η απλή κατάταξη συσχέτισης διατηρεί όλες τις σημαντικές συμμεταβλητές με μεγάλη πιθανότητα δόθηκε θεωρητικό αποτέλεσμα. Έτσι η μέθοδος ονομάστηκε σίγουρο ανεξάρτητο κρησάρισμα (sure independent screening) SIS. Για να χειριστεί δύσκολα προβλήματα όπως αυτό με κάποιες σημαντικές συμμεταβλητές που είναι οριακά ασυσχέτιστες με την απόκριση, προτάθηκε η επαναληπτική SIS, η οποία καλείται ISIS. Η ISIS αρχίζει με τη SIS, μετά υποστρέφει (*regresses*) την απόκριση στις συμμεταβλητές που έχει επιλέξει η SIS και στη συνέχεια χρησιμοποιεί τα υπόλοιπα (*residuals*) παλινδρόμησης ως «*working*» απόκριση για τη λήψη περισσότερων συμμεταβλητών με τη SIS. Αυτή η διαδικασία μπορεί να επαναληφθεί έως ότου κάποιο κριτήριο σύγκλισης να ικανοποιηθεί. Εμπειρική βελτίωση πάνω στη SIS έχει παρατηρηθεί για τη ISIS. Για να μπορέσει να αυξηθεί η δύναμη της τεχνικής του σίγουρου ανεξάρτητου κρησαρίσματος, οι SIS και ISIS έχουν επεκταθεί σε γενικότερα μοντέλα όπως τα γενικευμένα γραμμικά μοντέλα, η ανθεκτική (*robust*) παλινδρόμηση κι η ταξινόμηση και έγιναν αρκετά σημαντικές βελτιώσεις [Fan, J., Samworth, R. and Wu, Y. (2009).]. Θα επεκτείνουμε τώρα τη βασική ιδέα της SIS και ISIS για να χειριστούμε το μοντέλο αναλογικού κινδύνου του Cox.

Έστω M^* είναι το σύνολο των δεικτών του αραιού μοντέλου που υποκρύπτεται, δηλαδή $M^* = \{\mathrm{j} : \beta_j^* \neq 0 \text{ και } 1 \leq j \leq p\}$ όπου τα β_j^* είναι οι πραγματικοί συντελεστές παλινδρόμησης του αναλογικού μοντέλου του Cox (4.1).

4.4.1. Κατάταξη με οριακή χρησιμότητα

Αρχικά θα θυμηθούμε τον ορισμό της ιδιότητας του σίγουρου κρησαρίσματος.

Ορισμός 1. (Ιδιότητα σίγουρου κρησαρίσματος)

Λέμε ότι μια διαδικασία επιλογής μοντέλου ικανοποιεί την ιδιότητα του σίγουρου κρησαρίσματος αν το επιλεγμένο μοντέλο $\hat{M}$ με μέγεθος $o_p(\mathrm{n})$ περιέχει το πραγματικό μοντέλο M^* με πιθανότητα που τείνει στο 1.

Για κάθε συμμεταβλητή $X_m (1 \leq \mathrm{m} \leq \mathrm{p})$, ορίζουμε την οριακή της χρησιμότητα ως τη μέγιστη τιμή της μερικής πιθανοφάνειας της ενιαίας συμμεταβλητής

$$\mathrm{u}_m = \max_{\beta_m} \left(\sum_{i=1}^{n} \delta_i x_{im} \beta_m - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(\mathrm{y}_i)} \exp(\mathrm{x}_{jm} \beta_m) \right\} \right)$$

Εδώ, x_{im} είναι το *m*-οστό στοιχείο της x_i, δηλαδή $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$. Διαισθητικά, όσο μεγαλύτερη είναι η οριακή χρησιμότητα, τόσες περισσότερες πληροφορίες περιέχει η αντίστοιχη συμμεταβλητή σχετικά με την έκβαση επιβίωσης. Μόλις έχουμε συγκεντρώσει όλες τις οριακές χρησιμότητες u_m για $\mathrm{m} = 1, 2, ..., \mathrm{p}$, κατατάσσουμε όλες τις συμμεταβλητές σύμφωνα με τις αντίστοιχες οριακές χρησιμότητές τους από τη μεγαλύτερη προς τη μικρότερη και επιλέγουμε τις *d* κορυφαίες σε κατάταξη συμμεταβλητές. Δηλώνουμε με I το σύνολο των δεικτών των d συμμεταβλητών που έχουμε επιλέξει.

Το σύνολο των δεικτών I αναμένεται να καλύψει το πραγματικό σύνολο δεικτών M^* με υψηλή πιθανότητα, ειδικά όταν χρησιμοποιούμε ένα σχετικά μεγάλο d. Αυτό συνήθως δείχνεται για το γραμμικό μοντέλο με το *Gaussian* θόρυβο και τις κανονικές συμμεταβλητές και επεκτάθηκε σημαντικά στα γενικευμένα γραμμικά μοντέλα με μη Gaussian συμμεταβλητές [Fan, J. and Song, R. (2010)]. Η παράμετρος d συνήθως επιλέγεται να είναι αρκετά μεγάλη ώστε να διασφαλιστεί η ιδιότητα του σίγουρου κρησαρίσματος. Ωστόσο, το εκτιμώμενο σύνολο δεικτών I μπορεί επίσης να περιλαμβάνει πολλές μη σημαντικές συμμεταβλητές. Για να βελτιώσουμε τις επιδόσεις, η προσέγγιση της επιλογής μεταβλητών που βασίζεται στη ποινικοποίηση μπορεί να εφαρμοστεί στο επιλεγμένο υποσύνολο των μεταβλητών $\{\mathrm{X}_j, \mathrm{j} \in \mathrm{I}\}$ για να διαγράψει επιπλέον ασήμαντες μεταβλητές. Μαθηματικά, μπορούμε να λύσουμε το παρακάτω πρόβλημα ποινικοποιημένης μερικής πιθανοφάνειας

$$\min_{\beta_I} \left(-\sum_{i=1}^{n} \delta_i x_{I,i}^T \beta_I + \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(\mathrm{y}_i)} \exp(x_{I,\mathrm{j}}^T \beta_I) \right\} + \sum_{m \in I} p_\lambda(\beta_m) \right)$$

όπου $x_{I,i}$ είναι υπο-διάνυσμα του x_i με δείκτες στο I και ομοίως για το β_I. Θα οδηγήσει σε αραιά εκτίμηση των παραμέτρων παλινδρόμησης $\hat{\beta}_I$. Χαρακτηρίζει το σύνολο δεικτών μη μηδενικών συνιστωσών του $\hat{\beta}_I$ από το M, το οποίο θα μας χρησιμεύσει ως η τελική εκτίμησή μας M^*.

4.4.2. Κατάταξη χαρακτηριστικών υπό όρους και επαναληπτική επιλογή χαρακτηριστικών

Οι Fan και Lv (2008) επισήμαναν ότι η SIS μπορεί να αποτύχει άσχημα για κάποια προκλητικά σενάρια, όπως στην περίπτωση όπου υπάρχουν κοινές σχετιζόμενες αλλά οριακά ασυσχέτιστες συμμεταβλητές ή κοινές ασυσχέτιστες συμμεταβλητές που έχουν υψηλότερη οριακή συσχέτιση με την απόκριση από ορισμένες σημαντικές επεξηγηματικές μεταβλητές. Για να αντιμετωπιστούν τέτοια δύσκολα σενάρια, έχει προταθεί η επαναληπτική SIS (ISIS). Σε σύγκριση με τη SIS που βασίζεται σε οριακές πληροφορίες μόνο, η ISIS προσπαθεί να χρησιμοποιήσει περισσότερο τις πληροφορίες των κοινών συμμεταβλητών. Η επαναληπτική SIS

ξεκινά με τη χρήση της SIS, για την επιλογή ενός συνόλου δεικτών $\hat{I}_1$, κατά την οποία εφαρμόζεται μια ποινή που βασίζεται στο βήμα της επιλογής μεταβλητών για να πάρει μια εκτίμηση της παραμέτρου παλινδρόμησης $\hat{\beta}_{\hat{I}_1}$. Μια εκλεπτυσμένη εκτίμηση του αληθούς συνόλου δεικτών λαμβάνεται και συμβολίζεται με $\hat{M}_1$, όπου ο δείκτης αντιστοιχεί στα μη μηδενικά στοιχεία του $\hat{\beta}_{\hat{I}_1}$.

Ορίζουμε την υπό όρους χρησιμότητα της κάθε συμμεταβλητής *m* που δεν είναι στο $\hat{M}_1$, ως εξής:

$$u_{m|\hat{M}_1} = \max_{\beta_m, \beta_{\hat{M}_1}} \left(\sum_{i=1}^{n} \delta_i (x_{im}\beta_m + x_{\hat{M}_1,i}^T \beta_{\hat{M}_1}) - \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp(x_{jm}^T \beta_m + x_{\hat{M}_1,j}^T \beta_{\hat{M}_1}) \right\} \right)$$

Αυτή η υπό όρους χρησιμότητα μετρά την πρόσθετη συνεισφορά της m-οστής συμμεταβλητής, δεδομένου ότι όλες οι συμμεταβλητές με δείκτες στο $\hat{M}_1$ έχουν συμπεριληφθεί στο μοντέλο. Μόλις οι υπό όρους χρησιμότητες έχουν οριστεί για κάθε συμμεταβλητή που δεν είναι στο $\hat{M}_1$, τις κατατάσσουμε από τις μεγαλύτερες στις μικρότερες και επιλέγουμε εκείνες τις συμμεταβλητές που έχουν υψηλές ταξινομήσεις. Ορίζουμε το σύνολο δεικτών αυτών των επιλεγμένων συμμεταβλητών με $\hat{I}_2$. Έχοντας ορίσει το $\hat{I}_2$, ελαχιστοποιούμε

$$-\sum_{i=1}^{n} \delta_i (x_{\hat{M}_1 \cup \hat{I}_2,i}^T \beta_{\hat{M}_1 \cup \hat{I}_2}) + \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp(x_{\hat{M}_1 \cup \hat{I}_2,j}^T) \right\} + \sum_{m \in \hat{M}_1 \cup \hat{I}_2} p_\lambda(\beta_j) \qquad (4.5)$$

σε σχέση με το $\beta_{\hat{M}_1 \cup \hat{I}_2}$, για να πάρει αραιή εκτίμηση $\hat{\beta}_{\hat{M}_1 \cup \hat{I}_2}$. Ορίζουμε το σύνολο δεικτών που αντιστοιχούν στα μη μηδενικά στοιχεία του $\hat{\beta}_{\hat{M}_1 \cup \hat{I}_2}$ να είναι το $\hat{M}_2$, που είναι ο ανανεωμένος εκτιμητής του συνόλου δεικτών M^*. Παρατηρούμε ότι αυτό το βήμα μπορεί να διαγράψει κάποιες μεταβλητές $\left\{ X_j \in \hat{M}_1 \right\}$ που είχαμε επιλέξει προηγουμένως [Fan, J., Samworth, R.and Wu, Y. (2009)].

