ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Τομέας Μαθηματικών

Διπλωματική Εργασία

"Επιλογή της ρυθμιστικής παραμέτρου στις μεθόδους ποινικοποιημένης πιθανοφάνειας"

Μπουλουγούρη Χριστίνα

Επιβλέπων: Κουκουβίνος Χρήστος, Καθηγητής ΕΜΠ

Αθήνα, Ιανουάριος 2014

ΠΕΡΙΛΗΨΗ

Βασικός στόχος της στατιστικής είναι η εκτίμηση και η περιγραφή της σχέσης εξάρτησης μεταξύ μεταβλητών. Μάλιστα, τις περισσότερες φορές η σχέση αυτή είναι στοχαστική, γεγονός που δυσκολεύει την επιλογή του βέλτιστου μοντέλου. Τα τελευταία χρόνια, έχει προταθεί μια νέα μέθοδος ποινικοποιημένης πιθανοφάνειας, που βασίζεται στην εισαγωγή ενός όρου ποινής στην πιθανοφάνεια. Η βασική διαφορά της μεθόδου αυτής είναι το ισχυρό θεωρητικό υπόβαθρο που έχει. Ο όρος ποινής περιλαμβάνει μια παράμετρο την οποία συμβολίζουμε με "λ" και καλούμε "ρυθμιστική παράμετρο". Η βέλτιστη επιλογή της ρυθμιστικής παραμέτρου έχει σαν αποτέλεσμα την καλύτερη απόδοση των μεθόδων αυτών. Στην παρούσα εργασία παρουσιάζεται ο τρόπος της βέλτιστης επιλογής της ρυθμιστικής παραμέτρου στις διάφορες μεθόδους ποινής.

Στο πρώτο κεφάλαιο, κάνουμε μια εισαγωγή στο Γενικό Γραμμικό Μοντέλο και μια σύντομη αναφορα στη Μέθοδο Ελαχίστων Τετραγώνων και τη Μέθοδο Μέγιστης Πιθανοφάνειας. Στο δεύτερο κεφάλαιο, παρουσιάζεται ο τρόπος επιλογής μεταβλητών μέσω της μη-κοίλης ποινοκοποιημένης πιθανοφάνειας. Γίνεται εκτενής αναφορά στην προσέγγιση αυτή καθώς και στις διάφορες συναρτήσεις ποινής που περιλαμβάνουν τη ρυθμιστική παράμετρο. Στο τρίτο κεφάλαιο, αναλύεται η Μέθοδος της Γενικευμένης Διασταυρωμένης Επικύρωσης η οποία είναι μια αποδοτική μέθοδος που εξηγεί τα δεδομένα με θόρυβο, δηλαδή δεδομένα που δεν μπορούν να αναλυθούν. Η μέθοδος αυτή εισάγει μια νέα συνάρτηση η οποία περιέχει την ρυθμιστική παράμετρο. Παρουσιάζεται ο τρόπος επιλογής της βέλτιστης ρυθμιστικής παραμέτρου καθώς και οι θεωρητικές της ιδιότητες. Στο τέταρτο κεφάλαιο, παρουσιάζεται η Μέθοδος SCAD η εφαρμογή της οποίας βασίζεται στην κατάλληλη επιλογή ρυθμιστικής παραμέτρου. Παρουσιάζονται οι δύο βασικοί τρόποι εκτίμησης της παραμετρου "λ" και συγκρίνονται ως προς την αποδοτικότητα τους. Στο πέμπτο κεφάλαιο, αναλύεται το Κριτήριο Γενικευμένης Πληροφορίας που χρησιμοποιείται για την επιλογή της ρυθμιστικής παραμέτρου σε μη-κοίλες συναρτήσεις ποινικοποιημένης πιθανοφάνειας.

ABSTRACT

The main aim of statistics is to estimate and report the dependence n the relation among the variables. Due to the fact that the correlation is commonly stochastic, it is more difficult to choose the optimal model. Over the last years, a new penalized loglikelihood method has been proposed, which is based in the introduction of a penalty term in the likelihood function. The main difference of this method is the strong theoretical background that it imposes. The penalty term contains a parameter, which is represented as "λ" and is called "tuning parameter". As a result, the optimal choice of the tuning parameter has the best efficiency of the method. This paper presents the way choosing the best tuning parameter in various penalized methods.

In the first chapter, we provide an introduction to the General Linear Model and we do a brief report to the Least Squares Method and Likelihood Method. In the second chapter, we present the way of choosing the variable via nonconcave penalized likelihood. This approximation is extensively presented along with the different penalty functions which include the tuning parameter. In the third chapter we define the Generalized Cross Validation Method which is an efficient method for explaining noisy data, which are data that can not be analyzed. This method imports a new function which contains the tuning parameter. We present the way of choosing the optimal tuning parameter and its theoretical properties. In the fourth chapter, we define SCAD which implementation is based in the appropriate choice of tuning parameter. We introduce the two most basic ways of estimating "λ" which are compared in relation to their efficiency. In the last chapter, we elaborate the Generalized Information Criterion which implements the selection of the tuning parameter in nonconcave penalized loglikelihood function.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κύριο Χρήστο Κουκουβίνο για την δυνατότητα που μου έδωσε να ασχοληθώ με την επιστημονική περιοχή που με ενδιαφέρει, καθώς και για την καθοδήγησή του κατά τη διάρκεια εκπόνησης της εργασίας.

Επίσης θα ήθελα να ευχαριστήσω ιδιαιτέρως τον υποψήφιο διδάκτορα Εμμανουήλ Ανδρουλάκη για το ενδιαφέρον και την βοήθεια που μου προσέφερε.

Τέλος οφείλω ένα μεγάλο ευχαριστώ στην οικογένεια μου για την συμπαράσταση και την αδιάκοπη ενθάρρυνση σε κάθε μου βήμα.

Περιεχόμενα

Κεφάλαιο 5: Κριτήριο Γενικευμένης Πληροφορίας

Appendix

ΚΕΦΑΛΑΙΟ 1

Το Γενικό Γραμμικό Μοντέλο

1.1 Εισαγωγή

Η εκτίμηση και η περιγραφή της σχέσης εξάρτησης μεταξύ μεταβλητών, αποτελεί βασικό στόχο σε πολλές επιστήμες. Η σχέση αυτή, τις περισσότερες φορές είναι στοχαστική, με αποτέλεσμα η χρήση στατιστικών μοντέλων να είναι απαραίτητη για την περιγραφή της. Το βασικό εργαλείο για την ανάλυση και την επεξεργασία των δεδομένων που προκύπτουν από τη μελέτη στοχαστικών φαινομένων είναι τα μοντέλα παλινδρόμησης.

Έστω Y η μεταβλητή που μας ενδιαφέρει και $\chi_1, \chi_2, \dots \chi_\kappa$ ένα σύνολο επεξηγηματικών μεταβλητών. Το πρόβλημα έγκειται στην επιλογή των κατάλληλων επεξηγηματικών μεταβλητών, εκείνων δηλαδή που έχουν σημαντική επίδραση στην απόκριση Y. Στην πράξη, ο αριθμός των στατιστικά σημαντικών παραγόντων είναι μικρότερος από το αρχικό υποσύνολό τους, μια ιδιότητα γνωστή ως "αρχή της σποραδικότητας των επιδράσεων" (*sparsity of effects principle*). Το πρόβλημα της επιλογής μεταβλητών είναι αρκετά συνηθισμένο και γι'αυτό το λόγο έχουν αναπτυχθεί διάφορες μέθοδοι και αλγόριθμοι επιλογής μεταβλητών.

1.2 Το Γενικό Γραμμικό Μοντέλο

Στα περισσότερα μοντέλα που συναντάμε και μας ενδιαφέρει να μελετήσουμε, η τιμή της μεταβλητής Y εξαρτάται από περισσότερες από μια επεξηγηματικές μεταβλητές.

Το γενικό γραμμικό μοντέλο δίνεται από τη σχέση:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \qquad (1.2.1)$$

όπου:

- y_i , i=1,...,n είναι οι τιμές των παρατηρήσεων της εξαρτημένης μεταβλητής y

- x_{ij}, i=1,2,...,n, j=1,2,...,k είναι οι τιμές για την i-οστή παρατήρηση των ανεξάρτητων ή επεξηγηματικών μεταβλητών xj
- $\beta_0,\beta_1,...,\beta_\kappa$ οι άγνωστες παράμετροι του μοντέλου
- ε_i, i=1,...,n τα τυχαία σφάλματα τα οποία υποθέτουμε οτι ικανοποιούν τις παρακάτω υποθέσεις:
    - $E(\varepsilon_i)= 0$ για κάθε i
    - $V(\varepsilon_i)= \sigma^2$ για κάθε i, δηλαδή τα τυχαία σφάλματα ικανοποιούν την υπόθεση της ομοσκεδαστικότητας
    - $Cov(\varepsilon_i,\varepsilon_j)= 0$ για $i\neq j$, δηλαδή τα ei είναι ασυσχέτιστα μεταξύ τους

Υπό μορφή πινάκων, η σχέση (1.2.1) γράφεται:

$$\tilde{Y} = \tilde{X}*\tilde{\beta} + \tilde{\varepsilon} \quad (1.2.2)$$

Όπου:

$\tilde{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ το (n×1)-διάνυσμα της μεταβλητής απόκρισης y

$\tilde{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ & \vdots & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$ ο (n×p)-πίνακας με p=k+1, ο οποίος

ονομάζεται πίνακας σχεδιασμού ,

$\tilde{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$ το (p×1)-διάνυσμα των παραμέτρων

$\tilde{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ το (n×1)-διάνυσμα των τυχαίων σφαλμάτων

Το n-διάστατο διάνυσμα ε ακολουθεί την πολυδιάστατη κανονική κατανομή, δηλαδή:

$\tilde{\varepsilon} \sim N_n(0, \sigma^2 I)$, όπου $E(\tilde{\varepsilon})=0$ η αναμενόμενη μέση τιμή των ε

και:

$Var(\tilde{\varepsilon})= E\{(\varepsilon-E(\varepsilon)) *(\varepsilon-E(\varepsilon))'\}= E(\varepsilon*\varepsilon')= \sigma^2 I$, ο πίνακας διασποράς συνδιασποράς των ε, με I τον (n×n)-μοναδιαίο πίνακα.

Υπό την υπόθεση οτι $\tilde{\varepsilon}$ ~N(0,σ^2I), ο γραμμικός μετασχηματισμός $\tilde{Y}=\tilde{X}*\tilde{\beta}$ + $\tilde{\varepsilon}$ θα είναι της πολυμεταβλητής κατανομής,δηλαδή:

$$\tilde{Y} \sim N_n(\tilde{X}\tilde{\beta}\ , \sigma^2 I)$$

1.2.1 Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο ελαχίστων τετραγώνων

Η μέθοδος ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων $\tilde{\beta}$ βασίζεται, όπως και στο απλό γραμμικό μοντέλο, στην ελαχιστοποίηση της παράστασης:

$$S(\tilde{\beta}) = (\tilde{y}\text{-}E(\tilde{y}))' (\tilde{y}\text{-}E(\tilde{y}))= (\tilde{Y} -\tilde{X}\tilde{\beta})'(\tilde{Y}-\tilde{X}\tilde{\beta})$$

Παραγωγίζοντας ως προς $\tilde{\beta}$ έχουμε:

$$\frac{\partial S(\beta)}{\partial \beta} =-2*\tilde{X}'(\tilde{Y}-\tilde{X}\tilde{\beta}\ )$$

Και θέτοντας ίσο με μηδέν καταλήγουμε στη σχέση:

$$\tilde{X}\ *\tilde{y} = \tilde{X}'\ *\tilde{X}*\hat{\tilde{\beta}}$$

Εάν ο ($\tilde{X}'\tilde{X}$) αντιστέφεται, η εκτιμήτρια ελαχίστων τετραγώνων (*Ordinary Least Squares Estimator- OLS*) του διανύσματος $\tilde{\beta}$ δίνεται από τη σχέση:

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

1.2.2 Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο μέγιστης πιθανοφάνειας

Η μέθοδος μέγιστης πιθανοφάνειας προτάθηκε απο τον Fisher (1997). Συγκεκριμένα, έστω πληθυσμός με άγνωστη παράμετρο $\tilde{\theta}=(\theta_1, \theta_2, \dots, \theta_\kappa) \in \Theta$ και συνάρτηση πυκνότητας πιθανότητας f($\tilde{x}|\ \tilde{\theta}$). Σκοπός είναι η εκτίμηση της παραμέτρου $\tilde{\theta}$. Οπότε θεωρούμε ένα τυχαίο δείγμα $\chi_1, \chi_2, \dots, \chi_n$ από τον πληθυσμό. Αν:

f ($x_1|\tilde{\theta}$), f($x_2|\tilde{\theta}$), …, f($x_n|\tilde{\theta}$) είναι η συνάρτηση πυκνότητας πιθανότητας κάθε τιμής του τυχαίου δείγματος, τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας των τυχαίων μεταβλητών $\chi_1, \chi_2,\ldots, \chi_n$ είναι:

$$f(\chi_1, \chi_2,\ldots, \chi_n \,|\tilde{\theta}) = f(\chi_1|\tilde{\theta})* f(\chi_2|\,\tilde{\theta})*....*f(\chi_n|\tilde{\theta}) \qquad (1.2.2.1)$$

Στην περίπτωση συγκεκριμένων παρατηρήσεων $\chi_1, \chi_2,\ldots, \chi_n$ τυχαίου δείγματος, η (1.2.2.1) είναι συνάρτηση μόνο της παραμέτρου $\tilde{\theta}$ και συμβολίζεται ως:

$$L(\tilde{\theta} \mid x_1,x_2,\ldots,x_n) = f(x1|\tilde{\theta})* f(x2|\,\tilde{\theta})*....*f(\,xn|\tilde{\theta}) = \prod_{i=1}^{n} f(xi \mid \widetilde{\theta}) \quad (1.2.2.2)$$

Η (1.2.2.2) καλείται συνάρτηση πιθανοφάνειας (*Likelihood function*) του τυχαίου δείγματος $x_1,x_2,\ldots,x_n$ και εκφράζει το πόσο «πιθανοφανείς», δηλαδή πόσο σύμφωνες με το συγκεκριμένο δείγμα είναι οι διάφορες τιμές της παραμέτρου $\tilde{\theta}$. Η μέθοδος μέγιστης πιθανοφάνειας συνιστάται στην επιλογή της τιμής $\hat{\tilde{\theta}}$, η οποία μεγιστοποιεί τη συνάρτηση πιθανοφάνειας :

$$L(\hat{\tilde{\theta}}|\chi_1, \chi_2, \ldots, \chi_n) = sup_{\theta\in\Theta} L(\tilde{\theta}|\chi_1, \chi_2, \ldots, \chi_n)$$

Η τιμή $\hat{\tilde{\theta}}$ καλείται εκτιμήτρια μέγιστης πιθανοφάνειας της $\tilde{\theta}$. Μεγιστοποίηση της $L(\tilde{\theta} \mid x_1,x_2,\ldots,x_n)$ σημαίνει μεγιστοποίηση της πιθανότητας εμφάνισης των τιμών $x_1,x_2,\ldots,x_n$ στο δείγμα $X_1,X_2,\ldots, X_n$. Η τιμή $\hat{\tilde{\theta}} = (\widehat{\theta 1}, \widehat{\theta 2}, \ldots, \widehat{\theta\kappa})$ βρίσκεται με λύση της εξίσωσης:

$$\frac{\partial Log}{\partial \theta_r} L\left(\tilde{\theta}\middle|\chi_1, \chi_2, \ldots, \chi_n\right) = 0$$

Για να είναι η λύση αυτή σημείο μεγίστου, θα πρέπει ο Εσσιανός Πίνακας

$$\left[\frac{\partial^2 \theta Log L(\tilde{\theta})}{\partial\theta_i \partial\theta_j}\right]_{k*k}$$

να είναι γνήσια αρνητικός.

ΚΕΦΑΛΑΙΟ 2

Επιλογή Μεταβλητών Μέσω Μη Κοίλης Ποινικοποιημένης Πιθανοφάνειας

2.1 Εισαγωγή

Η επιλογή μεταβλητών είναι ένα σημαντικό κομμάτι στην ανάλυση παλινδρόμησης. Στην πράξη, ένας μεγάλος αριθμός μεταβλητων εισάγεται στο αρχικό στάδιο της μοντελοποίησης. Για να ενισχύσουμε την προβλεψιμότητα και να επιλέξουμε σημαντικές μεταβλητές χρησιμοποιούμε διάφορες μεθόδους, πιο σημαντικές από τις οποίες είναι η κατα βήματα επιλογή (*stepwise selection*) και η επιλογή καλύτερου υποσυνόλου (*best subset selection*). Παρ' ότι αυτές οι μέθοδοι είναι πρακτικά χρήσιμες, αγνοούν στοχαστικά σφάλματα που εμφανίζονται κατά το στάδιο επιλογής μεταβλητών. Έτσι, οι θεωρητικές τους ιδιότητες είναι δύσκολο να γίνουν κατανοητές. Επιπλέον, η επιλογή καλύτερου υποσυνόλου (*best subset selection*) παρουσιάζει διάφορα μειονεκτήματα, το σημαντικότερο από τα οποία είναι η έλλειψη ευστάθειας, όπως παρουσιάζεται από τον Breiman (1996).

Οι Fan&Li(2001) [15], στην προσπάθειά τους να αυτοματοποιήσουν και ταυτόχρονα να επιλέξουν μεταβλητές, προτείνουν μια προσέγγιση μέσω ποινικοποιημένων ελαχίστων τετραγώνων (*penalized least squares*). Η μέθοδος αυτή διατηρεί τα καλά χαρακτηριστικά της μεθόδου επιλογής καλύτερου υποσυνόλου (*best subset selection*), αλλά και της παλινδρόμησης κορυφογραμμής (*ridge regression*). Οι συναρτήσεις ποινής είναι μοναδικές καθώς παράγουν σποραδικές λύσεις (πολλοί από τους εκτιμημένους παράγοντες είναι μηδέν), ικανοποιούν συγκεκριμένες συνθήκες με αποτέλεσμα να παράγουν συνεχή μοντέλα και τέλος φράσσονται από μια σταθερά ώστε να παράγουν αμερόληπτους εκτιμητές για μεγάλες παραμέτρους.Η παλινδρόμηση κορυφογραμμής (*ridge regression*) που προτάθηκε από τους Frank&Friedman(1993) [19], αλλά και η μέθοδος Lasso (*Least Absolute Shrinkage & Selection Operator*) που προτάθηκε από τον Tibshirani (1996,1997) αποτελούν κομμάτι των ποινικοποιημένων ελαχίστων τετραγώνων, παρ'όλο που η συνάρτηση ποινής L_q δεν ικανοποιεί τις απαιτούμενες ιδιότητες.

Η ιδέα των ποινικοποιημένων ελαχίστων τετραγώνων μπορεί να επεκταθεί στα μοντέλα που βασίζονται στην πιθανοφάνεια σε πολλά στατιστικά πεδία. Οι συναρτήσεις ποινής που επιλέγονται είναι:

- Συμμετρικές
- Μη κοιλες στο $(0,\infty)$ (*nonconcave*)
- Διακατέχονται από ιδιομορφίες στην αρχή, ώστε να παράγουν σποραδικές λύσεις.Δηλαδή πολλοί εκ των εκτιμηθέντων συντελεστών να είναι ίσοι με μηδέν (*singularities*)

Η βασική διαφορά της μεθόδου αυτής σε σχέση με άλλες μεθόδους είναι το ισχυρό θεωρητικό υπόβαθρο που έχουν. Επιπλέον, η βελτιστοποίηση της ποινικοποιημένης πιθανοφάνειας είναι αρκετά δύσκολη, διότι πρόκειται για πολυ-διάστατη μη-κοίλη συνάρτηση με ιδιομορφίες.

Οι Fan&Li(2001) [15], πρότειναν ένα νέο αλγόριθμο, ο οποίος δίνει προτεραιότητα στη διαδικασία επιλογής παραμέτρου, καθώς επίσης και μία φόρμουλα για το τυπικό σφάλμα, μέσω της μεθόδου «sandwich». Η προτεινόμενη διαδικασία έχει ελεγχθεί και συγκριθεί με άλλες μεθόδους και τα αποτελέσματα δείχνουν οτι προτιμάται. Οι Fan&Li δείχνουν στην εργασία τους οτι ο εκτιμητής ποινικοποιημένης πιθανοφάνειας έχει «μαντική» ιδιότητα για την επιλογή του σωστού μοντέλου, όταν η παράμετρος κανονικοποίησης είναι σωστά επιλεγμένη. Με άλλα λόγια, όταν οι πραγματικές παράμετροι έχουν μηδενικά στοιχεία, εκτιμώνται σαν μηδέν, με πιθανότητα να τείνει στη μονάδα, ενώ τα μη μηδενικά στοιχεία εκτιμώνται σαν το σωστό υπομοντέλο να είναι γνωστό. Εν συντομία, ο εκτιμητής ποινικοποιημένης πιθανοφάνειας λειτουργεί σαν το σωστό υπομοντέλο (*submodel*) να είναι γνωστό εκ των προτέρων. Αυτό είναι πολύ σημαντικό διότι φαίνεται να ξεπερνά τον μέγιστο εκτιμητή πιθανοφάνειας και να λειτουργεί όσο καλά περιμένουμε. Θα ακολουθήσει εκτενής αναφορά για την μεθοδολογία αυτή.

2.2 Ποινικοποιημένα Ελάχιστα Τετράγωνα και Επιλογή Μεταβλητών

Θεωρούμε το γραμμικό μοντέλο παλινδρόμησης:

$$\tilde{Y} = \tilde{X} * \tilde{\beta} + \tilde{\varepsilon} \qquad (2.1.1)$$

Όπου το $\tilde{Y}$ είναι (n*1) διάνυσμα και το $\tilde{X}$ (n*d) πίνακας.

Όπως και στο παραδοσιακό γραμμικό μοντέλο παλινδρόμησης, υποθέτουμε οτι τα y_i είναι ανεξάρτητα. Υπάρχει ισχυρή σύνδεση ανάμεσα στα ποινικοποιημένα ελάχιστα τετράγωνα και στη επιλογή παραμέτρου στο γραμμικό μοντέλο παλινδρόμησης. Επιπλέον, υποθέτουμε οτι οι στήλες του πίνακα $\tilde{X}$ στην (2.1.1) είναι ορθοκανονικές. Η εκτίμηση των ελαχίστων τετραγώνων επιτυγχάνεται μέσω της ελαχιστοποίησης της :

$$\left\|\tilde{Y} - \tilde{X} * \tilde{\beta}\right\|^2$$

το οποίο είναι ισοδύναμο με το:

$$\left\|\tilde{\hat{\beta}} - \tilde{\beta}\right\|^2$$

Όπου το:

$$\tilde{\hat{\beta}} = \tilde{X}' * \tilde{Y}$$

Είναι η εκτιμήτρια ελαχίστων τετραγώνων (OLS).

Συμβολίζουμε με:

$$z = \tilde{X}' * \tilde{Y}$$

και:

$$\tilde{\hat{Y}} = \tilde{X} * \tilde{X}' \tilde{Y}$$

Μια μορφή των ποινικοποιημένων ελαχίστων τετραγώνων είναι:

$$\frac{1}{2}\left\|\tilde{Y} - \tilde{X} * \tilde{\beta}\right\|^2 + \lambda \sum_{j=1}^{d} p_j\left(|\beta_j|\right) = \frac{1}{2}\left\|\tilde{Y} - \tilde{\hat{Y}}\right\|^2 + \frac{1}{2}\sum_{j=1}^{d}\left(z_j - \beta_j\right)^2 + \lambda \sum_{j=1}^{d} p_j\left(|\beta_j|\right)$$

(2.1.2)

Οι συναρτήσεις ποινής $p_j(.)$ στην (2.1.2) δεν είναι απαραίτητα ίδιες για όλα τα j. Για απλότητα, υποθέτουμε οτι οι συναρτήσεις ποινής είναι ίδιες για όλους τους συντελεστές και συμβολίζουμε με p(|.|). Επιπλέον, συμβολίζουμε το λ* p(|.|) με $p_\lambda(|.|)$ και έτσι το p(|.|) εξαρτάται από το λ. Εξαιρέσεις σε περίπτωση με διαφορετικό λ δεν παρουσιάζουν ιδιαίτερη δυσκολία.

Η ελαχιστοποίηση του προβλήματος (2.1.2) είναι ισοδύναμη με την ελαχιστοποίηση των συνιστωσών. Αυτό μας οδηγεί να σκεφτούμε το πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων:

$$\frac{1}{2}(\mathrm{z}-\theta)^2+ p_\lambda(|\theta|) \quad (2.1.3)$$

Παίρνοντας την Hard συνάρτηση ποινής (Figure 1a) :

$$p_\lambda(|\theta|)= \lambda^2 - (|\theta| - \lambda)^2 * \mathrm{I}(|\theta| < \lambda) \quad (2.1.4)$$

Προκύπτει η Hard εκτιμήτρια:

$$\hat{\theta} = z * I(|z| > \lambda) (2.1.5)$$

Με άλλα λόγια, η λύση της (2.1.2) είναι απλά

$$z_j\, I(\left|z_j\right| > \lambda) \,,$$

το οποίο συμπίπτει με την επιλογή καλύτερου υποσυνόλου και με την κατά βήματα επιλογή, όπως επίσης και με τους ορθοκανονικούς σχεδιασμούς. Σημειώνουμε οτι η Hard συνάρτηση ποινής είναι πιο ομαλή (*smooth*) συνάρτηση ποινής από την ποινή εντροπίας :

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

η οποία επίσης καταλήγει στην (2.1.5).

Η μορφή αυτή διευκολύνει υπολογιστικούς σκοπούς.

Μια καλή συνάρτηση ποινής μας παρέχει εκτιμητές με τις εξής τρεις ιδιότητες:

1. Αμεροληψία (*Unbiasedness*): Ο εκτιμητής που προκύπτει είναι σχεδόν αμερόληπτος όταν η πραγματική άγνωστη παράμετρος είναι μεγάλη, για να αποφύγουμε ανεπιθύμητη μεροληψία του μοντέλου.
2. Σποραδικότητα (*Sparsity*): Ο εκτιμητής που προκύπτει αποτελεί έναν κανόνα περιορισμού (*thresholding rule*) , ο οποίος αυτόματα θέτει τους μικρούς εκτιμημένους συντελεστές ίσους με το μηδέν, με σκοπό να μειώσει την πολυπλοκότητα του μοντέλου.

3. Συνέχεια (*Continuity*): Ο εκτιμητής που προκύπτει είναι συνεχής στα δεδομένα z με σκοπό την αποφυγή της αστάθειας (*instability*) στην πρόβλεψη του μοντέλου.

Τώρα παραθέτουμε κάποιες λεπτομέρειες πάνω σε αυτές τις απαιτήσεις:

Η πρώτη παράγωγος της (2.1.3) ως προς θ είναι:

$$\text{sgn}(\theta) * \{|\theta| + p_\lambda'(|\theta|)\} - z$$

Είναι εύκολο να δούμε οτι όταν $p_\lambda'(|\theta|) = 0$ για μεγάλο $|\theta|$, ο εκτιμητής που προκύπτει είναι z, όταν το $|z|$ είναι σημαντικά μεγάλο. Έτσι, όταν η πραγματική παράμετρος $|\theta|$ είναι μεγάλη, η παρατηρούμενη τιμή $|z|$ είναι μεγάλη,με μεγάλη πιθανότητα. Έτσι, τα ποινικοποιημένα ελάχιστα τετράγωνα απλά είναι $\hat{\theta} = z$, δηλαδή σχεδόν αμερόληπτο. Έτσι, η συνθήκη $p_\lambda'(|\theta|)=0$ για μεγάλα $|\theta|$ είναι μια επαρκής συνθήκη για την αμεροληψία για μια μεγάλη πραγματική παράμετρο. Μια επαρκής συνθήκη για τον εκτιμητή που προκύπτει, ώστε να είναι ένας κανόνας περιορισμού (*thresholding rule*) είναι το ελάχιστο της συνάρτησης:

$$|\theta| + p_\lambda'(|\theta|)$$

να είναι θετικό (Εικόνα 3).

Όταν

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

η παράγωγος της (2.1.3) είναι θετική για όλα τα θετικά θ (και αρνητική για όλα τα αρνητικά θ).

Συνεπώς, ο εκτιμητής ελαχίστων τετραγώνων είναι μηδέν σε αυτή την περίπτωση, συγκεκριμένα $\hat{\theta} = 0$ για $|z| < \min_{\theta \neq 0}\{|\theta| + p_\lambda'(|\theta|)\}$.

Όταν $|z| > \min_{\theta \neq 0}\{|\theta| + p_\lambda'(|\theta|)\}$, δύο διασταυρώσεις (*Crossings*) υπάρχουν (Σχήμα 1). Η μεγαλύτερη είναι ο εκτιμητής ελαχίστων τετραγώνων (*PLS*). Αυτό σημαίνει οτι μια επαρκής και αναγκαία συνθήκη για τη συνέχεια είναι οτι η ελαχιστοποίηση της συνάρτησης $|\theta| + p_\lambda'(|\theta|)$ να φτάνει στο μηδέν. Η συνάρτηση ποινής που ικανοποιεί τις συνθήκες σποραδικότητας και συνέχειας πρέπει να είναι μοναδική από την αρχή.

Είναι γνωστό οτι η L_2 ποινή :

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

Οδηγεί στην παλινδρόμηση κορυφογραμμής (*ridge regression*).

Η L_1 ποινή :

$$p_\lambda(|\theta|)= \lambda*|\theta|$$

οδηγεί στον Soft οριακό κανόνα (*soft thresholding rule*):

$$\hat{\theta}_j = \text{sgn}(z_j)(|z_j - \lambda|)_+ \quad (2.1.6)$$

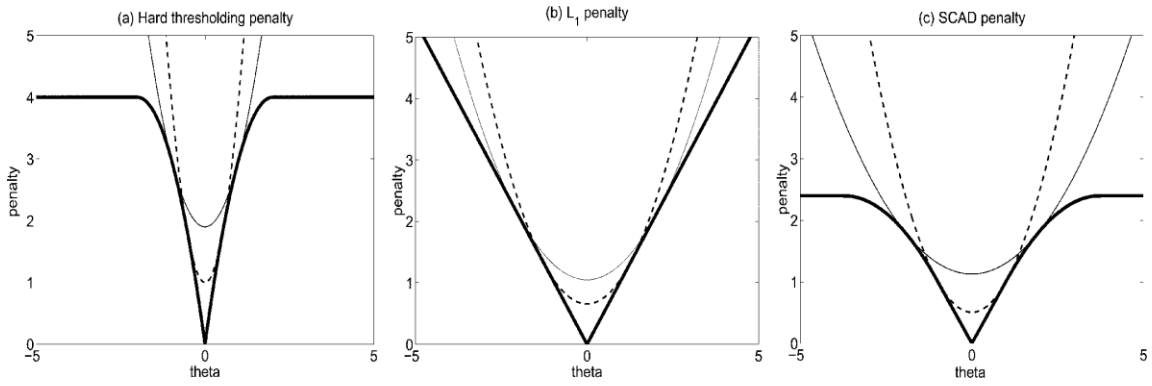
κάτι το οποίο προτάθηκε από τους Dohoro & Johnstone (1994).

Η μέθοδος LASSO που προτάθηκε από τον Tibshirani (1996,1997) είναι η εκτίμηση των ποινικοποιημένων ελαχίστων τετραγώνων με την L_1 ποινή.

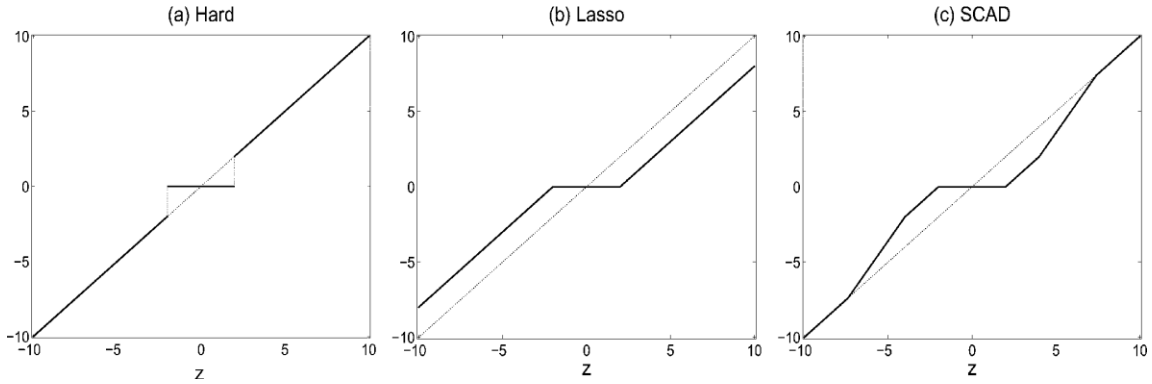
Η L_q ποινή:

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

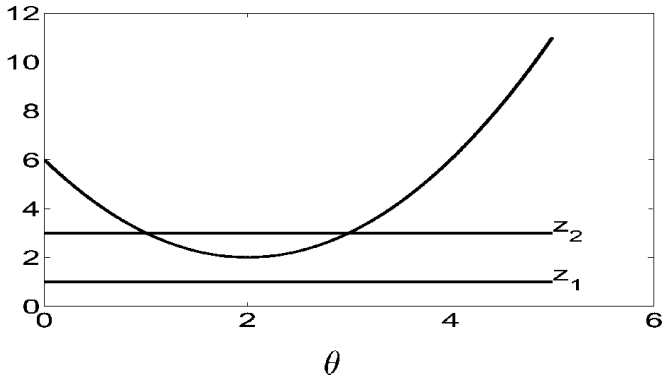
Οδηγεί στην παλινδρόμηση κορυφογραμμής (*bridge regression*). Η λύση είναι συνεχής μόνο όταν q≥ 1. Παρ'όλ'αυτά, όταν q>1 , το ελάχιστο του $|\theta|+p'_\lambda(|\theta|)$ είναι μηδέν και έτσι δεν παράγει σποραδικές λύσεις (Εικόνα 4α). Η μόνη συνεχής λύση με ένα κανόνα περιορισμού (*thresholding rule*) σ'αυτή την οικογένεια είναι η L_1 ποινή, αλλά αυτό προκύπτει με την μετακίνηση του εκτιμητή κατά μια σταθερά λ (Εικόνα 2β).



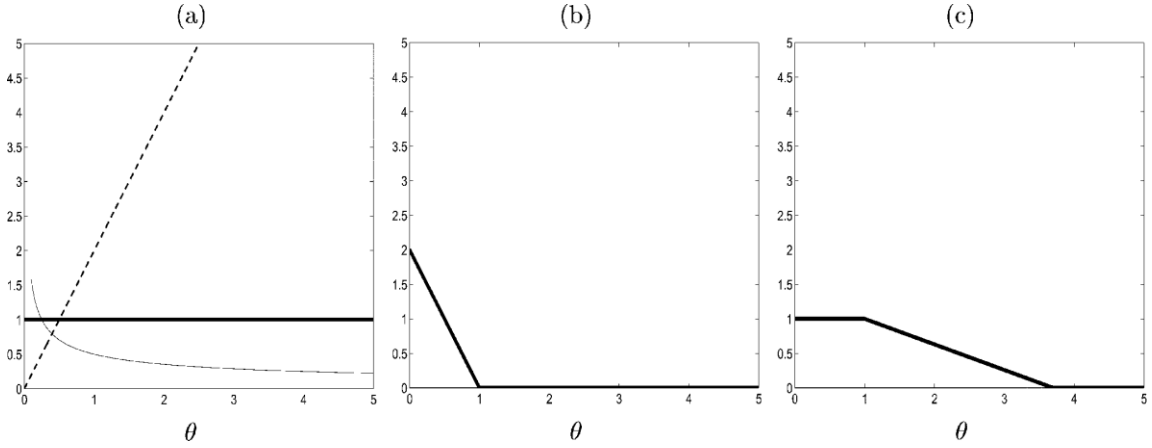
Εικόνα 1.Οι τρεις συναρτήσεις ποινής και οι τετραγωνικές τους προσεγγίσεις.



Εικόνα 2. Thresholding συναρτήσεις για α)Hard, β) Soft ή Lasso, γ)Scad.



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



Εικόνα 4. Οι συναρτήσεις $p'_\lambda(|\theta|)$ ως προς θ, για (a) τις συναρτήσεις ποινής L_q, (β) τη Hard συνάρτηση ποινής και (γ) 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.

2.2.1 Η Συνάρτηση Ποινής SCAD

Η L_q και η Hard συναρτήσεις ποινής δεν ικανοποιούν ταυτόχρονα τις μαθηματικές συνθήκες για αμεροληψία, σποραδικότητα και συνέχεια. Η συνάρτηση ποινής SCAD είναι μια συνεχής διαφορική συνάρτηση ποινής που βελτιώνει της ιδιότητες της L_1 ποινής, καθώς και της Hard συνάρτησης ποινής που ορίστηκε στην (2.1.4) και ορίζεται ως:

$$p'_\lambda(\theta)= \lambda\left\{I(\theta \le \lambda) + \frac{(\alpha-\lambda\theta)_+}{(\alpha-1)\lambda} I(\theta > \lambda)\right\} \text{ για } \alpha > 2, \theta > 0 \qquad (2.2.1.1)$$

Πρόκειται για μια τετραγωνική συνάρτηση ποινής, που δεν ποινικοποιεί υπερβολικά τις μεγάλες τιμές του θ, και κάνει τη λύση συνεχή.

Η λύση που προκύπτει είναι:

$$\hat{\theta} = \begin{cases} sgn(z)(|z| - \lambda)_+, & |z| \le 2\lambda \\ \frac{(\alpha-1)z - sgn(z)\alpha\lambda}{\alpha-2}, & 2\lambda \le z \le \alpha\lambda \\ z, & |z| > \alpha\lambda \end{cases} \qquad (2.2.1.2)$$

(Εικόνα 2c)

Η λύση αποδίδεται στον Fan(1997) [14].

Ο οριακός κανόνας (*thresholding rule*) στο (2.2.1.2) περιλαμβάνει δυο άγνωστες παραμέτρους λ και α.Στην πράξη, μπορούμε να βρούμε το καλύτερο ζευγάρι (α,λ) στον δι-διάστατο χώρο, χρησιμοποιώντας κάποια κριτήρια , όπως τη διασταυρωμένη επικύρωση (*cross-validation*) και τη γενικευμένη διασταυρωμένη επικύρωση (*generalized cross-validation*). Μια τέτοια διαδικασία είναι υπολογιστικά χρονοβόρα.

Οι Fan και Li, χρησιμοποιώντας εργαλεία Μπεϋζιανής ανάλυσης ρίσκου (*Bayesian risk analysis*), κατέληξαν στην επιλογή του $\alpha = 3.7$.

2.2.2 Απόδοση Οριακών Κανόνων

Οι Marron & Al. (1998) [42], εφάρμοσαν την ανάλυση ρίσκου για να κατανοήσουν τη διαφορά του hard και του soft οριακού κανόνα στα μικρά δείγματα. Η κλειστή μορφή της L_2 συνάρτησης ρίσκου $R(\hat{\theta}, \theta)=E(\hat{\theta} - \theta)^2$ προήλθε από το Γκαουσιανό μοντέλο

Z~ N(θ, σ^2) για τους hard και τους soft οριακούς κανόνες που προτάθηκε από τους Dohono & Johnstone(1994) [11].Οι Fan και Li (2000), 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 ιδιότητες τους.