Η παραπάνω επανάληψη μπορεί να επαναληφθεί μέχρι κάποιο κριτήριο σύγκλισης να ικανοποιηθεί. Υιοθετούμε το κριτήριο που είτε έχουν εντοπιστεί *d* συμμεταβλητές ή ισχύει $\hat{M}_j = \hat{M}_{j-1}$ για κάποιο j.

4.4.3 Νέες παραλλαγές της SIS και ISIS για τη μείωση της FSR

Οι Fan, Samworth και Wu (2009) παρατήρησαν ότι η ιδέα του διαχωρισμού του δείγματος μπορεί επίσης να χρησιμοποιηθεί για τη μείωση του ποσοστού της εσφαλμένης επιλογής (*FSR*). Χωρίς βλάβη της γενικότητας, υποθέτουμε ότι το μέγεθος του δείγματος είναι άρτιο. Χωρίζουμε τυχαία το δείγμα σε δύο ίσα μέρη. Στη συνέχεια, εφαρμόζουμε ξεχωριστά τη SIS ή ISIS στα δεδομένα του κάθε μέρους για να πάρουμε δύο εκτιμήσεις $\hat{I}^{(1)}$ και $\hat{I}^{(2)}$ του αρχικού συνόλου δεικτών M^*. Και οι δύο αυτές εκτιμήσεις θα μπορούσαν να έχουν υψηλή *FSR* επειδή βασίζονται σε μια απλή και ακατέργαστη μέθοδο κρησαρίσματος. Παρόλα αυτά, κάθε μία από αυτές θα πρέπει να περιλαμβάνει όλες τις σημαντικές συμμεταβλητές με υψηλές πιθανότητες. Δηλαδή, οι σημαντικές συμμεταβλητές θα πρέπει να εμφανίζονται και στα δύο σύνολα με πιθανότητα που τείνει στο 1 ασυμπτωτικά. Ορίζουμε μια νέα εκτίμηση από την τομή: $\hat{I} = \hat{I}^{(1)} \cap \hat{I}^{(2)}$. Ο νέος εκτιμητής $\hat{I}$ θα πρέπει να περιλαμβάνει όλες τις σημαντικές συμμεταβλητές με υψηλή πιθανότητα λόγω και των ιδιοτήτων του κάθε εκτιμητή. Ωστόσο, από την κατασκευή, ο αριθμός των ασήμαντων συμμεταβλητών στη νέα εκτίμηση $\hat{I}$ είναι πολύ μικρότερος. Ο λόγος είναι ότι, προκειμένου για μια ασήμαντη συμμεταβλητή να εμφανίζεται στο $\hat{I}$, θα πρέπει να συμπεριλαμβάνεται τυχαία και στα $\hat{I}^{(1)}$ και $\hat{I}^{(2)}$.

Για τη νέα μέθοδο παραλλαγής, που βασίζεται στον τυχαίο διαχωρισμό, οι Fan, Samworth και Wu (2009) έλαβαν κάποια μη-ασυμπτωτική πιθανότητα που δεσμεύεται στην περίπτωση που *r* ασήμαντες συμμεταβλητές περιέχονται στην τομή $\hat{I}$, για οποιοδήποτε φυσικό αριθμό *r* κάτω από κάποιες συνθήκες ισοτιμίας σε όλες τις ασήμαντες συμμεταβλητές. Το όριο της πιθανότητας μειώνει τη διάσταση, δείχνοντας μια «ευλογία» διάστασης. Πρέπει να επισημάνουμε ότι το θεωρητικό όριο ισχύει και στα δεδομένα επιβίωσης χρόνου, επειδή το θεωρητικό όριο βασίζεται στο

διαχωρισμό του δείγματος σε δύο μέρη και απαιτεί μόνο την ανεξαρτησία αυτών των δύο μερών.

Καθώς ορίζουμε νέες παραλλαγές, θα χρησιμοποιήσουμε το ίδιο *d* που χρησιμοποιήθηκε στην αρχική SIS και ISIS. Παρόλα αυτά, θα οδηγήσει σε ένα πολύ «επιθετικό» κρησάρισμα. Θα καλούμε την αντίστοιχη παραλλαγή ως την πρώτη παραλλαγή της (I)SIS. Εναλλακτικά, σε κάθε βήμα θα επιλέγουμε μεγαλύτερα $\hat{I}^{(1)}$ και $\hat{I}^{(2)}$ για να διασφαλίσουμε ότι η τομή τους $\hat{I}^{(1)} \cap \hat{I}^{(2)}$ έχει *d* συμμεταβλητές, την οποία θα καλούμε δεύτερη παραλλαγή. Η δεύτερη παραλλαγή διασφαλίζει ότι υπάρχουν τουλάχιστον *d* συμμεταβλητές που συμπεριλαμβάνονται πριν την εφαρμογή της ποινής σε κάθε βήμα και είναι λιγότερο «επιθετική». Αριθμητικά παραδείγματα θα χρησιμοποιηθούν για να ανακαλύψουμε την επίδοση και την επιλογή της πρώτης εναλλαγής.

4.5. Προσομοίωση

4.5.1 Σχεδιασμός προσομοίωσης

Σε αυτή την ενότητα, θα διεξάγουμε μελέτες προσομοίωσης για να δείξουμε τη δύναμη της (I)SIS και των παραλλαγών της συγκρίνοντάς τα με τη LASSO [Tibshirani, R. (1997).] στο αναλογικό μοντέλο κινδύνου του Cox. Εδώ, η παράμετρος κανονικοποίησης της LASSO είναι συντονισμένη με πενταπλάσια επικύρωση. Οι περισσότερες από τις ρυθμίσεις υιοθετούνται από τους Fan και Lv (2008) και Fan, Samworth και Wu (2009). Τέσσερις διαφορετικές περιπτώσεις θεωρούνται με n=300 και p=400. Και οι δύο από αυτές έχουν αναθεωρηθεί με ένα διαφορετικό ζευγάρι μεγέθους δείγματος n=400 και p=1000. Δημιουργούνται έτσι, συμμεταβλητές με διαφορετικές ρυθμίσεις, ως εξής:

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

Περίπτωση 2: $X_1,...,X_p$ είναι κανονικές πολυπαραγοντικές, οριακά N(0,1) και με συσχέτιση $corr(\mathrm{X}_i,\mathrm{X}_j)=\rho$, αν $i \neq j$. Εδώ δεχόμαστε $\rho = 0.5$

Περίπτωση 3: $X_1,...,X_p$ είναι κανονικές πολυπαραγοντικές, οριακά N(0,1) και με δομή συσχέτισης $corr(\mathrm{X}_i,\mathrm{X}_4)=1/\sqrt{2}$ για όλα τα $i\neq 4$ και $corr(\mathrm{X}_i,\mathrm{X}_j)=1/2$ αν τα i και j είναι διακριτά στοιχεία του $\{1,...,p\}\setminus\{4\}$.

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

Περίπτωση 5: Όπως η περίπτωση 2 εκτός από : n=400 καιp=1000.

Περίπτωση 6: Όπως η περίπτωση 4 εκτός από : n=400 καιp=1000.

Εδώ η υπόθεση 1 με ανεξάρτητες επεξηγηματικές μεταβλητές είναι η πιο απλή για επιλογή μεταβλητών. Στις περιπτώσεις 2-6, όμως, έχουμε συσχέτιση, έτσι ώστε το $corr(\mathrm{X}_i,\mathrm{X}_j)$ να μην φθίνει καθώς το $|i-j|$ αυξάνεται. Αργότερα, θα δούμε ότι για τις περιπτώσεις 3, 4 και 6 οι αληθινοί συντελεστές είναι ειδικά επιλεγμένοι, έτσι ώστε η απόκριση να είναι οριακά ανεξάρτητη αλλά να εξαρτάται από κοινού με το X_4. Συνεπώς, περιμένουμε η επιλογή μεταβλητών, σε αυτές τις περιπτώσεις, να είναι πιο δύσκολες, ειδικά για τις εκδόσεις (*versions*) της SIS που είναι μη επαναλαμβανόμενες. Παρατηρούμε ότι στην ασυμπτωτική θεωρία της SIS, αυτό το είδος εξάρτησης έχει αποκλειστεί από τις συνθήκες [Fan, J. and Li, R. (2001)].

Στην εφαρμογή μας, επιλέγουμε $d=\left\lfloor \frac{n}{4\log n} \right\rfloor$ τόσο για τη Vanilla εκδοχή της SIS (Van-SIS) όσο και για τη δεύτερη παραλλαγή (Var2-SIS). Για την πρώτη παραλλαγή (Var1-SIS), ωστόσο, χρησιμοποιούμε $d=\left\lfloor \frac{n}{\log n} \right\rfloor$ (παρατηρούμε ότι αφού οι επιλεγμένες μεταβλητές για την πρώτη παραλλαγή είναι η τομή των δύο συνόλων μεγέθους d, συνήθως καταλήγουμε με πολύ λιγότερες από d μεταβλητές

που επιλέγονται με αυτή τη μέθοδο). Για κάθε τύπο της SIS ή ISIS, εφαρμόζουμε τη SCAD με αυτές τις επεξηγηματικές μεταβλητές να παίρνουν μια τελική εκτίμηση των συντελεστών παλινδρόμησης στο τέλος του βαθμού κρησαρίσματος. Όποτε απαιτείται, χρησιμοποιείται η BIC για να επιλέξουμε την καλύτερη, στα πλαίσια της κανονικοποίησης, παράμετρο.

Σε όλες τις ρυθμίσεις, ο χρόνος αποκοπής παράγεται από την εκθετική κατανομή με μέση τιμή 10. Αυτό αντιστοιχεί στο να επιλέξουμε την αρχική συνάρτηση κινδύνου $h_0(\mathrm{t}) = 0.1$ για $t \geq 0$. Οι πραγματικοί συντελεστές παλινδρόμησης και το ποσοστό αποκοπής σε κάθε μία από τις έξι περιπτώσεις έχουν ως εξής:

Περίπτωση 1: $\beta_1 = -1.6328$, $\beta_2 = 1.3988$, $\beta_3 = -1.6497$, $\beta_4 = 1.6353$, $\beta_5 = -1.4209$, $\beta_6 = 1.7022$ και $\beta_j = 0$ για $j > 6$. Το αντίστοιχο ποσοστό αποκοπής είναι 33%.

Περίπτωση 2: Οι συντελεστές είναι ίδιοι όπως και στην περίπτωση 1. Το αντίστοιχο ποσοστό αποκοπής είναι 27%.

Περίπτωση 3: $\beta_1 = 4$, $\beta_2 = 4$, $\beta_3 = 4$, $\beta_4 = -6\sqrt{2}$ και $\beta_j = 0$ για $j > 4$. Το αντίστοιχο ποσοστό αποκοπής είναι 30%.

Περίπτωση 4: $\beta_1 = 4$, $\beta_2 = 4$, $\beta_3 = 4$, $\beta_4 = -6\sqrt{2}$, $\beta_5 = 4/3$ και $\beta_j = 0$ για $j > 5$. Το αντίστοιχο ποσοστό αποκοπής είναι 31%.

Περίπτωση 5: $\beta_1 = -1.5140$, $\beta_2 = 1.2799$, $\beta_3 = -1.5307$, $\beta_4 = 1.5164$, $\beta_5 = -1.3020$, $\beta_6 = 1.5833$ και $\beta_j = 0$ για $j > 6$. Το αντίστοιχο ποσοστό αποκοπής είναι 23%.

Περίπτωση 6: Οι συντελεστές είναι ίδιοι όπως και στην περίπτωση 4. Το αντίστοιχο ποσοστό αποκοπής είναι 36%.

Στις περιπτώσεις 1, 2 και 5, οι συντελεστές επιλέχθηκαν τυχαία, και παρήχθησαν $(4\log n/\sqrt{n} + |Z|/4)\,\mathrm{U}$ με $\mathrm{Z} \sim \mathrm{N}(0,1)$ και U=1 με πιθανότητα 0.5, και $U = -1$ με

πιθανότητα 0.5, ανεξάρτητα του Z. Για τις περιπτώσεις 3, 4 και 6 οι επιλογές εξασφαλίζουν ότι, μολονότι $\beta_4 \neq 0$, τα X_4 και Y είναι οριακά ανεξάρτητα. Το γεγονός ότι το X_4 είναι οριακά ανεξάρτητο της απόκρισης, έχει σχεδιαστεί για να καταστήσει δύσκολο για την κοινή ανεξάρτητη μάθηση να επιλέξει αυτή τη μεταβλητή. Στις περιπτώσεις 4 και 6, προσθέτουμε μια άλλη σημαντική μεταβλητή, την X_5, με ένα μικρό συντελεστή για να το καταστήσει ακόμα δυσκολότερο.

4.5.2 Τα αποτελέσματα των προσομοιώσεων

Αναφέρουμε τα αποτελέσματα της προσομοίωσής μας, τα οποία είναι βασισμένα σε 100 επαναλήψεις Monte Carlo για κάθε ρύθμιση στους πίνακες 4.1-4.7. Για να παρουσιάσουμε τα αποτελέσματα της προσομοίωσής μας, χρησιμοποιήσαμε πολλά διαφορετικά μέτρα απόδοσης. Στις γραμμές με την ένδειξη $\left\|\beta-\hat{\beta}\right\|_1$ και $\left\|\beta-\hat{\beta}\right\|_2^2$ αναφέρουμε το μέσο L_1 και το τετραγωνικό L_2 σφάλμα εκτίμησης $\left\|\beta-\hat{\beta}\right\|_1=\sum_{j=1}^{p}\left|\hat{\beta}_j-\beta_j\right|$ και $\left\|\beta-\hat{\beta}\right\|_2^2=\sum_{j=1}^{p}\left|\hat{\beta}_j-\beta_j\right|^2$, αντίστοιχα, όπου η διάμεσος είναι πάνω από τις 100 επαναλήψεις. Στις γραμμές με την ένδειξη P_1, αναφέρουμε το ποσοστό των 100 επαναλήψεων όπου η διαδικασία (I)SIS, κατόπιν εξέτασης, περιλαμβάνει όλες τις σημαντικές μεταβλητές στο μοντέλο, ενώ η γραμμή με ετικέτα P_2 αναφέρει το αντίστοιχο ποσοστό των φορών που οι τελικές επιλεγμένες μεταβλητές, μετά από περαιτέρω εφαρμογή της ποινής SCAD, περιλαμβάνουν όλες τις σημαντικές. Επίσης, αναφέρουμε το μέσο μέγεθος του τελικού μοντέλου μεταξύ των 100 επαναλήψεων στη σειρά με την ένδειξη MMS.

Αναφέρουμε τα αποτελέσματα των περιπτώσεων 1 και 2 στον πίνακα 4α. Υπενθυμίζουμε ότι οι συμμεταβλητές στην περίπτωση 1 είναι όλες ανεξάρτητες.

	Van-SIS	Van-ISIS	Var1-SIS	Var1-ISIS	Var2-SIS	Var2-ISIS	LASSO
Case 1: independent covariates							
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_1$	0.79	0.57	0.73	0.61	0.76	0.62	4.23
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2$	0.13	0.09	0.15	0.1	0.15	0.1	0.98
P_1	1	1	0.99	1	0.99	1	–
P_2	1	1	0.99	1	0.99	1	1
MMS	7	6	6	6	6	6	68.5
Case 2: Equi-correlated covariates with $\rho = 0.5$							
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_1$	2.2	0.64	4.22	0.8	3.95	0.78	4.38
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2$	1.74	0.11	4.71	0.29	4.07	0.28	0.98
P_1	0.71	1	0.42	0.99	0.46	0.99	–
P_2	0.71	1	0.42	0.99	0.46	0.99	1
MMS	7	6	6	6	7	6	57

ΠΙΝΑΚΑΣ 4α. Τα αποτελέσματα των περιπτώσεων 1 και 2. Το P_1 αντιστοιχεί στην πιθανότητα η (I)SIS να συμπεριλαμβάνει το αληθινό μοντέλο, το οποίο έχει υποστεί sure screening (σίγουρο κρησάρισμα). Το P_2 αντιστοιχεί στην πιθανότητα η (I)SIS να συμπεριλαμβάνει το τελικό μοντέλο, το οποίο έχει υποστεί sure screening (σίγουρο κρησάρισμα). Το MMS (Median Mode lSize) αντιστοιχεί στο μέσο μέγεθος του μοντέλου μεταξύ 100 επαναλήψεων. Το μέγεθος του δείγματος είναι n=300 κι ο αριθμός των μεταβλητών είναι p=400.

Σ' αυτή την περίπτωση, η Van-SIS εκτελείται αρκετά καλά. Ωστόσο, δεν αποδίδει καλά για την εξαρτημένη περίπτωση (περίπτωση 2). Παρατηρούμε ότι η μόνη διαφορά μεταξύ της περίπτωσης 1 και 2 είναι η δομή της συνδιακύμανσης των συμμεταβλητών. Και για τις δύο περιπτώσεις, τη Vanilla ISIS και τη δεύτερη παραλλαγή, τις αποδίδει πολύ καλά. Αξίζει να σημειωθεί ότι η ISIS βελτιώνει σημαντικά τη SIS, όταν οι συμμεταβλητές είναι εξαρτημένες, σε ότι αφορά την πιθανότητα να συμπεριληφθούν όλες οι αληθινές συμμεταβλητές και η μείωση του σφάλματος εκτίμησης. Αυτή η σύγκριση δείχνει ότι η ISIS αποδίδει πολύ καλύτερα όταν υπάρχει σοβαρή συσχέτιση μεταξύ των συμμεταβλητών.

Κατά την εφαρμογή της ποινής LASSO στο μοντέλο αναλογικού κινδύνου του Cox, υιοθετούμε τον πηγαίο κώδικα Fortran στο πακέτο R "glmpath". Υπενθυμίζουμε ότι η αντικειμενική συνάρτηση της ποινής LASSO στο μοντέλο αναλογικού κινδύνου του Cox είναι κυρτή και μη-γραμμική. Αυτό που κάνει η Fortran είναι να καλεί μια υπορουτίνα MINOS για να επιλύσει το αντίστοιχο μη γραμμικό κυρτό πρόβλημα βελτιστοποίησης. Εδώ η MINOS είναι μια βελτιστοποίηση του λογισμικού που αναπτύχθηκε από το Εργαστήριο Βελτιστοποίησης Συστημάτων στο Πανεπιστήμιο του Stanford. Αυτό το μη γραμμικό, κυρτό πρόβλημα βελτιστοποίησης είναι πιο πολύπλοκο από ότι το γενικό τετραγωνικό πρόβλημα προγραμματισμού. Έτσι, γενικά χρειάζεται πολύ περισσότερο

χρόνο για την επίλυση, ειδικά όταν η διάσταση είναι υψηλή, όπως επιβεβαιώνεται στον Πίνακα 4γ. Παρόλα αυτά, ο αλγόριθμος που χρησιμοποιήσαμε, συγκλίνει καθώς η αντικειμενική συνάρτηση είναι αυστηρά κυρτή.