2.3 Επιλογή μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας

Η μεθοδολογία που έχουμε αναπτύξει μέχρι στιγμής μπορεί να εφαρμοστεί σε πολλά στατιστικά κομμάτια, όπως στα γραμμικά μοντέλα παλινδρόμησης (*linear regression models*) , στα εύρωστα γραμμικά μοντέλα (*robust linear models*) και στα γενικευμένα γραμμικά μοντέλα που βασίζονται στην πιθανοφάνεια (*likelihood-based generalized linear models*).

Από εδώ και στο εξής, θα θεωρούμε οτι ο πίνακας σχεδιασμού $\tilde{X}$=(x_{ij}) είναι ορθοκανονικοποιημένος, έτσι ώστε κάθε στήλη του να έχει μέση τιμή 0 και διασπορά 1.

2.3.1 Ποινικοποιημένα ελάχιστα τετράγωνα και πιθανοφάνεια

Στο κλασικό γραμμικό μοντέλο παλινδρόμησης, η εκτίμηση των ελαχίστων τετραγώνων επιτυγχάνεται μέσω της ελαχιστοποίησης του αθροίσματος του τετραγώνου του σφάλματος (*squared residual errors*). Συνεπώς, η (2.1.2) μπορεί πολύ φυσικά να επεκταθεί στην περίπτωση που οι πίνακες σχεδιασμού δεν είναι ορθοκανονικοί.

Ισοδύναμα με το (2.1.2), μια μορφή ποινικοποιημένων ελαχίστων τετραγώνων είναι:

$$\frac{1}{2}(\tilde{Y} - \tilde{X}\,\tilde{\beta})'(\tilde{Y} - \tilde{X}\tilde{\beta}) + n^* \sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (2.3.1.1)$$

Η ελαχιστοποίηση ως προς $\tilde{\beta}$ της παραπάνω εξίσωσης οδηγεί στον εκτιμητή ποινικοποιημένων ελαχίστων τετραγώνων του β. Είναι γνωστό οτι ο εκτιμητής ποινικοποιημένων ελαχίστων τετραγώνων (*OLS*) του β δεν είναι εύρωστος (*robust*).

Τώρα μπορούμε να θεωρήσουμε την Loss- συνάρτηση L_1 ή γενικότερα την συνάρτηση ψ του Huber. Έτσι, αντί να ελαχιστοποιήσουμε την (2.3.1.1) ελαχιστοποιούμε την :

$$\sum_{i=1}^{n} \psi\left(\left|y_i - \widetilde{x_i}\tilde{\beta}\right|\right) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (2.3.1.2)$$

ως προς $\tilde{\beta}$.

Το αποτέλεσμα είναι ένας εύρωστος ποινικοποιημένος εκτιμητής ως προς β.

Για τα γενικευμένα γραμμικά μοντέλα, τα στατιστικά συμπεράσματα βασίζονται στις «υποκρίπτουσες» συναρτήσεις πιθανοφάνειας. Ο ποινικοποιημένος εκτιμητής μέγιστης πιθανοφάνειας χρησιμεύει στο να συλλέγει σημαντικές μεταβλητές. Υποθέτουμε οτι τα δεδομένα $\{(\widetilde{\chi_i}, y_i)\}$ συλλέχθηκαν ανεξάρτητα. Δεδομένων των x_i, τα y_i έχουν συνάρτηση πυκνότητας (ή μάζας, αν είναι διακριτά δεδομένα) πιθανότητας $f_i(g(\widetilde{x_i'}\tilde{\beta}),y_i)$, όπου η g είναι μια γνωστή συνάρτηση σύνδεσης (*link function*).

‘Εστω

$$l_i = \log f_i$$

η υπό συνθήκη λογαριθμοποιημένη πιθανοφάνεια του y_i.Τότε μια μορφή της λογαριθμοποιημένης πιθανοφάνειας είναι:

$$\sum_{i=1}^{n} l_i\left(g\left(\widetilde{x_i'}\tilde{\beta}\right), y_i\right) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

Μεγιστοποίηση της συνάρτησης της ποινικοποιημένης πιθανοφάνειας ισοδυναμεί με την ελαχιστοποίηση της :

$$-\sum_{i=1}^{n} l_i\left(g\left(\widetilde{x_i'}\tilde{\beta}\right), y_i\right) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (2.3.1.3)$$

ως προς β.

2.3.2 Δειγματοληπτικές και Προβλεπτικές Ιδιότητες

Σ’ αυτή την ενότητα θα καθιερώσουμε την ασυμπτωτική θεωρία για τον μη-κοίλο εκτιμητή ποινικοποιημένης πιθανοφάνειας.

Έστω :

$$\tilde{\beta}_0 = (\beta_{10}, ..., \beta_{d0})' = (\widetilde{\beta_{10}}', ..., \widetilde{\beta_{20}}')'$$

Χωρίς βλάβη της γενικότητας θεωρούμε οτι :

$$\tilde{\beta}_{20} = \tilde{0}$$

‘Εστω $I(\tilde{\beta}_0)$ ο πίνακας της πληροφορίας κατά Fisher (*Fisher Information Matrix*) και $I(\tilde{\beta}_0, \tilde{0})$ η πληροφορία κατά Fisher δεδομένου οτι $\tilde{\beta}_{20} = \tilde{0}$.

Αρχικά θα δείξουμε οτι υπάρχει ένας εκτιμητής ποινικοποιημένης πιθανοφάνειας ο οποίος συγκλίνει στο:

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

όπου:

$$a_n = \max\{p'_{\lambda_n}(|\beta_j|) : \beta_{j0} \neq 0\}$$

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

Επιπλέον, δείχνουμε οτι ένας εκτιμητής n-οστής τάξης πρέπει να ικανοποιεί:

$$\widehat{\widetilde{\beta_2}} = 0$$

και $\widehat{\widetilde{\beta_1}}$ ασυμπτωτικά με πίνακα συνδιασποράς I_1^{-1}, εάν $n^{1/2}\lambda_n \to \infty$.

Αυτό συνεπάγεται όπως εάν το $\tilde{\beta}_{20} = \tilde{0}$ να ήταν γνωστό. Σύμφωνα με τον Dohono&Johnstone(1994) ο εκτιμητής δουλεύει όπως και ο εκτιμητής «πρόβλεψης» ο οποίος γνωρίζει εκ των προτέρων οτι $\tilde{\beta}_{20} = \tilde{0}$.

Η προηγούμενη «μαντική» ικανότητα είναι στενά συνδεδεμένη με το φαινόμενο της υπεραποδοτικότητας (*superefficiency phenomenon*).

Θεωρούμε το απλούστερο γραμμικό μοντέλο παλινδρόμησης :

$$\tilde{Y} = \widetilde{1_n}\mu + \tilde{\varepsilon}\ ,\ \text{όπου}\ \tilde{\varepsilon} \sim N_n(\tilde{0}, I_n)$$

Ένας υπεραποδοτικός εκτιμητής για το μ είναι :

$$\delta_n = \begin{cases} \tilde{Y}, \varepsilon\acute{\alpha}\nu\ \left|\tilde{Y}\right| \geq n^{-1/4} \\ c\tilde{Y}\ , \varepsilon\acute{\alpha}\nu \left|\tilde{Y}\right| < n^{-1/4} \end{cases}$$

Εάν θέσουμε όπου c =0, τότε το δ_n συμπίπτει με τον εκτιμητή Hard με παράμετρο $\lambda_n = n^{-1/4}$. Αυτός ο εκτιμητής εκτιμά την παράμετρο στο σημείο μηδέν χωρίς να την εκτιμά σε κανένα άλλο σημείο.

Τώρα θέτουμε το αποτέλεσμα σε μια γενικότερη βάση. Για να διευκολύνουμε την παρουσίαση υποθέτουμε οτι η ποινικοποίηση εφαρμόζεται σε κάθε συνιστώσα του β. Παρ' όλ'αυτά δεν υπάρχει καμια δυσκολία στην περίπτωση που κάποιοι παράγοντες δεν ποινικοποιούνται.

Έστω $V_i = (x_i, y_i), i = 1, \ldots, n$. Έστω $L(\tilde{\beta})$ η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας και $Q(\tilde{\beta})$ η ποινικοποιημένη συνάρτηση πιθανοφάνειας :

$$Q\ (\tilde{\beta}) = L\ (\tilde{\beta}) - n\ p_{\lambda_n}(\left|\beta_j\right|)$$

Στη συνέχεια, θα αναφέρουμε κάποια θεωρήματα και λήμματα των Fan&Li, πρώτα όμως αναφέρουμε κάποιες υποθέσεις κανονικότητας (*regularity conditions*):

(A)Οι παρατηρήσεις $\underset{\sim}{V_i}$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (*independent and identically distributed*) με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V}, \underset{\sim}{\beta})$. Η $f(\underset{\sim}{V}, \underset{\sim}{\beta})$ έχει μια κοινή βάση και το μοντέλο είναι αναγνωρίσιμο (*identifiable*). Επίσης, η πρώτη και η δεύτερη λογαριθμημένη παράγωγος της f ικανοποιεί τις εξισώσεις:

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

Και:

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

(B) Ο πίνακας πληροφορίας του Fisher:

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

Είναι πεπερασμένος και θετικά ορισμένος στο $\tilde{\beta} = \widetilde{\beta_0}$. κ

(C) Υπάρχει ένα ανοιχτό υποσύνολο ω του Ω που περιέχει την πραγματική παράμετρο $\widetilde{\beta_0}$ τέτοιο ώστε για σχεδόν ολα τα $\tilde{V}$η συνάρτηση πυκνότητας πιθανότητας $f(\tilde{V},\tilde{\beta})$ να επιδέχεται παραγώγους τρίτης τάξης:

$$\frac{\partial f(\underset{\sim}{V},\underset{\sim}{\beta})}{\partial\beta_j\partial\beta_k\partial\beta_l}$$

Για $\forall\widetilde{\beta} \in \omega$.

Υπάρχουν συναρτήσεις M_{jkl} τέτοιες ώστε:

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

Για $\forall\widetilde{\beta} \in \omega$, όπου $m_{jkl} = E_{\underset{\sim}{\beta_0}}[M_{jkl}] < \infty$, $\forall\, j,k,l$.

Θεώρημα 2.3.2.1:Έστω $V_1,\ldots, V_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (*independent and identically distributed*) κάθε μια από τις οποίες έχει συνάρτηση πυκνότητας πιθανότητας $f(\tilde{V},\tilde{\beta})$. Εάν :

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

Τότε υπάρχει ένα τοπικό μέγιστο $\hat{\tilde{\beta}}$ του $Q(\tilde{\beta})$ τέτοιο ώστε :

$$\left\|\hat{\tilde{\beta}} - \widetilde{\beta_0}\right\| = O_p(n^{-1/2}+a_n)$$

Όπου a_n είναι γνωστό από την εξίσωση (2.3.2.1).

Είναι φανερό από το παραπάνω θεώρημα ότι επιλέγοντας ένα κατάλληλο λ_nυπάρχει $\sqrt{n}$ –συνεπής ποινοκοποιημένος εκτιμητής.

Τώρα θα δείξουμε οτι ο εκτιμητής θα πρέπει να κατέχει την ιδιότητα της σποραδικότητας $\widehat{\widetilde{\beta_2}}$=0.

Λήμμα 2.3.2.1: Έστω $V_1,\ldots, V_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (*independent and identically distributed*) κάθε μια από τις οποίες έχει συνάρτηση πυκνότητας πιθανότητας $f(\tilde{V}, \tilde{\beta})$. Υποθέτουμε οτι :

$$\lim_{n\to\infty} inf\left(\lim_{\theta\to 0^+} \inf p'_{\lambda_n}(\theta)|\lambda_n\right) > 0 \quad (2.3.2.2)$$

Εάν $\lambda_n \to 0$ και $\sqrt{n}\lambda_n \to \infty$ τότε με πιθανότητα να τείνει στη μονάδα, για κάθε δεδομένο $\widetilde{\beta_1}$ που ικανοποιεί :

$$\left\|\tilde{\beta}_1 - \tilde{\beta}_{10}\right\| = O_p(n^{-1/2})$$

Και για κάθε σταθερά C ισχύει :

$$Q\left\{\begin{pmatrix}\widetilde{\beta_1}\\ 0\end{pmatrix}\right\} = \max_{\|\beta_2\|\le Cn^{-1/2}} Q\left\{\begin{pmatrix}\widetilde{\beta_1}\\ \widetilde{\beta_2}\end{pmatrix}\right\}$$

Συμβολίζουμε:

$$\Sigma = \text{diag}\{p''_{\lambda_n}(|\beta_{10}|), \ldots, p''_{\lambda_n}(|\beta_{s0}|)\}$$

Και:

$$\tilde{b} = (p'_{\lambda_n}(|\beta_{10}|)sgn(\beta_{10}), \ldots, p'_{\lambda_n}(|\beta_{s0}|)sgn(\beta_{s0}))^T$$

Όπου s ο αριθμός των στοιχείων του β_{10}.

Θεώρημα 2.3.2.2 (Προβλεπτική Ιδιότητα): Έστω $V_1,\ldots, V_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (*independent and identically distributed)* κάθε μια από τις οποίες έχει συνάρτηση πυκνότητας πιθανότητας $f(\tilde{V}, \tilde{\beta})$. Υποθέτουμε οτι η συνάρτηση ποινής $p_{\lambda_n(|\theta|)}$ ικανοποιεί τη συνθήκη (2.3.2.2). Εάν $\lambda_n \to 0$ και $\sqrt{n}\lambda_n \to \infty$ όσο $n \to \infty$ τότε με πιθανότητα που τείνει στη μονάδα οι $\sqrt{n}$ – συνεπείς εκτιμητές $\hat{\tilde{\beta}} = \begin{pmatrix}\widehat{\beta_1}\\ \widehat{\beta_2}\end{pmatrix}$ που αναφέρθηκαν στο Θεώρημα 1 ικανοποιούν τα εξής :

a) Σποραδικότητα (*Sparsity*): $\widehat{\beta_2}$=0

b) Ασυμπτωτική Κανονικότητα (*Asymptotic 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 αναπτύχθησαν στην ενότητα 1.6.2, είναι προσεγγιστικά ίσος με:

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

Σημειώνουμε οτι για τις συναρτήσεις Hard και SCAD εάν $\lambda_n\to 0$ τότε a_n=0. Έτσι, από το θεώρημα 2.3.2.2 , όταν $\sqrt{n}\lambda_n\to\infty$, οι αντίστοιχοι εκτιμητές ποινικοποιημένης πιθανοφάνειας έχουν προβλεπτική ιδιότητα (*oracle property*) και συμπεριφέρονται εξίσου καλά με τους εκτιμητές μέγιστης πιθανοφάνειας όσον αφορά στην εκτίμηση του $\widetilde{\beta_1}$δεδομένου ότι $\widetilde{\beta_2}$=0. Παρ'όλ'αυτά για την L_1ποινή ισχύει $a_n=\lambda_n$. Έτσι, η $\sqrt{n}$-συνέπεια απαιτεί $\lambda_n=O_p(n^{-1/2})$. Από την άλλη μεριά η ιδιότητα του θεωρήματος 2.3.2.2 απαιτεί $\sqrt{n}\lambda_n\to\infty$.Αυτές οι δυο συνθήκες για τη LASSO δεν μπορούν να ικανοποιηθούν ταυτόχρονα. Πράγματι, για την L_1 δεν ισχύει η προβλεπτική ιδιότητα. Αντίθετα ισχύει για την L_q με q<1 εάν έχουμε επιλέξει σωστό λ_n.

Τώρα θα συζητήσουμε εκτενώς τις συνθήκες κανονικότητας (A)-(C) για τα γενικευμένα γραμμικά μοντέλα. Με μια *canonical link* η κατανομή του $\tilde{y}$ δεδομένου οτι $\tilde{X}=\tilde{x}$ ανήκει στην canonical εκθετική οικογένεια με συνάρτηση πυκνότητας πιθανότητας:

$$f(y;\tilde{x},\tilde{\beta}) = c(y)exp\left\{\frac{y\tilde{x}^T\tilde{\beta} - b(\tilde{x}^T\tilde{\beta})}{\alpha(\varphi)}\right\}$$

Οι συνθήκες κανονικότητας (A) ισχύουν.

Ο πίνακας πληροφορίας κατά Fisher είναι:

$$I(\tilde{\beta})=\mathrm{E}\left\{b''(\tilde{x}^T\tilde{\beta})\tilde{x}\tilde{x}^T\right\}\Big/\alpha(\varphi)$$

Εάν $\mathrm{E}\{b''(\tilde{x}^T\tilde{\beta})\tilde{x}\tilde{x}^T\}$ πεπερασμένο και θετικά ορισμένο η συνθήκη (B) ισχύει.

Εάν για κάθε $\tilde{\beta}$ στη γειτονιά του $\widetilde{\beta_0}$ ισχύει :

$$\left|b^{(3)}(\tilde{x}^T\tilde{\beta})\right| \le M_0(\tilde{x})$$

Για κάποια συνάρτηση $M_0(\tilde{x})$ που ικανοποιεί :

$$\mathrm{E}_{\beta_0}\{\mathrm{M}_0(\tilde{\mathrm{x}})\mathrm{x_j}\mathrm{x_k}\mathrm{x_l}\} < \infty \forall\, \mathrm{j,k,l}$$

Τότε η συνθήκη (C) ικανοποιείται.

Για γενικότερες συναρτήσεις σύνδεσης (*link functions*) παρόμοιες συνθήκες πρέπει να ισχύουν για τις συνθήκες (B), (C).

2.3.3 Προτεινόμενος Αλγόριθμος

Ο Tibshirani (1996) πρότεινε έναν αλγόριθμο για την επίλυση των ελαχίστων τετραγώνων της LASSO, ενώ ο Fu (1998) [20], πρότεινε έναν «shooting» αλγόριθμο για την LASSO. Θα παρουσιάσουμε τώρα έναν ενοποιημένο αλγόριθμο των Fan&Li για την ελαχιστοποίηση των προβλημάτων (2.3.1.1), (2.3.1.2) και (2.3.1.3) μέσω της τοπικής τετραγωνικής προσέγγισης (*local quadratic approximation*).

Ο $1^{ος}$ όρος στις (2.3.1.1), (2.3.1.2) και (2.3.1.3) μπορεί να αντικατασταθεί από μια συνάρτηση απώλειας (*Loss function*) του $\tilde{\beta}$. Συμβολίζουμε τον όρο αυτό με: l($\tilde{\beta}$). Τότε οι εκφράσεις (2.3.1.1), (2.3.1.2) και (2.3.1.3) μπορούν να γραφούν σαν ενιαία μορφή ως:

$$l(\tilde{\beta})+n\sum_{j=i}^{d} p_\lambda(|\beta_j|) \qquad (2.3.3.1)$$

Οι συναρτήσεις ποινής L_1, Hard και SCAD έχουν ιδιομορφίες στην αρχή και δεν έχουν συνεχή δεύτερη παράγωγο. Παρ'όλ'αυτά μπορούν τοπικά να προσεγγιστούν από μια τετραγωνική συνάρτηση ως ακολούθως: Υποθέτουμε οτι δίνεται μια αρχική τιμή $\tilde{\beta}_0$ που είναι κοντά στην τιμή που ελαχοστοποιεί την (2.3.3.1). Εάν το β_{j0} είναι πολύ κοντά στο μηδέν, τότε θέτουμε $\hat{\beta}_j$=0, οπότε διαγράφουμε το x_j από το αρχικό μοντέλο.

Διαφορετικά, μπορεί τοπικά να προσεγγιστεί από μια τετραγωνική συνάρτηση ως εξής:

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)sgn(\beta_j) \approx \left\{\frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}\right\}\beta_j \quad \text{για } \beta_j \neq 0$$

Με άλλα λόγια,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\left\{\frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}\right\}(\beta_j^2 - \beta_{j0}^2) \quad \text{για } \beta_j \approx \beta_{j0} \qquad (2.3.3.2)$$

Το σχήμα 1 που παρουσιάστηκε παραπάνω στην ενότητα 2.1 δείχνει την συνάρτηση L_1, Hard και SCAD καθώς και τις προσεγγίσεις τους για διάφορα β_{j0}.Ένα μειονέκτημα της μεθόδου αυτής είναι οτι εφόσον ένα συντελεστής συρρικνωθεί στο μηδέν, θα παραμείνει στο μηδέν. Παρ'όλ'αυτά, η μέθοδος αυτή μειώνει σημαντικά το υπολογιστικό φορτίο.

Εάν η $l(\tilde{\beta})$ είναι η L_1 loss συνάρτηση στην (2.3.1.2) τότε δεν έχει συνεχείς δεύτερες μερικές παραγώγους ως προς $\tilde{\beta}$.

Παρ'όλα αυτά, η:

$$\psi(|y - \tilde{x}'\tilde{\beta}|) \text{ στην (2.3.1.2)}$$

μπορεί ανάλογα να υπολογιστεί από την:

$$\left\{\frac{\psi(y - \tilde{x}'\tilde{\beta})}{(y - \tilde{x}'\widetilde{\beta_0})^2}\right\}(y - \tilde{x}'\tilde{\beta})^2$$

Εφόσον η αρχική τιμή $\tilde{\beta}_0$ του $\tilde{\beta}$ είναι κοντά στην τιμή ελαχιστοποίησης.Όταν μερικά από τα υπόλοιπα $|y - \tilde{x}'\tilde{\beta}_0|$ είναι μικρά, αυτή η προσέγγιση δεν είναι καλή.

Τώρα, υποθέτουμε οτι η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας έχει συνεχείς μερικές παραγώγους δεύτερης τάξης ως προς $\tilde{\beta}$. Έτσι, ο πρώτος όρος στην (2.3.3.1) μπορεί τοπικά να προσεγγιστεί από μια τετραγωνική συνάρτηση. Παρ'όλ'αυτά, η ελαχιστοποίηση της (2.3.3.1) μπορεί να μειωθεί σε μια τετραγωνική ελαχιστοποίηση του προβλήματος και έτσι μπορεί να χρησιμοποιηθεί ο αλγόριθμος Newton-Raphson. Πράγματι, η (2.3.3.1) μπορεί τοπικά να προσεγγιστεί από:

$$l(\tilde{\beta}_0) + \nabla l(\tilde{\beta}_0)'(\tilde{\beta} - \tilde{\beta}_0) + \frac{1}{2}(\tilde{\beta} - \tilde{\beta}_0)'\nabla^2 l(\tilde{\beta}_0)(\tilde{\beta} - \tilde{\beta}_0) + \frac{1}{2}n\tilde{\beta}'\Sigma_\lambda(\tilde{\beta}_0)\tilde{\beta} \quad (2.3.3.3)$$

όπου:

$$\nabla l(\tilde{\beta}_0) = \frac{\partial l(\tilde{\beta}_0)}{\partial \beta}$$

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

$$\Sigma_\lambda(\tilde{\beta}_0) = \text{diag}\left\{\frac{p_\lambda'(|\beta_{10}|)}{|\beta_{10}|}, \dots, \frac{p_\lambda'(|\beta_{d0}|)}{|\beta_{do}|}\right\}$$

Η τετραγωνική ελαχιστοποίηση του προβλήματος (2.3.3.3) έχει λύση:

$$\tilde{\beta}_1 = \tilde{\beta}_0 - \{\nabla^2 l(\tilde{\beta}_0) + n\Sigma_\lambda(\tilde{\beta}_0)\}^{-1}\{\nabla l(\tilde{\beta}_0) + nU_\lambda(\tilde{\beta}_0)\}$$

Όπου:

$$U_\lambda(\tilde{\beta}_0) = \Sigma_\lambda(\tilde{\beta}_0)\tilde{\beta}_0$$

Όταν ο αλγόριθμος συγκλίνει, ο εκτιμητής ικανοποιεί τη συνθήκη:

$$\frac{\partial l\left(\hat{\tilde{\beta}}_0\right)}{\partial \beta_j} + np_\lambda'(|\hat{\beta}_{j0}|)sgn(\hat{\beta}_{j0}) = 0$$

Δηλαδή την εξίσωση ποινικοποιημένης πιθανοφάνειας με μη μηδενικούς συντελεστές του $\hat{\tilde{\beta}}_0$. Ειδικά για το πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων (2.3.1.1), η

λύση μπορεί να βρεθεί επαναληπτικά, υπολογίζοντας την παλινδρόμηση κορυφογραμμής:

$$\tilde{\beta}_1 = \{\tilde{X}'\tilde{X} + n\Sigma_\lambda(\tilde{\beta}_0)\}^{-1}\tilde{X}'\tilde{Y}$$

Όμοια βρίσκουμε τη λύση της (2.3.1.2) επαναληπτικά:

$$\tilde{\beta}_1 = \left\{\tilde{X}'\widetilde{W}\tilde{X} + \frac{1}{2}n\Sigma_\lambda(\tilde{\beta}_0)\right\}^{-1}\tilde{X}'\widetilde{W}\tilde{Y}$$

Όπου:

$$\widetilde{W} = diag\{\frac{\psi(|y_1 - \tilde{x}_1{}'\tilde{\beta}_0|)}{(y_1 - \tilde{x}_1{}'\tilde{\beta}_0)^2}, \ldots\ldots, \frac{\psi(|y_n - \tilde{x}_n{}'\tilde{\beta}_0|)}{(y_n - \tilde{x}_n{}'\tilde{\beta}_0)^2}\}$$

Όπως και στην περίπτωση του εκτιμητή μέγιστης πιθανοφάνειας (*MLE*) , με μια καλή αρχική τιμή $\tilde{\beta}_0$ η διαδικασία ενός βήματος μπορεί να είναι τόσο αποδοτική όσο η διαδικασία πλήρους επανάληψης, ειδικά ο εκτιμητής μέγιστης πιθανοφάνειας όταν ο αλγόριθμος Newton-Raphson χρησιμοποιείται. Θεωρώντας το $\beta^{(\kappa-1)}$ σαν καλή αρχική τιμή στο κ-οστό βήμα, ο επόμενος υπολογισμός μπορεί να θεωρηθεί σαν μονοβηματική διαδικασία. Έτσι, ο εκτιμητής που θα προκύψει από τον αλγόριθμο κάνοντας λίγες επαναλήψεις, μπορεί να θεωρηθεί ως εκτιμητής ενός βήματος και να έχει την ίδια απόδοση. Οπότε δεν χρειάζεται να επαναλάβουμε τον αλγόριθμο μέχρι να επέλθει σύγκλιση, αρκεί οι αρχικές τιμές να είναι καλές. Ως αρχικές εκτιμήσεις τώρα, μπορούν να δωθούν αυτές του πλήρους μοντέλου, αρκεί να μην είναι υπερβολικά παραμετροποιημένες.

2.3.4 Υπολογισμός Τυπικού Σφάλματος

Τα τυπικά σφάλματα για τις εκτιμημένες παραμέτρους μπορούν να βρεθούν απ'ευθείας επειδή εκτιμούμε παραμέτρους και επιλέγουμε μεταβλητές ταυτόχρονα. Ακολουθώντας την συμβατική τεχνική της πιθανοφάνειας, ο αντίστοιχος τύπος «sandwich» μπορεί να χρησιμοποιηθεί ως εκτιμητής για τη συνδιασπορά των εκτιμητών $\hat{\beta}_1$, η μη εξαφανισμένη συνιστώσα του $\hat{\tilde{\beta}}$.

Οπότε:

$$\widehat{cov}\left(\hat{\tilde{\beta}}_1\right) = \left\{\nabla^2 l\left(\hat{\tilde{\beta}}_1\right) + n\Sigma_\lambda\left(\hat{\tilde{\beta}}_1\right)\right\}^{-1} \widehat{cov}\left\{\nabla l\left(\widehat{\widetilde{\beta}_1}\right)\right\}\left\{\nabla^2 l\left(\hat{\tilde{\beta}}_1\right) + n\Sigma_\lambda\left(\widehat{\widetilde{\beta}_1}\right)\right\}^{-1} \quad (3.4.1)$$

Αυτός ο τύπος φαίνεται να έχει καλή ακρίβεια για μέτρια μεγέθη δείγματος. Όταν η συνάρτηση απώλειας (*loss function*) L_1 χρησιμοποιείται στην εύρωστη παλινδρόμηση, κάποιες τροποποιήσεις χρειάζονται στον αλγόριθμο και στον αντίστοιχο τύπο «sandwich».

Στην περίπτωση που $\psi(\chi) = |\chi|$, τα διαγώνια στοιχεία του $\widetilde{W}$ είναι:

$$\{|r_i|^{-1}\}$$

Με $r_i = y_i - \tilde{x}_i{}'\widetilde{\beta_0}$

Για μια δεδομένη τιμή του $\widetilde{\beta_0}$ όταν κάποια από τα υπόλοιπα $\{r_i\}$ είναι κοντά στο μηδέν, αυτά τα σημεία αποκτούν πολύ βάρος.Αντικαθιστούμε το βάρος:

$$(a_n + |r_i|)^{-1}$$

Στους υπολογισμούς μας, παίρνουμα το a_n σαν $2n^{-1/2}$quantile των τιμών των υπολοίπων $\{|r_i|$,i=1,..,n$\}$. Έτσι, η σταθερά a_n αλλάζει από επανάληψη σε επανάληψη.

2.3.5 Έλεγχος της συνέπειας του αλγορίθμου

Οι Fan&Li έδειξαν οτι ο αλγόριθμος που πρότειναν συγκλίνει στη σωστή λύση. Χρησιμοποίησαν για το σκοπό αυτό ένα 100-διάστατο διάνυσμα $\tilde{\beta}$ που αποτελείται από 50 μηδενικά στοιχεία και άλλα 50 μη-μηδενικά, τα οποία ακολουθούν την κατανομή N(0, 5^2). Χρησιμοποιώντας έναν πίνακα σχεδιασμού X παρήγαγαν ένα διάνυσμα απόκρισης Y σύμφωνα με το γραμμικό μοντέλο παλινδρόμησης που παρουσιάστηκε νωρίτερα. Στην περίπτωση αυτή επιλέχθηκε ορθογώνιος πίνακας επειδή τα ποινικοποιημένα ελάχιστα τετράγωνα έχουν κλειστή μαθηματική μορφή και έτσι μπορεί να γίνει σύγκριση με την αλγοριθμική μέθοδο. Το πείραμα έδειξε οτι πράγματι επήλθε σύγκλιση στην σωστή λύση. Το Matlab χρειάστηκε 0.27, 0.39, 0.16 δευτερόλεπτα για να υπολογίσει τα ποινικοποιημένα ελάχιστα τετράγωνα για τις συναρτήσεις L_1, Hard και SCAD. Ο αριθμός των επαναλήψεων ήταν 30, 5 και 30

αντίστοιχα. Στην πραγματικότητα, μετά από 10 επαναλήψεις τα ποινικοποιημένα ελάχιστα τετράγωνα ήταν κοντά στην πραγματική λύση.

2.4 Αριθμητική Σύγκριση

Στο κομμάτι αυτό, θα συγκρίνουμε τις προτεινόμενες μεθόδους, καθώς και την μέθοδο για το τυπικό σφάλμα.

2.4.1 Σφάλμα Πρόβλεψης και Σφάλμα Μοντέλου

Το σφάλμα πρόβλεψης (*prediction error*) ορίζεται ως το μέσο σφάλμα στην πρόβλεψη του $\tilde{Y}$ για δεδομένο νέο $\tilde{\chi}$. Διακρίνουμε δύο περιπτώσεις: το X να είναι τυχαίο (*random*) και το X να είναι ελεγχόμενο (*controlled*). Στην περίπτωση που το X είναι τυχαίο, τότε το Y και το $\tilde{\chi}$ επιλέγονται τυχαία. Στην ελεγχόμενη περίπτωση, οι πίνακες σχεδιασμού επιλέγονται από τους πειραματιστές και μονο το Y είναι τυχαίο. Για ευκολία στην παρουσίαση θεωρούμε μόνο την περίπτωση που το X είναι τυχαίο. Σ' αυτή την περίπτωση, τα δεδομένα $(\tilde{\chi}_i, Y_i)$ θεωρούνται οτι είναι τυχαία από κάποια κατανομή. Τότε εάν το $\hat{\mu}(\tilde{\chi})$ είναι μια πρόβλεψη που κατασκευάστηκε χρησιμοποιώντας τα παρόντα δεδομένα, το σφάλμα πρόβλεψης ορίζεται ως:

$$\text{PE}(\hat{\mu})=\text{E}\{Y-\mu(\tilde{\chi})\}^2$$

Όπου η πρόβλεψη υπολογίζεται ως προς τη νέα παρατήρηση $(\tilde{\chi}, Y)$.

Το σφάλμα πρόβλεψης τροποποιείται ως:

$$\text{PE}\ (\hat{\mu})=\text{E}\{Y-E(Y|\tilde{x})\}^2+E\{E(Y|\tilde{x})-\hat{\mu}(\tilde{\chi})\}^2$$

Ο πρώτος όρος είναι το εγγενές σφάλμα πρόβλεψης εξαιτίας του θορύβου. Ο δεύτερος όρος είναι το σφάλμα εξαιτίας της έλλειψης προσαρμογής μοντέλου (*lack of fit*). Αυτός ο όρος καλείται σφάλμα μοντέλου (*model error*) και συμβολίζεται με $\text{ME}(\hat{\mu})$. Εάν $Y=\tilde{\chi}'\tilde{\beta}+\varepsilon$, όπου $E(\varepsilon|\tilde{\chi})=0$, τότε :

$$\text{ME}(\hat{\mu})=\left(\hat{\tilde{\beta}}-\tilde{\beta}\right)' E(\tilde{\chi}\tilde{\chi}')(\hat{\tilde{\beta}}-\tilde{\beta})$$

2.4.2 Επιλογή Οριακών Παραμέτρων

Για να εφαρμόσουμε τις μεθόδους που αναπτύξαμε παραπάνω, πρέπει να εκτιμήσουμε τη ρυθμιστική παράμετρο (*tuning parameter*) σε καθεμία περίπτωση. Συμβολίζουμε με $\tilde{\theta} = \lambda$ στην περίπτωση της μεθόδου Hard και LASSO και με $\tilde{\theta} = (\alpha, \lambda)$ στην περίπτωση της μεθόδου SCAD. ΟιFan&Li χρησιμοποίησαν δυο μεθόδους για την εκτίμηση του $\tilde{\theta}$: Την πενταπλή (*fivefold*) διασταυρωμένη επικύρωση και τη γενικευμένη διασταυρωμένη επικύρωση (*generalized cross validation*). Θα αναπτύξουμε τις δύο αυτές διαδικασίες για την περίπτωση των γραμμικών μοντέλων παλινδρόμησης. Η επέκταση των διαδικασιών αυτών σε εύρωστα γραμμικά μοντέλα παλινδρόμησης καθώς και γραμμικά μοντέλα βασισμένα στην πιθανοφάνεια, δεν περιλαμβάνει δυσκολία.

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 $\tilde{\theta}$ και ν, βρίσκουμε τον εκτιμητή $\hat{\tilde{\beta}}^{\nu}(\tilde{\theta})$ του $\tilde{\beta}$, χρησιμοποιώντας το σύνολο εκπαίδευσης $T - T^{\nu}$. Στη συνέχεια, εφαρμόζουμε το κριτήριο της διασταυρωμένης επικύρωσης:

$$CV(\tilde{\theta}) = \sum_{\nu=1}^{5} \sum_{(y_k, x_{k)} \in T^{\nu}} \{y_k - \tilde{x}_k' \hat{\beta}^{(\nu)}(\tilde{\theta})\}^2$$

και βρίσκουμε το $\tilde{\theta}$ που ελαχιστοποιεί το $CV(\tilde{\theta})$.

Στη δεύτερη μέθοδο, δηλαδή στην γενικευμένη διασταυρωμένη επικύρωση, μετατρέπουμε τη λύση ως:

$$\widetilde{\beta_1}(\tilde{\theta}) = \{\tilde{X}'\tilde{X} + n\Sigma_{\lambda}(\tilde{\beta}_0)\}^{-1}\tilde{X}'\tilde{Y}$$

Έτσι, η προσαρμοσμένη τιμή $\hat{Y}$ του Y είναι:

$$\tilde{X}\{\tilde{X}'\tilde{X} + n\Sigma_{\lambda}(\tilde{\beta}_0)\}^{-1}\tilde{X}'\tilde{Y}$$

Και θεωρούμε ως πίνακα προβολής τον:

$$\tilde{P}_x\{\hat{\tilde{\beta}}(\tilde{\theta})\} = \{\tilde{X}'\tilde{X} + n\Sigma_{\lambda}(\tilde{\beta})\}^{-1}\tilde{X}'$$

Το πλήθος των σημαντικών παραμέτρων στην προσαρμογή του ποινικοποιημένου μοντέλου ελαχίστων τετραγώνων είναι:

$$e(\tilde{\theta}) = tr[\{\tilde{P}_x \{\hat{\tilde{\beta}}(\tilde{\theta})\}\}]$$

Τότε το κριτήριο της γενικευμένης διασταυρωμένης επικύρωσης είναι:

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

Και:

$$\hat{\tilde{\theta}} = arg \min_{\theta} \{GCV(\tilde{\theta})\}$$

2.4.3 Προσομοίωση

- Γραμμική Παλινδρόμηση:

Δημιουργήθηκαν 100 σύνολα δεδομένων από n παρατηρήσεις σύμφωνα με το μοντέλο:

$$Y = \tilde{X}'\tilde{\beta} + \sigma\varepsilon$$

Όπου β=(3, 1.5, 0, 0, 2, 0, 0, 0)' και τα $\tilde{\chi}$ και $\tilde{\varepsilon}$ προέρχονται από την τυποποιημένη κανονική κατανομή. Τα στοιχεία χ_i και χ_j σχετίζονται μεταξύ τους με τη σχέση $\rho^{|i-j|}$ με ρ=0.5.

Αρχικά, επιλέγουμε $n = 40$ και $\sigma = 3$. Έπειτα, μειώνουμε το σ σε 1 και το n αυξάνεται στις 60 παρατηρήσεις. Το σφάλμα του μοντέλου συγκρίνεται με αυτό του εκτιμητή ελαχίστων τεταγώνων. Η διάμεσος των σχετικών σφαλμάτων του μοντέλου (*Median of Relative Model Errors –MRME)* από 100 προσομοιωμένα σύνολα δεδομένων, υπάρχει στον πίνακα 2.4.3.1. 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	Avg. No. of 0 Coefficients: 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 2.4.3.1: Αποτέλεσμα προσομοίωσης για το γενικό γραμμικό μοντέλο.

Από τον παραπάνω πίνακα, παρατηρούμε ότι όταν το επίπεδο του θορύβου είναι υψηλό και το μέγεθος του δείγματος μικρό, η *LASSO* συμπεριφέρεται καλύτερα, ενώ 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. Παρ'όλ'αυτά, όταν μειώθηκε ο θόρυβος, η *SCAD* φάνηκε 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. Σημειώνουμε ότι η *SCAD* έχει πολύ καλά αποτελέσματα με επιλογή του $\alpha = 3.7$, η οποία τιμή χρησιμοποιήθηκε και στις επόμενες προσομοιώσεις. 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 (2.4.3.1), παρατηρούμε τα εξής: Η διάμεσος των απολύτων τιμών της απόκλισης των 100 εκτιμηθέντων συντελεστών των 100 συνόλων δεδομένων, διαιρεμένη με 0.6745, συμβολιζόμενη ως *SD*, 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. Ο πίνακας (2.4.3.2)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* τύποςπου έχει αναφερθεί παραπάνω 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 2.4.3.2: Τυπικές αποκλίσεις των εκτιμητών στο γραμμικό μοντέλο παλινδρόμησης (n=60).