Ο Πίνακας 4α δείχνει ότι η LASSO έχει την ιδιότητα του σίγουρου κρησαρίσματος, όπως η SIS, ωστόσο, το μέσο μέγεθος μοντέλου είναι 10 φορές μεγαλύτερο από αυτό της SIS. Ως συνέπεια, έχει επίσης, μεγαλύτερο σφάλμα εκτίμησης στους όρους $\left\|\beta-\hat{\beta}\right\|_1$ και $\left\|\beta-\hat{\beta}\right\|_2^2$. Το γεγονός ότι το μέσο απόλυτο σφάλμα απόκλισης είναι πολύ μεγαλύτερο από το μέσο τετραγωνικό σφάλμα υποδηλώνει ότι η LASSO επιλέγει πολλούς μικρούς μη-μηδενικούς συντελεστές για εκείνες τις ασήμαντες μεταβλητές. Αυτό επαληθεύεται επίσης από το γεγονός ότι η LASSO έχει ένα πολύ μεγάλο μέσο μέγεθος μοντέλου. Η εξήγηση είναι το θέμα διαστρέβλωσης που έχει σημειωθεί από τους Fan και Li (2001). Για να έχει η LASSO μικρή διαστρέβλωση για μη-μηδενικούς συντελεστές, ένα μικρό λ θα πρέπει να επιλεγεί. Ωστόσο, ένα μικρό λ προσλαμβάνει πολλούς μικρούς συντελεστές για μη σημαντικές μεταβλητές. Για την περίπτωση 2, η LASSO έχει παρόμοια απόδοση όπως και στην περίπτωση 1, γεγονός που σημαίνει ότι περιλαμβάνει όλες τις σημαντικές μεταβλητές αλλά έχει ένα πολύ μεγαλύτερο μέγεθος μοντέλου.

Τα αποτελέσματα των περιπτώσεων 3 και 4 αναφέρονται στον Πίνακα 4β.

	Van-SIS	Van-ISIS	Var1-SIS	Var1-ISIS	Var2-SIS	Var2-ISIS	LASSO
Case 3: An important predictor that is independent of survival time							
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_1$	20.1	1.03	20.01	0.99	20.09	1.08	20.53
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2$	94.72	0.49	100.42	0.47	94.77	0.55	76.31
P_1	0	1	0	1	0	1	–
P_2	0	1	0	1	0	1	0.06
MMS	13	4	8	4	13	4	118.5
Case 4: Two very hard variables to be selected							
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_1$	20.87	1.15	20.95	1.4	20.96	1.41	21.04
$\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2$	96.46	0.51	102.14	1.77	97.15	1.78	77.03
P_1	0	1	0	0.99	0	0.99	–
P_2	0	1	0	0.99	0	0.99	0.02
MMS	13	5	9	5	13	5	118

ΠΙΝΑΚΑΣ 4β. Αποτελέσματα για τις Περιπτώσεις 3 και 4.

Παρατηρούμε ότι, και στις δύο περιπτώσεις, ο σχεδιασμός διασφαλίζει ότι το X_4 είναι οριακά ανεξάρτητο αλλά ταυτόχρονα εξαρτάται και από το Y. Αυτή η ειδική σχεδίαση δεν αφήνει τη SIS να συμπεριλάβει τη X_4 στο αντίστοιχο

ταυτοποιημένο μοντέλο, όπως επιβεβαιώθηκε από τα αριθμητικά μας αποτελέσματα. Παρόλα αυτά, χρησιμοποιώντας την ISIS, είμαστε σε θέση να επιλέξουμε τη X_4 για κάθε επανάληψη. Παραδόξως, η LASSO σπάνια περιλαμβάνει τη X_4, ακόμα και αν δεν είναι μια μέθοδος που βασίζεται στο οριακό κρησάρισμα. Στην περίπτωση 5, είναι ακόμα πιο δύσκολο. Επιπρόσθετα, με την ίδια πρόκληση της περίπτωσης 4, ο συντελεστής β_5 είναι τρεις φορές μικρότερος από τις τέσσερις πρώτες μεταβλητές. Μέσω της συσχέτισης με τις τέσσερις πρώτες μεταβλητές, μη σημαντικές μεταβλητές $\{X_j : j \geq 6\}$ έχουν μεγαλύτερη οριακή χρησιμότητα με την Y παρά με την X_5. Παρόλα αυτά, η ISIS λειτουργεί πολύ καλά και αποδεικνύει για μια ακόμα φορά ότι χρησιμοποιεί επαρκώς τη συμμεταβλητή πληροφορία.

Συγκρίνουμε επίσης το υπολογιστικό κόστος των Van-ISIS και LASSO στον Πίνακα 4γ για τις περιπτώσεις 1-4.

	Case 1	Case 2	Case 3	Case 4
Van-ISIS	379.29	213.44	402.94	231.68
LASSO	37730.82	26348.12	46847	28157.71

<u>ΠΙΝΑΚΑΣ 4γ</u>. Ο μέσος χρόνος «τρεξίματος» (σε δευτερόλεπτα) συγκριτικά για τη Van-ISIS και LASSO

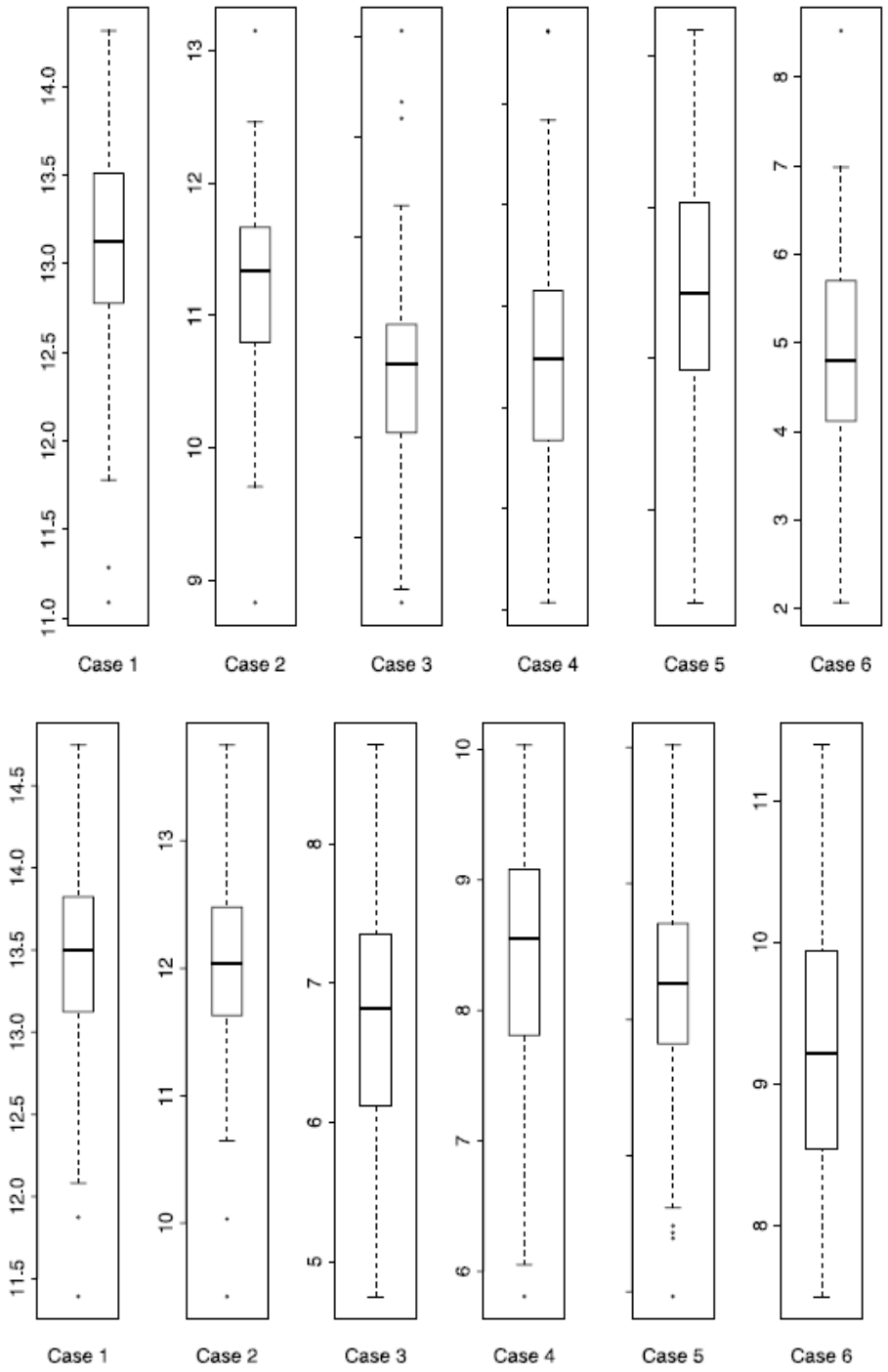
Ο Πίνακας 4γ δείχνει ότι η LASSO κάνει αρκετές ώρες για κάθε επανάληψη, ενώ η Van-ISIS μπορεί να τελειώσει σε μερικά μόλις λεπτά. Αυτό είναι μια τεράστια βελτίωση. Γι' αυτό το λόγο, για τις περιπτώσεις 5 και 6 όπου p=1000, αναφέρουμε μόνο τα αποτελέσματα της ISIS δεδομένου ότι η LASSO χρειάζεται αρκετές μέρες για να ολοκληρώσει μια μόνο επανάληψη. Τα αποτελέσματα των Περιπτώσεων 5 και 6 εμφανίζονται στον Πίνακα 4δ. Ο Πίνακας καταδεικνύει παρόμοιες επιδόσεις όπως και στις Περιπτώσεις 2 και 4 ακόμη και με περισσότερες συμμεταβλητές.

	Van-SIS	Van-ISIS	Var1-SIS	Var1-ISIS	Var2-SIS	Var2-ISIS
Case 5: The same as case 2 with $p = 1000$ and $n = 400$						
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	1.53	0.52	3.55	0.55	2.95	0.51
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	0.9	0.07	3.48	0.08	2.5	0.07
P_1	0.82	1	0.39	1	0.5	1
P_2	0.82	1	0.39	1	0.5	1
MMS	8	6	6	6	7	6
Case 6: The same as case 4 with $p = 1000$ and $n = 400$						
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1$	20.88	0.99	20.94	1.1	20.94	1.29
$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2$	93.53	0.39	104.76	0.44	94.02	1.35
P_1	0	1	0	1	0	0.99
P_2	0	1	0	1	0	0.99
MMS	16	5	8	5	16	5

<u>ΠΙΝΑΚΑΣ 4δ</u>. Αποτελέσματα για τις Περιπτώσεις 5 και 6.

Για να ολοκληρώσουμε την ενότητα της προσομοίωσης, θα αναφερθούμε στη δυσκολία των προσομοιωμένων μοντέλων μας να μας δείξουν την κατανομή μεταξύ 100 προσομοιώσεων, της ελάχιστης $|t|$-στατιστικής για τις εκτιμήσεις των πραγματικών μη-μηδενικών συντελεστών παλινδρόμησης στο *Oracle* μοντέλο, το οποίο περιλαμβάνει μόνο αληθινές επεξηγηματικές μεταβλητές. Πιο συγκεκριμένα, κατά τη διάρκεια κάθε επανάληψης κάθε προσομοίωσης, παριστάνουμε ότι γνωρίζουμε το σύνολο δεικτών M^*, το πραγματικό υποκείμενο αραιό μοντέλο που ταιριάζει στο μοντέλο αναλογικού κινδύνου του Cox, χρησιμοποιώντας μόνο τις επεξηγηματικές μεταβλητές με δείκτες στο M^* καλώντας τη συνάρτηση "coxph" του πακέτου «επιβίωσης» R, και αναφέρουμε τη μικρότερη απόλυτη τιμή της *t*-στατιστικής για τις εκτιμήσεις παλινδρόμησης. Για παράδειγμα, στην Περίπτωση 1, το μέγεθος μοντέλου είναι μόνο 6 και το ελάχιστο $|t|$-στατιστικής υπολογίζεται βασιζόμενο σ' αυτές τις έξι εκτιμήσεις της κάθε προσομοίωσης. Αυτό δείχνει τη δυσκολία να ανακτήσουμε όλες τις σημαντικές μεταβλητές ακόμα και στο μοντέλο *Oracle* με το ελάχιστο μέγεθος μοντέλου. Το αντίστοιχο boxplot για κάθε περίπτωση εμφανίζεται στο Σχήμα 4.2. Για να αποδείξουμε την επίδραση του να συμπεριλάβουμε τις μη-σημαντικές μεταβλητές, η μικρότερη/ ελάχιστη $|t|$-στατιστική για τις εκτιμήσεις των πραγματικών μη-μηδενικών συντελεστών παλινδρόμησης υπολογίζεται εκ νέου και φαίνεται από τα boxplots στο Σχήμα 4.1.(α) για το μοντέλο με τις πραγματικές σημαντικές μεταβλητές και 20 μη σημαντικές μεταβλητές.

Όπως ήταν αναμενόμενο, οι Περιπτώσεις 1 και 2 είναι σχετικά εύκολες υποθέσεις, ενώ οι Περιπτώσεις 3 και 4 είναι σχετικά πιο δύσκολες στο μοντέλο Oracle. Όταν δεν είμαστε στο μοντέλο *Oracle* με 20 μεταβλητές θορύβου να έχουν προστεθεί, η δυσκολία αυξάνεται, όπως φαίνεται στο Σχήμα 4.1.(β) Έχει μεγαλύτερη επίδραση στις περιπτώσεις 3, 4 και 6.



<u>Σχήμα 4.1.</u> Το boxplot της ελάχιστης $|t|$-στατιστική στα μοντέλα α) μεταξύ 100 προσομοιώσεων και β) όταν έχουν προστεθεί 20 μεταβλητές θορύβου.

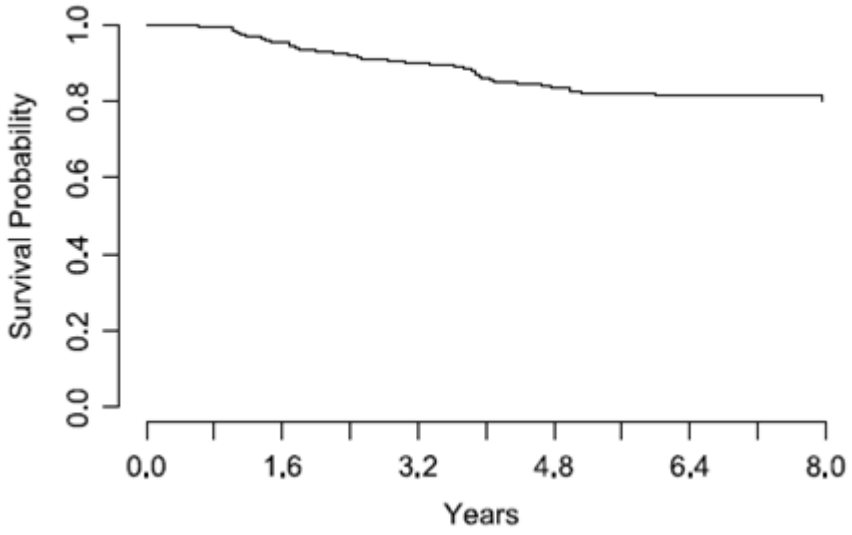
4.6. Πραγματικά δεδομένα

Σε αυτή την ενότητα, χρησιμοποιούμε ένα πραγματικό σύνολο δεδομένων για να δείξουμε την ισχύ της προτεινόμενης μεθόδου. Το σύνολο δεδομένων νευροβλαστώματος οφείλεται στον Oberthuer et al (2006). Χρησιμοποιήθηκε για μελέτες ταξινόμησης. Το νευροβλάστωμα είναι ένας εξωκρανιακός στερεός καρκίνος. Είναι πιο συχνό στην παιδική και νηπιακή ηλικία. Ο ετήσιος αριθμός περιστατικών είναι περίπου αρκετές εκατοντάδες στις Ηνωμένες Πολιτείες. Το νευροβλάστωμα είναι ένας κακοήθης όγκος της παιδιατρικής που προέρχεται από νευρικά ακρολοφία στοιχεία του σπλαχνικού νευρικού συστήματος. Η μελέτη περιλαμβάνει 251 ασθενείς των Γερμανικών δοκιμών νευροβλαστώματος NB90-NB2004, οι οποίοι είχαν διαγνωστεί μεταξύ 1989 και 2004. Οι ηλικίες των ασθενών ποικίλλουν από 0 μέχρι 296 μήνες στη διάγνωση με μέση ηλικία 15 μηνών. Τα δείγματα των νευροβλαστωμάτων των 251 ασθενών αναλύθηκαν με τη χρήση προσαρμοσμένων μικροσυστοιχιών ολιγονουκλεοτιδίων. Ο στόχος είναι να μελετηθεί η σύνδεση της γονιδιακής έκφρασης με μεταβλητές κλινικές πληροφορίες όπως ο χρόνος επιβίωσης και η επιβίωση ατόμων για τρία χρόνια χωρίς τη νόσο, μεταξύ άλλων.

Έχουμε λάβει τα δεδομένα για το νευροβλάστωμα από τον Ποιοτικό Έλεγχο Μικροσυστοιχιών φάσης II (MAQC-II) που διεξήχθη από την FDA (Food Drug Administration). Το πλήρες σύνολο δεδομένων περιλαμβάνει την έκφραση γονιδίων σε 10.707 έρευνες. Περιλαμβάνει επίσης τα στοιχεία επιβίωσης κάθε ασθενή. Σε αυτό το παράδειγμα, επικεντρωνόμαστε στη συνολική επιβίωση. Υπάρχουν πέντε διατάξεις στοιχείων ακραίων τιμών. Αφού αφαιρεθούν οι διατάξεις στοιχείων ακραίων τιμών από την εξέταση, υπάρχουν 246 ασθενείς. Οι πληροφορίες επιβίωσης είναι διαθέσιμες για όλους τους 246 ασθενείς. Το ποσοστό διαγραφής είναι 205/246, το οποίο είναι πολύ μεγάλο. Οι χρόνοι επιβίωσης των 246 ασθενών συνοψίζονται στο Σχήμα 4.2.

Επειδή τα πραγματικά δεδομένα είναι πάντα πολύπλοκα, μπορεί να υπάρχουν κάποια γονίδια που είναι οριακά μη σημαντικά αλλά δουλεύουν από κοινού με άλλα γονίδια. Έτσι, είναι σκόπιμο να εφαρμόζεται η επαναληπτική SIS αντί της SIS, δεδομένου ότι η πρώτη είναι πιο ισχυρή. Τυποποιούμε κάθε επεξηγηματική

μεταβλητή να έχει μέση τιμή μηδέν και τυπική απόκλιση 1 κι εφαρμόζουμε τη Van-ISIS στα τυποποιημένα δεδομένα με $d = \left\lfloor \frac{n}{\log n} \right\rfloor = 43$. Η ISIS ακολουθούμενη από τη SCAD ποινικοποιημένη παλινδρόμηση Cox επιλέγει 8 γονίδια με τα ονόματα έρευνας: A_23_P31816, A_23_P31816, A_23_P31816, A_32_P424973, A_32_P159651, Hs61272.2, Hs13208.1, και Hs150167.1.



Σχήμα 4.2. Η συνάρτηση εκτιμώμενης επιβίωσης για 246 ασθενείς.