- Λογιστική Παλινδρόμηση:

Δημιουργήθηκαν 100 σύνολα δεδομένων αποτελούμενα από 200 παρατηρήσεις, βάσει του μοντέλου:

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

Όπου:

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

Τα πρώτα 6 στοιχεία του $\tilde{\chi}$ και $\tilde{\beta}$ είναι ίδια όπως και στο παραπάνω παράδειγμα της γραμμικής παλινδρόμησης. Οι δύο τελευταίες συνιστώσες του $\tilde{\chi}$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές της κατανομής Bernoulli με ποσοστό επιτυχίας 0.5. Όλες οι μεταβλητές είναι κανονικοποιημένες. Το σφάλμα μοντέλου υπολογίστηκε μετά από 1000 MonteCarlo προσομοιώσεις. Τα αποτελέσματα παρουσιάζονται στους παρακάτω πίνακες.

Method	MRME (%)	Avg. No. of 0 Coefficients	
		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 2.4.3.3: Αποτελέσματα προσομοίωσης

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
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 2.4.3.4: 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.

ΚΕΦΑΛΑΙΟ 3

Η Μέθοδος της Γενικευμένης Διασταυρωμένης Επικύρωσης

(Generalized Cross-Validation Method)

3.1 Εισαγωγή

Σ' αυτό το κεφάλαιο θα προσπαθήσουμε να εξηγήσουμε μια αποδοτική και ταυτόχρονα πρακτική μέθοδο, η οποία εξηγεί τα δεδομένα με θόρυβο (*noisy data*). Θωρούμε το μοντέλο:

$$y(t) = g(t) + \varepsilon(t),\ t\in [0,1] \qquad (3.1.1)$$

όπου g(t) είναι μια λεία ($C^{\infty}-$ τάξης) κυρτή συνάρτηση και ε(t) μια τυχαία μεταβλητή με μέση τιμή μηδέν και διασπορά $\sigma^2\delta_{ij}$ *(white noise process)*, δηλαδή:

$$E(\varepsilon(t))=0 \quad \text{και} \quad E[\varepsilon(s),\varepsilon(t)]=\begin{cases}\sigma^2, & t=s\\ 0, & t\neq s\end{cases}$$

Οι παρατηρήσεις y(t) έχουν γίνει για t=$t_1, t_2, \dots, t_n$ με $0 \le t_1 < t_2 < \cdots < t_n \le 1$.

Σκοπός μας είναι να δημιουργήσουμε την συνάρτηση g(t) από τα δεδομένα $y(t_j) = y_j$, j=1,2,…,n.

Υποθέτουμε οτι $g \in W_2^{(m)}$,όπου :

$$W_2^{(m)} = \{g: g^{(\nu)} \alpha\pi\acute{o}\lambda\acute{\upsilon}\tau\omega\varsigma\ \sigma\upsilon\nu\varepsilon\chi\acute{\eta}\varsigma, \nu = 0,1,\dots,m-1,\ \ g^{(m)} \in L_2[0,1]\}.$$

Η εκτιμήτρια για το g είναι η $g_{n,\lambda}$ όπου $g_{n,\lambda}$ είναι η λύση του προβλήματος :

Βρες $f \in W_2^{(m)}$ που να ελαχιστοποιεί την παράσταση:

$$\frac{1}{n}\sum_{j=1}^{n}(f(t_j)-y_j)^2 + \lambda\int_0^1 (f^{(m)}(u))^2 du \qquad (3.1.2)$$

Η εκτιμήτρια $g_{n,\lambda}$ είναι μια απειροδιαφορίσιμη πολυωνυμική spline βαθμού (2m-1). Η παράμετρος λ η οποία πρέπει να επιλεγεί, είναι μια ρυθμιστική παράμετρος (*tuning*

parameter), η οποία ρυθμίζει την «ανταλλαγή» ανάμεσα στην «τραχύτητα» (*roughness*) της λύσης που ορίζεται ως:

$$\int_0^1 (f^{(m)}(u))^2 du$$

και την «ασυμφωνία» (*infidelity*) με τα δεδομένα που μετράται από :

$$\frac{1}{n}\sum_{j=1}^{n}(f(t_j)-y_j)^2 \qquad (3.1.3)$$

Θέλουμε λοιπόν να βρούμε μια καλή τιμή για το λ. Ο Reinsh πρότεινε οτι εάν το σ^2 είναι γνωστό, τότε το λ πρέπει να επιλεγεί ώστε η «ασυμφωνία» (εξίσωση (3.1.3)) να ικανοποιεί την εξίσωση:

$$\frac{1}{n}\sum_{j=1}^{n}(f(t_j)-y_j)^2=\sigma^2 \qquad (3.1.4)$$

Όταν επιβάλλονται επιπλέον συνθήκες για την ομαλότητα (*smoothness*) και την περιοδικότητα, ο Wahba επιτυγχάνει θεωρητικά αποτελέσματα για τη βέλτιστη επιλογή του λ.

Το βέλτιστο λ είναι εκείνο που ελαχιστοποιεί το πραγματικό μέσο τετραγωνικό σφάλμα (*Mean Square Error*) το οποίο υπολογίζεται από τα δεδομένα. Το πραγματικό μέσο τετραγωνικό σφάλμα συμβολίζεται με R(λ) και ορίζεται ως:

$$R(\lambda)=\frac{1}{n}\sum_{j=1}^{n}\left(g_{n,\lambda}(t_j)-g(t_j)\right)^2 \qquad (3.1.5)$$

Το λ πρέπει να επιλεγεί έτσι ώστε η «ασυμφωνία» που ορίζεται από το αριστερό τμήμα της (3.1.4) να είναι στην πραγματικότητα λίγο μικρότερο από σ^2. Παρ'όλ'αυτά, αυτό το αποτέλεσμα δεν είμαι πρακτικό διότι εξαρτάται από το n και από το άγνωστο g. Επιπλέον, το σ^2 ίσως να είναι και αυτό άγνωστο.

Εάν το σ^2 είναι γνωστό, η τιμή του λ μπορεί να επιλεγεί ως εξής:

Ορίζουμε τον (n×n) πίνακα A(λ) ο οποίος εξαρτάται από τα $\{t_i\}_{i=1}^{n}$και από το λ και ορίζεται ως:

$$\begin{pmatrix} g_{n,\lambda}(t_1) \\ \vdots \\ g_{n,\lambda}(t_n) \end{pmatrix} = A(\lambda) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Εάν η $g_{n,\lambda}(t)$ είναι μια γραμμική συνάρτηση των $y_1, \dots, y_n$ για κάθε t, τότε ο πίνακας A(λ) υπάρχει. Τότε ορίζουμε:

$$E(R(\lambda)) = E\left(\frac{1}{n}\|A(\lambda)y - g\|^2\right) \quad (3.1.6)$$

όπου:

$$y = (y_1, \dots, y_n)' \text{ και } g = (g(t_1), \dots, g(t_n))$$

Με τη νόρμα συμβολίζουμε την ευκλείδια νόρμα.

Μετά από στοιχειώδεις πράξεις στην (3.1.6) και χρησιμοποιώντας τη μέση τιμή και τη διασπορά του $\varepsilon = (\varepsilon(t_1), \dots, \varepsilon(t_n))'$ προκύπτει:

$$E\big(R(\lambda)\big) = \frac{1}{n}\big\|\big(I - A(\lambda)\big)g\big\|^2 + \frac{\sigma^2}{n} Trace A^2(\lambda) \quad (3.1.7)$$

Ακολουθούν δύο θεωρήματα που βασίζονται στα όσα έχουμε αναφέρει παραπάνω:

- Θεώρημα 3.1.1: Μια αμερόληπτη (*unbiased*) εκτιμήτρια του E(R(λ)) δίνεται από την $\hat{R}(\lambda)$ η οποία ορίζεται ως:

$$\hat{R}(\lambda) = \frac{1}{n}\big\|\big(I - A(\lambda)\big)y\big\|^2 - \frac{\sigma^2}{n} Tr\big(I - A(\lambda)\big)^2 + \frac{\sigma^2}{n} TrA^2(\lambda) \qquad (3.1.8)$$

και

$$E\left(\hat{R}(\lambda)\right) = E(R(\lambda))$$

Συνεπώς ελαχιστοποιώντας το $\hat{R}(\lambda)$ μπορούμε να βρούμε μια καλή τιμή για το λ. Μια εκτίμηση για αυτόν τον τύπο έχει κάνει ο Mallow.

Το ζητούμενο είναι να βρούμε μια καλή εκτιμήτρια για την ελαχιστοποίηση του E(R(λ)) από τα δεδομένα, χωρίς να είναι αναγκαία προυπόθεση η γνώση του σ^2.

Η εκτιμήτρια αυτή καλείται εκτιμήτρια γενικευμένης διασταυρωμένης επικύρωσης (*generalized cross validation*) ή εν συντομία εκτιμήτρια GCV. Η εκτιμήτρια αυτή ελαχιστοποιεί το $V(\lambda)$ το οποίο ορίζεται ως:

$$V(\lambda) = \frac{1}{n}\left\|\left(I - A(\lambda)\right)y\right\|^2 / \left[\frac{1}{n}Tr(I - A(\lambda))\right]^2 \qquad (3.1.9)$$

Όταν το λ ανήκει στη γειτονιά της ελαχιστοποίησης του E(R(λ)), τότε κάτω από γενικές συνθήκες για το g και την ακολουθία $\{t_i\}_{i=1}^n \equiv \{t_{in}\}_{i=1}^n$,n=1,2,… όταν το n είναι μεγάλο, ισχύει:

$$E\big(V(\lambda)\big) - \sigma^2 \approx E(R(\lambda))$$

Σαν συνέπεια από τα παραπάνω προκύπτει το παρακάτω θεώρημα:

- Θεώρημα 3.1.2: Για $g \in W_2^{(m)}$ και $\{t_{in}\}_{i=1}^n$ υπάρχει μια ακολουθία $\tilde{\lambda} = \tilde{\lambda}(\theta)$ η οποία ελαχιστοποιεί το $E(V(\lambda))$ και έχει την εξής ιδιότητα:

$$\lim_{n\to\infty} \frac{E(R(\tilde{\lambda}))}{\min_{\lambda} E(R(\lambda))} = 1 \qquad (3.1.10)$$

Το θεώρημα αυτό ουσιαστικά λέει οτι το αναμενόμενο μέσο τετραγωνικό σφάλμα (*Mean Square Error*) χρησιμοποιώντας το $\tilde{\lambda}$, τείνει στο ελάχιστο δυνατό, όσο το n→ ∞.

Τώρα θα περιγράψουμε την προέλευση της εκτιμήτριας GCV. Η ιδέα για την εκτιμήτρια GCVείναι αρκετά απλή και πηγαίνει ως εξής:

Έστω $g_{n,\lambda}$ μια C^∞-τάξης spline η οποία προκύπτει χρησιμοποιώντας όλα τα σημεία από τα δεδομένα εκτός από το κ-οστό. Παίρνουμε την πιθανότητα η $g_{n,\lambda}^k$να προβλέπει το χαμένο σημείο y_k σαν μέτρο προσαρμογής για το λ. Δηλαδή:

Έστω $g_{n,\lambda}^k$ να είναι μια συνάρτηση $f \in W_2^{(m)}$ η οποία ελαχιστοποιεί την εξίσωση:

$$\frac{1}{n}\sum_{\substack{j=1\\ j\neq k}}^{n}(f(t_j) - y_j))^2 + \lambda\int_0^1 (f^{(m)}(u))^2 du$$

Έστω επίσης:

$$V_o(\lambda) = \frac{1}{n}\sum_{\kappa=1}^{n}(g_{n,\lambda}^{[\kappa]}(t_k) - y_k)^2 \qquad (3.1.11)$$

Η εκτιμήτρια διασταυρωμένης επικύρωσης λ (*ordinary cross validation estimate of* λ) είναι εκείνη η εκτιμήτρια που ελαχιστοποιεί το $V_o(\lambda)$. Το ίδιο επιώθηκε από τους

Wahba και Word οι οποίοι εισήγαγαν την έννοια της εκτιμήτριας GCV. Διάφορα πειράματα Monte Carlo που έγιναν έδειξαν οτι η τιμή που ελαχιστοποιεί το $V_o(\lambda)$ είναι μια πολύ καλή εκτιμήτρια ελαχιστοποίησης και για την R(λ) για διαφορες τιμές του g και του σ^2.

Στην συμμετρική περίπτωση, η $g_{n,\lambda}$ είναι περιοδική και απαιτείται $t_j = {}^{j}/_{n}$, j=1,2,..,n.

Στην περίπτωση αυτή, όλα τα δεδομένα συμπεριφέρονται συμμετρικά. Αυτό σημαίνει οτι το σφάλμα πρόβλεψης στο t_k έχει το ίδιο βάρος με το σφάλμα πρόβλεψης στο t_j. Στη γενική περίπτωση γράφουμε :

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(g_{n,\lambda}^{[k]}(t_k) - y_k\right)^2 w_k(\lambda) \qquad (3.1.13)$$

Τα $w_k(\lambda)$ είναι τα βάρη, τα οποία αντισταθμίζουν τα σημεία που δεν καταλαμβάνουν τον ίδιο χώρο και την πιθανή μη-περιοδικότητα του g. Εάν:

$$w_k(\lambda) = [(1 - \alpha_{\kappa\kappa}(\lambda))/\frac{1}{n}Tr(I - A(\lambda))]^2 \quad \kappa=1,2,... \qquad (3.1.14)$$

όπου $\alpha_{\kappa\kappa}(\lambda)$ είναι τα διαγώνια στοιχεία του A(λ), τότε το V(λ) της (3.1.13) γίνεται το V(λ) της (3.1.9) και τότε ισχύει η (3.1.10).

Ένας διαφορετικός τρόπος για να εκτιμήσουμε το $V(\lambda)$ όπως στην (3.1.9), είναι να βρούμε μια διαφορετική μορφή της ευκλείδιας n-διάστατης νόρμας που θα μετατρέπει το γενικό πρόβλημα σε ισοδύναμο συμμετρικό και τότε να εφαρμόσουμε τον τύπο της διασταυρωμένης επικύρωσης.

Στη συνέχεια, θα ακολουθήσει εκτενής αναφορά στην μέθοδο της διασταυρωμένης επικύρωσης ενώ θα αναφερθούμε και στην περίπτωση των πολυωνύμων Bernoulli. Ένα ενδιαφέρον σημείο στα spline πολυώνυμα Bernoulli στην περίπτωση που τα δεδομένα έχουν ίσο βάρος, είναι οτι ο (n×n)-γραμμικός πίνακας που προκύπτει είναι κυκλικός.

3.2 Πολυώνυμα Bernoulli

Συμβολίζουμε τα πολυώνυμα Bernoulli με $B_r(t)$, t∈ [0,1]. Τα $\{B_r\}$ ορίζονται ως:

$$B_0(t) = 0$$

$$\frac{1}{(r+1)}\frac{d}{dt}B_{r+1}(t) = B_r(t)$$

Επιλέγουμε τη σταθερά ολοκλήρωσης έτσι ώστε:

$$\int_0^1 B_r(u)du = 0,\ \text{r=1,2,…}$$

Έστω $[x]$ το κλασματικό (*fractional*) μέρος του χ. Ορίζουμε:

$$k_r(t) = {B_r([t])}/{r!}$$

Εάν συμβολίσουμε με $L_k, k = 0,1,$ ...τα γραμμικά συναρτησοειδή:

$$L_0 f = \int_0^1 f(u)du$$

$$L_k f = f^{(k-1)}(1) - f^{(k-1)}(0) \equiv \int_0^1 f^{(k)}(u)du\,, k = 1,2,\ldots$$

Τότε:

$$L_k(k_r) = \begin{cases} 1, & k = r \\ 0, & k \neq r \end{cases} \text{k, r=0,1,2,…} \quad (3.2.1)$$

Ορίζουμε τώρα τον «Πυρήνα Bernoulli» («*Bernoulli kernel*») $k_r(s,t)$ως εξής:

$$k_r(s,t) = -\sum_{\substack{\nu=-\infty \\ \nu\neq 0}}^{\infty} \frac{1}{(2\pi i\nu)^r} e^{2\pi i\nu(s-t)}\text{,r=1,2,…} \qquad (3.2.2)$$

Γνωρίζουμε οτι:

$$k_r(s,t) = \frac{1}{r!}B_r([s-t]) = k_r([s-t]) \qquad (3.2.3)$$

Μπορεί να επαληθευτεί από τον ορισμό του $k_r(s,t)$ οτι:

$$\frac{\partial^p}{\partial s^p}k_r(s,t) = k_{r-p}(s,t) \qquad \text{p=1,2,…r-2}$$

$$\frac{\partial^p}{\partial t^p}k_r(s,t) = (-1)^p k_{r-p}(s,t)\text{s, t}\in[0,1] \qquad (3.2.4)$$

$$\frac{\partial^{r-1}}{\partial s^{r-1}} k_r(s,t) = k_1(s,t)$$

$$\frac{\partial^{r-1}}{\partial t^{r-1}} k_r(s,t) = (-1)^{r-1} k_1(s,t), \text{s, t} \in [0,1], \quad \text{s} \neq t \quad (3.2.5)$$

$$\int_0^1 \frac{\partial^m}{\partial s^m} k_{2m}(s,u) \frac{\partial^m}{\partial t^m} k_{2m}(t,u) du = (-1)^{m-1} k_{2m}(s,t) \quad (3.2.6)$$

Τώρα ακολουθεί ένα θεώρημα καθώς και η απόδειξή του σχετικά με τον τρόπο γραφής του $g_{n,\lambda}$ στην περίπτωση των κατά τμήματα οριζόμενων πολυωνύμων Bernoulli.

- Θεώρημα 3.2.1: Η λύση του προβλήματος:

Βρες $f \in W_2^{(m)}$ που να ελαχιστοποιεί την εξίσωση:

$$\frac{1}{n}\sum_{j=1}^{n}(f(t_j) - y_j)^2 + \lambda \int_0^1 (f^{(m)}(u))^2 du \quad (3.2.7)$$

είναι για $n \geq m$ μοναδική και έχει τη μορφή:

$$g_{n,\lambda}(t) = \sum_{r=0}^{m} \theta_r \, k_r(t) + (-1)^{m-1} \sum_{j=1}^{n} a_j \, k_{2m}(t, t_j) \quad (3.2.8a)$$

όπου:

$$\theta = (\theta_0, \theta_1, \dots, \theta_m)'$$

$$a = (a_1, a_2, \dots, a_n)'$$

και δίνονται από τους εξής τύπους:

$$\theta = (T' M^{-1} T + \Delta)' T' M^{-1} y$$

$$a = M^{-1}(y - T\theta) \quad (3.2.8b)$$

$$y = (y_1, y_2, \dots, y_n)'$$

Πιο συγκεκριμένα, οΤ είναι ένας n×(m+1) πίνακας με j-οστή είσοδο:

$$T_{jr} = k_r(t_j), \text{ r=0, 1 ,… , m}$$

$$\text{j=1, 2 , …, n} \quad (3.2.8c)$$

Ο Δ είναι ένας (m+1)×(m+1) πίνακας με όλα τα στοιχεία του μηδέν εκτός από ένα στη θέση (m+1),(m+1).

Το M δίνεται από τον τύπο:

$$M = K + n\lambda I$$

όπου K είναι ένας (n×n) πίνακας με j-οστή είσοδο k_{ij}:

$$k_{ij} = (-1)^{m-1} k_{2m}(t_j, t_k) \qquad (3.2.8d)$$

Ο I είναι ο (n×n) μοναδιαίος πίνακας. Ο πίνακας A(λ) δίνεται από τον τύπο:

$$A(\lambda) = KM^{-1}[I - T(T'M^{-1}T + \Delta)'T'M^{-1}] + T(T'M^{-1}T + \Delta)'T'M^{-1} \qquad (3.2.9)$$

Απόδειξη: Πρώτα θα δείξουμε ότι

$$g_{n,\lambda} \in span\{\{k_r(.)\}_{r=0}^{m} \cup \{k_{2m}(., t_j)\}_{j=1}^{n}\}$$

Αυτό μπορεί να αποδειχθεί με διάφορους τρόπους χρησιμοποιώντας γνωστά αποτελέσματα για τις συναρτήσεις splines. Εδώ βασιζόμαστε στα επιχειρήματα των Kimeldorf και Wahba [33]. Ο πυρήνας $Q(s,t)$που προκύπτει από το $W_2^{(m)}$εφοδιάζεται με το εσωτερικό γινόμενο:

$$< f, g >= \sum_{r=0}^{m-1} (L_r f)(L_r g) + \int_0^1 f^{(m)}(u) g^{(m)}(u) du$$

Και έτσι προκύπτει:

$$Q(s,t) = \sum_{r=0}^{m} k_r(s) k_r(t) + (-1)^{m-1} k_{2m}(s,t) \quad (3.2.10)$$

Έπειτα, προκύπτει οτι η $g_{n,\lambda}$ πρέπει να βρίσκεται στο:

$$L = span\{\{k_r(.)\}_{r=0}^{m-1} \cup \{Q_{t_j}\}_{j=1}^{n}\} \text{ όπου } Q_{t_j}(.) = Q(., t_j)$$

Παρ'όλ'αυτά, το L περιλαμβάνεται στο $\{\{k_r(.)\}_{r=0}^{m} \cup \{k_{2m}(., t_j)\}_{j=1}^{n}\}$ έτσι ώστε η $g_{n,\lambda}$ να έχει τη μορφή της (3.2.8α) για κάποια θ,α. Με αντικατάσταση της (3.2.8α) στην (3.2.7) και χρησιμοποιώντας την (3.2.6) παίρνουμε:

$$\sum_{j=1}^{n}(g_{n,\lambda}(t_j)-y_j)^2+n\lambda\int_0^1\left(g_{n,\lambda}^{(m)}(u)\right)^2du=$$

$$=\sum_{j=1}^{n}[\sum_{r=0}^{m}\theta_r k_r(t_j)+(-1)^{m-1}\sum_{k=1}^{n}a_k k_{2m}(t_j,t_k)-y_j]^2$$

$$+n\lambda[\sum_{j=1}^{n}\sum_{k=1}^{n}a_j a_k(-1)^{m-1}k_{2m}(t_j,t_k)+\theta_m^2]\equiv$$

$$\equiv\|T\theta+ka-y\|^2+n\lambda(a'ka+\theta_m^2) \qquad (3.2.11)$$

Τα διανύσματα θ,a επιλέγονται έτσι ώστε να ελαχιστοποιούν την έκφραση αυτή. Παραγωγίζοντας το δεξί μέλος της (3.2.11) ως προς θ και a και θέτοντας το αποτέλεσμα ίσο με μηδέν προκύπτει η απόδειξη του θεωρήματος.

Τονίζουμε ότι η $(-1)^{m-1}k_{2m}(t,t_k)$ θεωρείται συνάρτηση του t και είναι μια *monospline* βαθμού 2m, δηλαδή είναι το άθροισμα του μονονύμου (*monomial*) t^{2m}συν μια πολυωνυμική spline βαθμού 2m-1με ένα σημείο ενώσεως(*knot*). Όταν τα σημεία ένωσης $\{t_j\}$ καταλαμβάνουν ίσο χώρο, ο K και ο M είναι πίνακες κυκλικοί. Αυτό είναι και το περιεχόμενο του παρακάτω Λήμματος.

- Λήμμα 3.2.2: $\left\{(-1)^{m-1}k_{2m}(\frac{j}{n},\frac{k}{n})\right\}_{j,k=1,\dots,n}=WDW^*$

Όπου με «*» συμβολίζουμε τον μιγαδικό ερμιτιανό πίνακα και Wείναι ο $(n\times n)$ μοναδικός πίνακας με r-οστή είσοδο W_{rs} που δίνεται από τον τύπο:

$$W_{rs}=\frac{1}{\sqrt{n}}e^{2\pi irs/n}$$

Επίσης, Dείναι ένας διαγώνιος πίνακας με ν-οστή είσοδο$D_{\nu\nu}$ που ορίζεται ως:

$$D_{\nu\nu}=\lambda_{\nu n}^{2m}$$

$$\lambda_{\nu n}^{r}=n\sum_{\xi=-\infty}^{\infty}\frac{1}{[2\pi(\nu+\xi n)]^r}, \qquad \nu=1,2,\dots,n-1$$

$$(\lambda_{\nu n}^{r}\equiv\lambda_{n-\nu,n}^{r})$$

$$\lambda_{nn}^{r}=n\sum_{\substack{\xi=-\infty\\ \xi\neq 0}}^{\infty}\frac{1}{[2\pi\xi n]^r} \qquad (3.2.12)$$

Απόδειξη: $(-1)^{m-1}k_{2m}\left(\frac{j}{n},\frac{k}{n}\right)=\sum_{\substack{\nu=-\infty\\ \nu\neq 0}}^{\infty}\frac{1}{(2\pi\nu)^{2m}}e^{2\pi i\nu(j-k)}/n=$

$$=\sum_{\nu=1}^{n}\sum_{\xi=-\infty}^{\infty}\frac{1}{[2\pi(\nu+\xi n)]^{2m}}e^{2\pi i\nu(j-k+\xi n)/n}=$$
$$(\nu,\xi)\neq(n,-1)$$

$$=\sum_{\nu=1}^{n}\sum_{\xi=-\infty}^{\infty}\frac{1}{[2\pi(\nu+\xi n)]^{2m}}e^{2\pi i\nu(j-k)/n}$$
$$(\nu,\xi)\neq(n,-1)$$

3.3 Η συνάρτηση Γενικευμένης Διασταυρωμένης Επικύρωσης (GCV)

Έχουμε αναφέρει και παραπάνω ότι η κανονική (*ordinary*) συνάρτηση διασταυρωμένης επικύρωσης συμβολίζεται με $V_0(\lambda)$ και ορίζεται ως:

$$V_o(\lambda)=\frac{1}{n}\sum_{\kappa=1}^{n}(g_{n,\lambda}^{[\kappa]}(t_k)-y_k)^2$$

Επαναλαμβάνουμε ότι η $g_{n,\lambda}^{k}$ είναι η λύση του προβλήματος: Βρες $f\in W_2^{(m)}$ η οποία ελαχιστοποιεί την εξίσωση:

$$\frac{1}{n}\sum_{\substack{j=1\\ j\neq k}}^{n}(f(t_j)-y_j))^2+\lambda\int_0^1(f^{(m)}(u))^2du$$

Είναι χρήσιμο να γνωρίζουμε ότι αν αντικαταστήσουμε το σημείο y_kτων παρατηρήσεων με $g_{n,\lambda}^{k}(t_k)$ και λύσουμε το πρόβλημα ελαχιστοποίησης (3.1.2) με τα δεδομένα $y_1,y_2,\dots y_{k-1},g_{n,\lambda}^{[k]},y_{k+1},\dots,y_n$παίρνουμε τη λύση $g_{n,\lambda}^{k}$. Αυτό είναι και το περιεχόμενο του παρακάτω Λήμματος:

- Λήμμα 3.3.1: Έστω $n\geq m$ και παίρνουμε το $g_{n,\lambda}(t;k,z_k)$ να είναι η λύση για το πρόβλημα: Βρες $f\in W_2^{(m)}$ που να ελαχιστοποιεί:

$$\frac{1}{n}[(f(t_k)-z_k)^2+\sum_{\substack{j=1\\ j\neq k}}^{n}(f(t_j)-y_j)^2]+\lambda\int_0^1(f^{(m)}(u))^2du$$

Τότε ισχύει:

$$g_{n,\lambda}\left(t;k,g_{n,\lambda}^{[\kappa]}(t_k)\right)=g_{n,\lambda}^{[\kappa]}(t)$$

Απόδειξη:Έστω $h=g_{n,\lambda}^{[k]}$, $z_k=g_{n,\lambda}^{[k]}(t_k)$και έστω fνα είναι κάθε στοιχείο του $W_2^{(m)}$διαφορετικό του h. Τότε:

$$\frac{1}{n}[(h(t_k)-z_k)^2+\sum_{\substack{j=1\\ j\neq k}}^{n}(h(t_j)-y_j)^2]+\lambda\int_0^1(h^{(m)}(u))^2du=$$

$$\frac{1}{n}[\sum_{\substack{j=1\\ j\neq k}}^{n}(h(t_j)-y_j)^2+\lambda\int_0^1(h^{(m)}(u))^2du]<$$

$$<\frac{1}{n}[\sum_{\substack{j=1\\ j\neq k}}^{n}(f(t_j)-y_j)^2+\lambda\int_0^1(f^{(m)}(u))^2du]\leq$$

$$\leq\frac{1}{n}[(f(t_k)-z_k)^2+\sum_{\substack{j=1\\ j\neq k}}^{n}(f(t_j)-y_j)^2]+\lambda\int_0^1(f^{(m)}(u))^2du$$

Συγκρίνοντας το αριστερό με το δεξί μέλος της παραπάνω έκφρασης, παρατηρούμε ότι το hεπιλύει το n-διάστατο πρόβλημα ελαχιστοποίησης με το y_kνα αντικαθίσταται από το z_k.

- Λήμμα 3.3.2: $g_{n,\lambda}^{[k]}(t_k)-y_k=(g_{n,\lambda}(t_k)-y_k)/(1-\frac{\partial}{\partial y_k}g_{n,\lambda}(t_k))$

Απόδειξη: Έστω $z_k=g_{n,\lambda}^{[k]}(t_k)$. Τότε το από το Λήμμα 3.3.1 καθώς από το γεγονός ότι για κάθε ττο $g_{n,\lambda}^{[k]}(t_k)$ εξαρτάται γραμμικά από το y_kδίνει:

$$z_k=g_{n,\lambda}(t;k,z_k)=g_{n,\lambda}(t_k;k,y_k)+(z_k-y_k)\frac{\partial g_{n,\lambda}(t_k)}{\partial y_k}=$$

$$= g_{n,\lambda}(t_k) + (z_k - y_k)\frac{\partial g_{n,\lambda}(t_k)}{\partial y_k}$$

Το αποτέλεσμα ακολουθεί μετά από μερικές αλγεβρικές πράξεις.

Συμβολίζοντας τις εισόδους του πίνακα $A(\lambda)$ με a_{jk} έχουμε:

$$g_{n,\lambda}(t_k) = \sum_{j=1}^{n} a_{kj}\, y_j$$

Οπότε:

$$\frac{\partial g_{n,\lambda}^{[k]}}{\partial y_k} = a_{kk}$$

Και από το Λήμμα 3.3.2 προκύπτει ότι:

$$V_0(\lambda) = \frac{1}{n}\{\frac{(\sum_{j=1}^{n} a_{kj}\, y_j - y_k)^2}{(1-a_{kk})^2}\} \qquad (3.3.1)$$

Για να βρούμε το $V(\lambda)$ που είναι η συνάρτηση της γενικευμένης διασταυρωμένης επικύρωσης, θεωρούμε την περιοδική περίπτωση του προβλήματος εξομάλυνσης (*smoothing problem*): Βρες $g \in W_2^{(m)}$περιοδικό και με ολοκλήρωμα μηδέν που να ελαχιστοποιεί τη συνάρτηση:

$$\frac{1}{n}\sum_{\substack{j=1 \\ j\neq k}}^{n}(f(t_j) - y_j))^2 + \lambda\int_0^1 (f^{(m)}(u))^2 du$$

Όταν λέμε συνάρτηση g με ολοκλήρωμα μηδέν, εννοούμε ότι πρέπει να ικανοποιείται:

$$L_k g = 0\,, k = 0,1,\dots,m$$

Μπορεί να δειχθεί ότι η λύση $h_{n,\lambda}$ δίνεται από τη σχέση

$$h_{n,\lambda}(t) = \sum_{j=1}^{n} a_j(-1)^{m-1}k_{2m}(t, t_j) \qquad (3.3.2)$$

Όπου:

$$a = (k + n\lambda I)^{-1}y = M^{-1}y$$

Σ'αυτή την περίπτωση, το ρόλο του πίνακα Α παίζει το KM^{-1}. Εάν $t_j = {}^{j}/_{n}$, j=1,2,…nτότε το KM^{-1}είναι περιοδικό για κάθε λ και έτσι:

$$a_{kk} = \frac{1}{n}\sum_{j=1}^{n} a_{jj} \equiv \frac{1}{n} TraceA$$

Το $V_0(\lambda)$ γίνεται:

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{\sum_{j=1}^{n} a_{kj}\, y_j - y_k\right)^2}{\left(1 - \frac{1}{n}\sum_{k=1}^{n} a_{kk}\right)^2} =$$

$$= \frac{1}{n}\|(I-A)y\|^2 / [\frac{1}{n} Tr(I-A)]^2 \qquad (3.3.3)$$

Για να προκύψει η συνάρτηση γενικευμένης διασταυρωμένης επικύρωσης από την συνηθισμένη διασταυρωμένη επικύρωση, περιστρέφουμε το σύστημα έτσι ώστε ο πίνακας $\tilde{A}(\lambda)$ που προκύπτει και ο οποίος παίζει το ρόλο του εκτιμητή του Α(λ) να είναι κυκλικός. Εάν ο Α είναι συμμετρικός, γράφουμε:

$A(\lambda) = UD^2(\lambda)U'$, όπου D^2 διαγώνιος πίνακας και U ορθογώνιος. Τότε γράφοντας:

$\Gamma = WU'$ο $\tilde{A}(\lambda) = \Gamma A(\lambda)\Gamma'$ είναι κυκλικός.

Έστω $\tilde{y} = \Gamma y$. Τότε για το «λείο» (*smoothed*) $\tilde{y}$ είναι:

$$\Gamma\big(g_{n,\lambda}(t_1), \dots, g_{n,\lambda}(t_n)\big)' = \Gamma A(\lambda) y = \tilde{A}(\lambda)\tilde{y}$$

Εφαρμόζοντας στα δεδομένα τώρα την διασταυρωμένη επικύρωση, προκύπτει η $V(\lambda)$.

Σημειώνουμε ότι έλεγχος του $h_{n,\lambda}$ (εξ. 3.3.2) αποκαλύπτει τον «low pass filter» χαρακτήρα της απειροδιάστατης spline στην περίπτωση που τα δεδομένα είναι

συμμετρικά και καταλαμβάνουν τον ίδιο χώρο. Από το Λήμμα 3.3.2 και από την εξίσωση 3.3.2 βρίσκουμε ότι οι συντελεστές Fourier του δείγματος $\{h_{n,\lambda,\nu}\}$ του $h_{n,\lambda}$:

$$h_{n,\lambda,\nu} = \frac{1}{n}\sum_{j=1}^{n} h_{n,\lambda}(\frac{j}{n}) e^{-2\pi i \nu j / n}$$

Συνδέονται με τους συντελεστές Fourier $\{\hat{h}_n\}$των δεδομένων:

$$\hat{h}_n = \frac{1}{n}\sum_{j=1}^{n} y(\frac{j}{n}) e^{-2\pi i v j / n}$$

Μέσω των εξισώσεων:

$$h_{n,\lambda,v} = f_v \hat{h}_v \, , v = 1,2, \dots, n$$

Όπου:

$$f_v = \frac{1}{1 + n\lambda/\lambda_{vm}^{2m}}$$

Όταν $v \ll n$ προσεγγίζουμε το άθροισμα για το λ_{vn}^{r}της εξίσωσης (3.2.12) με $\xi = 0$και έτσι επιτυγχάνουμε:

$$f_v \approx \frac{1}{1 + n\lambda(2\pi v)^{2m}} = B(\frac{v}{v_0}) B^*(\frac{v}{v_0})$$

Όπου $B(\frac{v}{v_0})$ είναι το φίλτρο Butterworth.

3.4 Ιδιότητες της εκτιμήτριας της GCV

Όπως έχουμε αναφέρει και παραπάνω, το πραγματικό μέσο τετραγωνικό σφάλμα (*real mean square error*) δίνεται από τον τύπο:

$$R(\lambda) = \frac{1}{n}\sum_{j=1}^{n} \left(g_{n,\lambda}(t_j) - g(t_j) \right)^2 = \frac{1}{n} \|A(\lambda)y - g\|^2$$

Και η συνάρτηση διασταυρωμένης επικύρωσης δίνεται από:

$$V(\lambda) = \frac{1}{n} \left\| \left(I - A(\lambda)\right) y \right\|^2 / \left[\frac{1}{n} Tr(I - A(\lambda)) \right]^2$$

Σκοπός της ενότητας αυτής είναι να βρούμε εκείνο το λ που θα ελαχιστοποιεί το $R(\lambda)$. Αυτό δεν μπορεί να γίνει απ' ευθείας, αφού το $R(\lambda)$ περιλαμβάνει και το g το οποίο είναι άγνωστο.

Εάν το σ^2 είναι γνωστό, τότε το ζητούμενο λ μπορεί να είναι εκείνο το λ που ελαχιστοποιεί το $\hat{R}(\lambda)$ της εξίσωσης (3.1.8).

Εάν το σ^2 δεν είναι γνωστό, μπορούμε να χρησιμοποιήσουμε όπως θα δούμε και στη συνέχεια, το λ που ελαχιστοποιεί το $V(\lambda)$.

Για να δείξουμε τη χρησιμότητα του $V(\lambda)$ θα διακρίνουμε δύο περιπτώσεις: Στην πρώτη περίπτωση, $g(.) \in \pi_{m-1}$ όπου π_{m-1} είναι πολυώνυμο βαθμού (2m-1)ή και μικρότερου. Στην περίπτωση αυτή, τα $E(V(\lambda))$ και $E(R(\lambda))$ ελαχιστοποιούνται για $\lambda = \infty$. Εάν $\tilde{\lambda}$ είναι η τιμή που ελαχιστοποιεί το $E(V(\lambda))$, ορίζεται η «ανεπάρκεια» (*inefficiency*) της μεθόδου GCVως:

$$I^* = \frac{E(R(\tilde{\lambda}))}{\min_{\lambda} E(R(\lambda))} \qquad (3.4.1)$$

Η τιμή αυτού του μεγέθους τείνει στη μονάδα, όσο το $n \to \infty$.

Όταν η τιμή του λ έχει εκτιμηθεί ώστε να ελαχιστοποιεί το V, το μέσο τετραγωνικό σφάλμα (*MSE*) τείνει στο ελάχιστο δυνατό. Εάν $g(.) \in \pi_{m-1}$, τότε $I^* = 1$ οπότε τα $E(V(\lambda))$ και $E(R(\lambda))$ ελαχιστοποιούνται για την ίδια τιμή του λ. Στη δεύτερη περίπτωση, $g(.) \notin \pi_{m-1}$, $g \in W_2^{(m)}$. Θα δείξουμε στη συνέχεια ότι εάν $\tilde{\lambda}$,λ^*είναι οι τιμές που ελαχιστοποιούν τα $E(V(\lambda))$ και $E(R(\lambda))$ αντίστοιχα, τότε θα πρέπει να ικανοποιούνται οι εξής προϋποθέσεις:

$$\tilde{\lambda} \to 0$$

$$\lambda^* \to 0$$

$$\frac{1}{n\tilde{\lambda}^{1/2m}} \to 0$$

$$\frac{1}{n\lambda^{*1/2m} \to 0}$$

Στην πορεία θα αποδείξουμε πως προκύπτουν οι δύο τελευταίες σχέσεις, ενώ στο θεώρημα 3.4.1 που θα ακολουθήσει, θα αποδείξουμε ότι ισχύει η σχέση:

$$\left|\frac{E(R(\lambda))+\sigma^2-E(V(\lambda))}{E(R(\lambda))}\right| \leq h(\lambda) \qquad (3.4.2)$$

όπου h(λ) είναι μια μικρή ποσότητα που θα ορίσουμε.