Τώρα προσπαθούμε να κατανοήσουμε τη σημαντικότητα αυτών των επιλεγμένων γονιδίων στην πρόβλεψη των πληροφοριών επιβίωσης σε σύγκριση με άλλα γονίδια που δεν έχουν επιλεγεί. Αρχικά, εφαρμόζουμε το μοντέλο αναλογικού κινδύνου του Cox σε όλα αυτά τα 8 γονίδια. Οι εκτιμώμενοι συντελεστές που δίνονται στον Πίνακα 4ε, εκτιμούν την αρχική συνάρτηση επιβίωσης που παριστάνεται στο Σχήμα 4.2, και την αντίστοιχη λογαριθμική log-(μερική) πιθανοφάνεια, η οποία είναι -129.3517. Η log-πιθανοφάνεια που αντιστοιχεί στο null μοντέλο χωρίς καμία επεξηγηματική μεταβλητή είναι -215.4561. Μια δοκιμή χ^2 δείχνει την προφανή σημαντικότητα του μοντέλου με τα 8 επιλεγμένα γονίδια. Ο Πίνακας 4ε δείχνει ότι υπάρχουν δύο εκτιμημένοι συντελεστές που είναι στατιστικά μη-σημαντικοί σε επίπεδο σημαντικότητας α=1%.

Probe ID	Estimated coefficient	Standard error	p-value
A_23_P31816	0.864	0.203	2.1e−05
A_23_P31816	−0.940	0.314	2.8e−03
A_23_P31816	−0.815	1.704	6.3e−01
A_32_P424973	−1.957	0.396	7.8e−07
A_32_P159651	−1.295	0.185	2.6e−12
Hs61272.2	1.664	0.249	2.3e−11
Hs13208.1	−0.789	0.149	1.1e−07
Hs150167.1	1.708	1.687	3.1e−01

ΠΙΝΑΚΑΣ 4ε. Εκτιμημένοι συντελεστές για τα δεδομένα Νευροβλαστώματος.

Στη συνέχεια, για κάθε ένα από αυτά τα 8 γονίδια, το αφαιρούμε, εφαρμόζουμε το αναλογικό μοντέλο κινδύνου του Cox για τα υπόλοιπα 7 γονίδια και παίρνουμε την αντίστοιχη log-πιθανοφάνεια. Οι 8 λογαριθμικές πιθανοφάνειες είναι : -137.5785, -135.1846, -129.4621, -142.4066, -156.4644, -158.3799, -141.0432 και -129.8390. Ο μέσος όρος τους είναι -141.2948, μια μείωση της λογαριθμικής πιθανοφάνειας κατά 11.9431, που είναι πολύ σημαντική με τη μείωση ενός γονιδίου (μείωση του βαθμού ελευθερίας κατά 1). Σε σύγκριση με το μοντέλο των 8 επιλεγμένων γονιδίων, το χ^2 τέστ δείχνει τη σημαντικότητα για όλα τα επιλεγμένα γονίδια, εκτός από τα A_23_P31816 και Hs150167.1. Αυτό ταιριάζει με τις *p*-values που αναφέρθηκαν στον Πίνακα 4ε.

Τέλος, επιλέγουμε τυχαία 2 γονίδια από τα γονίδια που δεν έχουν επιλεγεί, εφαρμόζουμε το αναλογικό μοντέλο κινδύνου του Cox με τα παραπάνω 8 γονίδια μαζί με τα δύο γονίδια που έχουμε επιλέξει και καταγράφουμε την αντίστοιχη λογαριθμική–πιθανοφάνεια. Επαναλαμβάνουμε αυτή τη διαδικασία 20 φορές. Βρίσκουμε ότι ο μέσος όρος αυτών των 20 νέων λογαριθμικών–πιθανοφανειών είναι -128.3933, μια αύξηση της λογαριθμικής-πιθανοφάνειας μόνο κατά 0.9584 με δύο επιπλέον μεταβλητές να περιλαμβάνονται. Συγκρίνοντας με το μοντέλο των 8 επιλεγμένων μεταβλητών, το χ^2 τεστ δε δείχνει καμία σημαντικότητα για το μοντέλο που αντιστοιχεί σε οποιαδήποτε από τις 20 επαναλήψεις.

Τα παραπάνω πειράματα δείχνουν ότι τα επιλεγμένα 8 γονίδια είναι πολύ σημαντικά. Διαγράφοντας ένα, μειώνεται αρκετά η λογαριθμική-πιθανοφάνεια, ενώ προσθέτοντας δύο τυχαία γονίδια δεν αυξάνεται πολύ η λογαριθμική-πιθανοφάνεια.

4.7. Συμπεράσματα

Έχει αναπτυχθεί μια τεχνική επιλογής μεταβλητών για την ανάλυση επιβίωσης, με τη διάσταση να είναι πολύ μεγαλύτερη από το μέγεθος του δείγματος. Επικεντρωνόμαστε στην επαναληπτική SIS, η οποία εφαρμόζει επαναληπτικά μια μεγάλη κλίμακα κρησαρίσματος που φιλτράρει μη σημαντικές μεταβλητές, χρησιμοποιώντας την υπό όρους οριακή χρησιμότητα και μιας μέτριας κλίμακας επιλογή χρησιμοποιώντας τη μέθοδο ποινικοποιημένης μερικής πιθανοφάνειας, η οποία επιλέγει περαιτέρω τις αφιλτράριστες μεταβλητές. Η μεθοδολογική δύναμη της Vanilla ISIS έχει αποδειχθεί μέσω προσεχτικών σχεδιασμών για μελέτες προσομοίωσης. Έχει σίγουρα την ιδιότητα του ανεξάρτητου κρησαρίσματος με πολύ μικρή επιλογή σφάλματος. Συγκρίνοντας την με την έκδοση της LASSO που χρησιμοποιήσαμε, είναι πολύ πιο αποδοτική υπολογιστικά και πολύ πιο συγκεκριμένη στην επιλογή σημαντικών μεταβλητών. Ως αποτέλεσμα, έχει πολύ μικρότερη απόλυτη απόκλιση σφάλματος, καθώς και μικρότερο μέσο τετραγωνικό σφάλμα.

ΠΑΡΑΡΤΗΜΑ

1.Ανάλυση επιβίωσης

Η ανάλυση δεδομένων διάρκειας ζωής είναι ο κλάδος της Στατιστικής που ασχολείται με δεδομένα, τα οποία αντιπροσωπεύουν χρόνο μέχρι την εμφάνιση ενός γεγονότος. Όταν ασχολούμαστε με τεχνικά συστήματα είναι γνωστή ως θεωρία αξιοπιστίας (*Reliability Theory*), ενώ όσον αφορά τις βιοϊατρικές εφαρμογές ως ανάλυση επιβίωσης (*Survival Analysis*).

Το χαρακτηριστικό γνώρισμα των χρόνων επιβίωσης (*Survival Analysis*) είναι ότι δεν ακολουθούν σχεδόν ποτέ την κανονική κατανομή. Αυτή είναι μια από τις αιτίες που χρησιμοποιούνται διαφορετικές μέθοδοι στατιστικής ανάλυσης από τις συνηθισμένες.Γενικά, η ανάλυση επιβίωσης εστιάζεται στην εκτίμηση της πιθανότητας επιβίωσης ενός ατόμου για ένα δεδομένο χρονικό διάστημα.

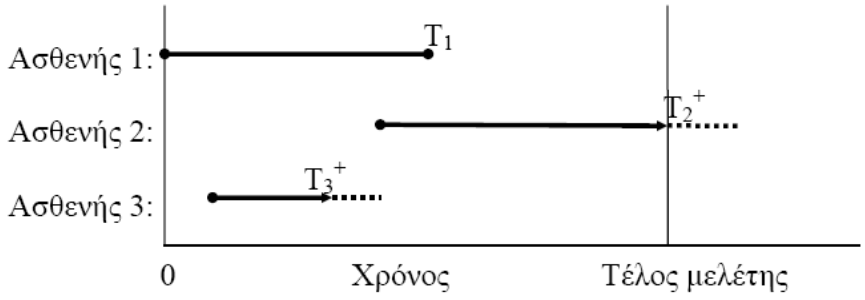
Ένας επιπλέον λόγος, εξαιτίας του οποίου, τα δεδομένα επιβίωσης δεν αναλύονται με την χρήση συνηθισμένων στατιστικών τεχνικών, είναι ότι οι χρόνοι επιβίωσης ορισμένων παρατηρήσεων είναι *αποκομμένοι* (*censored*).

2. Αποκομμένα δεδομένα

Όπως αναφέραμε προηγουμένως, ένα χαρακτηριστικό των δεδομένων επιβίωσης το οποίο και δεν επιτρέπει τη χρήση των συνηθισμένων στατιστικών τεχνικών, είναι η ύπαρξη αποκομμένων παρατηρήσεων. Η συνήθης αιτία που συμβαίνει αυτό, είναι ότι αρκετές φορές, δεν εισάγονται στη μελέτη όλα τα άτομα την ίδια χρονική στιγμή. Αυτό έχει ως αποτέλεσμα, ο χρόνος παρακολούθησης ορισμένων εξ αυτών να μην επαρκεί ώστε να παρουσιαστεί το υπό μελέτη γεγονός (π.χ. θάνατος). Πριν ορίσουμε τα είδη αποκοπής, θα δώσουμε ένα παράδειγμα για να γίνει περισσότερο κατανοητή η έννοια της αποκοπής.

Παράδειγμα 1

Έστω ότι διεξάγεται μια έρευνα όπου μελετάται η αποτελεσματικότητα μιας νέας θεραπείας για μια ασθένεια. Μας ενδιαφέρει ο χρόνος επιβίωσης (σε ημέρες) των συμμετεχόντων στη μελέτη. Η διάρκεια της είναι προκαθορισμένη, ενώ δεν εισήλθαν όλοι οι ασθενείς την ίδια χρονική στιγμή σε αυτήν. Έτσι, καταγράφεται για κάθε ασθενή, ο χρόνος από την είσοδό του στη μελέτη μέχρι το θάνατό του, οπότε και έχουμε πλήρη χρόνο (μη αποκομμένη παρατήρηση). Αν δεν παρουσιαστεί θάνατος για κάποιον ασθενή πριν λήξει η μελέτη ή αν για κάποιο λόγο σταματήσει η παρακολούθησή του, καταγράφεται η τελευταία χρονική στιγμή κατά την οποία είχαμε στη διάθεση μας πληροφορία για τον ασθενή. Οι τελευταίες αυτές περιπτώσεις αποτελούν αποκομμένες παρατηρήσεις. Παραθέτουμε και το παρακάτω γράφημα Α, για να γίνουν περισσότερο κατανοητά τα όσα προαναφέραμε. Σε αυτό φαίνονται οι χρόνοι επιβίωσης τριών ασθενών.