Χρησιμοποιούμε τους εξής συμβολισμούς:

$$b^2(\lambda) = \frac{1}{n} g'(I - A(\lambda))^2 g = \frac{1}{n} \|(I - A(\lambda)g\|^2$$

$$\mu_1(\lambda) = \frac{1}{n} TrA(\lambda)$$

$$\mu_2(\lambda) = \frac{1}{n} TrA^2(\lambda)$$

Τότε:

$$E\big(R(\lambda)\big) = b^2(\lambda) + \sigma^2 \mu_2(\lambda)$$

$$E\big(V(\lambda)\big) = \frac{b^2 + \sigma^2(1 - 2\mu_1(\lambda) + \mu_2(\lambda))}{[1 - \mu_1(\lambda)]^2}$$

Στην πρώτη περίπτωση, όπως έχουμε ήδη αναφέρει, θεωρούμε ότι $g(.) \in \pi_{m-1}$.Τότε το $g = (g(t_1), g(t_2), \dots, g(t_n))'$είναι ένας γραμμικός συνδιασμός από τις mπρώτες στήλες του πίνακα Τ,οπότε $\big(I - A(\lambda)\big)g = 0$για όλα τα λ και $b(\lambda) \equiv 0$. Έτσι, η ελαχιστοποίηση του $E(R(\lambda))$ μειώνεται στην ελαχιστοποίηση του $TrA^2(\lambda)$, το οποίο ελαχιστοποιείται για $\lambda = \infty$. Ομοίως, το $E(V(\lambda))$ γίνεται:

$$E\big(V(\lambda)\big) = \frac{(1 - 2\mu_1(\lambda) + \mu_2(\lambda))}{[1 - \mu_1(\lambda)]^2}$$

Τώρα το $(I - A(\lambda))$ έχει mμηδενικές ιδιοτιμές και οι εναπομείνουσες (n-m) ιδιοτιμές μπορεί να αποδειχθεί ότι είναι της μορφής

$$n\lambda(n\lambda + \xi_{\nu n})^{-1}, \nu = 1,2, \dots, n - m$$

και $\xi_{1n}, \xi_{2n}, \dots, \xi_{n-m}$είναι (n-m) θετικοί αριθμοί, όχι όλοι ίσοι μεταξύ τους. Οπότε, η έκφραση για το $E(V(\lambda))$ γίνεται:

$$E\big(V(\lambda)\big) = \frac{1}{n} \sum_{\nu=1}^{n-m} \left(\frac{n\lambda}{n\lambda + \xi_{\nu n}}\right)^2 / \left(\frac{1}{n} \sum_{\nu=1}^{n-m} \frac{n\lambda}{n\lambda + \xi_{\nu n}}\right)^2 =$$

$$= \frac{1}{\left(\frac{n-m}{n}\right)} \frac{\frac{1}{n-m} \sum_{\nu=1}^{n-m} \left(\frac{n\lambda}{n\lambda + \xi_{\nu n}}\right)^2}{\left[\frac{1}{n-m} \sum_{\nu=1}^{n-m} \left(\frac{n\lambda}{n\lambda + \xi_{\nu n}}\right)\right]^2} \geq \frac{1}{\left(\frac{n-m}{n}\right)}$$

Το ελάχιστο επιτυγχάνεται αν και μόνο αν $\lambda = \infty$.

Τώρα θα αναφερθούμε στην γενικότερη περίπτωση:

- Θεώρημα 3.4.1: $\frac{E(R(\lambda))+\sigma^2-E(V(\lambda))}{E(R(\lambda))} = \frac{-\mu_1(2-\mu_1)}{(1-\mu_1)^2} + \frac{\sigma^2}{b^2+\sigma^2\mu_2}\frac{{\mu_1}^2}{(1-\mu_1)^2}$

Και έτσι: $\frac{|E(R(\lambda))+\sigma^2-E(V(\lambda))|}{E(R(\lambda))} < h(\lambda)$

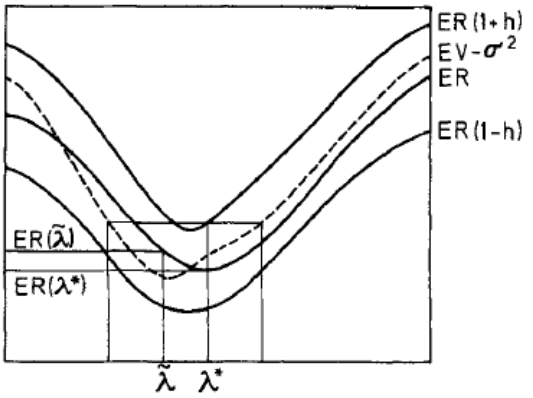
Όπου: $h(\lambda) = [2\mu_1(\lambda) + \frac{\mu_1(\lambda)^2}{\mu_2(\lambda)}]\frac{1}{(1-\mu_1(\lambda))^2}$

- Θεώρημα 3.4.2: Έστω λ^* η τιμή που ελαχιστοποιεί το $E(R(\lambda))$. Τότε το $E(V(\lambda))$ έχει ένα ελάχιστο $\tilde{\lambda}$ τέτοιο ώστε η αναμενόμενη «ανεπάρκεια» I^*ορίζεται ως:

$$I^* = \frac{E(R(\tilde{\lambda}))}{E(R(\lambda^*))}$$

Και ικανοποιεί:

$$I^* \leq \frac{1+h(\lambda^*)}{1-h(\tilde{\lambda})}$$



Εικόνα 1: Γραφική πρόταση για την απόδειξη του θεωρήματος 3.4.2

Απόδειξη: Έστω $\Lambda = \{\lambda: 0 \leq \lambda \leq \infty, E(V(\lambda)) - \sigma^2 \leq R(\lambda^*)(1+h(\lambda^*))\}$. Εάν

$$E(R(\lambda))(1-h(\lambda)) < E(V(\lambda)) - \sigma^2 < E(R(\lambda))(1+h(\lambda)), 0 \leq \lambda < \infty$$

Και οι $E\big(R(\lambda)\big), E\big(V(\lambda)\big), h$ είναι συνεχείς συναρτήσεις του λ, τότε το Λ είναι μη κενό κλειστό σύνολο. Εάν το μηδέν δεν είναι οριακό σημείο του Λ τότε το E(V(λ)) έχει ένα ελάχιστο στο εσωτερικό του Λ (ή πιθανά στο ∞), που το συμβολίζουμε $\tilde{\lambda}$ (Εικόνα1). Τώρα από το Θεώρημα 3.4.1προκύπτει:

$$E\left(R(\tilde{\lambda})\right)\left(1-h(\tilde{\lambda})\right) < E\left(V(\tilde{\lambda})\right) - \sigma^2 < E\big(R(\lambda^*)\big)\big(1+h(\lambda^*)\big)$$

Εάν το Λ περιλαμβάνει το μηδέν, τότε το $\tilde{\lambda}$ μπορεί να είναι οριακό του Λ, π.χ $\tilde{\lambda}=0$ αλλά το άνω όριο του I^* υπάρχει.

Στη συνέχεια, θα χρησιμοποιήσουμε κάποια λήμματα (οι αποδείξεις των οποίων βρίσκονται στο Appendix)για να αποδείξουμε ότι:

$$h(\lambda^*)\to 0 \text{ και } h(\tilde{\lambda})\to 0 \text{για } n\to\infty.$$

- Λήμμα 3.4.1: Εάν $g\in W_2^{(m)}$, $b^2(\lambda)\le \lambda\int_0^1 (g^{(m)}(u))^2 du$
- Λήμμα 3.4.2: Έστω $\{t_i\}_{i=1}^n \equiv \{t_{in}\}_{i=1}^n$ που ικανοποιεί:

$$\int_0^{t_{in}} w(u)du = \frac{i}{n},\ \ i=1,2,3,\dots,n \qquad\qquad n=1,2,\dots$$

Όπου $w(u)$είναι μια συνεχής, αυστηρά κυρτή συνάρτηση βάρους. Τότε εάν $g\notin \pi_{m-1}$και το λ είναι φραγμένο στο μηδέν όσο $n\to\infty$, τότε το $b(\lambda)^2$ είναι επίσης φραγμένο στο μηδέν.

- Λήμμα 3.4.3: Έστω $\{t_i\}_{i=1}^n$τέτοιο ώστε να ικανοποιούνται οι υποθέσεις του παραπάνω λήμματος με $0<a\le w(t)\le\beta<\infty$. Τότε:

$$\frac{k_m}{\beta^{1/2m}}+o(1)\le n\lambda^{1/2m}\mu_1(\lambda)\le\frac{k_m}{a^{1/2m}}+o(1)$$

$$\frac{l_m}{\beta^{1/2m}}+o(1)\le n\lambda^{1/2m}\mu_2(\lambda)\le\frac{l_m}{a^{1/2m}}+o(1)$$

Όπου:

$$o(1)=O(\lambda)+O(\tfrac{1}{n\lambda^{\frac{1}{2m}}}) \text{ καθώς } \lambda\to 0, n\lambda\to\infty \text{ και}$$

$$k_m = \int_0^\infty \frac{dx}{(1+x^{2m})}, \quad l_m = \int_0^\infty \frac{dx}{(1+x^{2m})^2}$$

Αντίστροφα, εάν το $n\lambda^{1/2m}$είναι φραγμένο από το μηδέν, τότε το ίδιο ισχύει και για τα $\mu_1(\lambda)$, $\mu_2(\lambda)$.

Τονίζουμε ότι μια συνέπεια του παραπάνω λήμματος είναι ότι:

$$\frac{\mu_1(\lambda)^2}{\mu_2(\lambda)} \to 0$$

- Θεώρημα 3.4.3: Έστω $g \in W_2^{(m)}$ και $\{t_i\}_{i=1}^n$ να ικανοποιεί $\frac{i}{n}\int_0^{t_{in}} w(u)du$, όπου w(u) είναι μια αυστηρά θετική και συνεχής συνάρτηση βάρους. Τότε, υπάρχει μια ακολουθία $\tilde{\lambda} = \tilde{\lambda}(n)$ στο ελάχιστο του $E(V(\lambda))$ τέτοιο ώστε:

$$\lim_{n\to\infty} \frac{E(R(\tilde{\lambda}))}{E(R(\lambda^*))} = 1$$

3.5 Αριθμητικά αποτελέσματα

Στην ενότητα αυτή, εφαρμόζουμε τη μέθοδο σε τεχνητά δεδομένα της μορφής:

$$y(t_i) = g(t_i) + \varepsilon_i$$

Όπου ε_i είναι ψευδο-τυχαίες μεταβλητές με μέση τιμή 0 και διασπορά σ^2. Όταν $m = 2$ τότε η $g_{n,\lambda}$ είναι μια κυβική C^∞τάξης spline. Σ'αυτή την περίπτωση, έχει καθιερωθεί (Reinsch)ότι:

$$I - A(\lambda) = \tilde{Q}(\tilde{Q}'\tilde{Q} + p\tilde{Q})\tilde{Q}' \qquad (3.5.1)$$

Όπου $p = \frac{1}{n}\lambda$.

Ο πίνακας $\tilde{Q}$τριδιαγώνιος $n \times (n-2)$με εισόδους $\tilde{q}_{ij}$, $i = 1,2,\dots,n$

$$j = 1,2,\dots,n-2$$

Που ορίζονται ως:

$$\tilde{q}_{i,i+1} = {}^{1}/_{h_{i+1}}$$

$$\tilde{q}_{ii} = -{}^{1}/_{h_i} - {}^{1}/_{h_{i+1}}$$

$$\tilde{q}_{i+1,i} = {}^{1}/_{h_{i+1}}$$

Όπου: $h_i = t_{i+1} - t_i$.

Ο πίνακας $\tilde{T}$είναι επίσης τριδιαγώνιος $(n-2) \times (n-2)$, αυστηρά θετικά ορισμένος, με εισόδους t_{ij} , $i, j = 1,2, \dots, n-2$

Που ορίζονται:

$$\tilde{t}_{ii} = 2(h_i + h_{i+1})/3$$

$$\tilde{t}_{i,i+1} = \tilde{t}_{i+1,i} = h_{i+1}/3$$

Έστω:

$$F = \tilde{Q}\tilde{T}^{-1/2}$$

Όπου $\tilde{T}^{-1/2}$ είναι η συμμετρική τετραγωνική ρίζα του $\tilde{T}^{-1}$.

Όταν ισχύει ότι $h_i \equiv {}^{1}/_{n}$, $i = 1,2, \dots, n$τότε ο $\tilde{T}^{-1/2}$ μπορεί να βρεθεί αναλυτικά με τον εξής τρόπο:

$$\begin{pmatrix} \alpha & \beta & \cdots & 0 & 0 \\ \beta & \alpha & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & & \alpha & \beta \\ 0 & 0 & \cdots & \beta & \alpha \end{pmatrix} = \Gamma D \Gamma' \qquad (3.5.2)$$

Όπου: $\Gamma_{jk} = \sqrt{\frac{2}{n+1}} sin \frac{jk\pi}{n+1}$ και Dείναι διαγώνιος πίνακας με j-οστή είσοδο:

$$\alpha + 2\beta cos \frac{j\pi}{n+1}$$

Και τελικά ο πίνακας $\tilde{T}^{-1/2}$υπολογίζεται από τον τύπο:

$$\tilde{T}^{-1/2} = \Gamma D^{-1/2} \Gamma'$$

Τότε:

$$I - A = F(F'F + pI)^{-1}F' \quad (3.5.3)$$

Έστω ότι ο πίνακας Fμπορεί να παραγοντοποιηθεί στη μορφή (*singular value decomposition*):

$$F = UDV'$$

Με U και Vορθογώνιους πίνακες διαστάσεων $n \times (n-2)$και $(n-2) \times (n-2)$αντίστοιχα και ο D έχει μη-μηδεμικές τιμές στη διαγώνιο που τις συμβολίζουμε $d_2, d_2, \dots, d_{n-2}$ και μηδέν στις υπόλοιπες θέσεις.Τότε:

$$I - A = U \begin{pmatrix} \frac{{d_1}^2}{{d_1}^2+p} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{{d_{n-2}}^2}{{d_{n-2}}^2+p} \end{pmatrix} U' \qquad (3.5.4)$$

$$V(p) = \frac{1}{n}\sum_{j=1}^{n-2}\left(\frac{{d_j}^2}{{d_j}^2+p}\right)^2 {z_j}^2 / [\frac{1}{n}\sum_{j=1}^{n-2}\left(\frac{{d_j}^2}{{d_j}^2+p}\right)]^2 \qquad (3.5.5)$$

Όπου: $z = (z_1, \dots, z_{n-2})' = U'y$

Τα αριθμητικά πειράματα έγιναν ως εξής: Για να συμβαδίσουν τα αριθμητικά αποτελέσματα με την παραπάνω μέθοδο, η οποία όπως αναφέραμε αποδίδεται στον Reinsch, το λ αντικαθίσταται παντού με το $p = \frac{1}{n}\lambda$. Για γνωστά g, σ^2 $και$ n , τα y_iπροκύπτουν από τον τύπο:

$$y_i = g\left(\frac{i-1}{n}\right) + \varepsilon_i, i = 1,2, \dots, n$$

Με ε_i ψευδο-τυχαίες μεταβλητές με μέση τιμή μηδέν και διασπορά σ^2. Το $V(p)$υπολογίστηκε από την εξίσωση (3.5.5) ενώ το $\log_{10} p$χρησιμοποιείται με την προσθήκη της ποσότητας $1/9$. Η ελαχιστοποίηση του pσυμβολίζεται με $\hat{p}$. Για $\hat{\lambda} = \frac{1}{n}\hat{p}$υπολογίζουμε το $g_{n,\lambda}$ ενώ το

$$R(p) = \frac{1}{n}\sum_{i=1}^{n}(g(t_i) - g_{n,\lambda}(t_i))^2$$

χρησιμοποιείται για σύγκριση.Επίσης, χρησιμοποιούνται συναρτήσεις ελέγχου (*testfunctions*)της μορφής:

$$g(t) = \sum_{j=1}^{r} w_j \beta_{p_j,q_j}(t)$$

Με $\beta_{pq}(t) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} t^{p-1}(1-t)^{q-1}$όπου Γ είναι η συνάρτηση Γάμμα.

Τα παραδείγματα είναι:

Παράδειγμα I: r=3 w_1=0,2 p_1=4 q_1=15

w_2=0,7 p_2=5 q_2=7

w_3=0,1 p_3=12 q_3=5

Παράδειγμα II: r=2 w_1=0,4 p_1=12 q_1=7

w_2=0,6 p_2=4 q_2=11

Παράδειγμα III: r=3 w_1=0,5 p_1=10 q_1=30

w_2=0,2 p_2=20 q_2=20

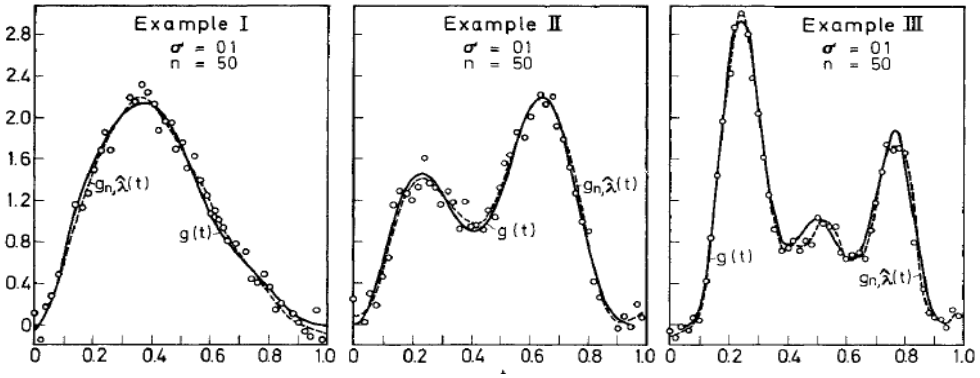
w_3=0,3 p_3=30 q_3=10

Το Διάγραμμα 2 δίνει τη γραφική παράσταση της συνάρτησης $g(t), της\ y_i = g\left(\frac{i}{n}\right) + \varepsilon_i, i = 1,2,\dots,n$ και της $g_{n,\hat{\lambda}}(t)$όπου $\hat{\lambda} = \frac{1}{n}\hat{p}$ και $\hat{p}$ είναι η τιμή που ελαχιστοποιεί την $V(p)$.

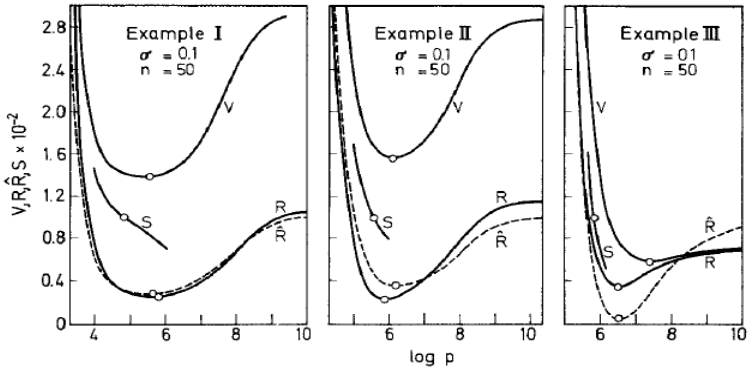
Στα παραδείγματα είναι σ=0,5 και n=50.

Το διάγραμμα 3 δίνει τις γραφικές παραστάσεις των $V(p)$, $R(p)$και $\hat{R}(p)$. Το $\hat{R}(p)$ έχει οριστεί με τον τύπο (3.1.8) παραπάνω και μπορεί επίσης να υπολογιστεί από:

$$\hat{R}(p) = \frac{1}{n}\sum_{j=1}^{n-2}\left(\frac{{d_j}^2}{{d_j}^2 + p}\right)^2 {z_j}^2 + \frac{2\sigma^2}{n}\sum_{j=1}^{n-2}\left(\frac{{d_j}^2}{{d_j}^2 + p}\right) - \sigma^2$$



Εικόνα 2. Παραδείγματα I, II, III, $g(t)$ και $g_{n,\hat{\lambda}}(t)$.



Εικόνα 3.

Table 1. Inefficiencies associated with $V, \hat{R}$ and S

	$\sigma = 0.1$			$\sigma = 0.01$		
	$\frac{R(\hat{p})}{\min_p R(p)}$	$\frac{R(\hat{p}_R)}{\min_p R(p)}$	$\frac{R(\hat{p}_S)}{\min_p R(p)}$	$\frac{R(\hat{p})}{\min_p R(p)}$	$\frac{R(\hat{p}_R)}{\min_p R(p)}$	$\frac{R(\hat{p}_S)}{\min_p R(p)}$
Example I	1.01	1.00	1.21	1.02	1.06	2.38
Example II	1.04	1.10	1.14	1.01	1.04	1.07
Example III	1.42	1.01	2.02	1.22	1.00	2.06
	$\sigma = 0.001$					
Example II	1.12	1.04	1.97			

Πίνακας 1: Ανεπάρκειες που συνδέονται με τα $V, S, \hat{R}$

Στην εικόνα 3, το ελάχιστο από κάθε μια καμπύλη σημειώνεται με ένα κύκλο. Διαλέγουμε το p με τέτοιο τρόπο, ώστε $S(p)/\sigma^2 = 1$. Στο διάγραμμα αναπαρίσταται και το $S(p)$ ενώ το σημείο που ισχύει οτι $S(p) = \sigma^2$ επίσης σημειώνεται με κύκλο. Σε κάθε ένα από τα παραδείγματα I, II και III το $\hat{R}(p)$τείνει στο $R(p)$ ενώ στη γειτονιά της ελαχιστοποίησης του $R(p)$ ισχύει ότι:

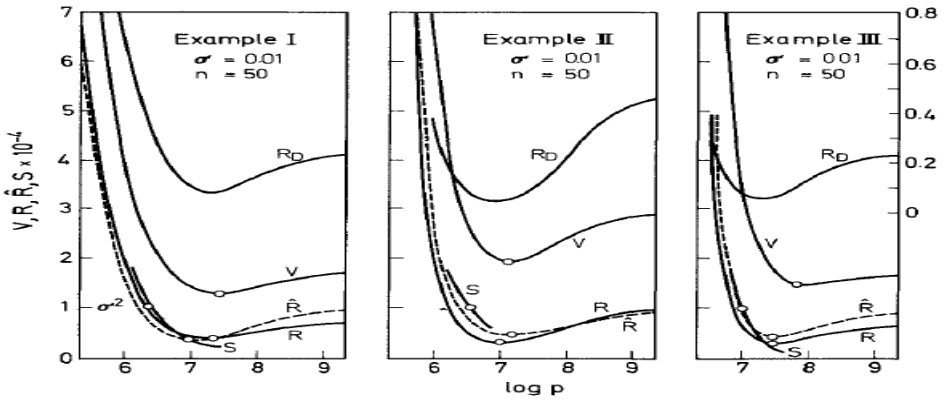
$$V(p) \approx R(p) + const.$$

με τη σταθερά να είναι γύρω από το σ^2. Όπως αναφέραμε και παραπάνω, το ρεπιλέχθηκε με τρόπο ώστε να ισχύει $S(p) = \sigma^2$ και καταλήγοντας σε μικρά p, επιβεβαιώνονται τα θεωρητικά αποτελέσματα.

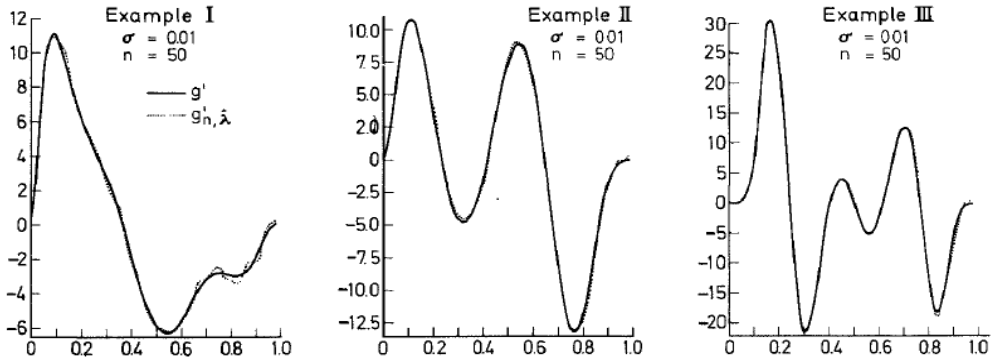
Συμβολίζοντας με $\hat{p}, \hat{p}_{\hat{R}}, \hat{p}_S$τους εκτιμητές του p (που προέκυψε από τη μέθοδο GCV), την τιμή που ελαχιστοποιεί το $\hat{R}(p)$ και την πρόταση του Reinschαντίστοιχα, οι πρώτες τρεις στήλες του πίνακα 1 δίνουν τις παρατηρημένες ανεπάρκειες (inefficiencies):

$$\frac{R(\hat{p})}{\min_p R(p)}, \quad \frac{R(\hat{p}_{\hat{R}})}{\min_p R(p)} \quad και \quad \frac{R(\hat{p}_S)}{\min_p R(p)}$$

Τα πειράματα επαναλήφθηκαν για σ=0,01 και σ=0,001. Για την περίπτωση που σ=0,01 οι γραφικές παραστάσεις του $V, \hat{R}, R \; και \; S$φαίνονται στο διάγραμμα 4, ενώ οι «ανεπάρκειες» φαίνονται στις στήλες 3-6 του πίνακα 1. Οι συναρτήσεις $g, g_{n,\lambda}$ στην περίπτωση που σ=0,01 είναι πανομοιότυπες και δεν αναπαρίστανται.



Εικόνα 4.

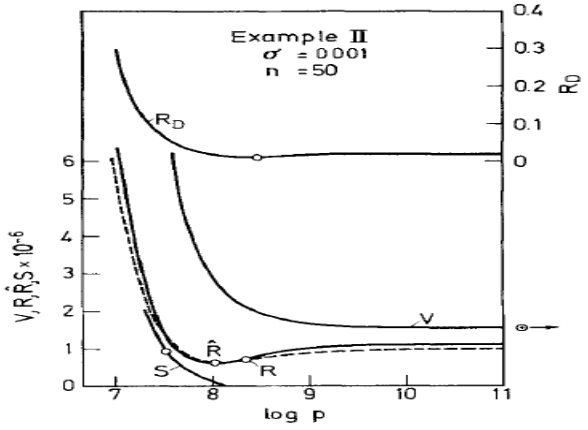


Εικόνα 5.Οι παράγωγοι $g^{'}$ και οι εκτιμητές τους $g^{'}_{n,\hat{\lambda}}$.

Οι καλές εκτιμήτριες της παραγώγου της gεπιτυγχάνονται παραγωγίζοντας την $g_{n,\lambda}$. Οι συναρτήσεις $g^{'}$και$g^{'}_{n,\hat{\lambda}}$. Οι συναρτήσεις $g^{'}$, $g^{'}_{n,\hat{\lambda}}$ αναπαρίστανται στο διάγραμμα 5. Το MSE $R_D(p)$ για την εκτίμηση της παραγώγου:

$$R_D(p) = \frac{1}{n}\sum_{j=1}^{n}(g^{'}\left(\frac{i}{n}\right) - g^{'}_{n,\lambda}(\frac{i}{n}))^2$$

Επίσης αναπαρίσταται στο διάγραμμα 4. Σημειώνουμε ότι το ελάχιστο του $R_D(p)$ είναι κοντά στο ελάχιστο του $R(p)$στα παραδείγματα αυτά, πράγμα που σημαίνει ότι η GCVεκτιμήτρια, καθώς και η τιμή που ελαχιστοποιεί το $\hat{R}(p)$έχουν επιλεγεί σωστά.



Εικόνα 6.Όσο $\sigma^2 \to 0$ το $R(p)$ «πλαταίνει» έτσι ώστε το βέλτιστο $p \to \infty$.

Συμπερασματικά, θα μπορούσαμε να πούμε ότι η μέθοδος γενικευμένης διασταυρωμένης επικύρωσης είναι μια μέθοδος αποδοτική, αφού τόσο σε θεωρητικό

επίπεδο, όσο και με βάση τα παραδείγματα, ανταποκρίνεται στην εύρεση της σωστής παραμέτρου για την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος.

ΚΕΦΑΛΑΙΟ 4

Επιλογή της ρυθμιστικής παραμέτρου στη Μέθοδο SCAD

4.1 Εισαγωγή

Στην ανάλυση παλινδρόμησης, κατά τη διαδικασία επιλογής κατάλληλου μοντέλου, μπορούμε πολύ εύκολα να οδηγηθούμε σε ένα μοντέλο που να περιέχει μικρότερο αριθμό σημαντικών μεταβλητών (*underfitted model*) ή να περιέχει περισσότερες μεταβλητές απ' ότι χρειάζεται (*overfitted model*). Στην πρώτη περίπτωση, το μοντέλο μπορεί να οδηγήσει σε μια αυστηρά αμερόληπτη εκτίμηση, ενώ στη δεύτερη περίπτωση μειώνεται η αποδοτικότητα της παραμέτρου εκτίμησης και πρόβλεψης. Επομένως, κύριος σκοπός μας είναι να παράγουμε ένα λιγότερο σποραδικό, αλλά περισσότερο αποδοτικό μοντέλο πρόβλεψης. Τα παραδοσιακά κριτήρια επιλογής μοντέλου, όπως το AIC (Akaike 1973 [1]) ή το BIC (Schwartz 1978) περιλαμβάνουν ένα μεγάλο αριθμό περιορισμών. Το βασικό μειονέκτημά τους είναι ότι απαιτούνται δύο διαφορετικές διαδικασίες για την επιλογή παραμέτρου εκτίμησης και μοντέλου, με αποτέλεσμα να οδηγούμαστε σε αστάθεια (*instability*) (Breiman 1996 [8]) και σε πολύπλοκες στοχαστικές ιδιότητες (Fan&Li 2001). Επιπλέον, σ' αυτές τις μεθόδους, ο συνολικός αριθμός των υποψήφιων μοντέλων αυξάνεται εκθετικά, όσο αυξάνεται και ο αριθμός των συμμεταβλητών. Οι Fan&Li, προτείνουν μια διαφορετική μέθοδο, την μέθοδο Smoothly Clipped Absolute Deviation, η οποία εκτιμά παραμέτρους και ταυτόχρονα επιλέγει σημαντικές μεταβλητές. Εάν τη συγκρίνουμε με άλλες δημοφιλείς μεθόδους επιλογής μοντέλου, π.χ. με την Lasso του Tibshirani (1996), η μέθοδος SCAD όχι μόνο επιλέγει σημαντικές μεταβλητές με συνέπεια, αλλά παράγει επίσης και εκτιμητές παραμέτρων τόσο αποτελεσματικά όσο εάν το πραγματικό μοντέλο ήταν εκ των προτέρων γνωστό. Τα παραπάνω χαρακτηριστικά της μεθόδου SCAD βασίζονται στην κατάλληλη επιλογή ρυθμιστικής παραμέτρου (*tuning parameter*) η οποία συνήθως επιλέγεται με τη βοήθεια της μεθόδου Γενικευμένης Διασταυρωμένης Επικύρωσης (GCV) (Craven&Wahba 1979 [9]).

Στη συνέχεια, θα δείξουμε ότι χρησιμοποιώντας τη μέθοδο GCV για την εύρεση της βέλτιστης παραμέτρου προσαρμογής καταλήγουμε σε ένα μοντέλο που περιέχει

περισσότερες μεταβλητές απ' όσες χρειάζονται (*overfitted model*) εάν το μέγεθος του δείγματος είναι πολύ μεγάλο. Επιπλέον, θα παρουσιαστεί και ένας άλλος τρόπος επιλογής της ρυθμιστικής παραμέτρου, που βασίζεται στο BIC. Από αυτή τη διαδικασία θα προκύψει ότι το μοντέλο που θα πάρουμε, θα ταυτίζεται με το πραγματικό μοντέλο.

4.2 Η μέθοδος SCAD

Θεωρούμε το γραμμικό μοντέλο παλινδρόμησης:

$$y_i = x_i^{'}\beta+\varepsilon_i \quad (4.2.1)$$

Όπου:

y_iείναι η απόκριση για την i-οστή μεταβλητή

$x_i = (x_{i1}, \dots, x_{id})^{'}$είναι η d-διάστατη επεξηγηματική μεταβλητή

$\beta = (\beta_1, \dots, \beta_d)^{'}$

ε_i είναι το τυχαίο σφάλμα με μέση τιμή μηδέν και διασπορά σ_ε^2

Έστω $(x_i, y_i), i = 1, \dots, n$ τυχαίο δείγμα.

Για να επιλέξουμε μεταβλητές και να εκτιμήσουμε παραμέτρους ταυτόχρονα, οι Fan&Li προτείνουν την μέθοδο SCAD, η οποία εκτιμά τα βελαχιστοποιώντας τη συνάρτηση των ποινικοποιημένων ελαχίστων τετραγώνων:

$$\frac{1}{2n}\|Y - X\beta\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|) \quad (4.2.2)$$

Όπου:$Y = (y_1, \dots, y_n)^{'}$

$X = (x_1, \dots, x_n)^{'}$

$\|.\|$ η ευκλείδια νόρμα

p_λ η ποινή SCAD με ρυθμιστική παράμετρο λ, η οποία επιλέγεται από μια μέθοδο που βασίζεται στα δεδομένα.

Η ποινή p_λ ικανοποιεί:

$$p_\lambda(0) = 0$$

ενώ η πρώτη παράγωγος δίνεται από:

$$p_\lambda'(\theta) = \lambda\left\{I(\theta \le \lambda) + \frac{(\alpha - \lambda\theta)_+}{(\alpha-1)\lambda} I(\theta > \lambda)\right\}$$

όπου α είναι μια σταθερά που συνήθως επιλέγεται να είναι ίση με α=3,7 (Fan&Li), ενώ $(t)_+ = tI\{t > 0\}$ είναι μια «εξαρτώμενη συνάρτηση ζημίας» (*hinge loss function*). Για δεδομένη παράμετρο προσαρμογής ο εκτιμητής επιτυγχάνεται ελαχιστοποιώντας την (4.2.2) με $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda 1}, \dots, \hat{\beta}_{\lambda d})'$. Οι Fan&Li (2001) στη μελέτη που έκαναν, έδειξαν ότι εάν $\lambda \to \infty$ και $\sqrt{n}\lambda \to \infty$ για $n \to \infty$τότε η μέθοδος αναγνωρίζει τις ασυσχέτιστες μεταβλητές παράγοντας μηδενικές λύσεις για τους σχετικούς συντελεστές παλινδρόμησης. Επίσης, η μέθοδος εκτιμά τους συντελεστές των μεταβλητών με την ίδια συνέπεια όπως εάν το πραγματικό μοντέλο να ήταν γνωστό, μια ιδιότητα που ονομάζεται ιδιότητα πρόβλεψης (*oracle property*).

Στην πράξη, το λ συνήθως επιλέγεται ελαχιστοποιώντας το κριτήριο γενικευμένης διασταυρωμένης επικύρωσης (GCV):

$$GCV_\lambda = \frac{\|Y - X\widehat{\beta_\lambda}\|^2}{n(1 - \frac{DF_\lambda}{n})^2} = \frac{\hat{\sigma}_\lambda^2}{(1 - \frac{DF_\lambda}{n})^2} \qquad (4.2.3)$$

Όπου: $\hat{\sigma}_\lambda^2 = n^{-1}\left\|Y - X\widehat{\beta_\lambda}\right\|^2$και DF_λ οι γενικευμένοι βαθμοί ελευθερίας (Fan&Li, 2001) που υπολογίζονται ως:

$$DF_\lambda = tr\{X(X'X + n\Sigma_\lambda)^{-1}X'\}$$

και:

$$\Sigma_\lambda = diag\{\frac{p_\lambda'(|\hat{\beta}_{\lambda 1}|)}{|\hat{\beta}_{\lambda 1}|}, \dots, \frac{p_\lambda'(|\hat{\beta}_{\lambda d}|)}{|\hat{\beta}_{\lambda d}|}\}$$

Τα διαγώνια στοιχεία του Σ_λείναι συμμεταβλητές του τετραγωνικού όρου του τετραγωνικού υπολογισμού της συνάρτησης ποινής $p_\lambda(.)$της SCAD. Εφόσον κάποιες μεταβλητές του εκτιμητή β είναι ακριβώς ίσες με μηδέν, το DF_λυπολογίζεται

αντικαθιστώντας το X με τον υποπίνακα απόκρισης με τις επιλεχθείσες μεταβλητές και αντικαθιστώντας το Σ_λ με τον αντίστοιχο υποπίνακα.

Η βέλτιστη παράμετρος προσαρμογής που προκύπτει είναι:

$$\hat{\lambda}_{GCV} = arg \min_{\lambda} GCV_\lambda$$

Ο λογαριθμοποιημένος μετασχηματισμός του GCV_λ είναι:

$$logGCV_\lambda = log\hat{\sigma}_\lambda^2 - 2\log\left(1 - \frac{DF_\lambda}{n}\right) \approx log\hat{\sigma}_\lambda^2 + 2\frac{DF_\lambda}{n} = AIC_\lambda$$

Έτσι, η $logGCV_\lambda$είναι πολύ όμοια με το παραδοσιακό κριτήριο επιλογής μοντέλου AIC που είναι αποδοτικό με την έννοια ότι επιλέγει το καλύτερο πεπερασμένης διάστασης υποψήφιο μοντέλο με όρους προβλεπτικής ακρίβειας, όταν το πραγματικό μοντέλο είναι άπειρης διάστασης. Παρ'όλ'αυτά το AIC δεν είναι ένα συνεπές κριτήριο επιλογής επειδή δεν επιλέγει το σωστό μοντέλο με πιθανότητα να τείνει στη μονάδα σε μεγάλα δείγματα, όταν το πραγματικό μοντέλο είναι άπειρης διάστασης. Συνεπώς, το μοντέλο που επιλέγεται από την $\hat{\lambda}_{GCV}$ ίσως δεν αναγνωρίσει το πεπερασμένης διάστασης μοντέλο με συνέπεια. Έτσι, θα εξετάσουμε στη συνέχεια ένα νέο κριτήριο επιλογής μοντέλου που είναι περισσότερο συνεπές, το BIC (Schartz, 1978). Σύμφωνα με αυτό το κριτήριο, το βέλτιστο λ επιλέγεται ελαχιστοποιώντας:

$$BIC_\lambda = log\hat{\sigma}_\lambda^2 + \frac{DF_\lambda \log(n)}{n} \quad (4.2.4)$$

Η βέλτιστη παράμετρος που προκύπτει συμβολίζεται με $\hat{\lambda}_{BIC}$.

4.3 Απαραίτητες Συνθήκες

Υποθέτουμε ότι υπάρχει ένας ακέραιος $0 \le d_0 \le d$ τέτοιος ώστε $\beta_{jk} \neq 0$ για $1 \le k \le d_0$, με όλα τα άλλα $\beta_j = 0$. Έτσι, το πραγματικό μοντέλο περιέχει τις $j_1, \dots, j_{d0}$ μεταβλητές, σαν σημαντικές μεταβλητές. Επιπλέον, με σκοπό να ορίσουμε τα «υποπροσαρμοσμένα» (*underfitted*) και τα «υπερπροσαρμοσμένα» (*overfitted*) μοντέλα, συμβολίζουμε με $S_F = \{1, \dots, d\}, S_T = \{j_1, \dots, j_{d0}\}$ το πλήρες και το πραγματικό υπομοντέλο αντίστοιχα. Τότε κάθε υποψήφιο μοντέλο $S \not\supset S_T$αναφέρεται σαν υποπροσαρμοσμένο με την έννοια ότι χάνει τουλάχιστον μια σημαντική

μεταβλητή. Αντίθετα, κάθε $S \supset S_T$ αναφέρεται σαν υπερπροσαρμσμένο με την έννοια ότι περιέχει όλες τις σημαντικές μεταβλητές αλλά και τουλάχιστον μια μη σημαντική μεταβλητή.

Για ένα αυθαίρετο μοντέλο $S = \{j_1, \dots, j_{d^*}\} \subset S_F$ συμβολίζουμε τον πίνακα συμμεταβλητών με X_s ο οποίος είναι ένας $n \times d^*$πίνακας με την i-οστή γραμμή να δίνεται ως$(x_{ij\,1}, \dots x_{ij\,d^*})$. Αφού προσαρμόσουμε τα δεδομένα με το μοντέλο S μέσω των ελαχίστων τετραγώνων, συμβολίζουμε τον εκτιμητή ελαχίστων τετραγώνων, το άθροισμα τετραγώνων των υπολοίπων, τον εκτιμητή της διασποράς και την τιμή της GCV ως:

$$\hat{\beta}_s = (X_s' X_s)^{-1}(X_s' Y) \qquad (4.3.1)$$

$$SSE_s = \left\| Y - X_s \hat{\beta}_s \right\|^2 \qquad (4.3.2)$$

$$\hat{\sigma}_s = SSE_s / n (4.3.3)$$

$$GCV_s = n^{-1} SSE_s / (1 - \frac{d^*}{n})^2 \qquad (4.3.4)$$

αντίστοιχα.

Ο εκτιμητής SCAD $\hat{\beta}_\lambda$προκύπτει ελαχιστοποιώντας τη συνάρτηση (4.2.2) και ταυτίζεται με το μοντέλο $S_\lambda = \{j : \hat{\beta}_{\lambda j} \neq 0\}$ για το οποίο ο εκτιμητής ελαχίστων τετραγώνων είναι $\hat{\beta}_{S_\lambda}$. Από τον ορισμό του εκτιμητή ελαχίστων τετραγώνων έχουμε:

$$SSE_\lambda = \left\| Y - X\hat{\beta}_\lambda \right\|^2 \geq \left\| Y - X_{S_\lambda} \hat{\beta}_{S_\lambda} \right\|^2 = SSE_{S_\lambda} (4.3.5)$$

Επιπλέον, η $\hat{\sigma}_\lambda$ που ορίζεται στην (4.2.3) μπορεί απλά να εκφραστεί ώς:

$$\hat{\sigma}_\lambda = SSE_\lambda / n.$$

Εάν λ=0, τότε ο όρος ποινής στην (4.2.2) είναι μηδέν και το $\hat{\beta}_0$είναι ακριβώς το ίδιο με τον εκτιμητή ελαχίστων τετραγώνων $\hat{\beta}_{S_F}$του πλήρους μοντέλου. Επιπλέον,

$$SSE_0 = SSE_{S_F}$$

$$\hat{\sigma}_0^2 = \hat{\sigma}_{S_F}^2$$

$$GCV_0 = GCV_{S_F}$$

Στην πράξη, παρ'όλ'αυτά, το λ είναι άγνωστο οπότε είναι απαραίτητο να ψάξουμε το βέλτιστο λ στο θετικό πραγματικο R^+ή μέσα στο φραγμένο ολοκλήρωμα $\Omega = [0, \lambda_{max}]$, όπου λ_{max} κάποιο άνω φράγμα.

Τώρα θα παρουσιάσουμε κάποιες τεχνικές συνθήκες που είναι απαραίτητες για τη μελέτη των θετικών ιδιοτήτων των εκτιμητών της ρυθμιστικής παραμέτρου (*tuning parameter*).

Συνθήκη 1. Για κάθε $S \subset S_F$, $\exists \hat{\sigma}_S^2 > 0$τέτοιο ώστε $\hat{\sigma}_S^2 \to \sigma_S^2$ με βεβαιότητα.

Συνθήκη 2. Για κάθε $S \not\supset S_T$, έχουμε $\sigma_S^2 > \sigma_{S_T}^2$, όπου $\sigma_{S_T}^2$ είναι μια θετική τιμή τέτοια ώστε $\hat{\sigma}_{S_T}^2 \to \sigma_{S_T}^2$ με βεβαιότητα.

Συνθήκη 3. Τα ε_iείναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ακολουθούν την κατανομή $N(0, \sigma_\varepsilon^2)$.

Συνθήκη 4. Για το άνω όριο είναι $\lambda_{max} \to 0$ για $n \to \infty$.

Συνθήκη 5. Ο πίνακας $Cov(x_i) = \Sigma_x$ είναι πεπερασμένος και θετικά ορισμένος.

Η πρώτη συνθήκη διευκολύνει την απόδειξη της ασυμπτωτικής συμπεριφοράς, ενώ η δεύτερη διευκρινίζει το «υποπροσαρμοσμένο» αποτέλεσμα. Και οι δύο συνθήκες ικανοποιούνται εάν τα (x_i, ε_i) έχουν από κοινού πολυδιάστατη κανονική κατανομή. Η τρίτη συνθήκη είναι απαραίτητη για την εκτίμηση του αποτελέσματος χρησιμοποιώντας τη μέθοδο GCV, που θα δουμε παρακάτω.Η τέταρτη συνθήκη υπονοεί ότι $\lambda \to \infty$ για $n \to \infty$.Τέλος, η πέμπτη συνθήκη επιβεβαιώνει την $\sqrt{n}$-συνέπεια του μη-ποινικοποιημένου εκτιμητή.

4.3.1 Το αποτέλεσμα της Γενικευμένης Διασταυρωμένης Επικύρωσης

Ορίζουμε:

$$\Omega_- = \{\lambda \in \Omega: S_\lambda \not\supset S_T\}$$

$$\Omega_0 = \{\lambda \in \Omega: S_\lambda = S_T\}$$

$$\Omega_+ = \{\lambda \in \Omega: S_\lambda \supset S_T \kappa\alpha\iota\, S_\lambda \neq S_T\}$$

Με άλλα λόγια, τα $\Omega_-, \Omega_0, \Omega_+$είναι τρία υποσύνολα του Ω, από τα οποία μπορούν να παραχθούν το υποπροσαρμοσμένο, το πραγματικό και το υπερπροσαρμοσμένο μοντέλο αντίστοιχα.

Αρχικά, θα δείξουμε ότι η μέθοδος SCAD με το GCV είναι μια μέθοδος συντηρητική, με την έννοια ότι δεν χάνει καμία σημαντική μεταβλητή, όταν το μέγεθος του δείγματος είναι αρκούντος μεγάλο.

- Λήμμα 4.3.1.1: Υπό τις συνθήκες 1 και 2 έχουμε:

$$pr(inf_{\lambda\in\Omega_-} GCV_\lambda > GCV_{S_F} = GCV_0) \to 1$$

Συμφωνα με το Λήμμα αυτό, η GCVπου υπολογίζει την παράμετρο προσαρμογής λ, είναι μεγαλύτερη απο το $GCV_{S_F} = GCV_0$. Σαν αποτέλεσμα, το βέλτιστο μοντέλο που επιλέγεται από την ελαχιστοποίηση της GCV, π.χ το $S_{\hat{\lambda}_{GCV}}$ πρέπει να περιέχει όλες τις σημαντικές μεταβλητές με πιθανότητα να τείνει στη μονάδα. Παρ' όλ'αυτά, δεν είναι απαραίτητο ότι το $S_{\hat{\lambda}_{GCV}}$ είναι το πραγματικό μοντέλο S_T.

Στο επόμενο λήμμα δείχνουμε ότι το βέλτιστο μοντέλο που επιλέγεται από την GCV υπερκαλύπτει το πραγματικό μοντέλο, με πιθανότητα να τείνει στη μονάδα.

- Λήμμα 4.3.1.2: Υπό τις συνθήκες 1-3, υπάρχει μη μηδενική πιθανότητα $\alpha > 0$τέτοια ώστε:

$$liminf_n pr(inf_{\lambda\in\Omega_0} GCV_\lambda > GCV_{S_F} = GCV_0) \geq \alpha$$

Σύμφωνα με αυτό το Λήμμα, υπάρχει μη μηδενική πιθανότητα ώστε η μικρότερη τιμή της GCVπου συνδέεται με το πραγματικό μοντέλο, να είναι μεγαλύτερη από την GCVτου πλήρους μοντέλου. Έτσι, υπάρχει μη μηδενική πιθανότητα ότι κάθε λ που συνδέεται με το πραγματικό μοντέλο δεν μπορεί να επιλεχθεί από την GCVόπως η βέλτιστη παράμετρος προσαρμογής.

Συνδιάζοντας τα δυο παραπάνω Λήμματα, προκύπτει το επόμενο θεώρημα.

- Θεώρημα 4.3.1.1: Εάν οι συνθήκες 1-3 ισχύουν, τότε υπάρχει μια μη μηδενική πιθανότητα $\alpha > 0$ τέτοια ώστε:

$$pr\left(S_{\hat{\lambda}_{GCV}} > S_T\right) \to 1$$

και:

$$\liminf_{n} pr(S_{\hat{\lambda}_{GCV}} \supset S_T \ \ και\, S_{\hat{\lambda}_{GCV}} \neq S_T) > \alpha$$

Το παραπάνω θεώρημα δείχνει ότι με πιθανότητα που τείνει στη μονάδα, το μοντέλο $S_{\hat{\lambda}_{GCV}}$ περιέχει όλες τις σημαντικές μεταβλητές, αλλά με μη μηδενική πιθανότητα περιλαμβάνει παραπανίσιες μεταβλητές, με αποτέλεσμα να οδηγεί σε υπερπροσαρμοσμένο μοντέλο.

4.3.2 Η συνέπεια της μεθόδου BIC

Για να δείξουμε τη συνέπεια της BIC, κατασκευάζουμε μια ακολουθία που αναφέρεται στην ρυθμιστική παράμετρο,

$$\lambda_n = \log n/\sqrt{n}$$

Έτσι,$\lambda_n \to 0$ και $\sqrt{n}\lambda_n \to \infty$. Οι Fan&Li (2001), είπαν ότι κάτω από συνθήκες ομαλότητας ισχύει ότι:

$$pr(S_{\lambda_n} = S_T) \to 1$$

Αυτό υπονοεί ότι το μοντέλο που προκύπτει από την ακολουθία της ρυθμιστικής παραμέτρου, συγκλίνει στο σωστό μοντέλο, όσο το δείγμα μεγαλώνει.

- Λήμμα 4.3.2.1: Υπό τη συνθήκη 5,

$$pr(BIC_{\lambda_n} = BIC_{S_T}) \to 1$$

Σύμφωνα με το Λήμμα, με πιθανότητα να τείνει στη μονάδα,

$$BIC_{\lambda_n} = BIC_{S_T} = \log \hat{\sigma}^2_{S_T} + d_0 \log(n)\, /n$$

Εφαρμόζοντας αυτό το αποτέλεσμα, τελικά προκύπτει ότι για κάθε λ που δεν μπορεί να ανιχνεύσει το πραγματικό μοντέλο, η τιμή της BICείναι μεγαλύτερη από την BIC_{λ_n}.

- Λήμμα 4.3.2.2: Υπό τις συνθήκες 1, 2, 4 και 5,

$$pr(inf_{\lambda \in \Omega_- \cup \Omega_+} BIC_\lambda > BIC_{\lambda_n}) \to 1$$

Σ'αυτό το σημείο, σημειώνουμε ότι αυτό το Λήμμα δεν υπονοεί απαραίτητα ότι $\lambda_n = \hat{\lambda}_{BIC}$. Παρ'όλ'αυτά, αποδεικνύει ότι εκείνα τα λ που αποτυγχάνουν στο να βρουν το σωστό μοντέλο, δεν μπορούν να επιλεχθούν από την BIC ασυμπτωτικά, διότι το πραγματικό μοντέλο που ανιχνεύεται από το λ_nείναι η καλύτερη επιλογή. Σαν αποτέλεσμα, η βέλτιστη τιμή $\hat{\lambda}_{BIC}$ μπορεί να είναι μόνο μιαπροκύπτει από εκείνα τα λ, των οποίων ο εκτιμητής SCADδίνει προτεραιότητα στο πραγματικό μοντέλο.

- Θεώρημα 4.3.2.1: Εάν ισχύουν οι συνθήκες 1, 2, 4, 5:

$$pr(S_{\hat{\lambda}_{BIC}} = S_T) \to 1$$

Εκτός από την GCV και την BIC υπάρχουν και άλλα κριτήρια που μπορούν να χρησιμοποιηθούν για την εύρεση της ρυθμιστικής παραμέτρου στη μέθοδο SCAD. Τέτοιες μέθοδοι είναι για παράδειγμα το AIC και το BIC (Shi&Tsai, 2002). Παρόμοιες τεχνικές με αυτές που αναφέραμε παραπάνω δείχνουν ότι το AIC λειτουργεί εξίσου καλά με την GCV, υπάρχει όμως πιθανότητα το μοντέλο που θα προκύψει να περιέχει παραπανίσιες μεταβλητές. Αντίθετα, η BIC μέθοδος βρίσκει με συνέπεια το πραγματικό μοντέλο.

4.4 Το Μερικώς Γραμμικό Μοντέλο

Οι Fan&Li(2004) στην εργασία τους αναφέρθηκαν στα μερικώς γραμμικά μοντέλα και επέκτειναν την μη-κοίλη μέθοδο ποινικοποιημένων ελαχίστων τετραγώνων με διαμήκη δεδομένα στην περίπτωση αυτή. Επίσης, έδειξαν ότι ο εκτιμητής που προκύπτει συμπεριφέρεται εξίσου καλά με τον συντελεστή πρόβλεψης. Για το σκοπό αυτό, χρησιμοποίησαν την μέθοδο GCV για τον προσδιορισμό της ρυθμιστικής παραμέτρου. Στη συνέχεια, θα δείξουμε ότι αυτός ο τρόπος οδηγεί σε μοντέλα με παραπανίσιες μεταβλητές, σε αντίθεση με την μέθοδο BIC, η οποία οδηγεί στο πραγματικό μοντέλο με συνέπεια.

Θεωρούμε το μερικώς γραμμικό μοντέλο:

$$y_i = a(u_i) + x_i'\beta + \varepsilon_i$$

Όπου:u_iείναι μια μεταβλητή

$a(u_i)$ είναι μια μη παραμετρική και λεία συνάρτηση του u_i

Και όλα τα υπόλοιπα όπως έχουν οριστεί προηγουμένως στην (4.2.1)

Για την εύρεση του καλύτερου εκτιμητή έχουν προταθεί διάφορες μέθοδοι (EngleetAl. 1986, Heckman 1986, Robinson 1988, Speckman 1988) , ενώ εκτενής ανάλυση για το μερικώς γραμμικό μοντέλο δίνεται από τον Hardle et Al. (2000).

Όμοια με την (4.2.2) υποθέτουμε ότι:

$$\frac{1}{2n}\|Y-\theta-X\beta\|^2+\sum_{j=1}^{d}p_\lambda(|\beta_j|) \qquad (4.4.2)$$

Είναι η συνάρτηση ποινικοποιημένων ελαχίστων τετραγώνων όπου:

$\theta=(\alpha(u_1),\dots,\alpha(u_n))'$

$p_\lambda(|\beta_j|)$η συνάρτηση ποινής όπως ορίστηκε στην (4.2.2)

$\alpha(.)$μια μη παραμετρική λεία συνάρτηση

Για να επιτύχουμε τον εκτιμητή ποινικοποιημένων ελαχίστων τετραγώνων πρώτα εφαρμόζουμε την τεχνική των Fan&Li (2004) για να εξαλείψουμε την ενοχλητική παράμετρο θ, για δεδομένο β. Σαν αποτέλεσμα, έχουμε:

$$y_i^*=a(u_i)+\varepsilon_i \qquad (4.4.3)$$

Οπου:$y_i^*=y_i-x_i'\beta$

Τότε, χρησιμοποιούμε την προσέγγιση της «τοπικής γραμμικής παλινδρόμησης» (*local linear regression*)των Fan&Gijbel (1996) για να εκτιμήσουμε το $a(.)$. Για u σε μια γειτονιά του u_i βρίσκουμε το $(\hat{a}_0,\hat{a}_1)$ ελαχιστοποιώντας το :

$$\sum_{i=1}^{n}\{y_i^*-a_0-a_1(u_i-u)\}^2K_h(u_i-u)$$

Όπου: $K(.)$είμαι η συνάρτηση Kernel

h είναι μια παράμετρος (*bandwidth*)

$K_h(.)=h^{-1}K(\frac{\cdot}{h})$

Ο γραμμικός εκτιμητής στο u είναι:$\hat{a}(u;\beta) = \hat{a}_0$.

Αφού ο τοπικός εκτιμητής είναι λείος γραμμικός , το $\hat{\theta}$παίρνει την έκφραση κλειστής μορφής:

$\hat{\theta} = S_h(Y - X\beta)$(4.4.4)

Όπου με S_h συμβολίζουμε τον λείο πίνακα που ανταποκρίνονται στον τοπικό γραμμικό εκτιμητή και εξαρτάται από το u_i και το $K_h(.)$.

Αντικαθιστώντας το θ στην (4.4.2) με $\hat{\theta}$, επιτυγχάνουμε τη μέθοδο ποινικοποιημένων ελαχίστων τετραγώνων:

$$\frac{1}{2n}\|(I - S_h)Y - (I - S_h)X\beta\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (4.4.5)$$

όπου: I είναι ο $(n \times n)$ μοναδιαίος πίνακας.

Κάτω από συγκεκριμένες συνθήκες ομαλότητας οι Fan&Li «ίδρυσαν» την ιδιότητα πρόβλεψης για τον εκτιμητή ποινικοποιημένων ελαχίστων τετραγώνων.

Βασιζόμενοι στην (4.4.5) ορίζουμε την GCV_λ για το παραπάνω πρόβλημα αντικαθιστώντας τα X και Y με $X_h = (I - S_h)X$ και $Y_h = (I - S_h)Y$αντίστοιχα. Με ανάλογο τρόπο, ορίζουμε την BIC_λαντικαθιστώντας τα Χκαι Υστον υπολογισμό του $\hat{\sigma}_\lambda^2$ στην (4.2.4) με X_h και Y_hαντίστοιχα.

Εφαρμόζοντας την τιμή που έχει προκύψει από την GCV_λ ή την BIC_λ μπορούμε τελικά να υπολογίσουμε τον εκτιμητή SCAD για το β. Όπως έδειξαν οι Fan&Liστην εργασία τους, ο εκτιμητής ποινικοποιημένων τετραγώνων $\hat{\beta}_\lambda$ είναι $\sqrt{n}$-συνεπής αφού $\lambda \to 0$ και $\sqrt{n}\lambda \to \infty$ όσο το $n \to \infty$.

Κάτω από συνθήκες ομαλότητας, η ασυμπτωτική εκτίμηση και διασπορά του $\hat{a}(u)$ είναι τάξης $O_p(h^2)$και $O_p(1/nh)$ αντίστοιχα, εφόσον η παραμετρική σύγκλιση της τιμής του $\hat{\beta}$ είναι ταχύτερη από την μη παραμετρική σύγκλιση της τιμής του $\hat{a}(u)$. Επιπρόσθετα, ισχύει ότι:

$$sup_{u\in U}|\hat{a}(u) - a(u)| = O_p(n^{-1/4})$$

όπου Uείναι το πεδίο τιμών του u.

Σημείωση: Στα παραδείγματα που θα ακολουθήσουν παρακάτω, για να διευκολύνουμε την εύρεση της ποσότητας h και της ρυθμιστικής παραμέτρου λ, θεωρούμε ότι το $\hat{\beta}$ είναι $\sqrt{n}$-συνεπής εκτιμητής του β. Αυτό μας βοηθά να αντικαταστήσουμε το β στην έκφραση των y_i^*με τον είναι $\sqrt{n}$-συνεπή εκτιμητή. Ακολουθώντας την προσέγγιση των Fan&Li, θα δούμε ότι οι αρχικοί όροι στην ασυμπτωτική εκτιμήτρια και τη διασπορά του $\hat{a}(u)$ είναι ίδιοι με αυτούς που προκύπτουν εάν αντικαταστήσουμε το β με την πραγματική του τιμή. Με αυτό τον τρόπο, επιλέγουμε το h και το λ ταυτόχρονα, επομένως επιταχύνουμε τους υπολογισμούς μας. Πιο συγκεκριμένα, εφαρμόζουμε τη μέθοδο των Fan&Li, αντικαθιστώντας το β στην έκφραση των y_i^*με τον εκτιμητή, έτσι ώστε η (4.4.3) να μετατραπεί σε μονοδιάστατο πρόβλημα εξομάλυνσης (*smoothing*). Στη συνέχεια, χρησιμοποιούμε έναν ομαλό εκτιμητή για να επιλέξουμε το h.Εδώ χρησιμοποιούμε μια «plug-in»μέθοδο (Ruppert et Al.,1995). Τελικά, μέσω της GCV_λ ή της BIC_λ βρίσκουμε το λ.

4.5 Αριθμητική Μελέτη

Σ' αυτό το κομμάτι, θα εξετάσουμε ένα πεπερασμένο δείγμα και θα προσαρμόσουμε τις παραμέτρους που έχουν προκύψει από την μέθοδο BIC και την GCV και στα δύο μοντέλα σφάλματος, δηλαδή στο μοντέλο έλλειψης προσαρμογής (*lack-of-fit*) και στο μοντέλο πολυπλοκότητας. Όμως, δεν θα συγκρίνουμε τη μέθοδο SCADμε την μέθοδο μέθοδο επιλογής καλύτερου υποσυνόλου μέσω της BIC όπως έκαναν οι Fan&Li στην εργασία τους. Για να διευκολύνουμε περισσότερο τη διαδικασία, εφαρμόζουμε τον τετραγωνικό αλγόριθμο προσέγγισης κατευθείαν για να βρούμε την λύση της SCAD. Θα θέσουμε την οριακή (*threshold*) συρρίκνωση (*shrinkage*) $\hat{\beta}_j$ από 0 σε 10^{-6}. Σαν αποτέλεσμα, ο μέσος όρος μηδενικών που προκύπτει από τηνρυθμιστική παράμετρο μέσω της μεθόδου GCV είναι πολύ μικρότερος από αυτόν που βρήκαν οι Fan&Listην εργασία τους (2001). Οι παρακάτω προσομοιώσεις έγιναν με το Matlab.

Κατ΄αρχήν, θεωρούμε το μοντέλο των Fan&Li (2001) για το μέτρο σφάλματος. Έστω (u, x, y) μια νέα παρατήρηση για το μοντέλο παλινδρόμησης με :

$$E(y|u,x) = \mu(u,x)$$

και έστω $\hat{\mu}(.,.)$ ένας εκτιμητής για τη συνάρτηση παλινδρόμησης που βασίζεται στα δεδομένα $\{(u_i, x_i, y_i), i = 1, \dots, n\}$. Τότε το μοντέλο σφάλματος ορίζεται ως:

$$E\{\hat{\mu}(u,x) - \mu(u,x)\}^2$$

όπου η αμεροληψία είναι η υπό συνθήκη αμεροληψία που προκύπτει από τα δεδομένα που χρησιμοποιούνται για τον υπολογισμό του $\hat{\mu}(.,.)$. Για ένα μερικώς γραμμικό μοντέλο:

$$\mu(u,x) = a(u) + x'\beta$$

Το μοντέλο σφάλματος είναι:

$$E\{\hat{a}(u) - a(u)\}^2 + E\left(x'\hat{\beta} - x'\beta\right) + 2E\{\hat{a}(u) - a(u)\}(x'\hat{\beta} - x'\beta) \qquad (4.5.1)$$

Ο πρώτος όρος στην (4.5.1) μετρά την μη παραμετρική προσαρμογή ενώ ο δεύτερος όρος αξιολογεί την παραμετρική προσαρμογή. Για να διερευνήσουμε την απόδοση της μεθόδου SCAD στο στοιχείο της παραμετρικής παλινδρόμησης επιλέγουμε την προσομοίωση ώστε το εξωτερικό γινόμενο (*cross product*) στην παραπάνω εξίσωση να ισούται με το μηδέν. Να σημειώσουμε ότι για το γραμμικό μοντέλο παλινδρόμησης (4.2.1) το μοντέλο σφάλματος ισούται ακριβώς με:

$$E\left(x'\hat{\beta} - x'\beta\right)^2$$

Για να συγκρίνουμε τις μεθόδους GCV και BIC ορίζουμε το μοντέλο σφάλματος ως:

$$ME\left(\hat{\beta}\right) = E\left(x'\hat{\beta} - x'\beta\right)^2 = \left(\hat{\beta} - \beta\right)'E(x'x)(\hat{\beta} - \beta)$$

Και ορίζουμε το σχετικό μοντέλο σφάλματος ως:

$$RME = ME/ME_{S_F}$$

όπου ME_{S_F} είναι το μοντέλο σφάλματος του πλήρους μοντέλου S_Fμε τον μη ποινικοποιημένο εκτιμητή ελαχίστων τετραγώνων $\hat{\beta}_{S_F}$.Επιπλέον, υπολογίζουμε το ποσοστό που το μοντέλο προσαρμόζεται ακριβώς από τις μεθόδους GCV και BICκαθώς και το μέσο όρο των μηδενικών μεταβλητών που παράγονται με τη μέθοδο SCAD.

Παράδειγμα 1: Προσομοιώνουμε 1000 δεδομένα κάθε ένα από τα οποία αποτελείται από τυχαίο δείγμα μεγέθους n, από το γραμμικό μοντέλο παλινδρόμησης:

$$y = x'\beta + \sigma_\varepsilon \varepsilon$$

Όπου:

$$\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$$

$$\varepsilon \sim N_8(0,1)$$

Και το (8×1) διάνυσμα $\chi \sim N_8(0, \Sigma_\chi)$για το οποίο $(\Sigma_\chi)_{ij} = \rho^{|i-j|} \forall\, i, j$

Οι τιμές που έχουν επιλεχθεί είναι $\sigma_\varepsilon = 3\ και\ 1$, n=50, 100 και 200 ενώ ρ=0.75, 0.5, 0.25. Σαν σταθερά h υπολογίσαμε τον εκτιμητή πρόβλεψης ο οποίος είναι ο εκτιμητής ελαχίστων τετραγώνων του πραγματικού υπομοντέλου:

$$y = \beta_1 \chi_1 + \beta_2 \chi_2 + \beta_5 \chi_5 + \varepsilon$$

Επειδή τα αποτελέσματα είναι ίδια και για τις τρεις συμμεταβλητές, θα παρουσιάσουμε το αποτέλεσμα για ρ=0.5 .Ο πίνακας που ακολουθεί (Πίνακας 1) δείχνει ότι όλοι οι διάμεσοι του RME πάνω από 1000 επαναλήψεις της μεθόδου BIC προσεγγίζει ταχέως τον εκτιμητή πρόβλεψης όσο το μέγεθος του δείγματος μεγαλώνει ή όσο το επίπεδο θορύβου μειώνεται. Αντίθετα, η τιμή της GCV παραμένει σχεδόν ίδια ανάμεσα σε διάφορα επίπεδα θορύβου και μεγέθη δείγματος. Έτσι, η μέθοδος BIC ξεπερνά τη μέθοδο GCVστο μοντέλο του μέτρου σφάλματος.

Table 1. *Example* 1. *Simulation results for the linear regression model*

σ_ε	n	Method	Under-fitted (%)	Correctly fitted (%)	Overfitted (%) 1	2	⩾ 3	No. of Zeros I	C	MRME (%)
3	50	$\hat{\lambda}_{GCV}$	6·4	16·9	23·0	31·6	22·1	0·064	3·279	64·17
		$\hat{\lambda}_{BIC}$	10·1	30·0	31·1	20·5	8·3	0·101	3·899	62·30
		Oracle	0	100	0	0	0	0	5	30·63
	100	$\hat{\lambda}_{GCV}$	0·2	24·0	20·4	29·8	25·6	0·02	3·369	57·72
		$\hat{\lambda}_{BIC}$	1·0	52·5	27·0	14·6	4·9	0·100	4·275	50·43
		Oracle	0	100	0	0	0	0	5	33·05
	200	$\hat{\lambda}_{GCV}$	0	25·4	35·8	25·4	13·4	0	3·300	55·18
		$\hat{\lambda}_{BIC}$	0	72·7	21·9	4·5	0·9	0	4·528	42·12
		Oracle	0	100	0	0	0	0	5	34·45
1	50	$\hat{\lambda}_{GCV}$	0	17·1	24·5	33·2	25·2	0	3·272	55·64
		$\hat{\lambda}_{BIC}$	0	45·6	23·9	21·1	9·4	0	4·042	40·97
		Oracle	0	100	0	0	0	0	5	30·62
	100	$\hat{\lambda}_{GCV}$	0	19·0	24·5	31·8	24·7	0	3·324	55·91
		$\hat{\lambda}_{BIC}$	0	54·9	23·6	16·9	4·6	0	4·277	40·53
		Oracle	0	100	0	0	0	0	5	33·05
	200	$\hat{\lambda}_{GCV}$	0	48·1	37·5	11·7	2·7	0	3·302	55·00
		$\hat{\lambda}_{BIC}$	0	81·8	16·4	1·3	0·5	0	4·405	38·36
		Oracle	0	100	0	0	0	0	5	34·42

I, the average number of the three truly nonzero coefficients incorrectly set to zero; C, the average number of the five true zero coefficients that were correctly set to zero; MRME, median of relative model error.

Πίνακας 1. Το RME για τις δύο μεθόδους για πάνω απο 1000 επαναλήψεις

Η στήλη «C» δείχνει τη μέση τιμή των πέντε πραγματικών μηδενικών συμμεταβλητών που σωστά τέθηκαν ίσες με μηδέν, ενώ η στήλη «I» δείχνει τη μέση τιμή των τριών πραγματικών μη μηδενικών συμμεταβλητών που λάθος είχαν τεθεί ίσες με μηδέν. Επιπλέον, ο πίνακας δείχνει την αναλογία των ελλειπών, των σωστά προσαρμοσμένων και των υπερπλήρων μοντέλων. Στην τελευταία περίπτωση, οι στήλες «1», «2», «≥3» είναι οι αναλογίες των μοντέλων που συμπεριλαμβάνουν 1, 2 και ≥3 ανεξάρτητες μεταβλητές αντίστοιχα. Παρατηρούμε ότι η μέθοδος BIC είναι καλύτερη στο να βρίσκει το πραγματικό μοντέλο από ότι η μέθοδος GCV. Επίσης, ανάμεσα στα "overfitted" μοντέλα, η BIC πειλαμβάνει μόνο μια ασύνδετη μεταβλητή, σε αντίθεση με την GCV που συνήθως περιλαμβάνει περισσότετρες από δύο. Παρατηρούμε δηλαδή ότι και οι δύο μέθοδοι ανταποκρίνονται στα θεωρητικά αποτελέσματα.

Παράδειγμα 2: Θεωρούμε το μερικώς γραμμικό μοντέλο:

$$y = a(u) + x'\beta + \sigma_\varepsilon \varepsilon$$

Όπου:

$$u \sim U_n(0,1)$$

$$a(u) = \exp\{2\sin(2\pi u)\}$$

Η προσομοίωση του προβλήματος αυτού είναι ακριβώς ίδια με την προσομοίωση του προηγούμενου παραδείγματος. Όπως έχουμε αναφέρει και παραπάνω, αντικαθιστούμε το β στην y_i^* με τον εκτιμητή και χρησιμοποιούμε την «plug-in»μέθοδο που προτάθηκε απο τον Ruppert et Al. (1995) [51], για την επιλογη του h. Ο πίνακας 2 παρουσιάζει τα αποτελέσματα της προσομοίωσης για ρ=0,5 και παρατηρούμε πάλι ότι η μέθοδος BIC υπερέχει της μεθόδου GCV ανιχνεύοντας το σωστό μοντέλο και μειώνοντας το λάθος, καθώς και την πολυπλοκότητα.

Table 2. *Example 2. Simulation results for the partially linear regression model*

σ_ϵ	n	Method	Under-fitted (%)	Correctly fitted (%)	Overfitted (%) 1	2	⩾ 3	No. of Zeros I	C	MRME (%)
3	50	$\hat{\lambda}_{GCV}$	10·9	15·9	24·6	25·8	22·8	0·112	3·263	66·78
		$\hat{\lambda}_{BIC}$	15·5	29·3	29·3	18·4	7·5	0·160	3·929	67·04
		Oracle	0	100	0	0	0	0	5	29·29
	100	$\hat{\lambda}_{GCV}$	0·8	23·1	22·6	29·7	23·8	0·08	3·368	58·15
		$\hat{\lambda}_{BIC}$	1·9	51·8	29·4	13·1	3·8	0·19	4·301	52·10
	100	Oracle	0	100	0	0	0	0	5	33·58
	200	$\hat{\lambda}_{GCV}$	0	22·9	21·5	30·5	25·1	0	3·352	54·47
		$\hat{\lambda}_{BIC}$	0	70·0	16·7	10·9	2·4	0	4·540	43·34
		Oracle	0	100	0	0	0	0	5	34·50
1	50	$\hat{\lambda}_{GCV}$	0	26·0	25·7	31·0	17·3	0	3·567	51·93
		$\hat{\lambda}_{BIC}$	0·1	60·3	20·6	13·9	5·1	0·01	4·356	38·31
		Oracle	0	100	0	0	0	0	5	29·30
	100	$\hat{\lambda}_{GCV}$	0	26·3	27·5	27·5	18·7	0	3·567	50·90
		$\hat{\lambda}_{BIC}$	0	67·9	18·9	9·9	3·3	0	4·509	39·10
		Oracle	0	100	0	0	0	0	5	33·42
	200	$\hat{\lambda}_{GCV}$	0	26·5	26·9	28·9	17·7	0	3·582	49·24
		$\hat{\lambda}_{BIC}$	0	75·7	15·7	7·2	1·4	0	4·656	39·01
		Oracle	0	100	0	0	0	0	5	34·77

I, the average number of the three truly nonzero coefficients incorrectly set to zero; C, the average number of the five true zero coefficients that were correctly set to zero; MRME, median of relative model error.

Πίνακας 2. Προσομοίωση για το Μερικώς Γραμμικό Μοντέλο

Παράδειγμα 3: Θεωρούμε τα δεδομένα σχετικά με την γυναικεία εργασία που συλλέχθηκαν από την Ανατολική Γερμανία το 1994. Το δείγμα αποτελείται από 607 παρατηρήσεις και έχει αναλυθεί από τους Fan&Alχρησιμοποιώντας προσθετικά μοντέλα. Χρησιμοποιούμε την μεταβλητή απόκρισης y για να συμβολίσουμε τον μισθό/ώρα. Η μεταβλητή uπου θα χρησιμοποιήσουμε στο μερικώς γραμμικό μοντέλο αναφέρεται στην ηλικία και αυτό γιατί η σχέση y και u δεν μπορεί να χαρακτηριστεί από μια απλή συναρτησιακή μορφή. Υπάρχουν εφτά επεξηγηματικές μεταβλητές :

$$\chi_1 = o\ \varepsilon\beta\delta o\mu\alpha\delta\iota\alpha\acute{\iota}os\ \alpha\rho\iota\theta\mu\acute{\omega}\nu\ \omega\rho\acute{\omega}\nu\ \varepsilon\rho\gamma\alpha\sigma\acute{\iota}\alpha s$$

$$\chi_2 = o\ \delta\varepsilon\acute{\iota}\kappa\tau\eta s\ \kappa\acute{\upsilon}\rho ous\ \tau\eta s\ \varepsilon\rho\gamma\alpha\sigma\acute{\iota}\alpha s$$

$$\chi_3 = \tau o\ \mu\eta\nu\iota\alpha\acute{\iota}o\ \kappa\alpha\theta\alpha\rho\acute{o}\ \kappa\acute{\varepsilon}\rho\delta os\ \tau ou\ \sigma\upsilon\zeta\acute{\upsilon}\gamma ou$$

$$\chi_4 = \begin{cases} 1, \varepsilon\acute{\alpha}\nu\ \tau\alpha\ \chi\rho\acute{o}\nu\iota\alpha\ \varepsilon\kappa\pi\alpha\acute{\iota}\delta\varepsilon\upsilon\sigma\eta s\ \varepsilon\acute{\iota}\nu\alpha\iota\ 13-16 \\ 0, \alpha\lambda\lambda\iota\acute{\omega}s \end{cases}$$

$$\chi_5 = \begin{cases} 1, \varepsilon\acute{\alpha}\nu\ \tau\alpha\ \chi\rho\acute{o}\nu\iota\alpha\ \varepsilon\kappa\pi\alpha\acute{\iota}\delta\varepsilon\upsilon\sigma\eta s\ \delta\varepsilon\nu\ \varepsilon\acute{\iota}\nu\alpha\iota\ \lambda\iota\gamma\acute{o}\tau\varepsilon\rho\alpha\ \alpha\pi\acute{o}\ 17 \\ 0, \alpha\lambda\lambda\iota\acute{\omega}s \end{cases}$$

$$\chi_6 = \begin{cases} 1, \varepsilon\acute{\alpha}\nu\ \delta\varepsilon\nu\ \pi\alpha\iota\delta\acute{\iota}\ \mu\iota\kappa\rho\acute{o}\tau\varepsilon\rho o\ \alpha\pi\acute{o}\ 16\ \varepsilon\tau\acute{\omega}\nu \\ 0, \alpha\lambda\lambda\iota\acute{\omega}s \end{cases}$$

$$\chi_7 = \tau o\ \pi o\sigma o\sigma\tau\acute{o}\ \alpha\nu\varepsilon\rho\gamma\acute{\iota}\alpha s\ \sigma\tau\eta\nu\ \pi\varepsilon\rho\iota o\chi\acute{\eta}$$

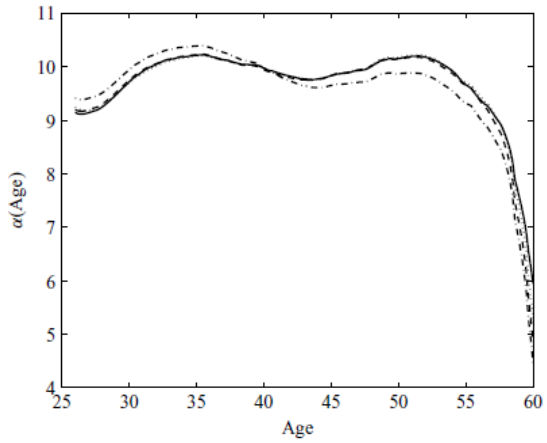
Μετά από μια αρχική ανάλυση, θεωρούμε το ακόλουθο μερικώς γραμμικό μοντέλο με 7 γραμμικές κύριες επιδράσεις και μερικά πρώτης τάξης αποτελέσματα αλληλεπίδρασης ανάμεσα στα χ_1, χ_2, χ_3:

$$y = a(u) + \sum_{j=1}^{7} \beta_j x_j + \sum_{k=1}^{3} \sum_{l=k}^{3} \beta_{kl} x_k x_l + \varepsilon$$

Ακολουθώντας τη μέθοδο των Fan&Li [1] (2004), υπολογίζουμε τον εκτιμητή για το β και εφαρμόζουμε μια «plug-in» μέθοδο, η οποία χρησιμοποιείται για την κατάλληλη επιλογή της παραμέτρου h (*bandwidth*) στην μη-παραμετρική παλινδρόμηση. Με αυτό τον τρόπο, επιλέγουμε το 4,6249 ως h για το $\hat{a}(.)$. Με αυτό το h επιλέγουμε τη ρυθμιστική παράμετρο ελαχιστοποιώντας τις GCV και BIC και προκύπτει αντιστοίχως:

$\hat{\lambda}_{GCV} = 0{,}0896$ και $\hat{\lambda}_{BIC} = 0{,}2655$

Έτσι, επιτυγχάνουμε τον μη ποινικοποιημένο εκτιμητή ελαχίστων τετραγώνων καθώς και τον εκτιμητή SCAD που βασίζεται στην GCVκαι BIC.