ΓΡΑΦΗΜΑ Α

Παρατηρούμε τα εξής: Ο ασθενής 1, εισήλθε στη μελέτη ακριβώς τη στιγμή της έναρξής της και πεθαίνει τη χρονική στιγμή T_1 οπότε δίνει μη αποκομμένη παρατήρηση. Ο ασθενής 2, εισήλθε αργότερα στη μελέτη και στο τέλος της βρίσκεται ακόμα εν ζωή. Οπότε πρόκειται για αποκομμένη παρατήρηση. Ο ασθενής 3, εισήλθε και αυτός μετά από κάποιο χρονικό διάστημα από την έναρξη της μελέτης, και τη χρονική στιγμή T_3 χάθηκε από παρακολούθηση, δίνοντας και αυτός αποκομμένη παρατήρηση (οι δύο αποκομμένοι χρόνοι στο σχήμα συμβολίζονται με +).

Να σημειωθεί ότι οι συνηθέστεροι λόγοι παύσης της παρακολούθησης (οπότε έχουμε αποκομμένη παρατήρηση) είναι οι εξής: Καταρχήν υπάρχει η περίπτωση της αυτόβουλης απόσυρσης του ασθενή από τη μελέτη, λόγω μετακόμισής του, λόγω αλλαγής του θεράποντα ιατρού, ή και για άλλους προσωπικούς λόγους. Οπότε αναγκαστικά χάνεται η επαφή μαζί του. Υπάρχει επίσης το ενδεχόμενο, μια θεραπεία να έχει πολλές παρενέργειες, οπότε και καθίσταται αναγκαία η απομάκρυνσή του ασθενή από τη θεραπεία και κατ' επέκταση από τη μελέτη. Τέλος, δε θα πρέπει να ξεχνάμε και την περίπτωση θανάτου του ασθενή από αίτια που δεν έχουν σχέση με την όλη μελέτη.

Όσον αφορά τα είδη αποκοπής, αυτά είναι τρια. Η δεξιά αποκοπή (*right-censoring*), η αριστερή (*left censoring*) και η αποκοπή διαστήματος (*interval censoring*). Επιπλέον, προκύπτουν τρεις μηχανισμοί αποκοπής παρατηρήσεων. Η αποκοπή τύπου I (*type I censoring*), η αποκοπή τύπου II (*type II censoring*) και η αποκοπή τύπου III (*type III censoring*).

Ορισμός 1: Δεξιά αποκοπή, έχουμε όταν ο χρόνος επιβίωσης είναι μεγαλύτερος από τον χρόνο λήξης της μελέτης ή γενικότερα μεγαλύτερος από κάποιο χρονικό όριο (π.χ. τη στιγμή που για κάποιο λόγο χάθηκε η επαφή μαζί του). Αυτό σημαίνει ότι προφανώς δεν είναι γνωστός, αλλά τουλάχιστον ίσος με τη διάρκεια παραμονής του ατόμου στη μελέτη.

Ορισμός 2: Η αριστερή αποκοπή συμβαίνει όταν ο πραγματικός χρόνος επιβίωσης είναι μικρότερος από τον παρατηρούμενο δηλαδή ξέρουμε πως το άτομο

είχε ήδη πεθάνει σε χρόνο $t_0 + c$ αλλά το ακριβές χρονικό σημείο $t_0 + t$ (όπου $t < c$ και t_0 η χρονική στιγμή ένταξής του στη μελέτη) που συνέβη αυτό είναι άγνωστο.

Ορισμός 3: Η αποκοπή διαστήματος, παρατηρείται όταν ξέρουμε πως το υπό μελέτη συμβάν έχει πραγματοποιηθεί σε ένα διάστημα και πάλι όμως χωρίς να είναι γνωστό το ακριβές σημείο. Αυτό παρατηρείται συνήθως όταν έχουμε περιοδική παρακολούθηση του ασθενή.

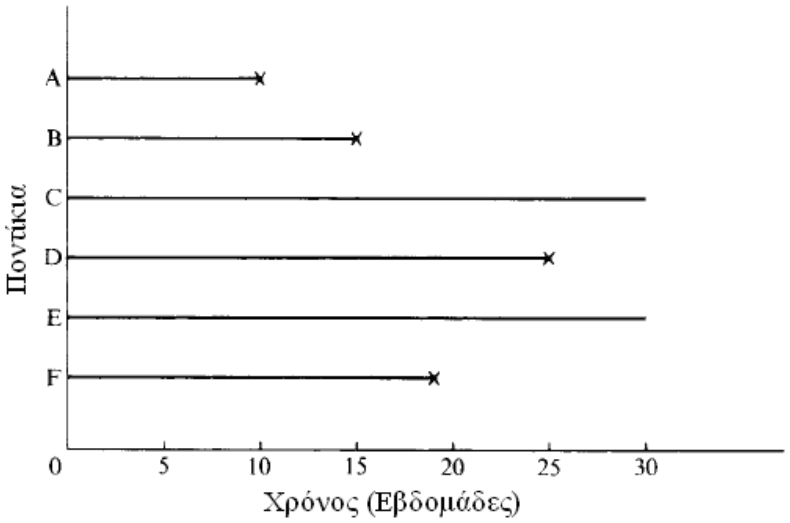
Ορισμός 4: Στην αποκοπή τύπου Ι, η διάρκεια της μελέτης είναι προκαθορισμένη και ίση έστω με t^*. Οπότε είτε γνωρίζουμε τη διάρκεια ζωής, αν $T < t^*$, αλλιώς το μόνο που ξέρουμε είναι ότι έχει υπερβεί το t^* ($T > t^*$).

Ορισμός 5: Στην αποκοπή τύπου ΙΙ, η μελέτη συνεχίζεται εως ότου παρουσιαστεί το γεγονός που μας ενδιαφέρει σε k το πλήθος άτομα, με k προκαθορισμένο.

Ορισμός 6: Ο τελευταίος μηχανισμός αποκοπής τύπου ΙΙΙ (ή τυχαίας αποκοπής), εμφανίζεται στις περισσότερες κλινικές και επιδημιολογικές μελέτες. Το κύριο χαρακτηριστικό τους είναι ότι η διάρκεια της μελέτης είναι προκαθορισμένη, ενώ οι ασθενείς δεν εισέρχονται την ίδια χρονική στιγμή σε αυτήν (βλ. προηγούμενο παράδειγμα). Αποτέλεσμα αυτού, είναι οι χρόνοι αποκοπής να είναι τυχαίοι (*random censoring).* Να σημειωθεί ότι και οι τρεις αυτοί μηχανισμοί ανήκουν στην κατηγορία της δεξιάς αποκοπής.

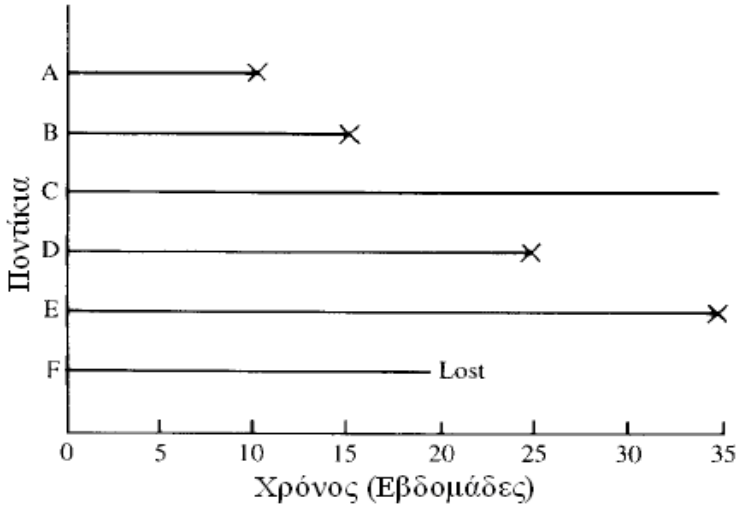
Παράδειγμα 2

Τελειώνουμε αναφέροντας ένα ακόμα παράδειγμα, όπου διαφαίνεται τόσο η περίπτωση της (δεξιάς) αποκοπής όσο και οι δύο πρώτοι μηχανισμοί αποκοπής. Έστω ότι 6 ποντίκια (A, B, C, D, E, F) υποβάλλονται σε διαδικασία καρκινογένεσης με εμβολιασμό καρκινικών κυττάρων την ίδια χρονική στιγμή. Αυτό που ενδιαφέρει τον ερευνητή είναι ο χρόνος που απαιτείται για την ανάπτυξη όγκου προκαθορισμένου μεγέθους. Οπότε το υπό μελέτη συμβάν το οποίο και πρέπει να παρουσιαστεί πριν το πέρας της μελέτης, ώστε να έχουμε πλήρη χρόνο αποτυχίας, είναι η δημιουργία όγκου. Η διάρκεια της μελέτης είναι προκαθορισμένη στις 30 εβδομάδες. Στο παρακάτω γράφημα Β, βλέπουμε ότι τα ποντίκια A, B και D ανέπτυξαν όγκο έπειτα από 10, 15 και 25 εβδομάδες αντίστοιχα (οι χρόνοι αυτοί είναι πλήρεις χρόνοι αποτυχίας), ενώ τα ποντίκια C και E δεν ανέπτυξαν όγκο κατά τη διάρκεια της μελέτης, άρα οι χρόνοι ανάπτυξης όγκου δεν είναι γνωστοί (δεξιά αποκομμένες παρατηρήσεις). Το ποντίκι F πέθανε ξαφνικά έπειτα από 19 εβδομάδες παρακολούθησης (χωρίς να έχει αναπτύξει κάποιον όγκο), άρα δίνει και αυτό αποκομμένη παρατήρηση. Έτσι, τα δεδομένα επιβίωσης είναι 10, 15, 30+, 25, 30+ και 19+ εβδομάδες. Τα αποκομμένα δεδομένα στην περίπτωση αυτή είναι τύπου I και συμβολίζονται με “+” [25].