Εικόνα 4.5.1:Γραφική παράσταση του $\hat{a}(u)$.

Variable	Profile LSE	SCAD $\hat{\lambda}_{GCV}$	SCAD $\hat{\lambda}_{BIC}$	Best-subset BIC	Best-subset BIC($n = 602$)
x_1	1·244(0·637)	1·281(0·636)	1·872(0·562)	1·343(0·496)	0
x_1^2	−1·451(0·563)	−1·446(0·559)	−1·841(0·517)	−2·192(0·497)	−0·853(0·119)
x_2	1·520(0·721)	1·602(0·704)	1·357(0·681)	0	0
x_2^2	1·162(0·617)	1·281(0·599)	1·341(0·601)	1·433(0·136)	1·410(0·137)
x_3	−1·229(0·692)	−1·063(0·549)	0	0	0
x_3^2	−0·011(0·276)	0	0	0	0
x_1x_2	−1·781(0·702)	−1·885(0·684)	−1·493(0·653)	0	0
x_1x_3	0·922(0·559)	0·995(0·549)	0	0	0
x_2x_3	0·313(0·485)	0	0	0	0
x_4	0·609(0·130)	0·593(0·130)	0·249(0·055)	0·605(0·129)	0·590(0·129)
x_5	1·194(0·140)	1·183(0·140)	1·030(0·131)	1·168(0·138)	1·172(0·139)
x_6	−0·290(0·189)	−0·028(0·019)	0	0	0
x_7	0·118(0·117)	0·005(0·006)	0	0	0

LSE, least squares estimate; SCAD, smoothly clipped absolute deviation .

Εικόνα 4.5.2

Επίσης θεωρούμε το μοντέλο που επιλέγεται από τη μέθοδο μη ποινικοποιημένων ελαχίστων τετραγώνων (εξίσωση (4.4.5)) χωρίς τον όρο ποινής και με το κριτήριο επιλογής καλύτερου υποσυνόλου:

$$BIC = log\hat{\sigma}_n^2 + d^* \log(n)\,/n$$

Όπου:

$$\hat{\sigma}_n^2 = SSE_{hs}/n$$

Και SSE_{hs} το άθροισμα τετραγώνων λάθους ιπό την Y_hμε $X_{hs} = (I - S_h)X_s$, d^*Η διάσταση του X_s.

Οι πρώτες τέσσερις στήλες του Πίνακα δείχνουν αθαρά ότι το μη ποινικοποιημένο μοντέλο ελαχίστων τετραγώνων που βασίζεται στη GCVτείνει να συμπεριλάβει μεταβλητές με μικρή επορροή. Αντίθετα, όλες οι μεταβλητές που επιλέγονται από την SCAD μέσω της BIC είναι σημαντικές με επίπεδο σημαντικότητας 0,5.

Το παραπάνω διάγραμμα δείχνει ότι οι τέσσερις εκτιμητές για το α(.) είναι περίπου ίσοι αλλά ο εκτιμητής από την μη ποινικοποιημένη μέθοδο ελαχίστων τετραγώνων είναι είναι διαφορετικός από τους υπόλοιπους.

Η τέταρτη στήλη του πίνακα δείχνει ότι το κριτήριο επιλογής καύτερου υποσυνόλου με την BICδείνει καλύτερο και πιο απλό μοντέλο σε σχέση με τη μέθοδο SCADμε την βοήθεια του BIC. Παρ'όλ'αυτά, ο Breiman (1996) έδειξε ότι το κριτήριο επιλογής

καλύτερου υποσυνόλου υποφέρει από έλλειψη ευστάθειας. Για να το δείξουμε αυτό, αφαιρούμε τις πέντα τελευταίες παρατηρήσεις και έτσι έχουμε ένα δείγμα με n=602. Η τελευταία στήλη του πίνακα 3 δείχνει ότι το κριτήριο BICμε τη μέθοδο επιλογής καλύτερου υποσυνόλου φτιάχνει διαφορετικό μοντέλο από αυτό που φτιάχνει για n=607. Αντίθετα, το μοντέλο που φτιάχνει η SCADμε την BICφαίνεται ότι δεν αλλάζει. Το τελευταίο είναιτης μορφής:

$$\hat{y} = \hat{a}(u) + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{22} x_2^2 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$

Έτσι, το ωρομίσθιο εξαρτάται πρωτίστως από τις ώρες εργασίας, από το κύρος της δουλειάς καθώς και από τα χρόνια εκπαίδευσης. Αντίθετα, οι απολαβές του συζύγου, το επίπεδο ανεργίας και το εάν μια γυναίκα έχει ή όχι μικρό παιδί δεν φαίνεται να παίζουν ρόλο.

Από το διάγραμμα, βγάζουμε το συμπέρασμα ότι ο μισθός παραμένει σχεδόν σταθερός μέχρι την ηλικία των 50, μειώνεται όμως δραματικά στη συνέχεια.

Κεφάλαιο 5

Κριτήριο Γενικευμένης Πληροφορίας

5.1 Εισαγωγή

Η χρήση των συναρτήσεων ποινικοποιημένης πιθανοφάνειας, που ταυτόχρονα επιλέγουν μεταβλητές και εκτιμούν άγνωστες παραμέτρους, έχει λάβει αξιόλογη προσοχή τα τελευταία χρόνια. Για να αποφύγουμε την αστάθεια της κλασικής διαδικασίας επιλογής καλύτερου υποσυνόλου, ο Tibshirani (1996) πρότεινε τη μέθοδο LASSO (*Least Absolute Shrinkage & Selection Operator*). Στο ίδιο πνεύμα με τη LASSO, οι Fan&Li (2001) (όπως έχουμε αναφέρει αναλύτικά στο κεφαλαιο 2), πρότειναν μια μοναδική προσέγγιση μέσω της μη κοίλης ποινικοποιημένης συνάρτησης πιθανοφάνειας και απέδειξαν τη χρησιμότητά της τόσο στα γραμμικά όσο και στα γενικευμένα γραμμικά μοντέλα. Στη συνέχεια, οι Fan&Li (2002) εργάστηκαν για την προσέγγιση της ποινικοποιημένης πιθανοφάνειας στο μοντέλο του Cox ενώ τα χρησιμοποίησαν (2004) και σε ημι-παραμετρικά μοντέλα παλινδρόμησης. Επιπλέον, οι Fan & Peng (2004)[18], ερεύνησαν τις θεωρητικές ιδιότητες της μη κοίλης ποινικοποιημένης πιθανοφάνειας, όταν ο αριθμός των παραμέτρων του μοντέλου τείνει στο άπειρο, όσο το μέγεθος του δείγματος μεγαλώνει. Πρόσφατα, πολλοί ερευνητές εφάρμοσαν τις μεθόδους ποινής για τη μελέτη της επιλογής μεταβλητών (Park & Hastie 2007, Wang & Leng 2007, Yuan & Lin 2007, Zhang & Lu 2007, Li & Liang 2008).

Καθώς εργαζόμαστε με τη μη κοιλη ποινικοποιημένη πιθανοφάνεια στην ανάλυση παλινδρόμησης, ερχόμαστε αντιμέτωποι με δύο προβλήματα. Η πρώτη δυσκολία είναι να υπολογίσουμε τον εκτιμητή της μη κοίλης ποινικοποιημένης πιθανοφάνειας. Αυτό είναι ένα κομμάτι που έχει μελετηθεί πολύ στη σύγχρονη βιβλιογραφία. Ενδεικτικά αναφέρουμε τους Fan & Li (2001) οι οποίοι πρότειναν τον αλγόριθμο τοπικής τετραγωνικής προσέγγισης (*Local Quadratic Approximation- LQA*) , ο οποίος αναλύθηκε στη συνέχεια από τους Hunter & Li (2005) [31]. Στη συνέχεια, οι Efron et Al. (2004) εισήγαγαν τον αλγόριθμο LARS που μπορεί να χρησιμοποιηθεί για την LASSO. Με τη βοήθεια του αλγορίθμου της τοπικής τετραγωνικής προσέγγισης

(*Local Linear Approximation-LLA*) (Zan & Li 2008), ο αλγόριθμος LARS υιοθετήθηκε για να λύσει προβλήματα βελτιστοποίησης σε μη κυρτές συναρτήσεις ποινικοποιημένης πιθανοφάνειας. Όμως, η υπολογιστική διαδικασία εξαρτάται από μια παράμετρο προσαρμογής. Έτσι, η επιλογή της παραμέτρου αυτής είναι η δεύτερη δυσκολία που έχουμε να αντιμετωπίσουμε. Στο κεφάλαιο αυτό, θα εξηγήσουμε με ποιους τρόπους μπορούμε να επιλέξουμε με βέλτιστο τρόπο αυτή την παράμετρο.

Στη βιβλιογραφία, τα κριτήρια επιλογής συνήθως ταξινομούνται σε δύο κατηγορίες: στα συνεπή (π.χ το Μπευζιανό κριτήριο πληροφορίας- BIC, Schwartz 1978) και στα αποτελεσματικά (π.χ το Akaike κριτήριο πληροφορίας-AIC , Akaike 1974 ή η Γενικευμένη Διασταυρωμένη Επικύρωση GCV, Craven & Wahba 1979).

Ένα συνεπές κριτήριο ανιχνεύει το πραγματικό μοντέλο με πιθανότητα που τείνει στη μονάδα σε μεγάλα δείγματα, όταν ένα σύνολο από υποψήφια μοντέλα περιλαμβάνει το πραγματικό μοντέλο. Ένα αποτελεσματικό κριτήριο επιλέγει το μοντέλο έτσι ώστε το μέσο τετραγωνικό σφάλμα να είναι ασυμπτωτικά ίσο με το αντίστοιχο ελάχιστο, όταν το πραγματικό μοντέλο έχει υπολογιστεί προσεγγιστικά από κάποια υποψήφια. Στα γραμμικά και στα γενικευμένα γραμμικά μοντέλα (*GLIM*) οι Fan & Li πρότειναν τη μέθοδο γενικευμένης διασταυρωμένης επικύρωσης (*Generalized Cross Validation*) για την επιλογή της ρυθμιστικής παραμέτρου. Όμως, οι Wang, Li & Tsai (2007) [74] διαπίστωσαν ότι το μοντέλο που προκύπτει από τη μέθοδο GCV περιλαμβάνει και επιπλέον μεταβλητες (*overfitted model*) ενώ το BIC ανιχνεύει το πεπερασμένης διάστασης πραγματικό μοντέλο, είτε είναι γραμμικό είτε είναι μερικώς γραμμικό, με συνέπεια. Οι Wang, Li&Tsai (2007) διαπίστωσαν επίσης ότι η GCV καταλήγει στο ίδιο αποτέλεσμα με την AIC.

Στη συνέχεια, θα υιοθετήσουμε την μέθοδο του Nishii (1984) για το κριτήριο γενικευμένης πληροφορίας (*Generalized Information Criterion*) για να επιλέξουμε την ρυθμιστική παράμετρο σε μη κοίλες συναρτήσεις ποινικοποιημένης πιθανοφάνειας.Το κριτήριο αυτό όχι μόνο περιλαμβάνει τα AIC και BIC σαν ειδικές περιπτώσεις, αλλά γεφυρώνει τη σχέση μεταξύ του κλασικού κριτηρίου καλύτερου υποσυνόλου και της μεθοδολογίας της μη κοίλης ποινικοποιημένης πιθανοφάνειας. Αυτή η σύνδεση, παρέχει περισσότερη ευελιξία ώστε να επιλέξει ο καθένας το κριτήριο που επιθυμεί. Όταν το πραγματικό μοντέλο είναι ανάμεσα σε ένα σύνολο υποψήφιων μοντέλων με δομή γενικευμένων γραμμικών μοντέλων, θα δείξουμε ότι η ρυθμιστική παράμετρος

που βασίζεται στη μέθοδο BIC (*BIC-type selector*) είναι ικανή να επιλέξει το πραγματικό μοντέλο με ακρίβεια. Αντίθετα, η ρυθμιστική παράμετρος που προκύπτει από την μέθοδο AIC (*AIC-type selector*) βρίσκει μοντέλα το οποία περιέχουν και περιττές μεταβλητές. Από την άλλη μεριά, εάν το πραγματικό μοντέλο έχει προσεγγιστεί από ένα σύνολο υποψήφιων μοντέλων με γραμμική μορφή, θα δούμε ότι η ρυθμιστική παράμετρος που προκύπτει από το AIC έχει ασυμπτωτικά απόδοση ζημίας (*asymptotically loss efficient*) ενώ το λ που βασίζεται στο BIC δεν έχει αυτή την ιδιότητα. Αυτά τα συμπεράσματα είναι σύμφωνα με τις ιδιότητες του BIC και AIC που χρησιμοποιούνται στην επιλογή καλύτερου υποσυνόλου.

5.2 Αναγκαίες Συνθήκες Ποινής

Θεωρούμε τα δεδομένα $(\chi_1, y_1), \dots (\chi_n, y_n)$τα οποία είναι ανεξάρτητα και ισόνομα , με y_i να είναι η απόκριση του i-οστού αντικειμένου και χ_iη d-διάστατη μεταβλητή πρόβλεψης. Έστω $l(\tilde{\beta})$ να είναι η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας (ή συνάρτηση ζημίας) του d-διάστατου διανύσματος $\tilde{\beta} = (\beta_1, \dots, \beta_d)'$. Τότε, υιοθετώντας την προσέγγιση των Fan & Li (2001) ορίζουμε την ποινικοποιημένη πιθανοφάνεια:

$$Q(\tilde{\beta}) = l(\tilde{\beta}) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (5.2.1)$$

Όπου $p_\lambda(.)$είναι μια συνάρτηση ποινής με ρυθμιστική παράμετρο λ. Διάφορες συναρτήσεις ποινής έχουν προταθεί στη βιβλιογραφία. Για παράδειγμα, η ποινή $L_q(0 < q < 2)$είναι: $p_\lambda(|\beta|) = q^{-1}\lambda|\beta|^q$ οδηγεί στην παλινδρόμηση κορυφογραμμής (Frank & Friedman 1993).Πιο συγκεκριμένα, η L_1 ποινή οδηγεί στον LASSO εκτιμητή (Tibshirani 1996).Οι Fan & Li (2001) πρότειναν την μη κοίλη ποινικοποιημένη μέθοδο πιθανοφάνειας και υποστήριξαν τη χρήση της ποινής SCAD της οποίας η πρώτη παράγωγος είναι:

$$p'_\lambda(|\beta_j|) = \lambda\{ I(|\beta_j| \le \lambda) + \frac{(\alpha\lambda - |\beta_j|)_+}{(\alpha-1)\lambda} I(|\beta_j| > \lambda)\} \text{ με } \alpha=3{,}7 \text{ και } p_\lambda(0) = 0.$$

Με κατάλληλα επιλεγμένη ρυθμιστική παράμετρο από το μηδέν έως ένα άνω όριο λ_{max}, ο προκύπτων ποινικοποιημένος εκτιμητής είναι σποραδικός οπότε και κατάλληλος για την επιλογή μεταβλητών. Για να εκτιμήσουμε τις ασυμπτωτικές

ιδιότητες των εκτιμητών της ρυθμιστικής παραμέτρου, παρουσιάζουμε τις απαραίτητες συνθήκες ποινής:

(C_1) Υποθέτουμε ότι το λ_{max} εξαρτάται από το n και ικανοποιεί $\lambda_{max} \to 0$ όσο $n \to \infty$.

(C_2) Υπάρχει μια σταθερά m τέτοια ώστε η ποινή $p_\lambda(\zeta)$ να ικανοποιεί:

$$p'(\zeta) = 0 \ \ για\ \zeta > m\lambda.$$

(C_3) Εάν $\lambda_n \to 0$για $n \to \infty$τότε η συνάρτηση ποινής ικανοποιεί:

$$lim_{n\to\infty} inf\, lim_{\zeta\to 0^+} inf \sqrt{n} p'_{\lambda_n}(\zeta) \to \infty$$

Η συνθήκη (C_1) δείχνει ότι μια μικρότερη ρυθμιστική παράμετρος χρειάζεται εάν το μέγεθος του δείγματος είναι μεγάλο. Η συνθήκη (C_2) μας διαβεβαιώνει ότι ο προκύπτων εκτιμητής ποινικοποιημένης πιθανοφάνειας είναι ασυμπτωτικά αμερόληπτος. Τόσο η SCAD όσο και η μη κοίλη ποινή του Zhang (*Minimax Concave Penalties-MCP*) ικανοποιούν αυτή τη συνθήκη. Η συνθήκη (C_3) προσαρμόστηκε από την εξίσωση (3.5) των Fan & Li (2001) που χρησιμοποιείται για τη μελέτη της ιδιότητας πρόβλεψης.

Σημείωση 5.2.1: Τόσο η ποινή SCAD όσο και η ποινή $L_q(0 < q < 2)$είναι μοναδικές στην προέλευση. Έτσι, γίνεται δελεαστικό να μεγιστοποιήσουμε την αντίστοιχη συνάρτηση ποινικοποιημένης πιθανοφάνειας. Οι Fan&Li (2001) πρότειναν τον LQA (*Local Quadratic Approximation*) αλγόριθμο για να βρουν τη λύση της μη κυρτής ποινικοποιημένης πιθανοφάνειας. Στον LQA η $p_\lambda(|\beta|)$ προσεγγίζεται τοπικά από μια τετραγωνική συνάρτηση $q_\lambda(|\beta|)$ της οποίας η πρώτη παράγωγος δίνεται από:

$$q'_\lambda(|\beta_j|) = \{\frac{p'_\lambda(|\beta_j|)}{|\beta_j|}\}\beta_j$$

Η παραπάνω εξίσωση υπολογίστηκε στο (K+1)-βήμα της επανάληψης της μεθόδου Newton-Raphson εάν το $\beta_j^{(\kappa)}$ δεν είναι πολύ κοντά στο μηδέν. Διαφορετικά, ο εκτιμητής παραμέτρου β_j τίθεται ίσος με μηδέν.

5.3 Κριτήριο Γενικευμένης Πληροφορίας

Αρχικά, θα δώσουμε έναν ορισμό για το υποψήφιο μοντέλο, το οποίο περιλαμβάνεται σε ένα σύνολο μοντέλων.

Ορισμός 5.3.1 (Υποψήφιο Μοντέλο): Ορίζουμε το α, ένα υποσύνολο του $\bar{\alpha} = \{1, \dots, d\}$, ως υποψήφιο μοντέλο, με την έννοια ότι οι αντίστοιχοι εκτιμητές που προέρχονται από το α συμπεριλαμβάνονται στο μοντέλο. Επακόλουθα, το $\bar{\alpha}$ είναι το πλήρες μοντέλο. Επιπλέον, ορίζουμε το μέγεθος του μοντέλου α (δηλαδή τον αριθμό των μη μηδενικών παραμέτρων του α) καθώς και του συντελεστές που σχετίζονται με τους εκτιμητές στο μοντέλο α και συμβολίζουμε με d_ακαι β_α αντίστοιχα. Επιπλέον συμβολίζουμε το σύνολο των υποψήφιων μοντέλων με A. Για έναν ποινικοποιημένο εκτιμητή $\hat{\beta}_\lambda$ που ελαχιστοποιεί την αντικειμενική συνάρτηση (5.2.1) ορίζουμε το μοντέλο που συνδέεται με το $\hat{\beta}_\lambda$ και το συμβολίζουμε με α_λ.

Στο γραμμικό μοντέλο παλιδρόμησης,

$y_i = \tilde{x}_i^{'}\beta_\alpha + \varepsilon_i$και $\varepsilon_i \sim N(0, \sigma^2) για\ i = 1, \dots, n$

ο Nishii (1984) πρότεινε για την επιλογή μεταβλητών το Κριτήριο Γενικευμένης Πληροφορίας (*Generalized Information Criterion*) το οποίο για λόγους συντομίας μπορούμε να συμβολίσουμε και ως GIC. Ορίζεται:

$$GIC_{k_n}(a) = log\hat{\sigma}_\alpha^2 + \frac{1}{n}k_n d_a$$

Όπου:

β_αείναι η παράμετρος του υποψήφιου μοντέλου a

$\hat{\sigma}_\alpha^2$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του σ^2

k_n είναι ένας θετικός αριθμός που ελέγχει τις ιδιότητες της επιλογής μεταβλητών.

Να σημειώσουμε σ'αυτό το σημείο ότι το GIC που ορίστηκε από τον Nishii είναι διαφορετικό από αυτό που προτάθηκε από τους Konishi & Kitagawa (1996) [34].

Όταν $k_n = 2$ τότε το GIC μετατρέπεται σε AIC ενώ όταν $k_n = \log(n)$οδηγεί στο BIC. Επειδή το GIC περιλαμβάνει ένα μεγάλο αριθμό από κριτήρια επιλογής,

προτείνουμε τον ακόλουθο εκτιμητή της ρυθμιστικής παραμέτρου λ που βασίζεται στο GIC:

$$GIC_{k_n}(\lambda) = \frac{1}{n}\{G(\tilde{y}, \hat{\beta}_\lambda) + k_n df_\lambda\} \qquad (5.3.1)$$

όπου:

$G(\tilde{y}, \hat{\beta}_\lambda)$μετρά την προσαρμογή του μοντέλου α_λ

$\tilde{y} = (y_1, \dots, y_n)'$

$\hat{\beta}_\lambda$ ο εκτιμητής της ποινικοποιημένης παραμέτρου που προκύπτει μεγιστοποιώντας την εξίσωση (5.2.1) ως προς β.

df_λ οι βαθμοί ελευθερίας του μοντέλου α_λ.

Για δεδομένο k_n επιλέγουμε το λ που ελαχιστοποιεί το $GIC_{k_n}(\lambda)$. Όσο μεγαλύτερο είναι το k_n τόσο μεγαλύτερη είναι και η ποινή για τα μοντέλα που περιλαμβάνουν πολλές μεταβλητές.Όμως, όσο το k_n αυξάνεται τόσο μειώνεται το μέγεθος του μοντέλου που προκύπτει όταν το προσαρμόζουμε σε δεδομένα.

Σημείωση 5.3.1: Για ένα δεδομένο μοντέλο α (συμπεριλαμβανομένου και του ποινικοποιημένου μοντέλου α_λκαι του πλήρους μοντέλου $\bar{\alpha}$) μπορούμε να βρούμε τον εκτιμητή $\hat{\beta}_\alpha^*$ της μη ποινικοποιημένης παραμέτρου, μεγιστοποιώντας την λογαριθμοποιημένη συνάρτηση πιθανοφάνειας $l(\beta)$ στην (5.2.1). Τότε, η εξίσωση (5.3.1) γίνεται:

$$GIC_{k_n}^*(a) = \frac{1}{n}\{G(\tilde{y}, \hat{\beta}_\alpha^*) + k_n d_a\} \qquad (5.3.2)$$

η οποία μπορεί να χρησιμοποιηθεί και στην κλασική επιλογή μεταβλητών. Επιπρόσθετα, η $GIC_{k_n}^*(a)$ μετατρέπεται στην $GIC_{k_n}(a)$ εάν αντικαταστήσουμε την $G(\tilde{y}, \hat{\beta}_\alpha^*)$ της (5.3.2) με την δυο φορές λογαριθμοποιημένη συνάρτηση πιθανοφάνειας του κανονικού μοντέλου παλινδρόμησης.

Στη συνέχεια, θα εξετάσουμε τους βαθμούς ελευθερίας του δεύτερου όρου της GIC. Στην επιλογή της ρυθμιστικής παραμέτρου, οι Fan&Li (2001,2002) πρότειναν οι βαθμοί ελευθερίας να είναι το ίχνος του γραμμικού πίνακα προβολής, δηλαδή:

$$df_\lambda \triangleq tr\{\left(\nabla_\lambda^{\otimes 2} Q^*\left(\tilde{\hat{\beta}}_\lambda\right)\right)^{-1} \nabla_\lambda^{\otimes 2} l\left(\tilde{\hat{\beta}}_\lambda\right)\} \qquad (5.3.3)$$

Όπου:

$$Q^*(\tilde{\beta}) = l(\tilde{\beta}) - n\sum_{j=1}^{d} q_\lambda\left(|\beta_j|\right)$$

$$[\nabla_\lambda^{\otimes 2} Q^*(\tilde{\beta})]_{jj'} = \frac{\partial^2}{\partial\beta_j\beta_{j'}} Q^*(\beta)$$

$$[\nabla_\lambda^{\otimes 2} l(\tilde{\beta})]_{jj'} = \frac{\partial^2}{\partial\beta_j\beta_{j'}} l(\tilde{\beta}) \ \ για\ j, j'\ τέτοια\ ώστε\ \hat{\beta}_j \neq 0\ ,\ \ \hat{\beta}_{j'} \neq 0.$$

Ακολουθεί μια πρόταση που αφορά στην ασυμπτωτική συμπεριφορά του df_λ στα μεγάλα δείγματα.

Πρόταση 5.3.2: Υποθέτουμε ότι ο εκτιμητής ποινικοποιημένης πιθανοφάνειας $\hat{\beta}_\lambda$ είναι σποραδικός (με πιθανότητα να τείνει στη μονάδα $\hat{\beta}_{\lambda j} = 0$ εάν η πραγματική τιμή του β_j είναι μηδέν) και συνεπής, όπου το $\hat{\beta}_{\lambda j}$ είναι το j-οστό στοιχείο του $\hat{\beta}_\lambda$. Υπό τις συνθήκες (C_1) και (C_2) που παρουσιάστηκαν παραπάνω ισχύει:

$$P\{df_\lambda(\lambda) = d_{\alpha_\lambda}\} \to 1$$

Όπου με d_{α_λ} συμβολίζουμε το μέγεθος του μοντέλου α_λ.

Απόδειξη: Μετά από αλγεβρικές απλοποιήσεις,

$$df_\lambda(\lambda) = tr\{\left(\nabla_\lambda^{\otimes 2} l_{\alpha_\lambda}\left(\tilde{\hat{\beta}}_\lambda\right) + n\Sigma_\lambda\right)^{-1} \nabla_\lambda^{\otimes 2} l\left(\tilde{\hat{\beta}}_\lambda\right)\}$$

Όπου: $\Sigma_\lambda = diag_{\hat{\beta}_{\lambda j}=0}\{\frac{p_\lambda'(|\hat{\beta}_{\lambda j}|)}{|\hat{\beta}_{\lambda j}|}\}$

Επειδή το $\hat{\beta}_\lambda$ είναι συνεπές και σποραδικό, το $\hat{\beta}_{\lambda j}$ συγκλίνει στο β_j με πιθανότητα να τείνει στη μονάδα για κάθε jτέτοιο ώστε $\hat{\beta}_{\lambda j} > 0$. Έτσι, εκείνα τα $\hat{\beta}_{\lambda j}$ φράσσονται από το μηδέν. Αυτό το αποτέλεσμα,σε συνδυασμό με τις συνθήκες (C_1) και (C_2) υπονοεί ότι $\Sigma_\lambda = \tilde{0}$ με πιθανότητα να τείνει στη μονάδα. Επακόλουθα, χρησιμοποιώντας το γεγονός ότι $n^{-1}\nabla_\lambda^{\otimes 2} l(\hat{\beta}_\lambda) = O_p(1)$ ολοκληρώνουμε την απόδειξη.

Από την παραπάνω πρόταση προκύπτει ότι η διαφορά ανάμεσα στο $df_\lambda(\lambda)$ και στο μέγεθος του μοντέλου d_{α_λ} είναι μικρή. Επειδή το d_{α_λ} είναι απλό στον υπολογισμό, το χρησιμοποιούμε ως τους βαθμούς ελευθερίας στην εξίσωση (5.3.1).

Στα γραμμικά μοντέλα παλινδρόμησης οι Efron & Al (2004) και οι Zou, Hastie & Tibshirani (2007) επίσης πρότειναν τη χρήση του d_{α_λ} σαν εκτιμητή για τους βαθμούς ελευθερίας της μεθόδου LASSO.

5.4 Η Συνέπεια του Κριτηρίου

Υποθέτουμε ότι το σύνολο των υποψήφιων μοντέλων περιλαμβάνει το πραγματικό μοντέλο και ότι ο αριθμός των παραμέτρων του πλήρους μοντέλου είναι πεπερασμένος. Υπό αυτές τις συνθήκες, είμαστε σε θέση να μελετήσουμε την ασυμπτωτική συνέπεια του GIC εισάγοντας τον ακόλουθο ορισμό και συνθήκες.

Ορισμός 5.4.1 **Υποπροσαρμοσμενα και υπερπροσαρμοσμενα μοντελα** *(Underfitted & Overfitted Models*): Υποθέτουμε ότι υπάρχει ένα μοναδικό πραγματικό μοντέλο α_0 στο A του οποίου οι αντίστοιχες συμμεταβλητές είναι μηδενικές. Έτσι, κάθε υποψήφιο $\alpha \not\supset \alpha_0$ μοντέλο αναφέρεται σαν υποπροσαρμοσμένο (*underfitted*) ενώ κάθε $\alpha \supset \alpha_0$ εκτός από το α_0 αναφέρεται ως υπερπροσαρμοσμένο (*overfitted*).

Βασιζόμενοι στον παραπάνω ορισμό, διαχωρίζουμε το διάστημα της ρυθμιστικής παραμέτρου (*tuning parameter*) σε τρία υποδιαστήματα, ένα για κάθε ένα μοντέλο (*υποπροσαρμοσμένο, πραγματικό και υπερπροσαρμοσμένο*). Συμβολίζουμε αντίστοιχα,

$$\Omega_- = \{\lambda: \alpha_\lambda \not\supset \alpha_0\}$$

$$\Omega_0 = \{\lambda: \alpha_\lambda = \alpha_0\}$$

$$\Omega_+ = \{\lambda: \alpha_\lambda \supset \alpha_0 \ \ και\ \alpha_\lambda \neq \alpha_0\}$$

Αυτή η διαμέριση μας επιτρέπει να εκτιμήσουμε την απόδοση των εκτιμητών της ρυθμιστικής παραμέτρου. Για να μελετήσουμε τις ασυμπτωτικές ιδιότητες των εκτιμητών της ρυθμιστικής παραμέτρου, εισάγουμε τις παρακάτω συνθήκες:

(C$_4$) Για κάθε υποψήφιο μοντέλο $\alpha \in A$, $\exists\, c_a > 0$τέτοιο ώστε $\frac{1}{n}\{G(\tilde{y}, \hat{\beta}_\alpha^*)\} \xrightarrow{P} c_a$. Επιπρόσθετα, για κάθε ελλειπές μοντέλο $\alpha \supset \alpha_0$, $c_a > c_{a0}$ όπου c_{a0} είναι το όριο του $\{G(\tilde{y}, \tilde{\beta}_{\alpha 0})\}$ και $\tilde{\beta}_{\alpha 0}$ είναι το διάνυσμα παραμέτρου του πραγματικού μοντέλου α_0.

Η παραπάνω συνθήκη μας διαβεβαιώνει ότι το υποπροσαρμοσμένο μοντέλο δίνει ένα μεγαλύτερο μέτρο προσαρμογής μοντέλου σε σχέση με το πραγματικό μοντέλο. Στην πορεία, θα ερευνήσουμε την ασυμπτωτική συμπεριφορά του GIC για τα γενικευμένα γραμμικά μοντέλα τα οποία χρησιμοποιούνται σε πολλές εφαρμογές.

Θεωρούμε τα γενικευμένα γραμμικά μοντέλα (GLIM)- Mc Cullagh & Nelder (1989)- των οποίων η συνάρτηση πυκνότητας των y_i δεδομένων των x_i είναι:

$$f_i(y_i; \theta_i, \varphi) = \exp\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\varphi)} + c(y_i, \varphi)\} \qquad (5.4.1)$$

όπου: $a(.), b(.)$ $και$ $c(.)$ είναι κατάλληλα επιλεγμένες συναρτήσεις

θ_i είναι η κανονικοποιημένη παράμετρος

$$E(y_i|x_i) = \mu_i = b'(\theta_i)$$

$g(\mu_i) = \theta_i$ με g μια συνάρτηση σύνδεσης (*link function*)

φ μία παράμετρος (*scale parameter*)

Σ'αυτό το σημείο, υποθέτουμε ότι το φ είναι γνωστό (όπως στο μοντέλο λογιστικής παλινδρόμησης ή στο λογαριθμοποιημένο μοντέλο Poisson) ή ότι το εκτιμούμε προσαρμόζοντας στα δεδομένα το πλήρες μοντέλο. Στη συνέχεια, εφαρμόζουμε κανονικά τις μεθόδους παλινδρόμησης. Βασιζόμενοι στην εξίσωση (5.4.1) η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας της εξίσωσης (5.2.1) είναι:

$$l(\tilde{\beta}) = l(\tilde{\mu}; \tilde{y}) = l(\tilde{\theta}) = \sum_{i=1}^{n} log f_i(y_i; \theta_i, \varphi) =$$

$$= \sum_{i=1}^{n}[\{y_i\widetilde{x_i'}\beta - b(\widetilde{x_i'}\beta)\}/a(\varphi) + c(y_i, \varphi)] \qquad (5.4.2)$$

όπου:

$$\mu = (\mu_1, \dots, \mu_n)'$$

$$\tilde{y} = (y_1, \dots, y_n)'$$

$$\tilde{\theta} = (\theta_1, \dots, \theta_n)'$$

Τότε, η απόκλιση (*deviance*) του ποινικοποιημένου εκτιμητή $\widehat{\beta_\lambda}$ είναι:

$$D(\tilde{y}; \widehat{\mu_\lambda}) = 2\{l(\tilde{y}; \tilde{y}) - l(\widehat{\mu_\lambda}; \tilde{y})\}$$

Όπου:

$$\widehat{\mu_\lambda} = (g^{-1}\left(\widetilde{x_1'}\widehat{\beta_\lambda}\right), \dots, g^{-1}(\widetilde{x_n'}\widehat{\beta_\lambda}))'$$

Για το μοντέλο α_λ χρησιμοποιούμε την απόκλιση $D(\tilde{y}; \widehat{\mu_\lambda})$ σαν το μέτρο καλής προσαρμογής (*goodness-of-fit)* και έτσι το GIC της εξίσωσης (5.3.1) για τα γενικευμένα γραμμικά μοντέλα (GLIM) είναι:

$$GIC_{k_n}(\lambda) = \frac{1}{n} D(\tilde{y}; \widehat{\mu_\lambda}) + \frac{1}{n} k_n df_\lambda \quad (5.4.3)$$

Επιπλέον, όταν προσαρμόζουμε τα δεδομένα με την εκτίμηση της μη ποινικοποιημένης πιθανοφάνειας για το μοντέλο α , το GIC γίνεται:

$$GIC^*_{k_n}(a) = \frac{1}{n} D(\tilde{y}; \widehat{\mu_a^*}) + \frac{1}{n} k_n da \quad (5.4.4)$$

όπου: $\widehat{\mu_a^*} = (g^{-1}\left(\widetilde{x_1'}\widehat{\beta_\alpha^*}\right), \dots, g^{-1}(\widetilde{x_n'}\widehat{\beta_\alpha^*}))'$

και β_α^* είναι ο εκτιμητής της μη ποινικοποιημένης μέγιστης πιθανοφάνειας του β. Κατά συνέπεια, το GIC^* μπορεί να χρησιμοποιηθεί στην κλασική επιλογή μεταβλητών. Στη συνέχεια, θα δείξουμε την ασυμπτωτική απόδοση του GIC.

Θεώρημα 5.4.1: Υποθέτουμε ότι η συνάρτηση πυκνότητας του γενικευμένου γραμμικού μοντέλου ικανοποιεί τις συνθήκες των Fan & Li (2001) καθώς και τη συνθήκη (C_4) που αναφέρθηκε παραπάνω.

(A) Εάν υπάρχει μια θετική σταθερά M τέτοια ώστε $k_n < M$ τότε η ρυθμιστική παράμετρος $\hat{\lambda}$ επιλέγεται ελαχιστοποιώντας την $GIC_{k_n}(\lambda)$ της εξίσωσης (5.4.3) και ικανοποιεί:

$P\{\hat{\lambda} \in \Omega_-\} \to 0$ και $P\{\hat{\lambda} \in \Omega_+\} \geq \pi$ όπου π είναι μια μη αρνητική πιθανότητα.

(B) Υποθέτουμε ότι οι συνθήκες (C_1)-(C_3) ικανοποιούνται. Εάν $k_n \to \infty$ και $k_n/\sqrt{n} \to 0$ τότε η ρυθμιστική παράμετρος $\hat{\lambda}$ που επιλέγεται ελαχιστοποιώντας την $GIC_{k_n}(\lambda)$ της εξίσωσης (5.4.3) ικανοποιεί:

$$P\{\hat{a}_\lambda = \alpha_0\} \to 1$$

Όπως παρατηρούμε, το θεώρημα (5.4.1) παρέχει έναν οδηγό για την επιλογή της ρυθμιστικής παραμέτρου. Το 5.4.1 (A) υπονοεί ότι όταν το k_n είναι φραγμένο, τότε το GIC οδηγεί σε ένα υπερπροσαρμοσμένο μοντέλο ανεξάρτητα από την συνάρτηση ποινής που χρησιμοποιείται.

Από εδώ και στο εξής, εάν $k_n \to 2$ θα ονομάζουμε τον εκτιμητή της GIC εκτιμητή τύπου-AIC. Εάν $k_n \to \infty$ και $k_n/\sqrt{n} \to 0$ θα ονομάζουμε τον εκτιμητή της GIC εκτιμητή τύπου BIC. Ο τελευταίος, σύμφωνα με το θεώρημα 5.4.1(B) ο εκτιμητής τύπου-BIC ανιχνεύει το πραγματικό μοντέλο με συνέπεια. Έτσι, αυτού του τύπου οι εκτιμητής κατέχει την προβλεπτική ιδιότητα (*oracle property*).