ΓΡΑΦΗΜΑ Β

Στην περίπτωση όμως, που ο ερευνητής ήθελε να σταματήσει τη μελέτη τη στιγμή που 4 ποντίκια εμφανίσουν όγκο (γράφημα Γ), τα δεδομένα που θα έπαιρνε θα ήταν 10, 15, 35+, 25, 35 και 19+ εβδομάδες και η αποκοπή θα ήταν τύπου II.



ΓΡΑΦΗΜΑ Γ

ΒΙΒΛΙΟΓΡΑΦΙΑ

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Statist. Math.*, **21**, pp. 243-247.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, pp. 716-723.

Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, **22**, pp. 162-176.

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables.*Technical Report No. 23*.Department of Statistics.University of Kentuky.

Andrews, D. F. and Herzberg, A. M. (1985).*Data*, Springer, New York.

Antoniadis, A. and Fan, J. (2001) Regularization of wavelets approximation (with discussion). *J. Am. Statist. Ass*, **96**, 939-967.

Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion). *J. Italian Statist. Assoc.*, **6**, pp. 97-144.

Barron, D., Wakin, M. B., Duarte, M. F., Sarvotham, S. and Baraniuk, R. G. (2005) Distributed compressed sensing. *Manuscript*.

Barron, A., Cohen, A., Dahmen, W. and DeVore, R. (2008) Approximation and learning by greedy algorithms.*Ann.Statist.*,**36**, 64-94.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2008) Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist., **36**, in the press.

Bickel, P. J. (1975), One-Step Huber Estimates in Linear Models, *Journal of the American Statistical Association*, **70**, pp. 428–433.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression.*Journal of the American Statistical Association*, **87**, pp. 738-754.

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote, *Technometrics*, **37**, pp.373–384.

Breslow, N.E. (1974). Covariance analysis of censored survival data.*Biometrics*, **20**, pp. 89-100.

Bryant, P.G. and Cordero-Brana, O.I. (2000). Model selection using the minimum description length principle.*The Amer. Statist*., **54**, pp. 257-268.

Eksioglu B., Demirer R. and Capar I. (2005). Subset selection in multiple linear regression: a new mathematical programming approach. Computers & Industrial Engineering 49(1), 155-167

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when *p* is much larger than *n* (with discussion). *Ann. Statist*., **35**, 2313-2404.

Cox, D. R. (1972) Regression models and life-tables (with discussion).*J. Roy.Statist. Soc. Ser. B*, **34**, 187-220.

Cox, D. R. (1975) Partial likelihood.*Biometrika*, **62**, 269-276.

Donoho, D. L. (2000) High-dimensional data analysis Q the curses and blessings of dimensionality.*American Mathematical Society Conf. Math Challenges of the 21st Century*.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaption by wavelet shrinkage.*Biometrika*, **81**, 425-455.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). *Ann Statist*., **32**, 407-499.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, pp. 557-565.

Fan, J. (1997) Comments on "Wavelets in statistics: a review," by Antoniadis, *J. Ital. Statist. Ass.*,**6**, 131-138.

Fan, J. and Fan, Y. (2008) High dimensional classification using features annealed independence rules. *Ann. Statist.*, to be published.

Fan, J., Feng, Y. and Song, R. (2010) Nonparametric independence screening in sparse ultra-high dimensional additive models. Submitted

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Statist. Ass*., **96**, 1348-1360.

Fan, J. and Li, R. (2002) Variable selection for Cox's proportional hazards model and frailty model.*Ann.Statist*., **30**, 74-99.

Fan, J. and Li, R. (2006) Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proc. Int. Congr. Mathematicians* (eds M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), vol.III, pp. 595-622. Freiburg: European Mathematical Society.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B*, **70**, 849-911.

Fan, J. and Peng, H. (2004) Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928-961.

Fan, J. and Ren, Y. (2006) Statistical analysis of DNA microarray data. *Clin. Cancer Res.*, **12**, 4469-4473.

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional variable selection: beyond the linear model. *J. March. Learn. Res*.Toappear.

Fan, J. and Song, R. (2010) Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist*. To appear.

Faraggi, D. and Simon, R. (1998) Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475-5

Frank, I. E., and Friedman, J. H. (1993), A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, pp. 109–148.

Freund, Y. and Schapire, R. E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting.*J.Comput. Syst. Sci.*,**55**, 119-139.

Fu, W. J. (1998). Penalized Regression: The Bridge Versus the LASSO.*Journal of Computational and Graphical Statistics*, **7**, pp. 397–416.

Grambsch, P.M. and Therneau, T.M. (1994). Proportional hazards tests and

diagnostics based on weigthed residuals. *Biometrika*, **81**, pp. 515-526.

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

Gribonval, R., Mailhe, B., Rauhut, H., Schnass, K. and Vandergheynst, P. (2007) Average case analysis of multi-channel thresholding.In*Proc. Int. Conf. Acoustic and Speech Signal Processing*. New York: Institute of Electrical and Electronics Engineers.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection.*Statist.Sin.*,**7**, 339-373.

Hannan, E.J. and Quinn, B. G. (1979).The determination of the order of an autoregression.*J. Roy. Statist. Soc*., **41**, pp. 190-195.

Hardin, J.W. and Hilbe, J.M. (2007) *Generalized Linear Models and Extensions*, 2nd Ed., College Station, TX: Stata Press.

Huang, J., Horowitz, J. and Ma, S. (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.,* **36**, 587-613.

Huber P. (1981). *Robust Estimation*.Wiley, New York.

Hunter, D. and Li, R. (2005) Variable selection using MM algorithms.*Ann.Statist*.,**33**, 1617-1642.

Hurvich, C. M. and Tsai, C-L.(1989). Regression and time series model selection in small samples.*Biometrika*, **76**, pp. 297-307.

Ibrahim, J. G., Chen, M.-H.andMaceachern, S. N. (1999) Bayesian variable selection for proportional hazards models.*Canand. J. Statist.*, **27**, 70117

Klein, J. P. and Moeschberger, M. L. (2005) *Survival Analysis*, 2nd ed. Springer

Konishi, S. andKitagama, G. (2008).*Information Criteria and Statistical Modeling*,Springer, New York.

Koukouvinos, C., Mylona, K. and Vonta, F. (2008). A Comparative study of variable selection procedures applied in high dimensional medical problems. *Journal of Applied Probability & Statistics*, **3**, pp. 195-209.

Lam, C. and Fan, J. (2007) Sparsistency and rates of convergence in large covariance matrices estimation. *Manuscript*.

Lawson, C. L. and Hanson, R. J. (1974).*Solving least-squares problems*. Prentice Hall, New Jersey.

Li, Y. and Dicker, L. (2009) Dantzig selector for censored linear regression. Technical report, Harvard Univ. Biostatistics.

Linhart, H. and Zucchini, W. (1986).*Model selection*, John Wiley, New York.

Mallows, C. L. (1973). Some comments on C_p. Technometrics, **15**, pp. 661-675.

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H., and Patil, P. (1998). Exact Risk Analysis of Wavelet Regression.*JournalComputationaland Graphical Statistics*, **7**, pp. 278–309.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd ed.,Chapman and Hall, London.

Meier, L., Van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *J. R. Statist. Soc*. B, **70**, 53-71.

Meinshausen, N. (2007) Relaxed Lasso. *Computnl Statist. Data Anal.*,**52**, 374-393.

Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso.*Ann. Statist.*,**34**, 1436-1462.

Meinshausen, N., Rocha, G. and Yu, B. (2007) Discussion of "The Dantzig selector: statistical estimation when *p* is larger than *n*". *Ann. Statist.*,**35**, 2373-2384.

Myers, R. H. and Montgomery, D. C. (1995) Response surface methodology: Process and product optimization using designed experiments. *New York, John Wiley & Sons*, 244-264.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., König, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006) Customized oligonucleotide microarray gene expression based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology*, **24**, 5070-5078.

Paul, D., Bair, E., Hastie, T. and Tibshirani, R. (2008) "Pre-conditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.,* to be published.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2007) Sparse additive models. *Manuscript*.

Rissanen, J. (1978). Modelling by shortest data description.*Automatica*, **14**, pp. 465-471.

Robinson, P. M. (1988). The Stochastic Difference BetweenEconometricsand Statistics.*Econometrica*, **56**, pp. 531–547.

Sauerbrei, W. and Schumacher, M. (1992) A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statist. Med.*,**11**, 2093-2109.

Schoenfeld, D. A. (1982).Partial residuals for the proportional hazards regression model.*Biometrika*, **69**, pp. 239-241.

Schwartz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, pp. 461-464.

Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genome-wide studies.*Proc.Natn.Acad.Sci. USA*, 100, 9940-9445.

Therneau, T. M., Grambsch, P.M. and Fleming, T. R. (1990).Martingale-based residuals for survival models.*Biometrika*, **77**, pp. 147-160.

Tibshirani, R. J. (1997). The Lasso method for variable selection in the Cox model.*StatisticsinMedicine*, **16**, pp. 385-395.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO.*J. Roy. Statist. Soc. Ser. B*., **58**, pp. 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natn. Acad. Sci. USA*, **99**, 6567-6572.

Wu, Y. and Liu, Y. (2009) Variable selection in quantileregression.*Statist.Sinica*, **19**, 801-817.

Zhang, C.-H. (2009) Penalized linear unbiased selection. *Ann. Statist*. To appear.

Zhang, C.-H. (2007) Penalized linear unbiased selection. *Technical Report 2007-003*.Department of Statistics, Rutgers University, Piscataway.

Zhang, C.-H.and Huang, J. (2008) Thesparsity and bias of the LASSO selection in high-dimensional linear regression.*Ann. Statist.*,**36**, 1567-1594.

Zhang, H. H. and Lu, W. (2007) Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691-703.

Zhao, P. and Yu, B. (2006) on model selection consistency of Lasso.*J. Mach. Learn. Res.*,**7**, 2541-2567.

Zou, H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Statist. Ass.*,**101**, 1418-1429.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*,**36**, 1509-1566.

Zou, H. and Hastie, T. (2005) Regylarization and variable selection via the elastic net.*J. Roy. Statist. Soc. Ser. B*, **67**, 301-320.