Σημείωση 5.4.1: Στα μοντέλα γραμμικής παλινδρόμησης οι Fan & Li (2001) εφάρμοσαν τον εκτιμητή GCV που δίνεται παρακάτω, για να επιλέξουν τη ρυθμιστική παράμετρο:

$$GCV^*(\lambda) = \frac{\|\tilde{y} - \tilde{x}\hat{\beta}_\lambda\|^2}{n\{1 - \frac{df_L(\lambda)}{n}\}^2} \quad (5.4.4)$$

όπου:

$\|.\|$ είναι η ευκλείδια νόρμα

$\chi = (\chi_1', \dots, \chi_n')'$

df_L όπως ορίστηκε στην (5.3.3)

Για να επεκτείνουμε την εφαρμογή της GCV αντικαθιστούμε το άθροισμα τετραγώνων των σφαλμάτων στην (5.4.4) από την $G(\tilde{y}, \hat{\beta}_\lambda)$ και τότε επιλέγουμε το λ που ελαχιστοποιεί την εξίσωση:

$$GCV(\lambda) = \frac{G(\tilde{y}, \hat{\beta}_\lambda)}{n\{1 - \frac{df_\lambda}{n}\}^2}$$

Χρησιμοποιώντας την επέκταση Taylor , έχουμε επιπλέον:

$$GCV(\lambda) \approx \frac{1}{n}\{G(\tilde{y}, \hat{\beta}_\lambda) + df_\lambda \left[\frac{2G(\tilde{y}, \hat{\beta}_\lambda)}{n}\right]\}$$

Επειδή η $\frac{G(\tilde{y},\hat{\beta}_\lambda)}{n}$ φράσσεται, η GCV δημιουργεί ένα υπερπροσαρμοσμένο μοντέλο με θετική πιθανότητα.

Σημείωση 5.4.2: Στα γραμμικά μοντέλα παλινδρόμησης, οι Wang, Li & Tsai (2007) έδειξαν ότι ο GCV- εκτιμητής των Fan & Li (2001) για τα ποινικοποιημένα ελάχιστα τετράγωνα της SCAD δεν μπορεί να επιλέξει τη ρυθμιστική παράμετρο με συνέπεια. Επιπλέον, πρότειναν την ακόλουθη εκτιμήτρια της ρυθμιστικής παραμέτρου μέσω της BIC:

$$BIC^*(\lambda) = \log(\hat{\sigma}_\lambda^2) + \frac{1}{n}\log(n)\, df_L(\lambda)$$

όπου:

$\hat{\sigma}_\lambda^2 = \sum_{i=1}^n (\tilde{y}_i - \tilde{x}_i'\hat{\beta}_\lambda)^2/n$

Λαμβάνοντας υπόψη ότι $\log(1+t) \approx t$ για μικρό t, η $BIC^*(\lambda)$ είναι ποσεγγιστικά ίση με:

$$BIC^{**}(\lambda) = \frac{1}{n}D^{**}(\tilde{y}; \hat{\mu}_\lambda) + \frac{1}{n}\log(n)\, df_L(\lambda)$$

όπου:

$D^{**}(\tilde{y}; \hat{\mu}_\lambda) = \hat{\sigma}_\lambda^2/\hat{\sigma}_{\bar{a}}^2$ είναι η απόκλιση για την κανονική κατανομή και $\hat{\sigma}_{\bar{a}}^2$ είναι ο εκτιμητής διασποράς του πλήρους μοντέλου. Είναι φανερό ότι η BIC^{**} είναι ένας εκτιμητής τύπου-BIC.

5.5 Αποδοτικότητα του Κριτηρίου

Υπό την υπόθεση ότι το πραγματικό μοντέλο περιλαμβάνεται σε μια οικογένεια υποψήφιων μοντέλων, θα ασχοληθούμε με την συνέπεια των εκτιμητών τυπου-BIC. Στην πράξη, η υπόθεση αυτή δεν είναι βάσιμη και έτσι, μελετούμε και την ασυμπτωτική συμπεριφορά των εκτιμητών τύπου-AIC. Στη βιβλιογραφία, η L_2 νόρμα χρησιμοποιείται για να εκτιμήσει την αποδοτικότητα της κλασικής διαδικασίας AIC

στα γραμμικά μοντέλα παλινδρόμησης. Έτσι, εστιάζουμε στην αποδοτικότητα του γραμμικού μοντέλου παλινδρόμησης μέσω της L_2 νόρμας:

Θεωρούμε το ακόλουθο μοντέλο:

$$y_i = \mu_i + \varepsilon_i, \qquad i = 1, \dots, n$$

όπου:

$\mu = (\mu_1, \dots, \mu_n)'$ είναι το άγνωστο διάνυσμα μέσης τιμής

ε_i ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με μέση τιμή μηδέν και διασπορά σ^2

Επιπλέον, υποθέτουμε ότι το $\tilde{X}\tilde{\beta}$ συνιστά την πλησιέστερη αναπαράσταση του διανύσματος του μέσου μ. Προσαρμόζοντας την μέθοδο του Li (1987) επιτρέπουμε στο d που είναι η διάσταση του β , να τείνει στο άπειρο και επίσης $d/n \to 0$. Για γνωστό σύνολο δεδομένων $\{(\tilde{x}_i, y_i): i = 1, \dots, n\}$ ακολουθούμε την (5.2.1) και ορίζουμε τη συνάρτηση ποινικοποιημένων ελαχίστων τετραγώνων :

$$Q^{LS}(\beta) = \sum_{i=1}^{n}(\tilde{y}_i - \tilde{x}_i'\beta)^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad (5.5.1)$$

Ο ποινικοποιημένος εκτιμητής του μ στο μοντέλο α_λ είναι $\hat{\mu}_\lambda = \tilde{X}\hat{\beta}_\lambda$. Επιπλέον, ο μη ποινικοποιημένος εκτιμητής του μ στο μοντέλο α είναι $\hat{\mu}_\alpha^* = \tilde{\chi}\hat{\beta}_\alpha^*$.

Η ρυθμιστική παράμετρος λ επιτυγχάνεται ελαχιστοποιώντας:

$$GIC_{k_n}^{LS}(\lambda) = \frac{1}{n}\{\sum_{i=1}^{n}(y_i - \tilde{x}_i'\hat{\beta}_\lambda)^2 + k_n\sigma^2 d_{a_\lambda}\} \qquad (5.5.2)$$

Εδώ το σ^2 είναι γνωστό. 'Οταν στην παραπάνω εξίσωση το $\hat{\beta}_\lambda$ αντικατασταθεί από τον εκτιμητή ελαχίστων τετραγώνων $\hat{\beta}_\alpha^*$ το $GIC_{k_n}^{LS}$ με $k_n = 2$ μετατρέπεται στο κριτήριο του C_p -Mallow.

Για να εκτιμήσουμε την απόδοση του εκτιμητή της ρυθμιστικής παραμέτρου, υιοθετούμε την μέθοδο του Shibata (1981) και ορίζουμε τη Μέση Τετραγωνική Ζημία ($L_2\ loss$) που συνδέεται με τον εκτιμητή $\hat{\beta}_\lambda$ και είναι:

$$L(\hat{\beta}_\lambda) = \frac{1}{n}\|\mu - \widehat{\mu_\lambda}\|^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \tilde{x}_i'\hat{\beta}_\lambda)^2 \qquad (5.5.3)$$

Ακολούθως, η συνάρτηση ρίσκου (Risk function) είναι:

$$R(\hat{\beta}) = E[L(\hat{\beta})]$$

Χρησιμοποιώντας το μέτρο της μέσης τετραγωνικής ζημίας ορίζουμε την ασυμπτωτική απόδοση ζημίας (*Asymptotic loss efficient*).

Ορισμός 5.5.1 (*Asymptotically Loss Efficient*): Μια διαδικασία επιλογής ρυθμιστικής παραμέτρου λέγεται ασυμπτωτική απόδοση ζημίας εάν:

$$\frac{L(\widehat{\beta_{\hat{\lambda}}})}{inf_{\lambda\in[0,\lambda_{max}]}L(\widehat{\beta_{\lambda}})} \to 1 \qquad (5.5.4)$$

με πιθανότητα, όπου $\widehat{\beta_{\hat{\lambda}}}$ σχετίζεται με τη ρυθμιστική παράμετρο $\hat{\lambda}$ που επιλέχθηκε από αυτή τη διαδικασία. Επίσης λέμε ότι το $\widehat{\beta_{\hat{\lambda}}}$ είναι *ALE* εάν ισχύει η παραπάνο σχέση.

Τώρα θα αναφέρουμε κάποιες τεχνικές συνθήκες για τη μελέτη της ασυμπτωτικής απόδοσης ζημίας για τους εκτιμητές τύπου-AIC στη γραμμική παλινδρόμηση:

(C_5) Ο πίνακας $(\frac{1}{n}X'X)^{-1}$ υπάρχει και η μεγαλύτερη ιδιοτιμή του φράσσεται από μια σταθερά C.

(C_6) $E\varepsilon_1^{4q} < \infty$ για θετικό ακέραιο q.

(C_7) Το ρίσκο των εκτιμητών ελαχίστων τετραγώνων $\hat{\beta}_\alpha^*$ $\forall \alpha \in A$ ικανοποιεί:

$$\sum_{a\in A}\left[nR(\hat{\beta}_\alpha^*)\right]^{-q} \to 0$$

(C_8) Έστω $b = (b_1, \dots, b_n)'$ όπου $b_j = p_\lambda'(|\hat{\beta}_{\lambda j}|)sgn(\hat{\beta}_{\lambda j})$ $\forall\, j$ *τέτοιο ώστε* $|\hat{\beta}_{\lambda j}| > 0$ *και* $b_j = 0$ *αλλιώς*, όπου $\hat{\beta}_{\lambda j}$ το j-οστό στοιχείο του ποινικοποιημένου εκτιμητή $\widehat{\beta_\lambda}$. Επιπλέον, έστω $\hat{\beta}_{\alpha_\lambda}$ ο εκτιμητής ελαχίστων τετραγώνων του β που επιτυγχάνεται από το μοντέλο α_λ. Τότε υποθέτουμε ότι με πιθανότητα

$$sup_{\lambda\in[0,\lambda_{max}]}\frac{\|b\|^2}{R(\hat{\beta}_{\alpha_\lambda})} \to 0$$

Θεώρημα 5.5.1: Υποθέτουμε ότι οι συνθήκες (C_5)-(C_8) ισχύουν. Τότε, η ρυθμιστική παράμετρος $\hat{\lambda}$ που επιλέγεται ελαχιστοποιώντας την $GIC_{k_n}^{LS}(\lambda)$ στην (5.5.2) με $k_n \to 2$ οδηγεί σε έναν εκτιμητή $\widehat{\beta_{\hat{\lambda}}}$ με την ιδιότητα της ασυμπτωτικής απόδοσης ζημίας με την έννοια της σχέσης (5.5.4).

Το θεώρημα αυτό δείχνει ότι ο εκτιμητής τύπου-AIC έχει την ιδιότητα της ασυμπτωτικής απόδοσης ζημίας. Χρησιμοποιώντας τώρα το γεγονός ότι $\log(1+t) \approx t$ για μικρά t, ο εκτιμητής AIC συμπεριφέρεται όμοια με τον εκτιμητή AIC*. Ο τελευταίος ορίζεται:

$$AIC^*(\lambda) = \log\left(\hat{\sigma}_\lambda^2\right) + \frac{2\sigma^2 d_{\alpha_\lambda}}{n}$$

Επακόλουθα, οι εκτιμητές AIC και AIC* έχουν την παραπάνω ιδιότητα.

Σ'αυτό το σημείο μπορούμε να αναφέρουμε ότι ο εκτιμητής τύπου-BIC δεν έχει αυτή την ιδιότητα, γεγονός που συμφωνεί με τα ευρήματα της κλασικής επιλογής μεταβλητών.

Στην πράξη, το σ^2 είναι άγνωστο, οπότε αντικαθιστούμε το σ^2 στον τύπο του GIC_2^{LS} με τον συνεπή του εκτιμητή $\tilde{\sigma}^2$.

Συμπέρασμα: Εάν η ρυθμιστική παράμετρος $\hat{\lambda}$ επιλέγεται ελαχιστοποιώντας την $GIC_{k_n}^{LS}$ με $k_n \to 2$ και το σ^2 αντικαθίσταται από τον συνεπή του εκτιμητή $\tilde{\sigma}^2$ τότε η διαδικασία είναι επίσης *asymptotically loss efficiency*.

Σημείωση 5.5.1: Μπορούμε επίσης να μελετήσουμε την ασυμπτωτική απόδοση ζημίας των γενικευμένων γραμμικών μοντέλων. Ακολουθώντας το πνεύμα του μέτρου διασποράς, υιοθετούμε το μέτρο απόστασης (distance measure) των Kullback-Leibler (KL) για να ορίσουμε την KL-ζημία (*KL-Loss*) ενός εκτιμητή $\hat{\beta}$ ως:

$$L_{KL}\left(\hat{\beta}\right) = \frac{2}{n} E_0\{l(\theta_0) - l\left(\hat{\theta}\right)\} \qquad (5.5.5)$$

όπου:

η $l(.)$ ορίστηκε στην (5.4.2)

$\theta_0 = (\theta_{01}, \ldots \theta_{0n})'$ η πραγματική άγνωστη παράμετρος κανονικοποίησης

$\hat{\theta} = \left(\hat{\theta}_1, \ldots \hat{\theta}_n\right)' = \tilde{X}\tilde{\hat{\beta}}$ υπό την κανονική συνάρτηση σύνδεσης

E_0 η αναμενόμενη τιμή (*expectation*) υπό το πραγματικό μοντέλο

Μπορεί να δειχθεί ότι η KL-ζημία είναι ίδια με την τετραγωνική ζημία (*squared loss*) για τα κανονικά γραμμικά μοντέλα παλινδρόμησης με άγνωστη διασπορά. Για να επιτύχουμε την ασυμπτωτική συνάρτηση ζημίας για τα γενικευμένα γραμμικά μοντέλα, είναι απαραίτητο να χρησιμοποιήσουμετην επέκταση Taylor για να επεκτείνουμε το $b(\hat{\theta}_i)$ στην πραγματική τιμή θ_{0i}.

5.6 Αριθμητική Μελέτη

Στη συνέχεια, θα παρουσιάσουμε δύο παραδείγματα.Και τα δύο είναι πειράματα Monte Carlo. Στο πρώτο παράδειγμα, θεωρούμε ότι το πραγματικό μοντέλο περιλαμβάνεται σε ένα σύνολο υποψήφιων μοντέλων τα οποία προσαρμόζονται με την λογιστική παλινδρόμηση. Αυτό μας επιτρέπει να εξετάσουμε την απόδοση της συνέπειας του κριτηρίου επιλογής, το οποίο αναμένεται να αποδίδειεξίσου καλά με το Θεώρημα 5.4.1. Αντίθετα, στο δεύτερο παράδειγμα, το πραγματικό μοντέλο δεν περιλαμβάνεται στο σύνολο των υποψήφιων μοντέλων, τα οποία ακολουθουν την Γκαουσιανή κατανομή.

Παράδειγμα 1: Υιοθετώντας το μοντέλο των Tibshirani (1996) και των Fan & Li (2001) προσομοιώνουμε τα δεδομένα από το μοντέλο λογιστικής παλινδρόμησης, $y|\tilde{x} \sim Bernoylli\{p(\tilde{x}'\beta)\}$ όπου:

$$p(\tilde{x}'\beta) = \mu(\tilde{x}'\beta) = \frac{\exp(\tilde{x}'\beta)}{1 + \exp(\tilde{x}'\beta)}$$

Επιπλέον, $\tilde{\beta} = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0)'$

και $\tilde{\chi}$ είναι ένα 12-διάστατο τυχαίο διάνυσμα, με τα πρώτα 9 στοιχεία του να ακολουθούν την κανονική κατανομή με πίνακα διασποράς $\Sigma = (\sigma_{ij})$ για τα οποία ισχύει $\sigma_{ij} = 0{,}5^{|i-j|}$,ενά τα τρία τελευταία στοιχεία του $\tilde{\chi}$ παράγονται από την κατανομή Bernoulli με ποσοστό επιτυχίας 0,5. Επιπλέον, κάνουμε 1000 επαναλήψεις με μέγεθος δείγματος n=200 και n=400.

Για να εκτιμήσουμε την απόδοση των προτεινόμενων μεθόδων, αναφέρουμε το ποσοστό των πλώρων, ελλειπών και περιττών μοντέλων με 1, 2, 3, 4 ,5 ή και περισσότερες παραμέτρους που προκύπτουν από τις SCAD-AIC ($k_n = 2$) , SCAD-BIC ($k_n = \log(n)$), SCAD-GCV, AIC και BIC όπως επίσης και μέσω της ιδιότητας πρόβλεψης.Τα αντίστοιχα σφάλματα υπολογίζονται από το $\sqrt{\frac{\hat{p}(1-\hat{p})}{1000}}$ όπου $\hat{p}$ είναι το παρατηρούμενο μέτρο (observed proportion) στις 1000 προσομοιώσεις. Επιπλέον, αναφέρουμε το μέσο όρο των μηδενικών συντελεστών που σωστά (C) και λανθασμένα (I) υπολογίστηκαν με διαφορετικές μεθόδους. Για να συγκρίνουμε την προσαρμογή του μοντέλου, υπολογίζουμε το ακόλουθο σφάλμα μοντέλου (model error) για τη νέα παρατήρηση $(\tilde{x}, y)$:

$$\mathrm{ME}(\hat{\tilde{\beta}}) = E_x\{\mu(\tilde{x}'\beta) - \mu(\tilde{x}'\hat{\beta})\}^2$$

όπου:

το E λαμβάνεται ως προς το νέο παρατηρούμενο διάνυσμα $\tilde{x}$ και $\mu(\tilde{x}'\beta) = E(y|\tilde{x})$. Στη συνέχεια, υπολογίζουμε τη διάμεσο του σχετικού μοντέλου σφάλματος (MRME):

$RME = ME/ME_{full}$ και ME_{full} το σφάλμα μοντέλου που υπολογίζεται από την προσαρμογή των δεδομένων με το πλήρες μοντέλο.

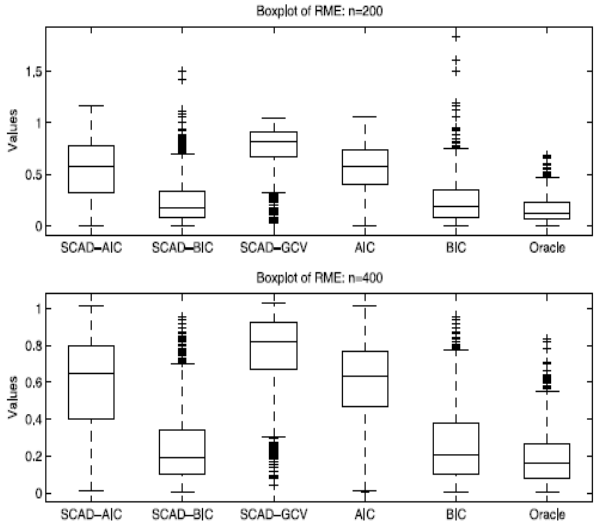
Ο πίνακας που παρουσιάζεται παρακάτω δείχνει ότι το MRME της μεθόδου SCAD-BIC είναι μικρότερο από το αντίστοιχο της SCAD-AIC. Όσο το μέγεθος του δείγματος αυξάνεται το MRME της SCAD-BIC προσεγγίζει αυτό του εκτιμητή

πρόβλεψης ενώ το MRME της SCAD-AIC παραμένει το ίδιο. Στο θηκογράφημα που ακολουθεί φαίνεται ακριβώς αυτό το αποτέλεσμα.

Στην αναγνώριση του μοντέλου, η SCAD-BIC έχει υψηλότερη απόδοση από την SCAD-AIC των σωστά βαλμένων 9 πραγματικών μηδενικών συμμεταβλητών σε μηδέν, ενώ η SCAD-BIC είναι πιο επιρρεπής στα να θέσει τις τρεις μη μηδενικές μεταβλητές ίσες με μηδέν, όταν το μέγεθος του δείγματος είναι μικρό. Επιπρόσθετα, η SCAD-BIC έχει μεγαλύτερο ποσοστό επιτυχίας στο να ανιχνεύσει το σωστό μοντέλο σε σχέση με την SCAD-AIC.

Method	MRME (%)	Zeros C	Zeros I	Under (%)	Exact (%)	Overfitted (%) 1	2	3	4	≥5
$n = 200$										
SCAD-AIC	58.03	7.4 (1.6)	0.0 (0.0)	0.2	30.2	23.2	19.6	13.6	7.5	5.9
SCAD-BIC	16.90	8.8 (0.5)	0.0 (0.1)	0.7	83.7	13.2	2.4	0.4	0.0	0.0
SCAD-GCV	81.91	5.8 (1.9)	0.0 (0.0)	0.0	7.8	12.0	18.1	18.8	17.6	25.7
AIC	57.05	7.4 (1.2)	0.0 (0.0)	0.1	19.9	31.5	26.7	14.8	4.3	2.8
BIC	18.27	8.8 (0.5)	0.0 (0.1)	0.8	80.4	17.0	2.3	0.1	0.0	0.0
Oracle	12.49	9.0 (0.0)	0.0 (0.0)	0.0	100.0	0.0	0.0	0.0	0.0	0.0
$n = 400$										
SCAD-AIC	64.45	7.4 (1.5)	0.0 (0.0)	0.0	29.4	21.9	21.0	15.1	8.5	4.1
SCAD-BIC	19.03	8.9 (0.4)	0.0 (0.0)	0.0	89.9	7.9	1.8	0.2	0.2	0.0
SCAD-GCV	82.13	6.0 (1.9)	0.0 (0.0)	0.0	9.2	14.7	18.6	19.8	15.9	21.8
AIC	63.17	7.4 (1.2)	0.0 (0.0)	0.0	21.7	29.2	27.3	14.6	5.5	1.7
BIC	20.46	8.8 (0.4)	0.0 (0.0)	0.0	86.3	12.1	1.4	0.2	0.0	0.0
Oracle	15.88	9.0 (0.0)	0.0 (0.0)	0.0	100.0	0.0	0.0	0.0	0.0	0.0

Πίνακας 1: Προσομοίωση αποτελεσμάτων για το λογιστικό μοντέλο παλινδρόμησης.



Γράφημα 1: Θηκογράφημα τοτ σχετικού μοντέλου σφάλματος (RME) για n=200 και n=400.

Σχετικά με τα overfitted μοντέλα, η μέθοδος SCAD-BIC είναι πιθανό να περιλαμβάνει μόνο μια ασυσχέτιστη μεταβλητή, ενώ η SCAD-AIC συχνα περιλαμβάνει δύο ή και περισσότερες. Όσο το μέγεθος του δείγματος αυξάνεται, η SCAD-BIC τείνει να φτιάχνει σωστά μοντέλα, σε αντλιθεση με την SCAD-AIC που οδηγεί σε overfitting. Αυτά τα αποτελέσματα συμφωνούν και με τα θεωρητικά ευρήματα. Σ'αυτό να ανφέρουμε ότι η SCAD-GCV Συμπεριφέρεται όμοια με την SCAD-AIC Ενώ τα κλασικά κριτήρια BIC και AIC συμπεριφέρονται όμοια με τις SCAD-BIC και SCAD-AIC αντίστοιχα. Είναι ενδιαφέρον ότι η SCAD-AIC προτείνει συνήθως ένα πιο

σποραδικό μοντέλο σε σχέση με την AIC. Σε γενικές γραμμές η SCAD-BIC συμπεριφέρεται καλύτερα από τις υπόλοιπες μεθόδους.

Παράδειγμα 2: Στην πραγματικότητα, το πλήρες μοντέλο αποτυγχάνει στο να συμπεριλάβει κάποιες σημαντικές επεξηγηματικές μεταβλητές. Αυτό μας δίνει το κίνητρο να μιμηθούμε εμείς μία κατάσταση και να εφαρμόσουμε κάποιες μεθόδους. Έτσι, θεωρούμε το γραμμικό μοντέλο παλινδρόμησης $y_i = \tilde{x}_i^{'}\tilde{\beta} + \varepsilon_i$ όπου $\tilde{x_i}$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές διάστασης 13, ενώ η σχέση ανάμεσα σε x_i και x_j είναι $0{,}5^{|i-j|}$. Τα ε_i είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ακολουθούν την $N(0, \sigma^2)$ με σ=4. Επιπλέον, διαχωρίζουμε το $x = (x_{full}^{'}, x_{exc}^{'})'$ όπου το x_{full} περιλαμβάνει d=12 μεταβλητές του πλήρους μοντέλου και το x_{exc} προκύπτει από την εφαρμογή του μοντέλου. Με τον ίδιο τρόπο διαχωρίζουμε το $\beta = (\beta_{full}^{'}, \beta_{exc}^{'})'$ όπου β_{full} είναι ένα (12×1) διάνυσμα ενώ το β_{exc} είναι μονοδιάστατο. Για να ερευνήσουμε την απόδοση των μεθόδων που θα εφαρμόσουμε έχουμε:

$$\beta_{full} = \beta_0 + \gamma\delta/\sqrt{n}$$

όπου: $\beta_0 = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0)'$

$\delta = (0, 0, 1.5, 1.5, 1, 1, 0, 0, 0, 0, 0.5, 0.5)'$

και το γ κυμαίνεται από 0-10

$$\beta_{exc} = 0.2$$

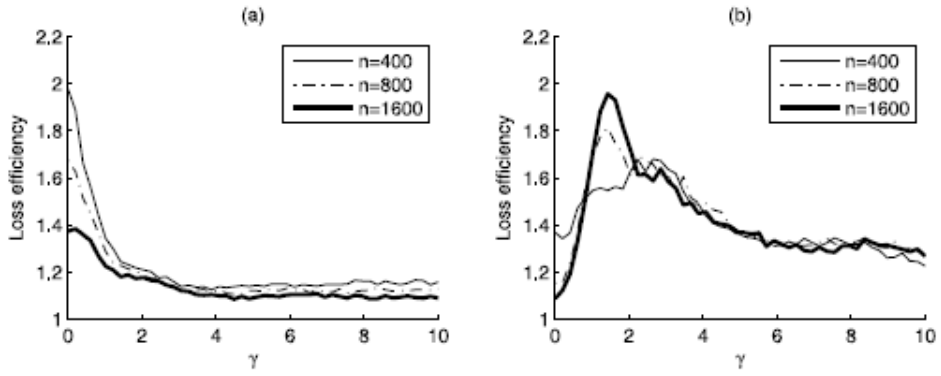
Επειδή το υποψήφιο μοντέλο είναι υποσύνολο του πλήρους μοντέλου που περιλαμβάνει 12 μεταβλητές στο x_{full} η παραπάνω προσαρμογή του μοντέλου μας διαβεβαιώνει ότι το πραγματικό μοντέλο δεν περιλαμβάνεται στο σύνολο των υποψήφιων μοντέλων. Προσομοιώνουμε 1000 δεδομένα με n=400, 800 και 1600. Για να εξετάσουμε την απόδοση των εκτιμητών, ορίζουμε την απόδοση ζημίας του πεπερασμένου δείγματος:

$$LE\left(\hat{\beta}_{\hat{\lambda}}\right) = \frac{L(\hat{\beta}_{\hat{\lambda}})}{inf_{\lambda} L(\hat{\beta}_{\lambda})}$$

όπου: $\hat{\lambda}$ επιλέγεται από την SCAD-GCV, SCAD-AIC και SCAD-BIC. Είναι αξιοσημείωτο, ότι το β_{full} εξαρτάται από το μέγεθος του δείγματος και το γ. Το γ

είναι μια ευαίσθητη μεταβλητή που συγκρίνει την απόδοση ζημίας των κριτηρίων επιλογής ανάμεσα σε διαφορετικά μεγέθη δείγματος. Επειδή το SCAD-GCV είναι πολύ όμοιο με το SCAD-AIC δεν γίνεται αναφορά εδώ. Το παρακάτω διάγραμμα δείχνει την απόδοση ζημίας της SCAD-AIC και της SCAD-BIC για διάφορες τιμές του γ. Σύμφωνα με το διάγραμμα 2(α), δείχνει ότι η απόδοση ζημίας της μεθόδου SCAD-AIC Τείνει στη μονάδα, ανεξάρτητα από την τιμή του γ, αντίθετα με την απόδοση ζημίας του SCAD-BIC που δεν δείχνει κάτι τέτοιο.

Όταν το γ είναι κοντά στο μηδέν, δηλαδή όταν το μοντέλο είναι σχεδόν σποραδικό, η μέθοδος SCAD-BIC καταλήγει σε μικρότερη ζημία εξαιτίας μεγαλύτερης συνάρτησης ποινής. Γενικά το διάγραμμα 2 δείχνει ότι η μέθοδος SCAD-AIC είναι αποδοτική ενώ η SCAD-BIC δεν είναι. Οι AIC, BIC συμπεριφέρονται ανάλογα με τις SCAD-AIC, SCAD-BIC αντιστοίχως και έτσι παραλείπονται.



Διάγραμμα 2: Απόδοση Ζημίας της SCAD-AIC (a) και της SCAD-BIC (b).

Appendix 2ου Κεφαλαίου

- Απόδειξη Θεωρήματος 2.3.2.1

Έστω $a_n = n^{-1/2} + a_n$. Θέλουμε να δείξουμε ότι για κάθε γνωστό $\varepsilon > 0$ υπάρχει μια μεγάλη σταθερά C ώστε:

$$P\{sup_{\|\tilde{u}\|=C} Q(\tilde{\beta}_0 + a_n\tilde{u}) < Q(\tilde{\beta}_0)\} \geq 1 - \varepsilon \qquad \text{(A.1)}$$

Αυτό υπονοεί ότι με πιθανότητα τουλάχιστον $1-\varepsilon$ υπάρχει τοπικό μέγιστο στη μπάλα $\{\tilde{\beta}_0 + a_n\tilde{u}: \|\tilde{u}\| \leq C\}$. Έστω ότι υπάρχει μια τιμή που τοπικά ελαχιστοποιεί τέτοια ώστε :

$$\left\|\hat{\tilde{\beta}} - \tilde{\beta}_0\right\| = O_p(a_n)$$

Χρησιμοποιώντας ότι $P_\lambda(0) = 0$ έχουμε:

$$D_n(\tilde{u}) = Q(\tilde{\beta}_0 + a_n\tilde{u}) - Q(\tilde{\beta}_0) \leq$$

$$\leq L(\tilde{\beta}_0 + a_n\tilde{u}) - L(\tilde{\beta}_0) - n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + a_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}$$

Έστω $L'(\tilde{\beta}_0)$ να είναι το ανάδελτα του L. Έχουμε:

$$D_n(\tilde{u}) \leq a_n L'(\tilde{\beta}_0)'\tilde{u} - \frac{1}{2}\tilde{u}'I(\tilde{\beta}_0)\tilde{u}na_n^2\{1 + O_p(1)\} -$$

$$-\sum_{j=1}^{s}[na_n p'_{\lambda_n}(|\beta_{j0}|)sgn(\beta_{j0})u_j +$$

$$+na_n^2 p''_{\lambda_n}(|\beta_{j0}|)u_j^2\{1 + O(1)\}] \qquad \text{(A.2)}$$

Σημειώνουμε ότι $n^{-1/2}L'(\tilde{\beta}_0)' = O_p(1)$. Έτσι, το δεξί μέλος της (A.2) είναι τάξης $O_p\left(n^{\frac{1}{2}}a_n\right) = O_p(na_n^2)$.

Επιλέγοντας ένα σημαντικά μεγάλο C ο δεύτερος όρος υπερισχύει του πρώτου με $\|u\| = C$.Αυτό ολοκληρώνει και την απόδειξη.

- Απόδειξη Λήμματος 2.3.2.1

Είναι σημαντικο να δείξουμε ότι με πιθανότητα να τείνει στη μονάδα όσο το $n \to \infty$ για κάθε $\tilde{\beta}_1$ που ικανοποιεί $\tilde{\beta}_1 - \tilde{\beta}_{10} = O_p(n^{-1/2})$ και για μικρό $\varepsilon_n = Cn^{-1/2}$ & $j = s+1, \dots, d$

$$\frac{\partial Q(\tilde{\beta})}{\partial \beta_j} = \begin{cases} < 0 \text{ για } 0 < \beta_j < \varepsilon_n & (A.3) \\ > 0 \text{ για } -\varepsilon_n < \beta_j < 0 & (A.4) \end{cases}$$

Για να δείξουμε την (A.3) από την επέκταση Taylor έχουμε:

$$\frac{\partial Q(\tilde{\beta})}{\partial \beta_j} = \frac{\partial L(\tilde{\beta})}{\partial \beta_j} - np'_{\lambda_n}(|\beta_j|)sgn(\beta_j) =$$

$$= \frac{\partial L(\widetilde{\beta_0})}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 L(\widetilde{\beta_0})}{\partial \beta_j \partial \beta_j}(\beta_l - \beta_{l_0}) + \sum_{l=1}^{d}\sum_{k=i}^{d} \frac{\partial^3 L(\tilde{\beta}^*)}{\partial \beta_j \partial \beta_i \beta_k} \times$$

$$\times (\beta_l - \beta_{l_0})(\beta_k - \beta_{k_0}) - np'_{\lambda_n}(|\beta_j|)sgn(\beta_j)$$

όπου το β^* βρίσκεται ανάμεσα στο $\tilde{\beta}$ και το $\widetilde{\beta_0}$. Σημειώνουμε ότι

$n^{-1}\frac{\partial L(\widetilde{\beta_0})}{\partial \beta_j} = O_p(n^{-1/2})$

και

$\frac{1}{n}\frac{\partial^2 L(\widetilde{\beta_0})}{\partial \beta_j \partial \beta_j} = E\left\{\frac{\partial^2 L(\widetilde{\beta_0})}{\partial \beta_j \partial \beta_j}\right\} + O_p(1)$

Υπό την υπόθεση ότι $\tilde{\beta} - \widetilde{\beta_0} = O_p(n^{-1/2})$ έχουμε:

$$\frac{\partial Q(\tilde{\beta})}{\partial \beta_j} = n\lambda_n\{-\lambda'_n p'_{\lambda_n}(|\beta_j|)sgn(\beta_j) + O_p(\frac{n^{-\frac{1}{2}}}{\lambda_n})\}$$

Ενώ $liminf_{\theta \to 0^+} {\lambda_n}^{-1} p'_{\lambda_n}(\theta) > 0$ και $\frac{n^{-\frac{1}{2}}}{\lambda_n} \to 0$, η παράγωγος εξαρτάται από αυτή του β_j. 'Ετσι, οι (A.3) και (A.4) ισχύουν. Αυτό ολοκληρώνει την απόδειξη.

- Απόδειξη Θεωρήματος 2.3.2.2

Από το παραπάνω Λήμμα προκύπτει ότι το (α) ισχύει. Θα ασχοληθούμε με την απόδειξη του (b). Μπορεί να δείξουμε εύκολα ότι υπάρχει ένα $\hat{\beta}_1$ στο θεώρημα 2.3.2.1 που είναι $\sqrt{n}$-συνεπής τοπικός "μεγιστοποιητής" $Q\{\binom{\beta_1}{0}\}$ το οποίο λαμβάνεται ως συνάρτηση του $\hat{\beta}_1$ και ικανοποιεί:

$$\frac{\partial Q(\widetilde{\beta_0})}{\partial \beta_j}_{\beta=\binom{\hat{\beta}_1}{0}} = 0 \text{ για } j = 1, \dots, s \qquad \text{(A.5)}$$

Σημειώνουμε ότι το $\hat{\beta}_1$ είναι συνεπής εκτιμητής ,

$$\frac{\partial L(\tilde{\beta})}{\partial \beta_j}_{\beta=\binom{\hat{\beta}_1}{0}} - np'_{\lambda_n}(|\hat{\beta}_j|)sgn(\hat{\beta}_j) =$$

$$= \frac{\partial L(\widetilde{\beta_0})}{\partial \beta_j} + \sum_{l=1}^{s}\{\frac{\partial^2 L(\widetilde{\beta_0})}{\partial \beta_j \partial \beta_l} + O_p(1)\}(\hat{\beta}_l - \beta_{l_0}) -$$

$$-n(p'_{\lambda_n}(|\beta_{j0}|)sgn(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + O_p(1)\}(\hat{\beta}_j - \beta_{j_0})$$

Προκύπτει από το Θεώρημα του Slutsky και από το Κ.Ο.Θ. ότι:

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

Appendix 3ου Κεφαλαίου

- Απόδειξη Λήμματος 3.4.1

Θεωρώντας ότι $A(\lambda)g$ είναι το διάνυσμα τιμών της συνάρτησης που συμβολίζεται με $g_{n,\lambda}^{*}$ και αποτελεί τη λύση του προβλήματος: Βρες $f \in W_2^{(m)}$ ώστε να ελαχιστοποιεί το:

$$\frac{1}{n}\sum_{j=1}^{n}(g(t_j)-f(t_j))^2+\lambda\int_0^1(f^{(m)}(u))^2du$$

Οπότε:

$$\frac{1}{n}\sum_{j=1}^{n}(g(t_j)-g_{n,\lambda}^{*}(t_j))^2+\lambda\int_0^1(g_{n,\lambda}^{*}{}^{(m)}(u))^2du \equiv$$

$$\equiv \frac{1}{n}\|I-A(\lambda)g\|^2+\lambda\int_0^1(g_{n,\lambda}^{*}{}^{(m)}(u))^2du \leq$$

$$\leq \frac{1}{n}\sum_{j=1}^{n}(g(t_j)-g(t_j))^2+\lambda\int_0^1(g^{(m)}(u))^2du=\lambda\int_0^1(g^{(m)}(u))^2du$$

- Απόδειξη Λήμματος 3.4.2

Θεωρούμε την $g_{n,\lambda}^{*}$ όπως ορίστηκε παραπάνω. Τότε το $g_{n,\lambda}^{*}$ συγκλίνει στο $W_2^{(m)}$ όσο το $n \to \infty$ στην g_{λ}^{*} (δηλαδή στην τιμή που ελαχιστοποιεί) και:

$$J_{\infty,g}(f)=\int_0^1\frac{(g(u)-f(u))^2}{\omega(u)}du+\lambda\int_0^1(f^{(m)}(u))^2du$$

Τώρα, εάν $g \in \pi_{m-1}$ είναι εύκολο να δούμε ότι $g_{\lambda}^{*}=g$ και σε αυτή την περίπτωση $J_{\infty,g}(g)=0$.

Εάν $g \notin \pi_{m-1}$ τότε:

$$J_{\infty,g}(\theta g)<J_{\infty,g}(g)$$

για:

$$\theta = \frac{\frac{(g(u))^2}{\omega(u)}du}{\frac{(g(u))^2}{\omega(u)}du + \lambda\int_0^1 (g^{(m)}(u))^2 du}$$

έτσι ώστε $g_\lambda^* \neq g$

Επιπλέον,

$$\frac{1}{n}g'(I - A(\lambda))^2 g = \frac{1}{n}\sum_{j=1}^{n}(g(t_i) - g_{n,\lambda}^*(t_i))^2 \to \int_0^1 \frac{(g(u) - g_\lambda^*(u))^2}{\omega(u)} > 0 \quad \gamma\iota\alpha\ \lambda > 0$$

- Απόδειξη Λήμματος 3.4.3

$A(\lambda) = (n\lambda P + k)M^{-1}$ όπου:

$M = k + n\lambda I \quad$ και $\quad P = M^{-1}T(T'^{M^{-1}}T + \Delta)^{-1}T'$

Έστω: $n\lambda P M^{-1} = E$ και $\quad kM^{-1} = A_0$

Τώρα, $0 \leq A_0 \leq A \equiv A_0 + E \leq I$ και $0 \leq E \leq I$ με E τάξης (m+1) έχουμε:

$$TrA_0 \leq TrA \equiv TrA_0 + (m+1)$$

$$TrA_0{}^2 \leq TrA^2 \equiv TrA_0{}^2 + 2TrA_0E + TrE^2 \leq TrA_0{}^2 + 3TrE \leq TrA_0{}^2 + 3(m+1)$$

Και έτσι:

$$\lambda^{1/2}\sum_{\nu=1}^{n}(\frac{\lambda_{\nu n}}{\lambda_{\nu n} + n\lambda}) \leq n\lambda^{1/2m}\left[\frac{1}{n}TrA(\lambda)\right] \leq \lambda^{\frac{1}{2m}}\sum_{\nu=1}^{n}\left(\frac{\lambda_{\nu n}}{\lambda_{\nu n} + n\lambda}\right) + \lambda^{\frac{1}{2m}}(m+1)$$

όπου $\lambda_{\nu n}$, $\nu = 1,2,\dots,n$ οι ιδιοτιμές του k. Συνεχίζουμε την απόδειξη με την υπόθεση ότι οι ιδιοτιμές $\lambda_{\nu n}$ ικανοποιούν:

$$a\frac{(\pi\nu)^{2m}}{n} \leq \lambda_{\nu n}{}^{-1} \leq \beta\frac{(\pi\nu)^{2m}}{n} \quad \text{(A 4.3.1)}$$

για κάποια α,β με $0 < \alpha \leq \beta < \infty$.

Επίσης:

$$\alpha = min_t w(t)(1 + o(1))$$

$$\beta = max_t w(t)(1 + o(1)) \qquad \text{(A 4.3.2)}$$

όπου $o(1) \to 0 \;\; \gamma\iota\alpha \;\; n \to \infty$.

Χρησιμοποιώντας την (A 4.3.1) έχουμε:

$$\lambda^{1/2m} \sum_{\nu=1}^{n} \left(\frac{1}{1 + \lambda\beta(\pi\nu)^{2m}}\right) \leq \lambda^{1/2m} \sum_{\nu=1}^{n} \left(\frac{1}{1 + n\lambda\lambda_{\nu n}{}^{-1}}\right)^2 \leq \lambda^{1/2m} \sum_{\nu=1}^{n} \left(\frac{1}{1 + \lambda\alpha(\pi\nu)^{2m}}\right)$$

Για κάθε γνωστό $\gamma > 0$ έχουμε:

$$\int_{\gamma\lambda^{1/2m}\pi}^{(n-1)\gamma\lambda^{1/2m}\pi} \frac{dx}{(1 + x^{2m})^2} \leq (\gamma\lambda)^{1/2m}\pi \sum_{\nu=1}^{n} \left(\frac{1}{(1 + \lambda\gamma(\pi)^{2m}\nu^{2m})^2}\right) \leq \int_{0}^{\infty} \frac{dx}{(1 + x^{2m})^2}$$

Και επιτυγχάνουμε:

$$\frac{1}{\beta\lambda^{1/2m}\pi} \int_{\gamma\lambda^{1/2m}\pi}^{(n-1)\beta\lambda^{1/2m}\pi} \frac{dx}{(1 + x^{2m})} \leq n\lambda^{1/2m} \left[\frac{1}{n} TrA(\lambda)\right] \leq$$

$$\leq \frac{1}{a\lambda^{1/2m}\pi} \int_{0}^{\infty} \frac{dx}{(1 + x^{2m})^2} + \lambda^{\frac{1}{2m}}(m + 1)$$

Έτσι:

$$\frac{k_m}{\beta^{1/2m}} + O(\lambda) + O\left(\frac{1}{n\lambda^{\frac{1}{2m}}}\right) \leq n\lambda^{\frac{1}{2m}} \left[\frac{1}{n} TrA(\lambda)\right] \leq \frac{k_m}{a^{\frac{1}{2m}}} + O(\lambda) + O\left(\frac{1}{n\lambda^{\frac{1}{2m}}}\right)$$

Η j-οστή είσοδος k_{jk} του K δίνεται από:

$$k_{jk} = (-1)^{m-1} k_{2m}(t_j, t_k) = \sum_{\nu=-\infty}^{\infty} \left(\frac{1}{(2\pi\nu)^{2m}} e^{2\pi i\nu(t_j - t_k)}\right) \cong$$

$$\cong \sum_{\nu=-n/2}^{n\neq 2} \left(\frac{1}{(2\pi\nu)^{2m}} e^{2\pi i\nu(t_j - t_k)}\right)$$

Και έτσι:

$K = \Phi D \Phi^*$ όπου Φ ο $(n \times n)$ πίνακας με j-οστή είσοδο:

$$\frac{1}{\sqrt{n}} e^{2\pi i v(t_{jn})}$$

Και D ο διαγώνιος πίνακας με ν-οστή είσοδο:

$$D_{vv} \approx \frac{n}{(2\pi v)^{2m}}, v = -\frac{n}{2}, \dots, \frac{n}{2}, \quad v \neq 0$$

Έτσι,

$(t_{j+1,n} - t_{jn} = \frac{1}{nw(t_*)}$ $για\ κάποιο\ t_* \in [t_{jn}, t_{j+1,n}]$ έχουμε:

$$\frac{1}{n} \sum_{j=1}^{n} e^{2\pi i v(t_{jn})} e^{-2\pi i \mu(t_{jn})} \frac{1}{w(t_{jn})} \approx \int_0^1 e^{2\pi i (v-\mu)s} ds = \begin{cases} 1, \mu = v \\ 0, αλλιώς \end{cases}$$

Και έτσι, θέτοντας D_w να είναι ο διαγώνιος πίνακας με j-οστή είσοδο $\frac{1}{w(t_{jn})}$ έχουμε:

$$\Phi^* D_w \Phi \approx 1$$

Θέτοντας $U = D_w^{1/2} \Phi$, έχουμε ότι το U είναι μοναδικό.

$$K \approx D_w^{-1/2} U D U^* D_w^{-1/2}$$

Εάν η ισότητα στο (Α 4.4.3) ισχύει, καθώς και η μοναδικότητα του U τότε οι ιδιοτιμές λ_{vn} του K ικανοποιούν:

$$min_t (\frac{1}{w(t)}) D_{vv} \leq \lambda_{vn} \leq max_t (\frac{1}{w(t)}) D_{vv}$$

ή

$$min_t\ w(t) D_{vv} \leq {\lambda_{vn}}^{-1} \leq max_t\ w(t) D_{vv}$$

Εφόσον το 2ν-οστό και το (2ν-1) στοιχεία του D_{vv} είναι $\frac{n}{(2\pi v)^{2m}}$ τότε έχουμε την (Α 4.3.1) με α, β όπως ακριβώς στην (Α 4.3.2).

Μένει να δείξουμε ότι εάν το $1/n\lambda^{1/2m}$ φράσσεται από το μηδέν όσο το $n \to \infty$ τότε το ίδιο ισχύει και για το $[\frac{1}{n} TrA^2(\lambda)]$. Έχουμε:

$$\frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{(1+\beta\pi^{2m}\lambda\nu^{2m})^2}\leq\frac{1}{n}TrA^2(\lambda)$$

Έστω $\lambda=\lambda(n)$ να ικανοποιεί: $n\lambda^{1/2m}=c^{\frac{1}{2m}}$ ή ισοδύναμα $\lambda=c/n^{2m}$

Τότε:

$$\frac{1}{(1+\beta\pi^{2m}c)^2}\leq\frac{1}{n}\sum_{\nu=1}^{n}\frac{1}{(1+\beta\pi^{2m}c\left(\frac{\nu^{2m}}{n^{2m}}\right))}\leq\frac{1}{n}TrA^2(\lambda)$$

Appendix 4ου Κεφαλαίου

- Απόδειξη Λήμματος 4.3.1.1

Όταν $\lambda = 0$ έχουμε $DF_0 = d \; και \; GCV_0 = GCV_{S_F}$.Τότε, εφαρμόζοντας τη συνθήκη 1 με $2\log\left(1 - \frac{d}{n}\right) = O(n^{-1})$ πετυχαίνουμε

$$logGCV_{S_F} = \log\left(\frac{SSE_{S_F}}{n}\right) - 2\log\left(1 - \frac{d}{n}\right) = log\hat{\sigma}^2{}_{S_F} + O(n^{-1}) \quad \text{(A.1)}$$

Επειδή $SSE_\lambda \geq SSE_{S_\lambda}$ καταλήγουμε σε:

$$logGCV_\lambda \geq \log\left(\frac{1}{n}SSE_\lambda\right) \geq \log\left(\frac{1}{n}SSE_{S_\lambda}\right) = log\hat{\sigma}^2{}_{S_\lambda}$$

Σαν αποτέλεσμα, $inf_{\lambda\in\Omega\_}logGCV_\lambda \geq min_{S\not\supset S_T} log\hat{\sigma}_s^2$ (A.2)

Επιπλέον, η Συνθήκη 2 υπονοεί ότι $\hat{\sigma}_s^2 > \hat{\sigma}^2{}_{S_F}$ για $S \not\supset S_T$. Έτσι,

$$min_{S\not\supset S_T} log\sigma_s^2 > log\sigma^2{}_{S_F}$$

Αυτό το αποτέλεσμα, σε συνδιασμό με τις συνθήκες 1 και 2 καθώς και οι εξισώσεις (A.1) και (A.2) οδηγούν:

$$pr\left(inf_{\lambda\in\Omega\_}logGCV_\lambda > logGCV_{S_F}\right) \to 1 \; για \; n \to \infty$$

και η απόδειξη ολοκληρώθηκε.

- Απόδειξη Λήμματος 4.3.1.2

Για κάθε $\lambda \in \Omega_0$ έχουμε $S_\lambda = S_T$. Έτσι, $GCV_\lambda \geq \left(\frac{1}{n}\right) SSE_{S_\lambda} = \left(\frac{1}{n}\right) SSE_{S_T}$. Αυτό μαζί με το γεγονός ότι

$$(1 - \frac{d}{n})^{-2} = 1 + \frac{2d}{n} + O(n^{-2})$$

οδηγεί στο:

$$pr\left(inf_{\lambda\in\Omega_0} GCV_\lambda > GCV_{S_F}\right) \geq pr\left\{\frac{SSE_{S_T}}{n} > \frac{SSE_{S_F}}{n\left(1-\frac{d}{n}\right)^{-2}}\right\} =$$

$$= pr\{\frac{SSE_{S_T} - SSE_{S_F}}{\hat{\sigma}^2{}_{S_F}} > 2d + O(n^{-1})\ \} \qquad \text{(A.3)}$$

εφόσον $\hat{\sigma}^2{}_{S_F} = SSE_{S_F}/n.$

Σύμφωνα με τις συνθήκες 1 και 2 έχουμε ότι:

$\hat{\sigma}^2{}_{S_F} \to \sigma^2{}_{S_F} = \sigma_\varepsilon^2$ με πιθανότητα.

Επιπλέον, υπό τη συνθήκη 3, το $\frac{SSE_{S_T} - SSE_{S_F}}{\sigma_\varepsilon^2}$ ακολουθεί την $\chi^2_{d-d_0}$ κατανομή. Ως αποτέλεσμα,

$pr\left(inf_{\lambda\in\Omega_0} GCV_\lambda > GCV_{S_F}\right) \geq pr[\chi^2_{d-d_0}\{1 + O_p(1)\} > 2d + O(n^{-1})\ ] \to$
$pr\left(\chi^2_{d-d_0} > 2d\right) \coloneqq a$

Αυτό ολοκληρώνει την απόδειξη.

- Απόδειξη Λήμματος 4.3.2.1

Έστω $\beta_S = (\beta_{j1}, \dots, \beta_{jd_0})'$ το διάνυσμα των σχετικών συμμεταβλητών και β_N το διάνυσμα των άχρηστων συμμεταβλητών. Χωρίς βλάβη της γενικότητας υποθέτουμε ότι $\beta_S = (\beta_1, \dots, \beta_{d_0})'$ και $\beta_N = (\beta_{d_0+1}, \dots, \beta_d)'$. Επιπλέον, έστω

$\hat{\beta}_{\lambda_n} = (\hat{\beta}'_{S_{\lambda_n}}, \dots, \hat{\beta}'_{N_{\lambda_n}})'$ όπου: $\hat{\beta}'_{S_{\lambda_n}}$ και $\hat{\beta}'_{N_{\lambda_n}}$ είναι οι εκτιμητές SCAD των $\beta_S{}'$ και $\beta_N{}'$ αντίστοιχα. Υπό τη συνθήκη 5, εφαρμόζουμε το θεώρημα των Fan & Li και επιτυγχάνουμε με πιθανότητα να τείνει στη μονάδα, ότι το $\hat{\beta}_{S_{\lambda_n}}$ ικανοποιεί:

$$\frac{1}{n} X'_{S_T}\left(Y - X_{S_T}\,\hat{\beta}_{S_{\lambda_n}}\right) + b_n\left(\hat{\beta}_{S_{\lambda_n}}\right) = 0 \qquad \text{(A.4)}$$

όπου $b_n(\beta_s) = (p'_{\lambda_n}(|\beta_1|)sgn(\beta_1), \dots, p'_{\lambda_n}(|\beta_{d_0}|)sgn(\beta_{d_0}))'$.

Σύμφωνα με το Θεώρημα 1 των Fan & Li, $\hat{\beta}_{S_{\lambda_n}} \to \beta_S \neq 0$ με πιθανότητα. Επιπλέον, επειδή $\lambda_n = \log(n)/\sqrt{n}$ έχουμε $\alpha\lambda_n \to 0$. Σαν αποτέλεσμα, $pr(|\hat{\beta}_{S_{\lambda_n}}| > \alpha\lambda_n) \to 1$ το

οποίο υπονοεί ότι $pr\{b_n\left(\hat{\beta}_{S_{\lambda_n}}\right) = 0\} \to 1$. Αυτό μαζί με το (Α.4) υπονοεί ότι με πιθανότητα να τείνει στη μονάδα, η κανονική εξίσωση (Α.4) είναι ακριβώς ίδια με την

$$\frac{1}{n} X_{S_T}^{'} \left(Y - X_{S_T}\, \hat{\beta}_{S_{\lambda_n}}\right) = 0$$

η οποία είναι η κανονική εξίσωση για τον εκτιμητή ελαχίστων τετραγώνων που βασίζεται στο πταγματικό μοντέλο. Σαν αποτέλεσμα, με πιθανότητα να τείνει στη μονάδα, ο εκτιμητής SCAD $\hat{\beta}_{S_{\lambda_n}}$ είναι ακριβώς ίδιος με τον $\hat{\beta}_{S_T} = \left(X_{S_T}^{'} X_{S_T}\right)^{-1} (X_{S_T} Y)$.

Ακολουθεί αμέσως ότι $pr(SSE_{\lambda_n} = SSE_{S_T}) \to 1$ και με πιθανότητα νε τείνει στη μονάδα $\hat{\beta}_{N_{\lambda_n}} = 0$ από τη σποραδικότητα του Θεωρήματος 2 των Fan & Li.

Χρησιμοποιώντας παρόμοια επιχειρήματα, μπορούμε να δείξουμε ότι με πιθανότητα να τείνει στη μονάδα τα διαγώνια στοιχεία του Σ_{λ_n} τείνουν στο μηδέν το οποίο υπονοεί ότι $pr(DF_{\lambda_n} = d_0) \to 1$. Σαν αποτέλεσμα, με πιθανότητα να τείνει στη μονάδα έχουμε ότι $BIC_\lambda = BIC_{S_T}$. Αυτό ολοκληρώνει και την απόδειξη.

- Απόδειξη Λήμματος 4.3.2.2

Για $S_\lambda \neq S_T$, π.χ $\lambda \in \Omega_- \cup \Omega_+$, βρίσκουμε δυο διαφορετικές περιπτώσεις, δηλαδή μια για υπερπροσαρμοσμένο μοντέλο και μια για υποπροσαρμοσμένο μοντέλο. Σε κάθε περίπτωση, δείχνουμε ότι το Λήμμα ισχύει:

Περίπτωση 1Υποπροσαρμοσμένο μοντέλο, $S_\lambda \not\supset S_T$. Εφαρμόζοντας τα παραπάνω και τη συνθήκη 1, έχουμε ότι:

$$BIC_{\lambda_n} = log\hat{\sigma}^2{}_{S_T} + \frac{d_0 \log(n)}{n} \to \log\left(\sigma^2{}_{S_T}\right) = \log(\sigma_\varepsilon^2) \qquad (A.5)$$

με πιθανότητα. Προκύπτει από το γεγονός ότι $S_\lambda \not\supset S_T$ και από τις συνθήκες 1 και 2 ότι:

$$BIC_\lambda = \log\left(\frac{1}{nSSE_{S_\lambda}}\right) + DF_\lambda \frac{\log(n)}{n} \geq \log\{\frac{1}{nSSE_{S_\lambda}}\} \geq$$

$$\geq min_{\{S:S\not\supset S_T\}} log\hat{\sigma}^2{}_S \rightarrow min_{\{S:S\not\supset S_T\}} log\sigma^2{}_S > log\sigma^2{}_\varepsilon \qquad \text{(A.6)}$$

με πιθανότητα. Τελικά, οι (A.5) και (A.6) υπονοούν ότι

$$pr\{inf_{\lambda\in\Omega\_} BIC_\lambda > BIC_{\lambda_n}\} \rightarrow 1$$

Περίπτωση 2: Υπερπροσαρμοσμένο μοντέλο, $S_\lambda \supset S_T$, αλλά , $S_\lambda \neq S_T$. Σύμφωνα με τη συνθήκη 1, με πιθανότητα $\hat{\sigma}_{S_T} \rightarrow \sigma^2{}_{S_T} = \sigma^2{}_\varepsilon > 0$. Έπειτα, έστω d_λ ο αριθμός των μεταβλητών που συμπεριλαμβάνονται στο μοντέλο S_λ. Τότε για το υποπροσαρμοσμένο μοντέλο $d_\lambda > d_0$. Επιπλέον, χρησιμοποιώντας το θεώρημα αθροίσματος τετραγώνων δείχνουμε εύκολα ότι $\frac{SSE_{S_T} - SSE_{S_\lambda}}{\sigma^2{}_\varepsilon} \rightarrow \chi^2_{d_\lambda - d_0}$ $\gamma\iota\alpha\, n \rightarrow \infty$. Υπό συνθήκες ομαλότητας, ισχύει και για πεπερασμένο μοντέλο. Για κάθε υπερπροσαρμοσμένο μοντέλο S έχουμε:

$$SSE_{S_T} - SSE_{S_\lambda} = O_p(1) \qquad \text{(A.7)}$$

Αυτό μαζί με το προηγούμενο Λήμμα και με τον ορισμό της BIC_λ υπονοεί ότι με πιθανότητα που τείνει στη μονάδα,

$$n\big(BIC_\lambda - BIC_{\lambda_n}\big) \geq nlog\left(\frac{SSE_{S_\lambda}}{SSE_{S_T}}\right) + \big(DF_{\lambda_n} - d_0\big)logn =$$

$$= \big\{\hat{\sigma}^2{}_{S_T}\big(SSE_{S_\lambda} - SSE_{S_T}\big) + O_p(1)\big\} + \big\{d_\lambda + O_p(\lambda) - d_0\big\}logn$$

και η τελευταία ποσότητα ακολουθεί διότι $DF_{\lambda_n} = d_\lambda + O_p(\lambda)$ από τις Συνθήκες 4 και 5. Έτσι,

$$inf_{\lambda\in\Omega_+} n\big(BIC_\lambda - BIC_{\lambda_n}\big) \geq \hat{\sigma}^2{}_{S_T} min_{\{S:S\supset S_T\}}\big(SSE_S - SSE_{S_T}\big) + \big\{1 + O_p(\lambda)\big\}logn + O_p(1) \qquad \text{(A.8)}$$

Επίσης από την (A.7) προκύπτει ότι: $min_{\{S:S\supset S_T\}}\big(SSE_S - SSE_{S_T}\big) = O_p(1)$

Αυτό μαζί με το γεγονός ότι $\hat{\sigma}_{S_T} \rightarrow \sigma^2{}_{S_T}$ με πιθανότητα, δείχνει ότι το δεξί μέλος της (A.8) συγκλίνει στο $+\infty$ όσο το $n \rightarrow \infty$ το οποίο σημαίνει ότι

$$pr\big\{inf_{\lambda\in\Omega_+} n\big(BIC_\lambda - BIC_{\lambda_n}\big) > 0\big\} = pr\left(inf_{\lambda\in\Omega_+}\big(BIC_\lambda - BIC_{\lambda_n}\big)\right) \rightarrow 1$$

- Συνθήκες Ομαλότητας για το Μερικώς Γραμμικό Μοντέλο

Υποθέτουμε ότι $\{(u_i; x_i, y_i), i = 1, \dots, n\}$ είναι τυχαίο δείγμα. Οι ακόλουθες συνθήκες ομαλότητας είναι απαραίτητες για τις αποδείξεις:

i. Η συνάρτηση Kernel είναι συμμετρική συνάρτηση πυκνότητας με συμπαγές πεδίο ορισμού.
ii. Η τυχαία μεταβλητή u_i ορίζεται στο U και η συνάρτηση πυκνότητάς της είναι Lipshitz συνεχής.
iii. Η συνάρτηση $\alpha(.)$ έχει συνεχή δεύτερη παράγωγο για $u \in U$.
iv. Η $E(x_1|u_1 = u)$ είναι Lipshitz συνεχής για $u \in U$.
v. $\exists\, s > 1$ τέτοιο ώστε $E\|x_1\|^{2s} < \infty$ και για κάποιο $n < 2 - s^{-1}$ τέτοιο ώστε $n^{2n-1}h \to \infty$.
vi. $h = O_p(n^{-1/s})$

Appendix 5ου Κεφαλαίου

- Απόδειξη Θεωρήματος 5.4.1

Το $GIC_{k_n}(\lambda)$ το οποίο παράγει το ελλειπές (*underfitted*) μοντέλο, είναι μεγαλύτερο από το $GIC_{k_n}{}^{*}(\bar{\alpha})$. Έτσι, το βέλτιστο μοντέλο που επιλέγεται ελαχιστοποιώντας το $GIC_{k_n}(\lambda)$ πρέπει να περιέχει όλες τις σημαντικές μεταβλητές με πιθανότητα που τείνει στη μονάδα. Επίσης, υπάρχει θετική πιθανότητα η μικρότερη τιμή του $GIC_{k_n}(\lambda)$ για $\lambda \in \Omega_0$ να είναι μεγαλύτερα από αυτή του πλήρους μοντέλου. Σαν αποτέλεσμα, υπάρχει θετική πιθανότητα ώστε κάθε λ που συνδέεται με το πραγματικό μοντέλο να μην μπορεί να επιλεχθεί σαν ρυθμιστική παράμετρος. Το μοντέλο που ανιχνεύεται από το λ_n συγκλίνει στο πραγματικό μοντέλο όσο το μέγεθος του δείγματος αυξάνεται. Επιπλέον, το λ που αποτυγχάνει να ανιχνεύσει το πραγματικό μοντέλο δεν μπορεί να επιλεχθεί από την $GIC_{k_n}(\lambda)$.

- Απόδειξη Θεωρήματος 5.5.1

Για να δείξουμε την ασυμπτωτική συνέπεια του εκτιμητή τύπου AIC αρκεί να δείξουμε ότι ελαχιστοποιώντας το $GIC_{k_n}^{LS}(\lambda)$ με $k_n \to 2$ είναι το ίδιο με την ελαχιστοποίηση του $L(\hat{\beta}_\lambda)$ ασυμπτωτικά. Γι' αυτό χρειάζεται να δείξουμε ότι με πιθανότητα,

$$sup_{\lambda \in [0,\lambda_{max}\,]} \left| \frac{GIC_{k_n}^{LS}(\lambda) - \frac{\|\varepsilon\|^2}{n} - L(\hat{\beta}_\lambda)}{L(\hat{\beta}_\lambda)} \right| \to 0 \qquad \text{(B.1)}$$

Έστω ο πίνακας προβολής που συνδέεται με το μοντέλο α να είναι:

$$\tilde{H}_a = \tilde{X}_a (\tilde{X}_a' \tilde{X}_a)^{-1} \tilde{X}_a'$$

Τότε:

$$GIC_{k_n}^{LS}(\lambda) = \frac{\left\| \tilde{y} - \tilde{x}\hat{\beta}_\alpha \right\|^2}{n} + \frac{k_n \sigma^2 d_{a_\lambda}}{n} =$$

$$= \frac{\left\|\tilde{y} - \hat{\mu}^*_{\alpha_\lambda}\right\|^2}{n} + \frac{\left\|\hat{\mu}^*_{\alpha_\lambda} - \hat{\mu}_\lambda\right\|^2}{n} + \frac{k_n \sigma^2 d_{a_\lambda}}{n} =$$

$$= \frac{\|\varepsilon\|^2}{n} + L\left(\hat{\beta}_\lambda\right) + \left[L\left(\hat{\beta}^*_{\alpha_\lambda}\right) - L\left(\hat{\beta}_\lambda\right)\right] + \frac{1}{n}\left\|\hat{\mu}^*_{\alpha_\lambda} - \hat{\mu}_\lambda\right\|^2 + \frac{2}{n}\varepsilon'\left(\tilde{I} - \tilde{H}_{\alpha_\lambda}\right)\mu +$$
$$\frac{2}{n}\left(\sigma^2 d_{a_\lambda} - \varepsilon' \tilde{H}_{\alpha_\lambda}\varepsilon\right) + \frac{1}{n}(k_n - 2)\sigma^2 d_{a_\lambda} \qquad \text{(B.2)}$$

Έστω:

$$J_1 = L\left(\hat{\beta}^*_{\alpha_\lambda}\right) - L\left(\hat{\beta}_\lambda\right)$$

$$J_2 = \frac{1}{n}\left\|\hat{\mu}^*_{\alpha_\lambda} - \hat{\mu}_\lambda\right\|^2$$

$$J_3 = \frac{2}{n}\varepsilon'\left(\tilde{I} - \tilde{H}_{\alpha_\lambda}\right)\mu$$

$$J_4 = \frac{2}{n}\left(\sigma^2 d_{a_\lambda} - \varepsilon' \tilde{H}_{\alpha_\lambda}\varepsilon\right)$$

$$J_5 = \frac{1}{n}(k_n - 2)\sigma^2 d_{a_\lambda}$$

Τότε σύμφωνα με τους Fan & Li (1987) επιτυγχάνουμε με πιθανότητα:

$$sup_{\lambda \in [0, \lambda_{max}]}\left|\frac{J_i}{L\left(\hat{\beta}_\lambda\right)}\right| \to 0 \;\; \gamma\iota\alpha\; j = 1, \ldots, 4$$

Επειδή $k_n \to 2$ και $n^{-1}\sigma^2 d_{a_\lambda}/R(\hat{\beta}_\lambda)$ φράσσεται από το 1, μπορούμε να δείξουμε ότι

$$sup_{\lambda \in [0, \lambda_{max}]}\left|\frac{J_5}{L\left(\hat{\beta}_\lambda\right)}\right| \to 0$$

Επακόλουθα, το (B.1) ισχύει, το οποίο υπονοεί ότι η διαφορά μεταξύ $GIC^{LS}_{k_n}(\lambda) - \frac{\|\varepsilon\|^2}{n}$ και $L\left(\hat{\beta}_\lambda\right)$ είναι αμελητέα συγκρίσιμη με το $L\left(\hat{\beta}_\lambda\right)$. Αυτό ολοκληρώνει την απόδειξη.

- Απόδειξη Συνέπειας 1

Όταν το σ^2 είναι άγνωστο, το $GIC^{LS}_{k_n}(\lambda)$ γίνεται:

$$GIC_{k_n}^{LS}(\lambda) = \frac{\|\varepsilon\|^2}{n} + L\left(\hat{\beta}_\lambda\right) + J_1 + J_2 + J_3 + J_4 + J_5 + \frac{2(\tilde{\sigma}^2 - \sigma^2 d_{a_\lambda})}{n}$$

Χρησιμοποιώντας απαραίτητες συνθήκες προκύπτει:

$$sup_{\lambda\in[0,\lambda_{max}\,]} \left| \frac{2(\tilde{\sigma}^2 - \sigma^2 d_{a_\lambda})}{nL\left(\hat{\beta}_\lambda\right)} \right| \to 0$$

με πιθανότητα και η απόδειξη ολοκληρώθηκε.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Akaike, H. (1974). A New Look at the Statistical Model Identification, IEEE Transactions on Automatic Control, **19**, pp.716–723.

[2] Abramowitz, M., Stegun, I.: Handbook of mathematical functions with formulas, graphs, and mathematical tables. U.S. Department of Commerce, National Bureau of Standards Applied Mathematics Series, **No. 55**, pp. 803-819, 1964

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

[4] Antoniadis, A., and Fan, J. (2001). Regularization of Wavelets Approximations. Journal of the American Statistical Association. **96**, pp. 939–967.

[5] Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. 68, 337-404 (1950)

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

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

[8] Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection, The Annals of Statistics, **24**, pp. 2350–2383.

[9] Craven, P., and Wahba, G. (1979). Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation, Numerische Mathematik, **31**, pp. 377–403.

[10] Donoho, D. L., and Johnstone, I. M. (1994a). Ideal Spatial Adaptation by Wavelet Shrinkage. Biometrika, **81**, pp. 425–455.

[11] Donoho, D. L., and Johnstone, I. M. (1994b). Minimax Risk Over l_p Balls for l_q Error. Probability Theory and Related Fields, **99**, pp. 277–303.

[12] Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. (2004). Least Angle Regression, The Annals of Statistics, **32**, 407–499.

[13] Elter, M., Schulz-Wendtland, R., and Wittenberg, T. (2007). The Prediction of Breast Cancer Biopsy Outcomes Using Two CAD Approaches That Both Emphasize an Intelligible Decision Process, Medical Physics, **34**, pp. 4164– 4172.

[14] Fan, J. (1997). Comments on "Wavelets in statistics a review" by A. Antoniadis. J. Italian Statist. Assoc., **6**, pp. 131-138.

[15] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc., **96**, pp. 1348-1360.

[16] Fan, J. & Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Fratile Model," The Annals of Statistics, 30, 74–99.

[17] Fan, J. & Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J. Am. Statist. Assoc. **99**, 710–23.

[18] Fan, J., and Peng, H. (2004). Nonconcave Penalized Likelihood With a Diverging Number of Parameters, The Annals of Statistics, **32**, 928–961.

[19] Frank, I. E., and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. Technometrics, **35**, pp. 109–148.

[20] Fu, W. J. (1998). Penalized Regression: The Bridge Versus the LASSO. Journal of Computational and Graphical Statistics, **7**, pp. 397–416.

[21] Gao, H. Y., and Bruce, A. G. (1997). WaveShrink With Firm Shrinkage, Statistica Sinica, **7**, pp. 855–874.

[22] Golomb, M.: Approximation by periodic spline interpolants on uniform meshes. J. Approximation Theory 1, 26~5 (1968)

[23] Golub, G., Heath, M., Wahba, G.: Generalized cross validation as a method for choosing a good ridge parameter, to appear, Technometrics

[24] Golub, G., Reinsch, C.: Singular value decomposition and least squares solutions. Numer. Math. 14, 403-420 (1970)

[25] Green, P. J., and Silverman, B. W. (1994). Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, London: Chapman and Hall.

[26] Hardle,W., Liang, H. & Gao, J. (2000). Partially Linear Models. Heidelberg: Springer Physica-Verlag.

[27] Heckman, N. E. (1986). Spline smoothing in partly linear models. J. R. Statist. Soc. B **48**, 244–8.

[28] Huang, J. & Yang, L. (2004). Identification of non-linear additive autoregressive models. J. R. Statist. Soc. B **66**, 463–77.

[29] Huber, P. (1981). Robust Estimation, New York: Wiley.

[30] Hudson, H.M.: Empirical Bayes estimation. Technical Report 4~58, Stanford University, Department of Statistics, Stanford, Cal., 1974

[31] Hunter, D., and Li, R. (2005). Variable Selection Using MM Algorithms, The Annals of Statistics, **33**, 1617–1642.

[32] Καρώνη Χ. (2005). Μοντέλα Αξιοπιστίας και Επιβίωσης. Ε.Μ.Π.

[33] Kimeldorf, G., Wahba, G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Inst. Statist. Math. **41**, pp. 495-502 (1970)

[34] Konishi, S., and Kitagawa, G. (1996). Generalised Information Criteria in Model Selection, Biometrika, **83**, 875–890.

[35] Κουκουβίνος Χ. "Γραμμικά μοντέλα και Σχεδιασμοί", Ε.Μ.Π. 2005.

[36] Lehmann, E. L. (1983), Theory of Point Estimation. Paci. c Grove, CA: Wadsworth and Brooks Cole.

[37] Li, R. (2000). High-Dimensional Modeling via Nonconcave Penalized Likelihood and Local Likelihood, unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill, Dept. of Statistics.

[38] Li, K.-C. (1987). Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set, The Annals of Statistics, **15**, pp. 958–975.

[39] Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Regression Modeling," The Annals of Statistics, **36**, pp. 261–286.

[40] Mack,Y. P. & Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. Z. Wahr. Verw. Geb. **61**, pp.405–15.

[41] Mallows, C.L.: Some comments on Cp. Technometrics 15, pp.661-675 (1973)

[42] Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H., and Patil, P. (1998). Exact Risk Analysis of Wavelet Regression, Journal Computational and Graphical Statistics, **7**, pp. 278–309.

[43] McCullagh, P., and Nelder, J. A. (1989). Generalized Linear Models. 2nd ed., London: Chapman and Hall.

[44] Mcquarrie, D. R. & Tsai, C. L. (1998). Regression and Time Series Model Selection. Singapore: World Scientific.

[45] Nishii, R. (1984). Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression, The Annals of Statistics, **12**, 758–765.

[46] Park, M.-Y., and Hastie, T. (2007). An L1 Regularization-Path Algorithm for Generalized Linear Models, Journal of the Royal Statistical Society, Ser. B, 69, pp. 659–677.

[47] Reinsch, C.M.: Smoothing by spline functions. Numer. Math. 10, pp.177-183 (1967)

[48] Reinsch, C.M.: Smoothing by spline functions, II. Numer. Math. 16, pp.451-454 (1971)

[49] Robinson, P. M. (1988). The Stochastic Difference Between Econometrics and Statistics, Econometrica, **56**, pp. 531–547.

[50] Robinson, P. M. (1988). Root-n-consistent semiparametric regression. Econometrica **56**, pp.931–54.

[51] Ruppert, D., Sheather, S. J. & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. J. Am. Statist. Assoc. **90**, pp. 1257–70.

[52] Schoenberg, I.J.: Spline functions and the problem of graduation. Proc. Nat. Acad. Sci. (USA) 52, pp. 947-950 (1964)

[53] Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. **6**, 461–4.

[54] Shao, J. (1997). An asymptotic theory for linear model selection. Statist. Sinica **7**, pp. 221–64.

[55] Shi, P. & Tsai, C. L. (2002). Regression model selection–a residual likelihood approach. J. R. Statist. Soc. B **64**, pp.237–52.

[56] Shi, P. & Tsai, C. L. (2004). A joint regression variable and autoregressive order selection criterion. J. Time Ser. Anal. **25**, pp.923–41.

[57] Shibata, R. (1980). Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process, The Annals of Statistics, 8, pp.147–164.

[58] Shibata, R. (1981). An Optimal Selection of Regression Variables, Biometrika,68, pp.45–54.

[59] Shibata, R. (1984). Approximation Efficiency of a Selection Procedure for the Number of Regression Variables, Biometrika, **71**, pp.43–49.

[60] Speckman, P. (1988). Kernel smoothing in partially linear models. J. R. Statist. Soc. B **50**, pp.413–36.

[61] Stone, C. J., Hansen, M., Kooperberg, C., and Truong, Y. K. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling (with discussion), The Annals of Statistics 25, pp. 1371–1470.

[62] Tibshirani, R. J. (1996). Regression Shrinkage and Selection via the LASSO, Journal of the Royal Statistical Society, Ser. B, **58**, pp. 267–288.

[63] Tibshirani, R. J. (1997). The LASSO Method for Variable Selection in the Cox Model, Statistics in Medicine, **16**, pp. 385–395.

[64] Wahba, G. (1990). Spline Models for Observational Data, Philadelphia: SIAM

[65] Wahba, G.: Convergence rates for certain approximate solutions to first kind integral equations. J. Approximation Theory 7, pp. 167-185 (1973)

[66] Wahba, G.: Smoothing noisy data with spline functions. Numer. Math. 24, 383-393 (1975)

[67] Wahba, G.: Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal. 14,pp. 651-667 (1977)

[68] Wahba, G., Wold, S.: A completely automatic French curve: Fitting spline functions by crossvalidation. Comm. Statist. 4, pp.1-17 (1975)

[69] Wahba, G., Wold, S.: Periodic splines for spectral density estimation: The use of cross-validation for determining the degree of smoothing. Comm. Statist. 4, pp.125-142 (1975)

[70] Wahba, G.: A survey of some smoothing problems and the method of generalized cross validation for solving them. University of Wisconsin-Madison, Statistics Dept.

[71] Wahba, G.: Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. Roy. Statist. Soc., Set. B.

[72] Wang, H., and Leng, C. (2007). Unified LASSO Estimation via Least Squares Approximation, Journal of the American Statistical Association, 102, pp.1039–1048.

[73] Wang, H., Li, G., and Tsai, C.-L. (2007a). Regression Coefficient and Autoregressive Order Shrinkage and Selection via LASSO, Journal of the Royal Statistical Society, Ser. B, **69**, pp.63–78.

[74] Wang, H., Li, R., and Tsai, C.-L. (2007b). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method, Biometrika, **94**, pp.553–568.

[75] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika **92**, pp.973–50.

[76] Yuan, M., and Lin, Y. (2007). On the Non-Negative Garrotte Estimator, Journal of the Royal Statistical Society, Ser. B, **69**, pp.143–161.

[77] Yatchew, A. (1997). An elementary estimator for the partially linear model. Economet. Lett. **57**, pp.135–43.

[78] Zhang, C.-H. (2007). Penalized Linear Unbiased Selection, Technical Report 2007-003, Rutgers University, Dept. of Statistics.

[79] Zhang, H. H., and Lu, W. (2007). Adaptive LASSO for Cox's Proportional Hazards Model, Biometrika, **94**, pp.691–703.

[80] Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," Journal of the American Statistical Association, 101, 1418–1429.

[81]Zou, H., and Li, R. (2008). One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models, The Annals of Statistics, **36**, pp.1509–1533.

[82] Zou, H., Hastie, T., and Tibshirani, R. (2007). On the Degrees of Freedom of the LASSO, The Annals of Statistics, **35**, pp.2173–2192.